Study notes on the Image formation model in Cryo-Electron Microscopy

Carlos Oscar S. Sorzano

Contents

1 Introduction **5**
1.1 Cryo-electron microscopy as a structural-biology measurement 6
1.2 Following the electron from emission to measurement 7
1.3 The source: establishing the initial electron state 9
1.4 Wave diffraction: establishing the language of propagation . . 10
1.5 Electron optics: from electromagnetic fields to the CTF . . . 11
1.6 The specimen: creating the exit wave 13
1.7 Detection: from electron probability to digital data 14
1.8 The forward model has several physical layers 15
1.9 A hierarchy of approximations 17
1.10 Why quantum mechanics appears 18
1.11 How the pieces fit together 19
1.12 The purpose of understanding the forward model 21

2 Electron sources **23**
2.1 Electron emission from a solid 23
2.1.1 Thermionic emission 24
2.1.2 Lowering of the surface barrier by an electric field . . 26
2.1.3 Cold field emission 28
2.2 Electron-gun configurations 30
2.2.1 Thermionic guns 31
2.2.2 Schottky field-emission guns 31
2.2.3 Cold field-emission guns 31
2.2.4 Comparison of the principal electron sources 32
2.3 Relativistic electron kinematics in electromagnetic fields . . . 33
2.3.1 Conservation of energy in static fields 33
2.3.2 The acceleration potential 35
2.4 Electron energy, velocity, and wavelength 39
2.4.1 Velocity of an accelerated electron 40

2.4.2 Relativistic electron wavelength 40
2.4.3 Non-relativistic limit 41
2.4.4 Numerical values relevant to cryo-electron microscopy 42
2.4.5 Energy and temporal frequency 43
2.4.6 Orders of magnitude in a cryo-EM exposure 43
2.5 Beam properties determined by the source 49
2.5.1 Effective and virtual source size 49
2.5.2 Brightness 50
2.5.3 Energy spread and temporal coherence 52
2.5.4 Finite source size and spatial coherence 53
2.5.5 Source-current stability 54
2.6 Summary 54

3 Wave diffraction theory 56
3.1 Fourier transform convention 57
3.2 Wave propagation in homogeneous, isotropic, and non-dispersive media 59
3.2.1 Wave velocity and refractive index 63
3.2.2 From vector to scalar wave theory 63
3.3 Solutions of the wave equation 64
3.3.1 Monochromatic waves and the phasor representation . 67
3.3.2 The Helmholtz equation 68
3.3.3 Integral representation of a Helmholtz field 69
3.4 Connecting classical optics to quantum mechanics 76
3.4.1 The stationary free electron 81
3.4.2 The non-relativistic limit 83
3.4.3 Why a stationary description is useful in electron microscopy 87
3.5 Wave diffraction by an aperture 88
3.5.1 Diffraction of a spherical wave 89
3.5.2 The Huygens–Fresnel interpretation 91
3.5.3 Plane-wave illumination 92
3.5.4 Propagation as a linear integral operator 93
3.5.5 Limitations of the Kirchhoff boundary conditions . . . 93
3.5.6 Extension to non-monochromatic waves 95
3.6 The angular spectrum of a wave 95
3.6.1 Plane waves and spatial frequency 97
3.6.2 The angular-spectrum decomposition 99
3.6.3 Propagation of the angular spectrum 99
3.6.4 Propagating and evanescent components 101

3.6.5 Spatial frequency and the wavelength scale 102
3.6.6 Free-space propagation as a linear filter 103
3.6.7 Apertures and phase-modifying elements 111
3.6.8 Relation to the diffraction-integral formulation 113
3.7 Fresnel and Fraunhofer approximations 113
3.7.1 The Fresnel or paraxial approximation 114
3.7.2 The Fraunhofer or far-field approximation 117
3.7.3 The Fresnel number 118
3.7.4 Fraunhofer diffraction by a circular aperture 121
3.7.5 From optical resolution to resolution in cryo-EM . . . 127
3.8 Phase transformation produced by a thin lens 128
3.8.1 The back focal plane as a Fourier plane 135
3.8.2 Limits of the ideal thin-lens approximation 137
3.8.3 Relation to electron lenses 138
3.9 Impulse response of a thin-lens imaging system 138
3.9.1 Defocus as a quadratic pupil phase 146
3.10 Complex optical configurations 148
3.10.1 Elementary paraxial operators 150
3.10.2 A lens converts angular spectrum into position 151
3.10.3 ABCD representation of a paraxial lens cascade 156
3.10.4 Alternating image and Fourier planes in a focal relay . 164
3.10.5 Interpretation of the main lens systems of a TEM . . 166
3.10.6 Why an electron microscope contains both image and Fourier planes 168
3.11 Detection and partial coherence 169
3.11.1 From wave amplitude to detected intensity 169
3.11.2 Statistical description of partial coherence 173
3.11.3 Coherent and incoherent limits 175
3.11.4 Temporal and spatial partial coherence in electron microscopy 177
3.12 Summary 183

4 Electron optics 191
4.1 Electron trajectories in a uniform axial magnetic field 191
4.1.1 Exact helical motion 192
4.1.2 Representation with the optical coordinate z 195
4.1.3 Paraxial form 196
4.1.4 Numerical scale for a 300 kV electron 198
4.1.5 What the uniform-field model does and does not describe 199
4.2 Paraxial trajectories in a rotationally symmetric magnetic lens 200

4.2.1 Paraxial trajectory equations 201
4.2.2 Larmor rotation emerges from the trajectory equations 211
4.2.3 Conservation of axial angular momentum 213
4.2.4 Why is the Larmor frequency one half of the cyclotron frequency? 218
4.2.5 Focal length in the thin-lens approximation 221
4.2.6 An analytically solvable model of a round magnetic lens 222
4.3 Geometrical optics: trajectories and wavefronts 231
4.3.1 Wavefronts emitted by a point source 232
4.3.2 Huygens' construction 232
4.3.3 The plane-wave limit of a spherical wavefront 233
4.3.4 Fermat's principle in an inhomogeneous medium . . . 236
4.3.5 The ray equation 237
4.3.6 The eikonal and the wavefront 241
4.3.7 From trajectories to wavefronts 241
4.3.8 The eikonal equation 244
4.3.9 Why this becomes more subtle for electrons 244
4.4 Analytical mechanics 245
4.4.1 Action and the principle of stationary action 246
4.4.2 A familiar example: recovery of Newton's equation . . 246
4.4.3 Canonical momentum 247
4.4.4 The Legendre transformation and the Hamiltonian . . 248
4.4.5 Hamilton's equations 249
4.4.6 Conservation of the Hamiltonian 252
4.4.7 The reduced action in a time-independent system . . . 254
4.4.8 From the action functional to a characteristic function 257
4.4.9 Hamilton's central equation 257
4.4.10 The Hamilton–Jacobi equation 259
4.4.11 Static systems and Hamilton's characteristic function 259
4.4.12 Connection with geometrical optics 263
4.4.13 A z-parameterized formulation of geometrical optics . 266
4.4.14 Geometrical optics as the short-wavelength limit of wave optics 268
4.5 Reconnecting geometrical optics to quantum mechanics . . . 272
4.5.1 The classical action as the phase of the quantum wave 272
4.5.2 From the electron wave to the Hamilton–Jacobi equation 275
4.5.3 Feynman's path integral and the emergence of classical rays 278
4.5.4 Meaning and validity of the geometrical-ray description 283
4.6 Hamiltonian dynamics of electrons in electromagnetic fields . 285

4.6.1 Electromagnetic potentials 285
4.6.2 The relativistic electromagnetic Lagrangian 285
4.6.3 The relativistic electron Hamiltonian 290
4.6.4 Static fields and the reduced electron action 293
4.7 Electron lenses and multipole elements 299
4.7.1 Round lenses 299
4.7.2 Cylindrical lenses 300
4.7.3 Dipoles 302
4.7.4 The transverse multipole hierarchy 307
4.7.5 Quadrupoles: focusing one direction while defocusing the other 313
4.7.6 Hexapoles: nonlinear control with six-pole symmetry . 314
4.7.7 Octupoles: cubic transverse forces 315
4.7.8 Multipole order is not aberration order 315
4.7.9 How these elements enter the image-formation model . 316
4.7.10 Summary 318
4.8 Wave mechanics in an electromagnetic field 319
4.8.1 The relativistic scalar wave equation 319
4.8.2 Stationary electromagnetic fields 321
4.8.3 Electrostatic and field-free regions 323
4.8.4 The effect of a magnetic vector potential 325
4.8.5 Gauge transformations and the phase of the electron wave 326
4.9 The eikonal approximation 329
4.9.1 Exact separation of amplitude and phase 329
4.9.2 The eikonal limit 333
4.9.3 Meaning of the approximation 334
4.9.4 Phase fronts and electron trajectories 337
4.9.5 Simple free-space examples 338
4.9.6 Where the eikonal approximation fails 339
4.10 Paraxial approximation 340
4.10.1 How small are paraxial angles in cryo-EM? 341
4.10.2 Why the paraxial approximation produces Gaussian optics 341
4.10.3 The paraxial equation of a round magnetic lens 344
4.10.4 Two fundamental rays describe the Gaussian system . 344
4.10.5 Image planes and angular planes 346
4.10.6 The same approximation in wave optics 348
4.10.7 The back focal plane and the Fourier transform 358

4.10.8 Defocus as propagation away from the Gaussian image plane . . . 364
4.10.9 From coherent wave transfer to the detected electron image . . . 371
4.10.10 First contrast transfer model: defocus phase contrast . 373
4.11 Aberrations . . . 376
4.11.1 Overview and families of aberrations . . . 377
4.11.2 The aberration function as excess action and wavefront error . . . 382
4.11.3 Symmetry, pupil coordinates, and aberration order . . 390
4.11.4 Twofold astigmatism as anisotropic defocus . . . 406
4.11.5 Spherical aberration . . . 416
4.11.6 Coma and odd/antisymmetric aberrations . . . 430
4.11.7 General higher-order aberrations and Zernike polynomials . . . 445
4.11.8 The classical field-dependent Seidel aberrations . . . 462
4.11.9 Chromatic aberration . . . 471
4.12 Amplitude contrast . . . 477
4.12.1 A weak object can modify both amplitude and phase . 477
4.12.2 Transfer of amplitude and phase contrast . . . 478
4.12.3 Why amplitude contrast survives at focus . . . 480
4.12.4 The conventional one-parameter amplitude-contrast model . . . 481
4.12.5 Normalized amplitude contrast . . . 481
4.12.6 Amplitude contrast as an effective CTF phase shift . . 482
4.12.7 The phase shift by amplitude contrast is not an aberration . . . 483
4.12.8 How good is the constant amplitude-contrast approximation? . . . 484
4.13 CTF envelopes . . . 484
4.13.1 The general origin of an envelope . . . 485
4.13.2 Temporal coherence . . . 488
4.13.3 Finite source size and spatial coherence . . . 489
4.13.4 Averaging caused by specimen motion . . . 494
4.13.5 Why independent envelopes multiply . . . 507
4.13.6 A single effective Gaussian envelope . . . 507
4.14 Phase plates . . . 512
4.14.1 Why low spatial frequencies are poorly transferred in focus . . . 512
4.14.2 A phase plate as a pupil-plane transfer function . . . 513

4.14.3 Only the phase relative to the direct beam matters . . 514
4.14.4 The ideal quarter-wave phase plate 516
4.14.5 The cut-on spatial frequency 517
4.14.6 Combination with amplitude contrast 518
4.14.7 A phase plate does not remove the other aberrations . 519
4.14.8 Real phase plates are not necessarily pure phase elements 519
4.14.9 Physical implementations 522
4.14.10 Phase-plate fluctuations lead naturally to an envelope 524
4.14.11 What a phase plate does—and does not do 525
4.15 The contrast transfer function 525
4.15.1 Level 1: the defocus-only CTF 526
4.15.2 Level 2: deterministic coherent aberrations 527
4.15.3 Level 3: symmetric and antisymmetric aberrations . . 527
4.15.4 Adding amplitude contrast 529
4.15.5 What does "coherence" mean here? 530
4.15.6 Level 4: statistical averaging and envelope functions . 532
4.15.7 The physically motivated envelopes 534
4.15.8 The complete factorized CTF model 535
4.15.9 What is coherent and what is incoherent in the final model? 537
4.15.10 What is visible in the power spectrum? 538
4.16 The background of the image power spectrum 539
4.16.1 Why the transfer function appears squared in the power spectrum 539
4.16.2 Why independent image components add in power . . 542
4.16.3 A layered physical model 544
4.16.4 Power spectrum of the ice 548
4.17 Summary 563

5 Interaction with the specimen 582
5.1 From the microscope to the specimen 582
5.2 The specimen as a quantum scatterer 587
5.2.1 The charge-density description of the specimen 587
5.2.2 Elastic and inelastic charge-density channels 589
5.3 The electrostatic potential of matter 591
5.3.1 Charge density, Coulomb potential, and Poisson's equation 591
5.3.2 The screened potential of an atom 592

5.3.3 Electron scattering factors as a Fourier representation of the atomic potential . . . 596
5.3.4 The Mott–Bethe relation: why electrons are especially sensitive to charge . . . 600
5.3.5 Atomic scattering factors in Fourier and real space . . 609
5.3.6 The projected atomic potential . . . 628
5.3.7 From atoms to a biological macromolecule: the independent-atom approximation . . . 635
5.3.8 What does the independent-atom approximation mean for a protein? . . . 641
5.3.9 The specimen potential as the interface to image formation . . . 643
5.4 Elastic scattering by a static potential . . . 645
5.4.1 Incident and scattered electron waves . . . 646
5.4.2 The scattering state and the scattering amplitude . . 648
5.4.3 From scattering amplitude to differential cross section 654
5.4.4 The first Born approximation . . . 657
5.5 The fast-electron equation inside a three-dimensional specimen 666
5.5.1 Starting from the scalar electron wave equation 666
5.5.2 Separating the rapidly varying carrier . . . 667
5.5.3 The interaction constant . . . 672
5.5.4 Physical meaning of the fast-electron equation 674
5.5.5 What has been approximated? . . . 675
5.6 The phase-object and projection approximations as limiting cases . . . 677
5.6.1 Neglecting transverse propagation inside the specimen 677
5.6.2 Why this is also a projection approximation 679
5.6.3 When may propagation through the thickness be neglected? . . . 679
5.6.4 The weak-phase-object approximation . . . 681
5.6.5 The microscopic origin of the specimen phase 682
5.6.6 Three approximations that should not be confused . . 683
5.7 From the specimen exit wave to the standard Single Particle Analysis image model . . . 684
5.7.1 Propagation of the exit wave through the microscope . 685
5.7.2 From intensity to detected electrons . . . 687
5.7.3 The standard single-particle forward model 688
5.7.4 What changes for a phase object that is not weak? . . 689
5.7.5 What changes under the first Born approximation? . . 690

5.7.6 How the three specimen models relate to the SPA equation . . . 691
5.8 The multislice method . . . 692
5.8.1 Propagation and interaction as operators . . . 692
5.8.2 Dividing the specimen into slices . . . 693
5.8.3 The transmission function of one slice . . . 698
5.8.4 Propagation between consecutive slices . . . 699
5.8.5 Why multislice contains multiple elastic scattering . . . 699
5.8.6 Where the multislice approximation enters . . . 701
5.8.7 Choosing the slice thickness . . . 702
5.8.8 A symmetric splitting . . . 702
5.8.9 Spatial sampling, bandwidth, and aliasing . . . 703
5.8.10 Periodic boundaries and the simulation box . . . 704
5.8.11 From the incident wave to the exit wave . . . 704
5.8.12 What multislice has changed . . . 705
5.9 The Ewald sphere and genuinely three-dimensional elastic scattering . . . 706
5.9.1 The Ewald sphere in the spatial-frequency space . . . 706
5.9.2 Connection with the scattering angle . . . 710
5.9.3 The paraxial curvature of the Ewald sphere . . . 713
5.9.4 What the projection approximation removes . . . 714
5.9.5 Depth appears as a phase . . . 715
5.9.6 Why a very flat Ewald sphere can still matter . . . 716
5.9.7 Ewald curvature and Fresnel propagation are the same paraxial phase . . . 719
5.9.8 From the Ewald sphere to a real-valued cryo-EM image 722
5.9.9 Ewald-sphere curvature and absolute handedness. . . . 738
5.10 Depth is defocus: the three-dimensional origin of depth-dependent phase . . . 741
5.10.1 Different specimen planes propagate different distances 742
5.10.2 A defocus value refers to a reference plane . . . 743
5.10.3 What the single-defocus image model neglects . . . 745
5.10.4 Multislice contains the depth dependence automatically 746
5.10.5 The same longitudinal phase appears in the Ewald sphere . . . 749
5.10.6 When is one defocus and the projection approximation adequate for Single Particle Analysis? . . . 752
5.11 Atomic disorder, thermal motion, occupancies, and B-factors 756
5.11.1 Atomic displacements and the Debye–Waller factor . . . 757
5.11.2 Occupancy and positional disorder . . . 759

5.11.3 Frozen configurations and the frozen-phonon construction . . . 763
5.11.4 Static conformational heterogeneity in SPA . . . 774
5.11.5 Radiation-induced motion and radiation damage . . . 781
5.11.6 What does a B-factor mean in Single Particle Analysis? 782
5.12 Inelastic scattering: transitions of the specimen . . . 784
5.12.1 From the static elastic potential to transition potentials 786
5.12.2 The combined electron–specimen transition . . . 787
5.12.3 Energy and momentum transfer . . . 788
5.12.4 The first-Born transition matrix element . . . 792
5.12.5 Transition charge density and transition form factor . 797
5.12.6 From transition amplitudes to a measurable cross section 805
5.12.7 Why different inelastic final states do not interfere, but an inelastic electron can still carry phase . . . 812
5.12.8 Small momentum transfer and the characteristic inelastic angle . . . 816
5.12.9 From microscopic cross section to the inelastic mean free path . . . 828
5.12.10 Inelastic mean free path, zero-loss electrons, and effective absorption . . . 837
5.12.11 Quantitative balance of elastic and inelastic scattering 852
5.12.12 Contribution of inelastic scattering to image formation and the power spectrum . . . 857
5.13 Radiation damage and beam-induced specimen evolution . . . 864
5.13.1 Irradiation changes the specimen while we are imaging it 864
5.13.2 Electron exposure, absorbed dose, and dose rate . . . 866
5.13.3 Radiolysis: from an inelastic event to chemical damage 874
5.13.4 Why cooling helps: radical trapping and the cage effect 885
5.13.5 Damage is chemically and spatially selective . . . 894
5.13.6 Resolution-dependent fading of structural information 900
5.13.7 The intrinsic spectral decay of the structural signal . . 913
5.13.8 Radiation damage of vitreous ice itself . . . 924
5.13.9 Elastic nuclear recoil: knock-on displacement and sputtering . . . 935
5.13.10 Beam heating . . . 942
5.13.11 Beam-induced mechanical motion . . . 947
5.13.12 Charging: radiation-induced specimen modification or electron-optical artifact? . . . 957
5.13.13 The low-dose acquisition principle . . . 968
5.13.14 From low-dose images to dose-fractionated movies . . 974

5.14 Summary 982

6 Electron detection **992**
6.1 From an arriving electron to a recorded measurement 992
6.2 From photographic film to direct electron detection 999
6.3 Interaction of the imaging electron with the detector 1006
6.3.1 Collisional stopping power. 1007
6.3.2 Radiative stopping power. 1032
6.3.3 Total stopping power 1048
6.4 Where did the electron hit? 1057
6.4.1 Transport and spatial spreading of energetic electrons in silicon 1059
6.4.2 Physical construction of a direct electron detector . . 1066
6.4.3 Spatial spreading from energy deposition to collected charge 1076
6.4.4 Pixel readout and localization of electron events . . . 1091
6.5 Spatial response of the detector 1101
6.5.1 The detector point-spread function 1101
6.5.2 The modulation transfer function 1102
6.5.3 Pixel integration 1105
6.5.4 Nyquist frequency and aliasing 1112
6.6 Physical sources of noise in a direct electron detector 1120
6.6.1 Incident-electron shot noise 1120
6.6.2 Fluctuations of the single-electron response 1129
6.6.3 Energy-deposition and charge-generation fluctuations 1141
6.6.4 Charge-sharing and localization fluctuations 1146
6.6.5 Coincidence loss in electron counting 1155
6.6.6 Electronic readout and reset noise 1162
6.6.7 Dark-current and thermal fluctuations 1173
6.6.8 Gain variations and fixed-pattern noise 1179
6.6.9 Analog-to-digital quantization 1188
6.7 Noise power spectrum 1194
6.7.1 Why a single variance is not enough 1195
6.7.2 Definition and estimation of the NPS 1201
6.8 Spectral signatures of the physical noise mechanisms 1209
6.8.1 Incident-electron shot noise 1210
6.8.2 Fluctuations of the single-electron response 1211
6.8.3 Energy-deposition and charge-generation fluctuations 1213
6.8.4 Charge-sharing and localization fluctuations 1214
6.8.5 Coincidence loss 1217

6.8.6 Electronic readout and reset noise 1222
6.8.7 Dark-current fluctuations 1225
6.8.8 Gain variations and fixed-pattern structure 1225
6.8.9 Analog-to-digital quantization 1226
6.8.10 Combination of noise sources and the origin of the background PSD 1227
6.9 Detective quantum efficiency 1237
6.9.1 DQE in the single-electron-response framework 1243
6.9.2 DQE at zero spatial frequency 1246
6.9.3 DQE near the sampling limits and beyond the physical Nyquist frequency 1249
6.10 Noise-equivalent quanta 1254
6.11 Nonideal detector behavior 1264
6.11.1 Saturation 1265
6.11.2 Dead time 1266
6.11.3 Persistence 1267
6.11.4 Temporal response and temporal sampling 1268
6.11.5 Bad pixels and columns 1272
6.11.6 Radiation damage of the detector 1272
6.11.7 Spatially varying response 1273
6.12 Experimental characterization of an electron detector 1275
6.12.1 Measurement of the gain 1276
6.12.2 Measurement of the MTF 1278
6.12.3 Measurement of the NPS 1281
6.12.4 Measurement of the DQE 1282
6.12.5 Dose-rate dependence 1285
6.12.6 How to interpret a reported detector specification . . . 1286
6.13 Detector contribution to the complete cryo-EM forward model 1287
6.13.1 The mean image incident on the detector 1288
6.13.2 Mean detector transfer 1288
6.13.3 The recorded image is a stochastic quantity 1290
6.14 The power spectrum of a recorded cryo-EM image 1291
6.14.1 Structural signal power 1292
6.14.2 Specimen background power 1293
6.14.3 Electron shot-noise power 1293
6.14.4 Detector-added noise power 1294
6.14.5 The complete observed power spectrum 1295
6.14.6 Recovering the empirical CTF-plus-background model 1296
6.14.7 Sampling, aliasing, and nonlinear detection 1297
6.14.8 The complete hierarchy of cryo-EM image formation . 1298

6.15 Summary 1299

A Minimal quantum mechanics for electron microscopy 1306
A.1 Quantum mechanics as geometry in a Hilbert space 1307
A.1.1 From ordinary vectors to Hilbert spaces 1307
A.1.2 Functions can also be vectors 1308
A.1.3 The abstract state and its representations 1309
A.1.4 A connection with Fourier analysis 1310
A.1.5 Superposition is vector addition 1311
A.1.6 Normalization and physical states 1311
A.1.7 Operators are transformations of states 1312
A.1.8 Eigenvectors are special directions in Hilbert space . . 1313
A.1.9 Inner products measure overlap 1313
A.1.10 The continuous position basis requires one qualification 1314
A.1.11 Why Hilbert spaces are useful for the rest of the book 1314
A.2 The state of a quantum system 1315
A.2.1 The position representation of a state 1315
A.3 Probability amplitudes and probabilities 1319
A.3.1 The wave function is an amplitude 1320
A.3.2 Interference of probability amplitudes 1321
A.4 Bras, kets, and inner products 1322
A.4.1 From a ket to a bra 1322
A.4.2 The inner product in Dirac notation 1323
A.4.3 Conjugation and linearity 1323
A.4.4 Norm and orthogonality in Dirac notation 1324
A.4.5 The inner product in the position representation . . . 1325
A.4.6 Outer products 1326
A.4.7 How to read bra–ket expressions 1327
A.4.8 A compact dictionary 1328
A.5 Momentum states, Fourier transforms, and uncertainty 1328
A.5.1 The momentum basis 1328
A.5.2 What does a state of definite momentum look like in real space? 1329
A.5.3 The position and momentum representations are Fourier transforms 1331
A.5.4 Connection with the spatial-frequency notation of this book 1331
A.5.5 Position and momentum states are opposite limiting cases 1333
A.5.6 Quantifying localization 1334

A.5.7 The Fourier uncertainty relation 1334
A.5.8 Heisenberg's position–momentum uncertainty principle 1338
A.5.9 What the uncertainty principle means 1339
A.5.10 The Gaussian reaches the minimum uncertainty . . . 1340
A.5.11 Does an electron have a preferred wave-packet shape? 1343
A.5.12 Connection with electron microscopy 1346
A.5.13 Summary: three descriptions of the same state 1346
A.6 Operators 1347
A.7 Eigenstates, observables, and measurement 1347
A.7.1 Hermitian operators 1348
A.7.2 Eigenvalues of a Hermitian operator are real 1348
A.7.3 Eigenstates with different eigenvalues are orthogonal . 1349
A.7.4 Expanding a state in eigenstates 1350
A.7.5 From eigenvalues to measurement outcomes 1351
A.7.6 The Born rule 1351
A.7.7 Expected value 1352
A.7.8 Spread of the measurement outcomes 1353
A.7.9 Degenerate eigenvalues 1357
A.7.10 What is derived and what is postulated 1358
A.8 Discrete and continuous measurement probabilities 1358
A.9 Expectation values in a chosen representation 1359
A.9.1 Example: mean position 1360
A.9.2 The imaging electron interacting with the specimen . 1361
A.10 The Hamiltonian 1378
A.10.1 The Hamiltonian determines time evolution 1379
A.10.2 The Hamiltonian is the energy operator 1379
A.10.3 Time evolution of an energy eigenstate 1380
A.10.4 Kinetic and potential energy 1381
A.10.5 A local scalar potential in position representation . . . 1382
A.10.6 The Hamiltonian in position representation 1382
A.10.7 Relation with the electron–specimen interaction 1383
A.10.8 Electrostatic interaction with the specimen 1384
A.10.9 Relativistic electrons 1384
A.11 Unitary time evolution 1389
A.11.1 The time-evolution operator 1389
A.11.2 Why quantum evolution preserves normalization . . . 1392
A.11.3 Time evolution in the energy eigenstate basis 1393
A.11.4 Why an interaction can produce transitions 1394
A.11.5 Why energy eigenstates are called stationary 1395
A.12 Matrix elements and transitions between states 1396

A.12.1 Why they are called matrix elements 1397
A.12.2 Diagonal and off-diagonal matrix elements 1397
A.12.3 Interaction matrix elements 1398
A.12.4 From coupling to a transition amplitude 1399
A.12.5 Weak-interaction approximation 1400
A.12.6 Fermi's golden rule 1401
A.12.7 Connection with elastic and inelastic electron scattering 1402
A.13 Why orthogonal specimen channels add as probabilities . . . 1404
A.14 Relativistic electron–electron scattering 1406
A.14.1 Four-momentum and Lorentz-invariant products . . . 1407
A.14.2 Covariant Dirac notation 1408
A.14.3 Invariant kinematics for two-particle scattering 1413
A.14.4 Møller scattering: two indistinguishable alternatives . 1416
A.14.5 Unpolarized scattering and spin sums 1423
A.14.6 From the invariant amplitude to a cross section 1436
A.14.7 Laboratory frame and transferred energy 1446
A.14.8 Restoring SI units 1451
A.15 Summary of notation and main quantum-mechanical results . 1455
A.15.1 Dictionary of notation 1455
A.15.2 The main results to remember 1455
A.15.3 A compact conceptual map 1461

Preface

I arrived at the laboratory twenty-nine years ago. At the time, my institute worked exclusively with negative-stain electron microscopy, which produces images with considerably higher contrast than cryo-electron microscopy. I had already worked with images before joining the laboratory, and when I first saw the electron micrographs alongside the structures people hoped to recover from them, I remember thinking the problem was either impossible or that my stay in the laboratory would be very short. Neither turned out to be true.

Soon after I arrived, I was given Joachim Frank's *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, which had been published the previous year. I read it from beginning to end. In fact, I read it three times in succession. Each time, I understood things I had failed to grasp the previous time. That experience has stayed with me: in this field, understanding often comes in layers.

Some aspects of the electron microscope fascinated me from the beginning, while others bothered me because I could use them without really understanding them.

I was fascinated that electrons travel down the microscope column at more than half the speed of light. At those velocities, relativistic effects are already significant. I liked to imagine the electrons crossing the specimen while, in their own frame of reference, time passed much more slowly (in reality, not so much). I was equally fascinated by the fact that most of matter is empty space, and that the electrons can nevertheless "see" the specimen at a distance, through the electromagnetic field associated with its charge distribution.

Other things were much less satisfying.

It bothered me to use the CTF formula without understanding where it came from. It bothered me that, under the weak-phase-object approximation, cryo-EM images could often be described by a comparatively simple projection model, while images of other materials clearly showed that this

picture could not always be correct. I did not understand the ghost-like shadows illustrated by Christopher Russo and Richard Henderson in their 2018 work on Ewald-sphere curvature. I was also bothered by the power spectral density of the noise beneath the Thon rings and by not knowing which physical processes produced its shape.

This book is, to a large extent, an attempt to resolve all of those things that bothered me.

It represents study time that I owe to the many outstanding researchers who worked in this field before me—and also to many who have worked alongside me, since twenty-nine years is a sufficiently long time for several generations of ideas to appear. Through their papers I repeatedly encountered beautiful pieces of theory and physical insight. The difficulty was that I did not know how to fit them all into a single coherent picture.

A scientific article has little space for this. Ten pages may be enough to choose a starting point and develop a short segment of theory. A book lets you follow the thread much further. There are, of course, excellent books on electron microscopy, diffraction, electron optics, quantum mechanics, scattering, and detector physics, and many of them develop individual parts of this story in considerable depth. Nevertheless, I never found the entire chain presented in exactly the way I was looking for.

The question that concerned me was very specific: how is a cryo-EM image physically formed, in enough detail that we can invert that process correctly with computational algorithms?

I wanted to follow the electron through the entire experiment: from its creation at the source, through acceleration and electron optics, through its interaction with the specimen, and finally to its detection. At every stage, I wanted to understand which mathematical transformations took place and which approximations were required to obtain the image-formation models we routinely use in cryo-EM image processing.

I have also tried not to let the discussion remain entirely at the level of formal theory. My background is in engineering, and I find it difficult to develop intuition without knowing the order of magnitude of the quantities involved. Wherever it seemed useful, I have therefore inserted concrete numerical values. Knowing that a quantity is "small" is much less informative than knowing whether it is 10^{-2}, 10^{-6}, or 10^{-12} in the units relevant to the experiment.

For similar reasons, I have tried to avoid the phrase "it can be shown that". If something can be shown, I would rather show it. I have deliberately used relatively small steps in many of the derivations so that equations do not appear suddenly, as if by magic. I have placed longer mathematical

developments in dedicated derivation boxes. A reader who is primarily interested in the physical meaning of a result can therefore skip them, while a reader who wants to understand where the formula comes from can follow the argument step by step.

I should also make clear what this book is not.

I am not an expert in every subject discussed here. Indeed, I would have liked to have internalized much more of this material than I have. Perhaps another twenty-nine years would be enough for some of these arguments to become completely natural to me. Instead, I have studied what others have written, tried to understand it as deeply as I can, connected the different pieces, and explained what I believe I understood.

For that reason, I cannot guarantee that everything written here is completely correct. There will undoubtedly be mistakes, omissions, and places where a specialist would argue more precisely. I apologize for those in advance.

This is also why I have called these *Study Notes*. I would have liked to call them *Teaching Notes*, but that would suggest a degree of mastery that I do not claim to have. These notes instead record a process of learning.

There may nevertheless be an advantage to this for the reader. I tend not to be satisfied until I understand the details of an argument, and I often need to proceed through it slowly. Readers who share this difficulty may therefore find that the book takes sufficiently small steps for them to follow the reasoning without too many gaps. The price is that the resulting book is considerably longer than a more concise treatment would have been.

I would like to express my gratitude to all the researchers whose work has helped me understand the physics of cryo-EM image formation a little better. I shall resist the temptation to name individuals here, because any such list would inevitably omit someone who deserves to be included.

I am equally grateful to all those who have passed through our laboratory over the years. Together we have spent a great deal of time trying to disentangle this remarkable puzzle: how electrons travel through an electron microscope, acquire information about molecular structure, and finally leave that information behind in images whose raw form may contain roughly a thousand times more noise than signal.

I would also like to acknowledge the contribution of generative large language models. They are, in my view, a remarkable invention and have proved extraordinarily useful in the preparation of this book. They write English more fluently than I do, are highly capable assistants in mathematical derivations, can keep track tirelessly of thousands of LaTeX references to equations, figures, and sections, and are excellent tools for summarizing,

organizing, and analysing large amounts of documentation. Used critically, they have been invaluable in helping me refine explanations, check the internal consistency of long derivations, and maintain coherence across a text of this length.

I hope that the few readers who decide to follow this book in sufficient depth will enjoy the process as much as I have enjoyed writing it.

And I hope that, when I return to these pages myself, I will experience again what I felt when reading Frank's book twenty-nine years ago: that with every new reading, some part of the mystery contained in the images we process becomes a little clearer.

Chapter 1

Introduction

Cryo-electron microscopy has become one of the central experimental techniques of structural biology. Its purpose is deceptively simple: to use electrons to obtain information about the three-dimensional structure of biological matter. The experiment itself, however, does not produce an atomic structure. Nor does it produce a direct map of the electrostatic potential of the specimen. What is recorded is a collection of noisy, discretely sampled measurements generated after high-energy electrons have been emitted, accelerated, transported through electromagnetic fields, scattered by the specimen, transformed by the electron-optical system, and finally detected.

Understanding the relationship between biological structure and the recorded image is the subject of this book.

The central question is therefore not merely

$$\text{How does an electron microscope work?} \tag{1.1}$$

but rather

$$\boxed{\begin{array}{c}\text{Given a physical specimen, what statistical model describes}\\ \text{the image recorded by the microscope?}\end{array}} \tag{1.2}$$

This distinction determines the book's scope. We shall discuss the physical components of the microscope only to the extent that they affect the image-formation model. Conversely, whenever a physical effect modifies the relationship between the specimen and the recorded data, we shall attempt to follow that effect to its mathematical consequence.

This is why some apparently important parts of the microscope receive less space than others. The electron source, for example, is essential: without it there is no experiment. Nevertheless, from the point of view of image

formation, much of the engineering of an electron gun can eventually be summarized by a comparatively small set of beam parameters—electron energy, energy spread, effective source size, angular distribution, brightness, and stability. Chapter 2 is therefore intentionally more compact than the chapters devoted to electron optics, specimen interaction, and detection, where the transformations that determine the recorded image must be developed in considerably greater detail.

The objective throughout the book is not simply to state the equations routinely used in cryo-EM image processing. It is to understand where they come from, which physical approximations are hidden inside them, which quantities can be interpreted independently, and under what conditions the usual image-formation model ceases to be adequate.

1.1 Cryo-electron microscopy as a structural-biology measurement

Structural biology seeks to determine how biological organization in three dimensions gives rise to molecular function. Cryo-electron microscopy contributes to this objective by probing specimens preserved in vitreous ice with high-energy electrons and inferring structural information from the resulting images.

The important word in the preceding sentence is *inferring*. A cryo-EM micrograph should not be interpreted as a photograph of the molecular structure. Between the atomic configuration and the recorded numbers lies a long sequence of physical transformations.

Even in the standard single-particle approximation, the quantity of interest is usually represented by a three-dimensional electrostatic potential,

$$V_0(\mathbf{r}), \tag{1.3}$$

whereas the experiment records a two-dimensional digital image,

$$y_i(\mathbf{n}). \tag{1.4}$$

Connecting these two objects requires a forward model.

One familiar version of that model is derived later in Sec. 5.7. In Fourier space it takes the form

$$\boxed{\widehat{y}_i(\mathbf{R}) = a_i \, \mathrm{CTF}_i(\mathbf{R}) e^{-j2\pi \mathbf{R}\cdot\mathbf{t}_i} \widehat{\mathcal{P}_{\mathcal{R}_i} V_0}(\mathbf{R}) + \widehat{n}_i(\mathbf{R}),} \tag{1.5}$$

which is derived explicitly in Eq. (5.608).

Equation (1.5) is extraordinarily useful. It lies behind much of single-particle cryo-EM image processing. It is not, however, a fundamental law of the experiment.

Every factor in it contains physics.

The projection operator $\mathcal{P}_{\mathcal{R}_i}$ assumes that a three-dimensional specimen can, under appropriate conditions, be represented by a two-dimensional projected potential. The CTF contains the coherent action of the objective lens, aberrations, partial coherence, and amplitude contrast. The apparently simple noise term $\widehat{n}_i$ can contain contributions generated by several physically different processes occurring at different stages of the experiment. Even the assumption that the specimen is described by one fixed potential V_0 becomes imperfect once radiation damage and beam-induced motion are taken into account.

A principal goal of this book is therefore to arrive at Eq. (1.5) rather than to assume it. Once we understand how it arises, we can also understand how to extend it.

1.2 Following the electron from emission to measurement

The organization of the book follows, as closely as possible, the physical path of an imaging electron.

At the most schematic level, this path is

$$\boxed{\begin{array}{c}\text{electron emission}\\ \Downarrow\\ \text{acceleration and illumination}\\ \Downarrow\\ \text{propagation through electromagnetic fields}\\ \Downarrow\\ \text{interaction with the specimen}\\ \Downarrow\\ \text{electron-optical image formation}\\ \Downarrow\\ \text{electron detection}\\ \Downarrow\\ \text{digital image.}\end{array}} \tag{1.6}$$

There is a small organizational subtlety in this sequence. Physically, electron optics occurs both before and after the specimen. Condenser lenses prepare the illumination, whereas the objective and projector systems act on the electron wave after its interaction with the specimen. It would be unnecessarily repetitive to develop the physics of electromagnetic electron lenses twice. We therefore develop the general theory of electron optics together in Chap. 4, and subsequently insert the specimen transformation into that framework in Chap. 5.

There is also one important mathematical detour. Before specializing to electron lenses, Chap. 3 develops the general theory of wave propagation, diffraction, Fourier optics, lenses, and partial coherence. This chapter does not correspond to a new physical component placed somewhere along the electron column. Rather, it provides the mathematical language with which the rest of the electron trajectory will be described.

The resulting logical organization is therefore

$$\boxed{\begin{array}{c}\text{electron source}\\ \text{Chap. 2}\\ \Downarrow\\ \text{general wave-propagation framework}\\ \text{Chap. 3}\\ \Downarrow\\ \text{electron optics}\\ \text{Chap. 4}\\ \Downarrow\\ \text{electron–specimen interaction}\\ \text{Chap. 5}\\ \Downarrow\\ \text{electron detection}\\ \text{Chap. 6.}\end{array}} \tag{1.7}$$

The following sections summarize what each of these stages contributes to the final image-formation model.

1.3 The source: establishing the initial electron state

The trajectory begins at the electron source, Chap. 2. Its first function is obvious: it must produce electrons. For image formation, however, the important question is not only whether electrons are emitted, but which physical distribution of electrons enters the microscope.

The microscopic emission mechanism is developed in Sec. 2.1. Thermionic emission, Schottky-assisted emission, and cold field emission provide different mechanisms for overcoming or tunnelling through the surface potential barrier. Their practical realizations in electron guns are discussed in Sec. 2.2.

Once emitted, the electrons are accelerated to the high kinetic energies used in cryo-EM. Their motion is already relativistic. The relation between acceleration potential, energy, momentum, velocity, and wavelength is therefore derived in Secs. 2.3 and 2.4. In particular, the electron wavelength is not introduced as an independent parameter: it follows from the relativistic momentum through the de Broglie relation of Eq. (2.104).

This wavelength will subsequently enter essentially every spatial-frequency relation in the book.

The source is also not a mathematical point emitting electrons of one exact energy. Its finite effective size, energy spread, angular distribution, brightness, and stability are introduced in Sec. 2.5. These quantities later become important because a distribution of electron energies produces temporal partial coherence, whereas a distribution of illumination directions produces spatial partial coherence. Their statistical interpretation is developed in Sec. 3.11.2, with the correspondence between the two mechanisms summarized in Eq. (3.753).

Thus, from the perspective of the rest of the book, the complicated physics of the source eventually supplies a relatively compact set of initial conditions:

$$\boxed{\begin{array}{c}\text{mean energy and wavelength,}\\ \text{energy distribution,}\\ \text{effective source size and angular distribution,}\\ \text{brightness and current,}\\ \text{temporal stability.}\end{array}} \tag{1.8}$$

This is precisely why the source chapter can be shorter than the chapters that follow. We are not attempting a comprehensive theory of electron-gun

engineering. We retain those properties of the source that survive into the image-formation model.

1.4 Wave diffraction: establishing the language of propagation

An electron is a quantum particle, not an electromagnetic wave. Nevertheless, much of electron-microscope image formation can be described using the mathematical machinery of classical wave optics. Chapter 3 establishes why.

For a monochromatic scalar wave, the spatial field satisfies the Helmholtz equation derived in Sec. 3.3.2. Starting from the relativistic Dirac equation, the stationary free electron is shown in Sec. 3.4.1 to satisfy the same spatial Helmholtz operator, with

$$k = \frac{p}{\hbar} = \frac{2\pi}{\lambda}, \tag{1.9}$$

as derived in Eqs. (3.159)–(3.163).

This does *not* identify electrons with electromagnetic waves. Rather, it establishes that their spatial propagation through a homogeneous field-free region obeys the same mathematical equation. Consequently, diffraction theory and Fourier optics can be transferred to electron waves.

The Fourier-transform convention is fixed at the beginning of Chap. 3, in Sec. 3.1. The distinction between the wavevector $\mathbf{k}$, measured in radians per unit length, and the spatial-frequency vector $\mathbf{R}$, measured in cycles per unit length, is established through Eq. (3.6).

The angular-spectrum construction of Sec. 3.6 then gives spatial frequency a direct physical interpretation. A transverse spatial frequency corresponds to a propagation direction. Thus,

$$\boxed{\text{spatial frequency} \longleftrightarrow \text{propagation angle} \longleftrightarrow \text{position in a Fourier plane.}} \tag{1.10}$$

This equivalence will recur throughout the rest of the book.

The Fresnel and Fraunhofer approximations developed in Sec. 3.7 show how free propagation becomes, under the appropriate conditions, multiplication by a quadratic phase or a Fourier transformation. Thin lenses, introduced in Sec. 3.8, also contribute a quadratic phase. The cancellation between propagation and lens phases explains why a lens can form both images and Fourier planes.

This point is particularly important in electron microscopy. Because the electron wavelength is extremely short, reaching the natural far-field

regime by free propagation alone can require impractically large distances. Section 3.10.2 shows how a lens brings the corresponding Fourier plane to a finite distance. The mapping

$$\mathbf{r}_{\mathrm{BFP}} = \lambda f\, \mathbf{R} \simeq f\, \boldsymbol{\theta} \tag{1.11}$$

is derived in Eqs. (3.608) and (3.610).

The chapter finally introduces detection at the level of electron probability flux and develops the statistical language of partial coherence in Sec. 3.11. These concepts are needed before energy spread, source size, chromatic effects, and illumination-angle effects can be incorporated consistently into the CTF.

Chapter 3 therefore establishes the common mathematical vocabulary of the book:

$$\boxed{\begin{array}{l}\text{wave amplitude} \longrightarrow \text{Fourier decomposition} \longrightarrow \text{propagation phase} \\ \longrightarrow \text{lenses} \longrightarrow \text{image and Fourier planes} \longrightarrow \text{partial coherence.}\end{array}} \tag{1.12}$$

1.5 Electron optics: from electromagnetic fields to the CTF

Having established the general language of wave propagation, we must specialize it to electrons moving through the electromagnetic fields of a transmission electron microscope. This is the purpose of Chap. 4.

The development begins at the most microscopic classical level. The Lorentz force bends the trajectory of a charged particle. An electron in a uniform axial magnetic field follows a helix, as derived in Sec. 4.1. A real rotationally symmetric magnetic lens has an axial field that varies along the optical axis, and the resulting paraxial trajectory equations lead to focusing, as derived in Sec. 4.2.

But a theory of individual trajectories is not yet a theory of image formation. A wave is determined not by one ray but by the phase relations among a family of possible trajectories.

This motivates the progression through geometrical optics and analytical mechanics in Secs. 4.3 and 4.4. The action and the Hamilton–Jacobi equation provide the conceptual bridge between trajectory mechanics and wave phase. This connection is made explicit for electrons in Sec. 4.5: in the semiclassical limit, the classical action appears as the phase of the electron wave.

The electromagnetic potentials are then incorporated into the relativistic electron Hamiltonian in Sec. 4.6. Round lenses, dipoles, quadrupoles, hexapoles, octupoles, and related electron-optical elements are discussed in Sec. 4.7. Their detailed fields may be complicated, but their relevance to image formation is ultimately condensed into how they transform the wavefront.

The paraxial approximation of Sec. 4.10 produces Gaussian imaging as the leading-order description. Departures from the ideal quadratic wavefront then appear as aberrations, developed systematically in Sec. 4.11.

This is one of the major reductions in the book:

$$\boxed{\begin{array}{c}\text{electromagnetic fields}\\ \Downarrow\\ \text{electron trajectories}\\ \Downarrow\\ \text{action}\\ \Downarrow\\ \text{wavefront phase}\\ \Downarrow\\ \chi(\mathbf{R}).\end{array}} \tag{1.13}$$

The complicated deterministic action of the objective lens is therefore largely summarized by its pupil $P(\mathbf{R})$ and aberration function $\chi(\mathbf{R})$.

Defocus, twofold astigmatism, spherical aberration, coma, trefoil, and higher-order aberrations are not introduced as unrelated correction terms. They are organized as particular forms of wavefront error. Once the familiar low-order aberrations have been understood, Zernike polynomials provide a systematic basis for representing more general smooth aberrations over the circular pupil.

The next step is to understand how the coherent electron-wave transfer becomes image intensity contrast. For a weak phase object, this produces the familiar contrast transfer function. The complete local CTF derived in Sec. 4.15 is

$$\begin{aligned}\mathrm{CTF}(\mathbf{R}) = {} & E_{\mathrm{tot}}(\mathbf{R})P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\\ & \times\left[\sqrt{1-A^2}\,\sin\chi_{\mathrm{sym}}(\mathbf{R}) + A\,\cos\chi_{\mathrm{sym}}(\mathbf{R})\right],\end{aligned} \tag{1.14}$$

as derived in Eq. (4.2117).

Equation (1.14) already reveals several physically distinct layers. The pupil selects spatial frequencies; symmetric aberrations produce the familiar

CTF oscillations; antisymmetric aberrations introduce Fourier phase shifts; amplitude contrast mixes the phase and amplitude quadratures; and statistical averaging over source and microscope fluctuations produces the envelope E_{tot}.

The CTF is therefore not an empirical sinusoid attached to a micrograph. It is the end of a long chain beginning with electromagnetic forces and ending with weak-object intensity formation.

1.6 The specimen: creating the exit wave

After developing the electron-optical system, one crucial quantity is still missing: the wave that the microscope must image. Chapter 5 supplies it.

The entrance and exit waves are related abstractly by the specimen scattering operator introduced in Eq. (5.4),

$$\boxed{\psi_{\text{exit}} = \mathcal{S}\,\psi_{\text{in}}.} \tag{1.15}$$

A substantial part of the chapter is devoted to understanding what the apparently compact symbol $\mathcal{S}$ contains.

The first step is to connect matter with an electrostatic potential. The specimen charge distribution produces the potential $V(\mathbf{r})$ through Poisson's equation, as developed in Sec. 5.3. Atomic scattering factors provide an equivalent reciprocal-space description and connect atomic structural models with the potential required for electron-wave calculations.

The propagation of a fast electron through a three-dimensional potential is then described by the paraxial fast-electron equation derived in Eq. (5.486). Its two terms have an especially useful interpretation:

$$\boxed{\begin{array}{c}\text{transverse diffraction}\\ + \\ \text{phase evolution produced by the specimen potential.}\end{array}} \tag{1.16}$$

The familiar phase-object and projection approximations emerge as limiting cases of this more general equation in Sec. 5.6. Under the phase-object approximation, the specimen is represented by the transmission function of Eq. (5.548). Under the additional weak-phase approximation, the exit wave can be linearized, eventually leading to the standard SPA model of Eq. (5.608).

An important objective of Chap. 5 is to keep these approximations conceptually separate. The first Born approximation, the projection approxi-

mation, and the weak-phase-object approximation are not three names for the same assumption.

When propagation through specimen depth cannot be neglected, Sec. 5.8 develops the multislice description. Interaction with successive slices alternates with free propagation, and the depth ordering of the three-dimensional potential is retained. The relation to Ewald-sphere curvature and genuinely three-dimensional scattering is developed in Sec. 5.9.

The specimen also introduces a distinction that cannot be understood purely within classical wave optics. Elastic scattering leaves the internal specimen state unchanged and can preserve coherent interference. Inelastic scattering leaves the specimen in a different internal state and generally produces distinguishable quantum channels. This distinction is developed in Sec. 5.12.

Finally, the specimen is not necessarily static during imaging. Radiation damage, beam-induced motion, charging, and related processes are developed in Sec. 5.13. The important physical message is

$$\boxed{\text{the electron beam probes the specimen while simultaneously changing it.}} \tag{1.17}$$

Thus the exit wave need not be generated by one immutable specimen potential throughout an exposure. The structure, the local electrostatic environment, and the effective imaging conditions can all evolve while the movie is being recorded.

1.7 Detection: from electron probability to digital data

After the electron wave has left the specimen and passed through the electron-optical system, the experiment is still not finished.

At the detector plane, electron optics determines a probability flux. It does not directly produce a digital pixel value.

Chapter 6 follows the remaining transformation, which begins with the chain introduced in Eq. (6.1):

$$\boxed{\begin{gathered}\text{electron wave} \longrightarrow \text{probability flux} \longrightarrow \text{electron arrival} \\ \longrightarrow \text{interaction with detector material} \longrightarrow \text{charge generation and transport} \\ \longrightarrow \text{electronic readout} \longrightarrow \text{sampled digital value.}\end{gathered}} \tag{1.18}$$

This sequence introduces another important conceptual transition. The wave calculation determines the probability distribution from which electron arrivals are drawn. The arrivals themselves are discrete stochastic events.

Furthermore, a high-energy electron does not generally deposit its energy at one mathematical point in the detector. It produces a spatial track of interactions and a distributed charge response. Consequently, the position at which an electron entered the detector and the position assigned to the event by the readout system need not coincide exactly. The microscopic origin of this distinction is developed in Secs. 6.3 and 6.4.

The mean spatial response is characterized by the detector point-spread function and modulation transfer function in Sec. 6.5. The MTF therefore describes the preservation of mean spatial modulation.

Noise requires a separate description. Incident-electron shot noise, fluctuations in the response to individual electrons, electronic noise, gain variations, quantization, and other processes are developed in Sec. 6.6. Their spatial correlations are characterized through the noise power spectrum in Sec. 6.7.

This distinction is crucial. Detector noise is not, in general, equivalent to adding independent white Gaussian noise after a perfect image has already been produced.

The detective quantum efficiency introduced in Sec. 6.9 combines transfer and noise into an information-related measure:

$$\mathrm{DQE}(\mathbf{R}) = \frac{\mathrm{SNR}_{\mathrm{out}}(\mathbf{R})}{\mathrm{SNR}_{\mathrm{in}}(\mathbf{R})}, \tag{1.19}$$

as defined in Eq. (6.1283). The noise-equivalent quanta of Sec. 6.10 express the same information preservation in units of ideal incident electrons.

The chapter culminates by combining specimen signal, specimen background, electron-optical transfer, detector transfer, electron shot noise, detector-added fluctuations, and sampling into the complete recorded-image model in Sec. 6.13. Its power-spectrum form is derived in Eq. (6.1593).

The detector therefore cannot be regarded as a passive device appended to an otherwise finished physical image. It is itself part of the image formation process.

1.8 The forward model has several physical layers

We can now state more precisely what is meant by an *image-formation model*.

For one coherent realization of the experiment, it is useful schematically to imagine an initial electron state $\psi_0^{(\omega)}$ transformed according to

$$\boxed{\psi_0^{(\omega)} \xrightarrow{\mathcal{P}_{\mathrm{pre}}^{(\omega)}} \psi_{\mathrm{in}}^{(\omega)} \xrightarrow{\mathcal{S}^{(\omega)}} \psi_{\mathrm{exit}}^{(\omega)} \xrightarrow{\mathcal{P}_{\mathrm{post}}^{(\omega)}} \psi_{\mathrm{det}}^{(\omega)}.} \tag{1.20}$$

Here $\mathcal{P}_{\mathrm{pre}}$ represents propagation and illumination before the specimen, $\mathcal{S}$ represents the specimen interaction, and $\mathcal{P}_{\mathrm{post}}$ represents the post-specimen electron-optical system.

Equation (1.20) is deliberately schematic. It is not yet the complete experiment.

First, the source is only partially coherent, so different electron energies and illumination directions correspond to an ensemble of realizations rather than one unique wave. The statistical framework for this averaging is developed in Sec. 3.11.2.

Second, inelastic interaction can correlate the outgoing electron with different orthogonal specimen states. Those alternatives cannot in general be collapsed into one coherent scalar exit wave, as discussed in Sec. 5.12.

Third, the specimen itself may evolve with accumulated exposure, as developed in Sec. 5.13.

Fourth, detection converts probability flux into discrete random electron arrivals and subsequently into a fluctuating detector response, as developed in Chap. 6.

The complete experiment therefore contains at least four conceptually different levels:

I.	coherent propagation of complex amplitudes,
II.	statistical averaging over physical realizations and channels,
III.	stochastic arrival of a finite number of electrons,
IV.	stochastic detector response and digital sampling.

(1.21)

Many apparent confusions in image-formation theory arise when these levels are mixed.

A deterministic aberration phase, for example, is not a loss of coherence. An envelope caused by an ensemble of defocus values is not the same physical process as the detector MTF. Electron shot noise is not the same as electronic detector noise. An inelastic channel is not simply another coherent aberrated wave. Radiation damage cannot always be represented as one fixed blur applied after imaging.

One of the objectives of the detailed derivations in the following chapters is precisely to keep these mechanisms separate until a well-defined approximation allows them to be combined.

1.9 A hierarchy of approximations

The standard cryo-EM image model is useful because an enormously complicated physical experiment can, under suitable conditions, be reduced to a comparatively simple linear model. The word *reduced* is important.

The chapters that follow repeatedly begin from a more general physical description and identify the assumptions under which a simpler one emerges. The overall hierarchy is approximately

$$\boxed{\begin{array}{c}\text{relativistic electron dynamics}\\ \Downarrow\\ \text{stationary wave description}\\ \Downarrow\\ \text{scalar wave approximation}\\ \Downarrow\\ \text{paraxial propagation}\\ \Downarrow\\ \text{Gaussian electron optics plus aberrations}\\ \Downarrow\\ \text{phase-object and projection approximations}\\ \Downarrow\\ \text{weak-object linearization}\\ \Downarrow\\ \text{CTF-based image model}\\ \Downarrow\\ \text{linear presampling detector model}\\ \Downarrow\\ \text{sampled stochastic digital image.}\end{array}} \tag{1.22}$$

None of these reductions should be regarded as universally valid.

The scalar approximation neglects spin-dependent effects when they do not materially affect conventional image formation. The paraxial approx-

imation assumes sufficiently small propagation angles. Gaussian optics is only the leading-order behavior of real electron lenses, whose departures are described by aberrations. The projection approximation depends simultaneously on specimen thickness, wavelength, and the spatial frequency of interest, as developed in Sec. 5.9. The weak-phase-object approximation requires sufficiently small phase modulation. The usual fixed-specimen model becomes incomplete when radiation damage and motion during the exposure are important. Finally, linear detector transfer must be qualified when coincidence loss, saturation, dead time, sampling, and other nonlinearities become appreciable, as discussed in Sec. 6.11.

Our aim is therefore not to avoid approximations. A useful physical model necessarily contains them. The aim is to know *which* approximation has been made, *where* it entered the derivation, and *what physical effect has been discarded*.

1.10 Why quantum mechanics appears

Most of the image-formation theory developed in this book can be followed using wave optics, Fourier analysis, electromagnetism, mechanics, and probability. Nevertheless, there are several points at which a genuinely quantum-mechanical description is unavoidable.

An electron must be described by a quantum state. Its position-space wave function is a probability amplitude rather than a classical field. Momentum and position representations are Fourier duals. The electrostatic potential of the specimen acts through the electron Hamiltonian. Most importantly, inelastic scattering involves a transition of the specimen into a different internal quantum state. Whether different scattering alternatives interfere or add as probabilities depends on the distinguishability and orthogonality of those final states.

The required background is collected in Appendix A. That appendix is intentionally different in character from most of the main text. It introduces Hilbert spaces, bras and kets, operators, eigenstates, measurement probabilities, Hamiltonian evolution, matrix elements, and joint electron–specimen states. These concepts are mathematically more abstract than the scalar Fourier-optical framework used through much of the main image-formation derivation. It is not necessary to master the entire appendix before beginning the book.

Readers already familiar with quantum mechanics may use it mainly as a reference for notation. Readers less familiar with the subject may

read it selectively when the corresponding concepts first become necessary. The Hilbert-space interpretation is introduced in Sec. A.1; wave functions, Fourier representations, and momentum are developed in Sec. A.5; the Hamiltonian is introduced in Sec. A.10; matrix elements and transitions are developed in Sec. A.12; and the reason orthogonal inelastic specimen channels add as probabilities is developed in Sec. A.13. A compact collection of the notation and principal results is provided in Sec. A.15.

Two places in the main text make particularly explicit use of this background. Section 3.4 connects the relativistic Dirac equation to the scalar Helmholtz description used in wave diffraction, while Sec. 4.5 connects the phase of the quantum electron wave with classical action and geometrical electron optics. The quantum description becomes important again in Sec. 5.2 and Sec. 5.12, where the imaging electron and the specimen must be treated as parts of a joint quantum system.

The appendix should therefore be regarded as supporting machinery. It allows the deeper steps of the image-formation model to be understood without requiring the entire book to become a general textbook on quantum mechanics.

1.11 How the pieces fit together

The preceding discussion can be summarized by identifying the principal mathematical object contributed by each part of the book.

At the end of the complete development, the structural information recorded in a cryo-EM image has therefore passed through a hierarchy that

Table 1.1: Principal contribution of each chapter to the cryo-EM image-formation model.

Part	**Principal object**	**Role in image formation**
Chap. 2	λ, $p(E)$, source size, angular distribution	Establishes the energy, wavelength, intensity, and statistical properties of the incident beam.
Chap. 3	propagation operators and Fourier representation	Establishes diffraction, angular-spectrum propagation, Fresnel propagation, lenses, Fourier planes, and partial coherence.
Chap. 4	$P(\mathbf{R})$, $\chi(\mathbf{R})$, $E_{\text{tot}}(\mathbf{R})$, CTF	Reduces electromagnetic electron-lens physics to coherent transfer, aberration phase, envelopes, and image contrast.
Chap. 5	$\mathcal{S}$ and ψ_{exit}	Converts the three-dimensional specimen into elastic and inelastic outgoing electron channels and establishes the approximations leading to the standard SPA model.
Chap. 6	H_{det}, MTF, NPS, DQE, NEQ	Converts electron probability flux into stochastic, spatially transferred, sampled digital measurements.
Appendix A	$\lvert\psi\rangle$, $\hat{H}$, matrix elements, quantum channels	Provides the quantum-mechanical foundations required for electron wave functions, transitions, and elastic versus inelastic scattering.

can be represented schematically as

$$
\boxed{
\begin{array}{cc}
\multicolumn{2}{c}{\text{atomic and molecular structure}} \\
\multicolumn{2}{c}{\Downarrow} \\
V(\mathbf{r}) & \text{specimen electrostatic potential} \\
\multicolumn{2}{c}{\Downarrow} \\
\mathcal{S} & \text{electron–specimen scattering} \\
\multicolumn{2}{c}{\Downarrow} \\
\psi_{\text{exit}} & \text{exit wave and scattering channels} \\
\multicolumn{2}{c}{\Downarrow} \\
P(\mathbf{R})e^{-j\chi(\mathbf{R})} & \text{coherent electron-optical transfer} \\
\multicolumn{2}{c}{\Downarrow} \\
\mathrm{CTF}(\mathbf{R}) & \text{intensity contrast transfer} \\
\multicolumn{2}{c}{\Downarrow} \\
H_{\text{det}}(\mathbf{R}) & \text{detector spatial transfer} \\
\multicolumn{2}{c}{\Downarrow} \\
\multicolumn{2}{c}{\text{stochastic electron and detector processes}} \\
\multicolumn{2}{c}{\Downarrow} \\
\multicolumn{2}{c}{\text{sampling}} \\
\multicolumn{2}{c}{\Downarrow} \\
y_i & \text{recorded cryo-EM image.}
\end{array}
}
\tag{1.23}
$$

Equation (1.23) is the roadmap for the rest of the book.

1.12 The purpose of understanding the forward model

Cryo-EM structure determination is ultimately an inverse problem. We observe images and wish to infer the biological structure that generated them. The quality of that inference depends on how well we understand the forward process.

A forward model tells us which features of an image arise from the specimen, which arise from electron optics, which are attenuated by partial coherence, which are modified by the detector, and which are stochastic. It tells us why some spatial frequencies are transferred with one sign and others

with the opposite sign, why CTF zeros occur, why aberrations alter phases, why finite source size and energy spread produce envelopes, why the projection approximation eventually fails, why radiation damage is frequency dependent, why motion must be treated through the exposure, and why detector noise need not be white. It also tells us what information has been lost.

The inverse problem cannot recover information that the forward experiment never transferred, cannot uniquely distinguish physical effects that produce identical observations without additional information, and cannot correctly weight measurements unless their statistics are understood.

For this reason, the emphasis of the book is deliberately on derivation. We repeatedly begin from a more fundamental physical description and work toward the model used in data analysis. Some derivations are longer than would be required merely to state the final formula. Their purpose is to expose the assumptions and to make visible the connections among quantities that are often introduced separately.

By the end of the book, the familiar cryo-EM image will therefore have been decomposed into its physical origins:

$$\boxed{\begin{array}{c}\text{source statistics}\\ +\\ \text{wave propagation}\\ +\\ \text{electron-lens fields and aberrations}\\ +\\ \text{elastic and inelastic specimen interaction}\\ +\\ \text{radiation-induced specimen evolution}\\ +\\ \text{stochastic electron detection}\\ +\\ \text{detector transfer, noise, and sampling.}\end{array}} \tag{1.24}$$

The aim is not to make the image appear more complicated than it is. It is to understand why, under the conditions normally achieved in a cryo-electron microscope, this complicated physical chain can so often be reduced to a tractable mathematical model—and to know when it cannot.

That understanding is the foundation on which the remainder of the book is built.

Chapter 2

Electron sources

The electron source determines the initial physical properties of the beam that propagates through the microscope. Some of these properties, such as the mean electron energy, are imposed primarily by the accelerating potential. Others, including the energy spread, effective source size, brightness, angular distribution, and stability, depend strongly on the mechanism by which the electrons are emitted.

These properties will appear repeatedly in the image-formation theory developed later in the book. The accelerating potential determines the electron momentum and wavelength; the finite energy distribution contributes to temporal partial coherence and chromatic aberration; and the finite effective size of the source contributes to spatial partial coherence. The rigorous statistical description of these two forms of partial coherence will be developed in Sec. 3.11.2, and their relationship to energy and illumination angle is summarized in Eq. (3.753).

The role of an electron gun is therefore not simply to provide electrons. It must produce a sufficiently intense beam while simultaneously controlling its phase-space extent and its energy distribution. Modern cryo-electron microscopes achieve this using field-emission sources, but it is useful to begin with the physical mechanisms by which electrons can leave a solid.

2.1 Electron emission from a solid

Inside a metal, conduction electrons occupy quantum states up to the Fermi energy. An electron close to the surface cannot leave the material freely, because the electrostatic interaction with the solid produces an energy barrier between the occupied states inside the material and a free electron in

vacuum.

The minimum energy required to remove an electron at the Fermi level from the solid into vacuum is the *work function*, which we denote by φ. Thus,

$$\varphi = E_{\text{vac}} - E_F, \tag{2.1}$$

where E_F is the Fermi energy and E_{vac} is the vacuum energy immediately outside the surface.

Typical work functions are of the order of a few electronvolts. This is enormous compared with the thermal energy $k_B T$ at room temperature, which explains why an ordinary metal surface does not spontaneously emit a useful electron current.

There are two main ways of overcoming this barrier. The first is to heat the emitter until a significant fraction of the electrons have sufficient thermal energy to pass over the surface barrier. This is *thermionic emission*. The second is to apply a sufficiently intense electric field to make the surface barrier narrow enough for electrons to tunnel quantum mechanically through it. This is *field emission*. Between these limiting regimes lies thermally assisted field emission, which is the principle exploited by the Schottky field-emission gun.

2.1.1 Thermionic emission

Consider a planar metallic surface normal to the z direction. Electrons inside the solid possess a distribution of momenta. An electron can escape only if the component of its energy associated with motion normal to the surface is sufficiently large to cross the vacuum barrier.

For a free-electron model of the metal, the thermionic current density is described by the Richardson–Dushman equation,

$$\boxed{J_{\text{th}} = A_R T^2 \exp\left(-\frac{\varphi}{k_B T}\right),} \tag{2.2}$$

where T is the emitter temperature and

$$\boxed{A_R = \frac{4\pi m_0 e k_B^2}{h^3}} \tag{2.3}$$

is the Richardson constant in the ideal free-electron model. In real materials, surface structure and electronic band structure modify the effective value of this constant.

The most important feature of Eq. (2.2) is its exponential dependence on the ratio $\varphi/(k_B T)$. A moderate increase in emitter temperature can therefore produce a very large increase in emission current.

Richardson–Dushman equation.

In the free-electron approximation, the number of electronic states per unit volume in the momentum-space element $dp_x\, dp_y\, dp_z$, including the two spin states, is

$$\frac{2}{h^3} dp_x\, dp_y\, dp_z. \tag{2.4}$$

The occupation probability is the Fermi–Dirac distribution

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/(k_B T)]}. \tag{2.5}$$

Only the very high-energy tail of this distribution contributes to thermionic emission. In this region,

$$f(E) \simeq \exp\left(-\frac{E - E_F}{k_B T}\right). \tag{2.6}$$

For a planar surface, the parallel momentum is conserved on crossing the surface. The condition for escape is therefore that the normal kinetic energy satisfies

$$\frac{p_z^2}{2m_0} > E_F + \varphi. \tag{2.7}$$

Define

$$p_{\min} = \sqrt{2m_0(E_F + \varphi)}. \tag{2.8}$$

The contribution to the current density from an occupied state is its charge multiplied by its normal velocity, $v_z = p_z/m_0$. Hence,

$$J_{\text{th}} = \frac{2e}{h^3} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{p_{\min}}^{\infty} \frac{p_z}{m_0} \exp\left[-\frac{p_x^2 + p_y^2 + p_z^2}{2m_0 k_B T} + \frac{E_F}{k_B T}\right] dp_z\, dp_y\, dp_x. \tag{2.9}$$

The two transverse Gaussian integrals give

$$\int_{-\infty}^{\infty} \exp\left(-\frac{p_x^2}{2m_0 k_B T}\right) dp_x = \sqrt{2\pi m_0 k_B T}, \tag{2.10}$$

and similarly for p_y. Their product is therefore

$$2\pi m_0 k_B T. \tag{2.11}$$

The remaining integral is

$$\int_{p_{\min}}^{\infty} \frac{p_z}{m_0} \exp\left(-\frac{p_z^2}{2m_0 k_B T}\right) dp_z = k_B T \exp\left(-\frac{p_{\min}^2}{2m_0 k_B T}\right) \tag{2.12}$$

$$= k_B T \exp\left(-\frac{E_F + \varphi}{k_B T}\right). \tag{2.13}$$

Multiplication by the factor $\exp(E_F/k_B T)$ from the occupation probability cancels the Fermi energy. Consequently,

$$J_{\text{th}} = \frac{2e}{h^3}(2\pi m_0 k_B T)(k_B T) \exp\left(-\frac{\varphi}{k_B T}\right) \tag{2.14}$$

$$= \frac{4\pi m_0 e k_B^2}{h^3} T^2 \exp\left(-\frac{\varphi}{k_B T}\right), \tag{2.15}$$

which gives Eqs. (2.2) and (2.3).

2.1.2 Lowering of the surface barrier by an electric field

An external electric field modifies the surface barrier. Let F denote the magnitude of an electric field directed so as to extract electrons from the surface. At a distance x outside an ideal planar conductor, the potential-energy barrier can be approximated by

$$U(x) = \varphi - eFx - \frac{e^2}{16\pi\varepsilon_0 x}. \tag{2.16}$$

The second term is the energy change produced by the external field. The last term is the attractive image-charge potential generated by the conducting surface.

The resulting barrier maximum is lower than the zero-field vacuum level

by

$$\boxed{\Delta\varphi_S = \sqrt{\frac{e^3 F}{4\pi\varepsilon_0}}.} \tag{2.17}$$

Thus the effective barrier becomes

$$\varphi_{\text{eff}} = \varphi - \Delta\varphi_S. \tag{2.18}$$

Schottky barrier lowering.

The maximum of the barrier in Eq. (2.16) satisfies

$$\frac{dU}{dx} = -eF + \frac{e^2}{16\pi\varepsilon_0 x^2} = 0. \tag{2.19}$$

Therefore its position is

$$x_m = \sqrt{\frac{e}{16\pi\varepsilon_0 F}}. \tag{2.20}$$

The reduction of the barrier relative to φ is

$$\Delta\varphi_S = eFx_m + \frac{e^2}{16\pi\varepsilon_0 x_m}. \tag{2.21}$$

At the maximum, the two terms are equal, because the stationary condition implies

$$eF = \frac{e^2}{16\pi\varepsilon_0 x_m^2}. \tag{2.22}$$

Hence,

$$\Delta\varphi_S = 2eFx_m \tag{2.23}$$

$$= 2eF\sqrt{\frac{e}{16\pi\varepsilon_0 F}} \tag{2.24}$$

$$= \sqrt{\frac{e^3 F}{4\pi\varepsilon_0}}, \tag{2.25}$$

which proves Eq. (2.17).

If thermal excitation remains the dominant mechanism, the principal effect of the field can be incorporated by replacing φ in Eq. (2.2) with the

lowered barrier. This gives the Richardson–Schottky form

$$\boxed{J_{\mathrm{RS}} = A_R T^2 \exp\left[-\frac{\varphi - \Delta\varphi_S}{k_B T}\right].} \tag{2.26}$$

A Schottky field-emission gun operates in a regime in which both high temperature and strong electric field are important; a more exact treatment also includes quantum tunnelling through the remaining barrier.

2.1.3 Cold field emission

If the extraction field is sufficiently strong, electrons near the Fermi level can tunnel through the surface barrier even when their thermal energy is far below the work function. This is the basis of a cold field-emission gun.

To expose the essential physics, first neglect the image-charge correction and approximate the vacuum barrier by a triangle,

$$U(x) = \varphi - eFx. \tag{2.27}$$

The classical turning point is then

$$x_t = \frac{\varphi}{eF}. \tag{2.28}$$

Within the WKB approximation, the tunnelling probability for an electron at the Fermi level is

$$\boxed{D_F \simeq \exp\left[-\frac{4\sqrt{2m_0}\,\varphi^{3/2}}{3e\hbar F}\right].} \tag{2.29}$$

The exponential dependence on $1/F$ explains why field emission becomes appreciable only for the extremely large local electric fields produced near a very sharp metallic tip.

Integrating over the electrons available near the Fermi surface gives, for an ideal triangular barrier at zero temperature,

$$\boxed{J_{\mathrm{FN}} \simeq \frac{e^3 F^2}{16\pi^2 \hbar\varphi} \exp\left[-\frac{4\sqrt{2m_0}\,\varphi^{3/2}}{3e\hbar F}\right].} \tag{2.30}$$

This is the simplest Fowler–Nordheim expression. In a quantitative description of a real field emitter, the image-charge term in Eq. (2.16), the detailed geometry of the tip, finite temperature, and the electronic structure of the emitting surface modify both the exponent and the prefactor.

Tunnelling exponent and Fowler–Nordheim current.

For a one-dimensional barrier $U(x)$, the WKB transmission probability is approximately

$$D \simeq \exp\left[-\frac{2}{\hbar}\int_{x_1}^{x_2}\sqrt{2m_0\left[U(x)-E\right]}\,dx\right]. \tag{2.31}$$

For an electron at the Fermi level, take $E = 0$ as the reference and use $U(x) = \varphi - eFx$. The turning points are $x_1 = 0$ and $x_2 = \varphi/(eF)$. Hence,

$$-\ln D_F = \frac{2}{\hbar}\int_0^{\varphi/(eF)}\sqrt{2m_0(\varphi - eFx)}\,dx. \tag{2.32}$$

With $u = \varphi - eFx$,

$$-\ln D_F = \frac{2\sqrt{2m_0}}{\hbar eF}\int_0^{\varphi} u^{1/2}\,du \tag{2.33}$$

$$= \frac{2\sqrt{2m_0}}{\hbar eF}\frac{2}{3}\varphi^{3/2} \tag{2.34}$$

$$= \frac{4\sqrt{2m_0}\varphi^{3/2}}{3e\hbar F}, \tag{2.35}$$

which gives Eq. (2.29).

To obtain the current density, let E_z denote the kinetic energy associated with motion normal to the surface. At zero temperature, the supply of electrons with a given E_z is proportional to $E_F - E_z$. The emitted current can be written

$$J = \frac{4\pi m_0 e}{h^3}\int_0^{E_F}(E_F - E_z)D(E_z)\,dE_z. \tag{2.36}$$

Introduce

$$\xi = E_F - E_z. \tag{2.37}$$

An electron a distance ξ below the Fermi level sees an effective barrier height $\varphi + \xi$. For the triangular barrier,

$$D(\xi) \simeq \exp\left[-\frac{4\sqrt{2m_0}}{3e\hbar F}(\varphi+\xi)^{3/2}\right]. \tag{2.38}$$

Because the exponential strongly restricts the contribution to $\xi \ll \varphi$, we linearize

$$(\varphi + \xi)^{3/2} \simeq \varphi^{3/2} + \frac{3}{2}\varphi^{1/2}\xi. \tag{2.39}$$

Thus,

$$D(\xi) \simeq D_F \exp(-a\xi), \tag{2.40}$$

with

$$a = \frac{2\sqrt{2m_0\varphi}}{e\hbar F}. \tag{2.41}$$

The upper integration limit can now be extended to infinity because of the rapid exponential decay:

$$J \simeq \frac{4\pi m_0 e}{h^3} D_F \int_0^\infty \xi e^{-a\xi}\, d\xi \tag{2.42}$$

$$= \frac{4\pi m_0 e}{h^3} D_F \frac{1}{a^2}. \tag{2.43}$$

Since

$$\frac{1}{a^2} = \frac{e^2\hbar^2 F^2}{8m_0\varphi} \tag{2.44}$$

and $h = 2\pi\hbar$, we obtain

$$J = \frac{e^3 F^2}{16\pi^2\hbar\varphi} D_F, \tag{2.45}$$

which, after substitution of Eq. (2.29), gives Eq. (2.30).

2.2 Electron-gun configurations

The emission mechanism does not by itself define an electron source. Electrodes surrounding the emitter establish the extraction field, collect the emitted electrons, and form the first beam crossover or virtual source. The electron gun therefore combines an emitter with an electrostatic system that extracts and accelerates the electrons into the microscope column.

The source must simultaneously provide sufficient current, small effective source size, small energy spread, high stability, and long operating lifetime. Different source technologies represent different compromises between these requirements.

2.2.1 Thermionic guns

A thermionic gun uses a heated emitter. Traditional transmission electron microscopes employ either a tungsten filament or a lanthanum hexaboride (LaB6) crystal. Heating raises the high-energy tail of the electronic distribution and produces the thermionic current described by Eq. (2.2).

The emitter is surrounded by an electrode conventionally called the *Wehnelt cylinder*, whose potential shapes the electrostatic field near the cathode. Together with the anode, it extracts the electrons and forms a crossover below the gun. The crossover acts as the effective source for the condenser system.

Thermionic emitters are mechanically simple and robust, but their emitting area is relatively large. Consequently their effective source size is larger and their brightness is substantially lower than that of modern field-emission sources. Their thermal energy distribution is also broader.

2.2.2 Schottky field-emission guns

A Schottky field-emission gun combines heating with a strong extraction field. The emitter is a sharply pointed crystalline tip, so that the local electric field at the apex can be extremely large even for an experimentally manageable extraction voltage.

The electric field lowers and narrows the surface barrier, as described in Sec. 2.1.2. Heating then supplies electrons close enough to the top of the barrier that emission becomes very efficient. Thus a Schottky emitter operates between the purely thermionic regime of Sec. 2.1.1 and the cold tunnelling regime of Sec. 2.1.3.

Because emission occurs from a very small region near the tip apex, the effective source is much smaller than in a thermionic gun. Schottky FEGs therefore provide high brightness and good spatial coherence. Heating also helps maintain a comparatively clean and stable emitting surface, which is important for the long acquisition sessions used in cryo-electron microscopy.

The emitted electrons nevertheless have a finite energy distribution. This source energy spread will later enter the temporal-coherence theory developed in Sec. 3.11.4.

2.2.3 Cold field-emission guns

A cold field-emission gun operates with little or no continuous heating of the emitting tip. The extraction field is sufficiently intense that electrons close

to the Fermi level tunnel directly through the surface barrier according to the mechanism described in Sec. 2.1.3.

Because the occupied states contributing to the current lie close to the Fermi level, the intrinsic energy distribution can be narrower than for a thermally assisted source. The emitting region is also extremely small. Cold FEGs therefore provide particularly high brightness, small effective source size, and a narrow energy distribution.

The price is a much greater sensitivity to the condition of the emitter surface. Adsorbed molecules modify the local work function and hence the emission current, whose exponential dependence on the field is evident in Eq. (2.30). Cold field-emission guns consequently require very high vacuum in the gun region and periodic cleaning or *flashing* of the tip.

For high-resolution cryo-electron microscopy, the narrow energy distribution of a cold FEG is especially attractive because it reduces one of the contributions to temporal partial coherence. However, the final performance depends on the complete microscope, not on the source alone: accelerating-voltage fluctuations, lens instabilities, aberrations, and specimen motion can all dominate once the intrinsic source properties have been sufficiently improved.

2.2.4 Comparison of the principal electron sources

The qualitative differences among the most common source types can be summarized as follows:

Source	Emission mechanism	Effective source size	Brightness	Energy spread
Tungsten thermionic	Thermal	Large	Low	Relatively large
LaB6 thermionic	Thermal	Intermediate	Intermediate	Relatively large
Schottky FEG	Thermal + field	Small	High	Small
Cold FEG	Tunnelling	Very small	Very high	Very small

These entries should be understood as trends rather than universal numerical specifications. The actual beam characteristics depend on the emitter material, tip geometry, extraction field, operating temperature, gun optics, vacuum conditions, and the definition used for quantities such as source size and energy width.

2.3 Relativistic electron kinematics in electromagnetic fields

Once an electron has been emitted, its subsequent acceleration and motion are governed by the electromagnetic fields of the electron gun and the microscope column. The relativistic kinematic relations introduced here will be used repeatedly in the electron-optics chapters. The treatment follows the formulation of Hawkes and Kasper (2017, Chap. 2), while using the momentum notation adopted elsewhere in this book.

We denote by m_0 the electron rest mass, by $-e$ its electric charge, with $e > 0$, and by

$$\mathbf{p} = \gamma m_0 \mathbf{v} \tag{2.46}$$

its relativistic kinetic momentum. Here

$$\beta \triangleq \frac{v}{c}, \qquad \gamma \triangleq \frac{1}{\sqrt{1-\beta^2}}, \tag{2.47}$$

where $v = \|\mathbf{v}\|$ and c is the speed of light in vacuum.

The trajectory is governed by the Lorentz equation

$$\boxed{\frac{d\mathbf{p}}{dt} = -e\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right).} \tag{2.48}$$

The momentum in Eq. (2.48) is the *kinetic* momentum. It should not be confused with the canonical momentum that will appear later in the variational and Hamiltonian formulations of electron optics.

Equation (2.48) is valid for electric and magnetic fields that depend on both position and time. Most of the externally applied fields in a transmission electron microscope can, however, be treated as static. This immediately leads to an important conservation law.

2.3.1 Conservation of energy in static fields

Let

$$\mathbf{E} = \mathbf{E}(\mathbf{r}), \qquad \mathbf{B} = \mathbf{B}(\mathbf{r}), \tag{2.49}$$

and introduce the electrostatic scalar potential through

$$\boxed{\mathbf{E}(\mathbf{r}) = -\nabla\Phi(\mathbf{r}).} \tag{2.50}$$

The total relativistic energy of a free electron is

$$E = \gamma m_0 c^2, \tag{2.51}$$

and its kinetic energy is therefore

$$\boxed{K = m_0 c^2(\gamma - 1).} \tag{2.52}$$

For static electromagnetic fields,

$$\boxed{K(\mathbf{r}) - e\Phi(\mathbf{r}) = \mathcal{E}_0 = \text{constant}} \tag{2.53}$$

along every electron trajectory.

Conservation of electron energy.

Take the scalar product of the Lorentz equation, Eq. (2.48), with $\mathbf{v}$:

$$\mathbf{v} \cdot \frac{d\mathbf{p}}{dt} = -e\mathbf{v} \cdot \mathbf{E} - e\mathbf{v} \cdot (\mathbf{v} \times \mathbf{B}). \tag{2.54}$$

The magnetic contribution vanishes identically,

$$\mathbf{v} \cdot (\mathbf{v} \times \mathbf{B}) = 0. \tag{2.55}$$

Thus a magnetic field can change the direction of the electron momentum, but it performs no work on the electron.

For the left-hand side,

$$\mathbf{v} \cdot \frac{d\mathbf{p}}{dt} = \frac{d}{dt}\left(\gamma m_0 c^2\right) = \frac{dK}{dt}, \tag{2.56}$$

because the rest energy $m_0 c^2$ is constant.

Using Eq. (2.50),

$$-e\mathbf{v} \cdot \mathbf{E} = e\mathbf{v} \cdot \nabla\Phi \tag{2.57}$$

$$= e\frac{d\Phi}{dt}, \tag{2.58}$$

where the last equality follows because Φ has no explicit time dependence. Therefore,

$$\frac{dK}{dt} = e\frac{d\Phi}{dt}, \tag{2.59}$$

or

$$\frac{d}{dt}\left(K - e\Phi\right) = 0. \tag{2.60}$$

Integration gives Eq. (2.53).

The kinetic energy in Eq. (2.52) has the low-velocity expansion

$$\boxed{K = \frac{1}{2}m_0 v^2 \left(1 + \frac{3}{4}\beta^2 + \frac{5}{8}\beta^4 + \cdots\right).} \tag{2.61}$$

Thus the familiar classical kinetic energy is recovered as the leading term when $\beta \ll 1$.

Low-velocity expansion of the kinetic energy.

Using Eq. (2.47),

$$\gamma = (1-\beta^2)^{-1/2} = 1 + \frac{1}{2}\beta^2 + \frac{3}{8}\beta^4 + \frac{5}{16}\beta^6 + \cdots. \tag{2.62}$$

Substitution into Eq. (2.52) gives

$$K = m_0 c^2 \left(\frac{1}{2}\beta^2 + \frac{3}{8}\beta^4 + \frac{5}{16}\beta^6 + \cdots\right) \tag{2.63}$$

$$= \frac{1}{2}m_0 v^2 \left(1 + \frac{3}{4}\beta^2 + \frac{5}{8}\beta^4 + \cdots\right), \tag{2.64}$$

which proves Eq. (2.61).

2.3.2 The acceleration potential

The electrostatic potential is defined only up to an additive constant. In electron optics it is convenient to choose the cathode surface as the zero of potential,

$$\Phi = 0 \qquad \text{at the cathode.} \tag{2.65}$$

Let K_0 denote the kinetic energy with which a particular electron leaves the cathode. Equation (2.53) then gives

$$\boxed{K(\mathbf{r}, K_0) = K_0 + e\Phi(\mathbf{r}).} \tag{2.66}$$

Retaining K_0 is important. Real electrons are emitted with a distribution of initial kinetic energies, and this distribution contributes to the final

energy spread of the beam. Setting $K_0 = 0$ is appropriate only when these chromatic effects are not under consideration.

Using Eq. (2.52), the local Lorentz factor is

$$\boxed{\gamma(\mathbf{r}, K_0) = 1 + \frac{K_0 + e\Phi(\mathbf{r})}{m_0 c^2}.} \tag{2.67}$$

The relativistic energy–momentum relation is

$$\boxed{E^2 = p^2c^2 + m_0^2c^4.} \tag{2.68}$$

Relativistic energy–momentum relation.

From $E = \gamma m_0 c^2$ and $p = \gamma m_0 v$,

$$E^2 - p^2c^2 = \gamma^2 m_0^2 c^4 - \gamma^2 m_0^2 v^2 c^2 \tag{2.69}$$
$$= \gamma^2 m_0^2 c^4 (1 - \beta^2). \tag{2.70}$$

Since $\gamma^2(1 - \beta^2) = 1$ by Eq. (2.47),

$$E^2 - p^2c^2 = m_0^2c^4, \tag{2.71}$$

which gives Eq. (2.68).

Since $E = m_0c^2 + K$, Eq. (2.68) gives the magnitude of the kinetic momentum as

$$\boxed{p(\mathbf{r}, K_0) = \sqrt{2m_0\left(K_0 + e\Phi(\mathbf{r})\right)\left[1 + \frac{K_0 + e\Phi(\mathbf{r})}{2m_0c^2}\right]}.} \tag{2.72}$$

Derivation.

Substituting $E = m_0c^2 + K$ into Eq. (2.68),

$$p^2c^2 = (m_0c^2 + K)^2 - m_0^2c^4 \tag{2.73}$$
$$= 2m_0c^2K + K^2. \tag{2.74}$$

Hence,

$$p = \sqrt{2m_0 K \left(1 + \frac{K}{2m_0 c^2}\right)}. \tag{2.75}$$

Using Eq. (2.66) gives Eq. (2.72).

Following the standard electron-optical notation, define

$$\begin{aligned} \Phi_0 &\triangleq K_0/e, && [\mathrm{V}], \\ \epsilon &\triangleq \frac{e}{2m_0 c^2} \simeq 0.9785 \times 10^{-6}, && [\mathrm{V}^{-1}], \\ \eta &\triangleq \sqrt{\frac{e}{2m_0}} \simeq 2.965 \times 10^5, && [\mathrm{C}^{1/2}\mathrm{kg}^{-1/2}]. \end{aligned} \tag{2.76}$$

We then introduce the *relativistic acceleration potential*

$$\boxed{\widehat{\Phi}(\mathbf{r}, \Phi_0) \triangleq [\Phi_0 + \Phi(\mathbf{r})] \{1 + \epsilon [\Phi_0 + \Phi(\mathbf{r})]\}.} \tag{2.77}$$

Since $K = e[\Phi_0 + \Phi(\mathbf{r})] \geq 0$ along an accessible electron trajectory, $\widehat{\Phi} \geq 0$.

All the principal scalar kinematic quantities can now be written in terms of this single function:

$$\boxed{\begin{aligned} p &= \sqrt{2m_0 e \widehat{\Phi}}, \\ \gamma &= \sqrt{1 + 4\epsilon \widehat{\Phi}}, \\ K &= \frac{2e\widehat{\Phi}}{1+\gamma}, \\ v &= \frac{2\eta}{\gamma}\sqrt{\widehat{\Phi}}. \end{aligned}} \tag{2.78}$$

Derivation.

Define

$$u = \Phi_0 + \Phi(\mathbf{r}). \tag{2.79}$$

Then $K = eu$ and Eq. (2.77) gives

$$\widehat{\Phi} = u(1 + \epsilon u). \tag{2.80}$$

Equation (2.72) therefore becomes

$$p = \sqrt{2m_0 e\widehat{\Phi}}. \tag{2.81}$$

From Eq. (2.67),

$$\gamma = 1 + 2\epsilon u. \tag{2.82}$$

Furthermore,

$$\begin{aligned} 1 + 4\epsilon\widehat{\Phi} &= 1 + 4\epsilon u(1 + \epsilon u) && (2.83)\\ &= (1 + 2\epsilon u)^2 && (2.84)\\ &= \gamma^2, && (2.85) \end{aligned}$$

and therefore

$$\gamma = \sqrt{1 + 4\epsilon\widehat{\Phi}}. \tag{2.86}$$

Since

$$2\widehat{\Phi} = u(1 + \gamma), \tag{2.87}$$

we have

$$K = eu = \frac{2e\widehat{\Phi}}{1 + \gamma}. \tag{2.88}$$

Finally,

$$\begin{aligned} v &= \frac{p}{\gamma m_0} && (2.89)\\ &= \frac{\sqrt{2m_0 e\widehat{\Phi}}}{\gamma m_0} && (2.90)\\ &= \frac{2\eta}{\gamma}\sqrt{\widehat{\Phi}}, && (2.91) \end{aligned}$$

which completes the derivation of Eq. (2.78).

Even the electric field can be written in terms of the acceleration potential:

$$\boxed{\mathbf{E} = -\frac{1}{\gamma}\nabla\widehat{\Phi}.} \tag{2.92}$$

Derivation.

Differentiation of Eq. (2.77) gives

$$\nabla\widehat{\Phi} = [1 + 2\epsilon(\Phi_0 + \Phi)]\,\nabla\Phi \tag{2.93}$$
$$= \gamma\nabla\Phi, \tag{2.94}$$

where Eq. (2.67) has been used. Since $\mathbf{E} = -\nabla\Phi$ from Eq. (2.50),

$$\mathbf{E} = -\frac{1}{\gamma}\nabla\widehat{\Phi}, \tag{2.95}$$

which proves Eq. (2.92).

The importance of $\widehat{\Phi}$ will become clearer in the electron optics chapters. Once the electrostatic potential and the initial emission energy are specified, momentum, velocity, kinetic energy, and relativistic factor are all determined by a single scalar function. The trajectory equations themselves, however, belong to electron optics rather than to the description of the source and will be derived there from Eq. (2.48).

2.4 Electron energy, velocity, and wavelength

We now specialize the general results of Sec. 2.3 to the nearly uniform final accelerating potential of a transmission electron microscope.

Let $V > 0$ denote the potential difference between the cathode and the region after acceleration. If the initial kinetic energy is neglected, $K_0 = 0$, Eq. (2.66) gives

$$\boxed{K = eV.} \tag{2.96}$$

If the initial emission energy is retained instead,

$$\boxed{K = K_0 + eV.} \tag{2.97}$$

Equation (2.97) is important when the finite energy width of the source is considered: ideal electrostatic acceleration adds essentially the same energy eV to every electron and therefore does not remove the initial energy distribution.

The total relativistic energy after acceleration from rest is

$$\boxed{E = m_0c^2 + eV.} \tag{2.98}$$

2.4.1 Velocity of an accelerated electron

From Eq. (2.67), with $K_0 = 0$ and $\Phi = V$,

$$\gamma = 1 + \frac{eV}{m_0 c^2}. \tag{2.99}$$

Combining this with Eq. (2.47) gives

$$\boxed{v(V) = c\sqrt{1 - \left(1 + \frac{eV}{m_0 c^2}\right)^{-2}}.} \tag{2.100}$$

No additional dynamical assumption is involved in this equation: it is simply the specialization of the general kinematic relations derived in Sec. 2.3.

At the voltages commonly used in transmission electron microscopy, the electron velocity is a substantial fraction of the speed of light. Relativistic mechanics is therefore essential.

2.4.2 Relativistic electron wavelength

Wave mechanics associates the kinetic momentum with the de Broglie wavelength

$$\boxed{\lambda = \frac{h}{p},} \tag{2.101}$$

where h is Planck's constant.

Using Eq. (2.68) with Eq. (2.98), the momentum of an electron accelerated from rest through V is

$$p = \frac{1}{c}\sqrt{eV\,(2m_0 c^2 + eV)}. \tag{2.102}$$

Consequently,

$$\boxed{\lambda(V) = \frac{hc}{\sqrt{eV\,(2m_0 c^2 + eV)}}.} \tag{2.103}$$

Equivalently,

$$\boxed{\lambda(V) = \frac{h}{\sqrt{2m_0 eV\left(1 + \frac{eV}{2m_0 c^2}\right)}}.} \tag{2.104}$$

Derivation.

Substitution of Eq. (2.98) into Eq. (2.68) gives

$$(m_0c^2 + eV)^2 = p^2c^2 + m_0^2c^4. \tag{2.105}$$

After cancellation of the rest-energy term,

$$p^2c^2 = eV(2m_0c^2 + eV), \tag{2.106}$$

and therefore

$$p = \frac{1}{c}\sqrt{eV(2m_0c^2 + eV)}. \tag{2.107}$$

Using Eq. (2.101) gives Eq. (2.103).

Factoring $2m_0c^2$ inside the square root gives

$$\lambda = \frac{hc}{\sqrt{2m_0c^2eV\left(1 + \frac{eV}{2m_0c^2}\right)}} \tag{2.108}$$

$$= \frac{h}{\sqrt{2m_0eV\left(1 + \frac{eV}{2m_0c^2}\right)}}, \tag{2.109}$$

which is Eq. (2.104).

2.4.3 Non-relativistic limit

If

$$eV \ll m_0c^2, \tag{2.110}$$

then the relativistic correction in Eq. (2.104) approaches unity and

$$\boxed{\lambda_{\mathrm{NR}} \simeq \frac{h}{\sqrt{2m_0eV}}.} \tag{2.111}$$

This is the familiar result obtained from $K = p^2/(2m_0)$.

The same small-energy condition appears from a different perspective in Sec. 3.4.2, where the relativistic wave equation for the electron is reduced to the Schrödinger equation. Modern cryo-EM does not satisfy this condition particularly well. Since

$$m_0c^2 \simeq 511 \text{ keV}, \tag{2.112}$$

a 300 kV electron has

$$\frac{eV}{m_0c^2} \simeq \frac{300}{511} \simeq 0.59. \tag{2.113}$$

The relativistic correction is therefore substantial.

2.4.4 Numerical values relevant to cryo-electron microscopy

The principal constants are approximately

$$e \simeq 1.602 \times 10^{-19}\ \mathrm{C}, \tag{2.114}$$

$$m_0 \simeq 9.109 \times 10^{-31}\ \mathrm{kg}, \tag{2.115}$$

$$c \simeq 2.998 \times 10^{8}\ \mathrm{m\,s^{-1}}, \tag{2.116}$$

$$h \simeq 6.626 \times 10^{-34}\ \mathrm{J\,s}, \tag{2.117}$$

and

$$m_0c^2 \simeq 511\ \mathrm{keV}. \tag{2.118}$$

Using Eqs. (2.100) and (2.103), one obtains

Accelerating voltage	v/c	λ
100 kV	0.548	3.70 pm
200 kV	0.695	2.51 pm
300 kV	0.777	1.97 pm

Thus the electron wavelength in a modern cryo-electron microscope is of the order of only a few picometres. A carbon–carbon bond has a characteristic length of approximately 120–154 pm, whereas the wavelength of a 300 kV electron is approximately 2 pm.

The short electron wavelength is a necessary condition for atomic resolution, but it should not be confused with the resolution actually attained by the microscope. The latter is affected by electron–specimen scattering, lens aberrations, partial coherence, radiation damage, specimen motion, detector response, sampling, and the reconstruction procedure.

The corresponding angular wavenumber is

$$\boxed{k = \frac{2\pi}{\lambda}.} \tag{2.119}$$

In Chapter 3, the stationary relativistic electron is treated directly from the Dirac equation. In particular, Sec. 3.4.1 shows that its spatial wavefunction satisfies the Helmholtz equation Eq. (3.159) and derives

$$k = \frac{p}{\hbar} = \frac{2\pi}{\lambda} \tag{2.120}$$

in Eq. (3.163). This establishes the precise connection between the particle kinematics developed here and the wave-optical propagation theory used throughout the remainder of the book.

2.4.5 Energy and temporal frequency

For a stationary electron state, the Planck–Einstein relation is

$$\boxed{E = h\nu = \hbar\omega.} \tag{2.121}$$

The temporal dependence of a definite-energy state and its reduction to a stationary spatial wave are derived in Sec. 3.4.1, in particular in Eq. (3.155). We therefore do not repeat that derivation here.

The important distinction for the present chapter is that Eq. (2.98) specifies the mean energy of the beam, whereas the physical source produces a distribution of energies around that value. The consequences of this distribution are considered in Sec. 2.5.

2.4.6 Orders of magnitude in a cryo-EM exposure

The electron velocity calculated above is so large that it is difficult to develop an intuitive feeling for the actual time and particle scales involved in a cryo-electron microscopy experiment. It is therefore useful to consider a representative single-particle acquisition and translate its usual experimental parameters into numbers of electrons and characteristic times.

For this purpose, consider a 300 kV microscope and adopt the following representative values:

$$\begin{aligned} L &= 3 \text{ m}, \\ N_x = N_y &= 4096, \\ a &= 0.8 \text{ Å/pixel}, \\ D &= 50\ e^-/\text{Å}^2, \\ T_{\text{exp}} &= 2 \text{ s}, \\ N_f &= 40, \end{aligned} \tag{2.122}$$

where L is an approximate electron flight distance through the microscope, a is the pixel size referred to the specimen, D is the total incident electron dose, T_{exp} is the movie exposure time, and N_f is the number of dose fractions into which the movie is divided. These numbers should be regarded as representative rather than universal: the precise values depend on the microscope, detector, magnification, specimen, and acquisition protocol.

How long does an electron spend in the microscope?

At 300 kV, Eq. (2.100) gives

$$\frac{v}{c} \simeq 0.7765, \tag{2.123}$$

or

$$\boxed{v \simeq 2.33 \times 10^8 \text{ m s}^{-1}.} \tag{2.124}$$

Thus an electron travels more than two hundred million metres per second.

If we approximate its path through the microscope by $L = 3$ m and neglect the very short initial acceleration region, the flight time measured in the laboratory is

$$\boxed{t_{\text{lab}} = \frac{L}{v} \simeq 1.29 \times 10^{-8} \text{ s} = 12.9 \text{ ns}.} \tag{2.125}$$

For comparison, light would traverse the same 3 m distance in approximately 10 ns.

The time elapsed along the electron's own trajectory is its *proper time*. For motion at constant speed,

$$d\tau = dt\sqrt{1 - \frac{v^2}{c^2}} = \frac{dt}{\gamma}. \tag{2.126}$$

At 300 kV,

$$\gamma = 1 + \frac{eV}{m_0 c^2} \simeq 1 + \frac{300}{511} \simeq 1.587, \tag{2.127}$$

and therefore

$$\boxed{\tau = \frac{t_{\text{lab}}}{\gamma} \simeq 8.12 \text{ ns}.} \tag{2.128}$$

Strictly speaking, an electron travelling through the microscope is accelerated and deflected and therefore does not possess a single global inertial "electron reference frame." The proper time in Eq. (2.128), however, is well-defined along its trajectory. The estimate assumes that the electron has essentially its final 300 keV kinetic energy during most of the 3 m flight.

Flight time of a 300 kV electron.

From Eq. (2.100),

$$v = c\sqrt{1 - \left(1 + \frac{eV}{m_0 c^2}\right)^{-2}}. \tag{2.129}$$

Using

$$V = 300 \text{ kV}, \qquad m_0 c^2 \simeq 511 \text{ keV}, \tag{2.130}$$

we obtain

$$\gamma = 1 + \frac{300}{511} \simeq 1.587, \tag{2.131}$$

and

$$\frac{v}{c} = \sqrt{1 - \frac{1}{\gamma^2}} \simeq 0.7765. \tag{2.132}$$

Consequently,

$$t_{\text{lab}} = \frac{3 \text{ m}}{(0.7765)(2.998 \times 10^8 \text{ m s}^{-1})} \tag{2.133}$$

$$\simeq 12.9 \text{ ns}. \tag{2.134}$$

The corresponding proper time is

$$\tau = \frac{t_{\text{lab}}}{\gamma} = \frac{12.9 \text{ ns}}{1.587} \simeq 8.12 \text{ ns}. \tag{2.135}$$

How many electrons form one micrograph?

A 4096×4096 image sampled at $a = 0.8$ Å/pixel corresponds at the specimen to a linear field of view

$$W = 4096 \times 0.8 \text{ Å} \tag{2.136}$$

$$= 3276.8 \text{ Å} \tag{2.137}$$

$$\simeq 328 \text{ nm}. \tag{2.138}$$

The corresponding area is

$$\boxed{A = W^2 \simeq 1.074 \times 10^7 \text{ Å}^2.} \tag{2.139}$$

At a total dose of $D = 50\ e^-/\text{Å}^2$, the number of electrons incident on

this area during one movie is therefore

$$N_e = DA \tag{2.140}$$
$$= (50)(1.074 \times 10^7) \tag{2.141}$$
$$\simeq 5.37 \times 10^8. \tag{2.142}$$

Thus,

$$\boxed{N_e \simeq 5.4 \times 10^8 \quad \text{electrons per movie.}} \tag{2.143}$$

More than half a billion electrons therefore traverse the recorded specimen area during a single two-second exposure.

How far apart in time are successive electrons?

Although the total number in Eq. (2.143) is very large, these electrons are distributed over the full exposure time. The mean rate of electrons crossing the recorded specimen area is

$$\dot{N}_e = \frac{N_e}{T_{\text{exp}}} \simeq \frac{5.37 \times 10^8}{2} \simeq 2.68 \times 10^8 \text{ s}^{-1}. \tag{2.144}$$

The equivalent incident beam current over this field of view is

$$\boxed{I = e\dot{N}_e \simeq 4.30 \times 10^{-11} \text{ A} = 43 \text{ pA.}} \tag{2.145}$$

The mean temporal separation between successive electrons in this useful incident beam is therefore

$$\boxed{\Delta t_e = \frac{1}{\dot{N}_e} \simeq 3.73 \text{ ns.}} \tag{2.146}$$

At $v \simeq 2.33 \times 10^8$ m s^{-1}, this temporal interval corresponds to a mean longitudinal separation

$$\boxed{\Delta z_e \simeq v\Delta t_e \simeq 0.87 \text{ m.}} \tag{2.147}$$

This is a striking scale. Although hundreds of millions of electrons contribute to a movie, consecutive electrons in the useful imaging beam are, on average, separated by a distance comparable with one metre along the optical axis.

Combining Eqs. (2.125) and (2.146), the mean number of electrons simultaneously travelling within a 3 m useful beam is only

$$\boxed{\langle N_{\text{flight}} \rangle = \dot{N}_e t_{\text{lab}} \simeq 3.5.} \tag{2.148}$$

This number should be interpreted carefully. The dose specifies the electrons incident on the recorded area of the specimen. It does *not* specify the total number of electrons emitted by the gun. Electrons outside the selected illumination area, electrons intercepted by apertures, and other losses are not included. If the total gun current I_{gun} were known, the mean separation between electrons actually emitted by the cathode would instead be

$$\boxed{\Delta t_{\text{gun}} = \frac{e}{I_{\text{gun}}}.} \tag{2.149}$$

How many electrons reach one pixel?

The specimen area represented by one image pixel is

$$A_{\text{pix}} = a^2 = (0.8\ \text{Å})^2 = 0.64\ \text{Å}^2. \tag{2.150}$$

At a dose $D = 50\ e^-/\text{Å}^2$, the mean number of incident electrons associated with one pixel during the complete movie is therefore

$$\boxed{\lambda_{\text{movie}} = DA_{\text{pix}} = 50 \times 0.64 = 32.} \tag{2.151}$$

If the arrival of electrons is modelled as an independent Poisson process, then the number N_{pix} reaching a particular pixel is

$$\boxed{N_{\text{pix}} \sim \text{Poisson}\left(\lambda_{\text{movie}}\right), \qquad \lambda_{\text{movie}} = 32.} \tag{2.152}$$

The standard deviation due to counting statistics alone is consequently

$$\sigma_{\text{pix}} = \sqrt{\lambda_{\text{movie}}} \simeq 5.66, \tag{2.153}$$

and the relative Poisson fluctuation is

$$\boxed{\frac{\sigma_{\text{pix}}}{\lambda_{\text{movie}}} = \frac{1}{\sqrt{32}} \simeq 0.177.} \tag{2.154}$$

Thus even before considering scattering, detector noise, specimen motion, or any other source of uncertainty, the finite number of incident electrons implies fluctuations of order 18% at the scale of one pixel.

The movie is much sparser frame by frame

The same exposure divided into $N_f = 40$ equally dosed frames has a mean dose per frame

$$D_f = \frac{D}{N_f} = \frac{50}{40} = 1.25\ e^-/\text{Å}^2. \tag{2.155}$$

The corresponding mean number of electrons per pixel and per frame is only

$$\boxed{\lambda_f = \frac{\lambda_{\text{movie}}}{N_f} = \frac{32}{40} = 0.8.} \tag{2.156}$$

For a Poisson random variable,

$$P(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}. \tag{2.157}$$

With $\lambda_f = 0.8$,

$$P(N = 0) = e^{-0.8} \simeq 0.449, \tag{2.158}$$

$$P(N = 1) = 0.8e^{-0.8} \simeq 0.359, \tag{2.159}$$

$$P(N \geq 2) = 1 - e^{-0.8}(1 + 0.8) \simeq 0.191. \tag{2.160}$$

Thus, in this simplified model, approximately 45% of the pixels in an individual dose fraction receive no electron at all, approximately 36% receive one electron, and only approximately 19% receive two or more.

This sparse-event picture helps explain why direct electron counting is so natural for cryo-electron microscopy. It should not, however, be interpreted too literally at the level of the raw detector pixels. An incident electron produces a spatially extended detector event, the detector has a finite detective quantum efficiency, and sufficiently high instantaneous rates can lead to coincidence effects. Equation (2.152) is therefore a model of the statistics of the *incident electron fluence referred to the specimen plane*, rather than a complete model of the detector output.

Putting the scales together

The orders of magnitude obtained above can be summarized as

$$\boxed{\begin{aligned} \text{electron speed} &\sim 2.3 \times 10^8\ \mathrm{m\,s^{-1}}, \\ \text{flight time through 3 m} &\sim 13\ \mathrm{ns}, \\ \text{proper time during the flight} &\sim 8\ \mathrm{ns}, \\ \text{electrons in one movie} &\sim 5 \times 10^8, \\ \text{useful beam current} &\sim 40\ \mathrm{pA}, \\ \text{mean time between useful electrons} &\sim 4\ \mathrm{ns}, \\ \text{mean longitudinal electron separation} &\sim 1\ \mathrm{m}, \\ \text{electrons per pixel per movie} &\sim 30, \\ \text{electrons per pixel per frame} &\sim 1. \end{aligned}} \tag{2.161}$$

These numbers reveal two apparently contradictory aspects of cryo-electron microscopy. A complete movie contains hundreds of millions of electrons, yet at the spatial and temporal scale of an individual pixel in an individual movie frame, the signal is generated by only of order one electron. Cryo-electron microscopy is therefore simultaneously a high-throughput electron experiment and an extremely sparse counting experiment.

2.5 Beam properties determined by the source

The mean accelerating voltage determines the central electron energy and wavelength, but it does not by itself determine the quality of the illumination. Two electron guns operating at the same accelerating voltage may produce beams with very different effective source sizes, angular distributions, brightnesses, energy spreads, and temporal stability.

These properties provide the physical connection between the electron gun and the partial-coherence formalism developed later in Sec. 3.11.2.

2.5.1 Effective and virtual source size

The physical region from which electrons are emitted is not necessarily identical to the source that is seen by the rest of the microscope. The electrostatic field immediately surrounding the emitter bends the electron trajectories. If these trajectories are extrapolated backwards, they generally appear to originate from a smaller region called the *virtual source*.

The effective source size is therefore an electron-optical quantity rather than simply the physical diameter of the emitter. A sharply pointed field-emission tip can produce a virtual source much smaller than the geometric dimensions of the tip itself.

This quantity is important because different points of an extended source give rise, after the condenser system, to slightly different illumination directions at the specimen. The formal treatment of this effect is given in Sec. 3.11.4. In the paraxial regime, the illumination angle is related to transverse spatial frequency by the angular-spectrum result Eq. (3.367). Thus the finite source ultimately produces a distribution of incident transverse wavevectors.

2.5.2 Brightness

A source is useful not merely when it emits a large total current, but when a large current can be concentrated into a small transverse area and a small angular range. This is quantified by the *brightness*.

The differential brightness is defined as

$$\boxed{B \triangleq \frac{dI}{dA_\perp \, d\Omega},} \tag{2.162}$$

where dI is the current carried by rays crossing the transverse area $dA_\perp$ within the solid-angle element $d\Omega$. Its SI units are

$$[B] = \mathrm{A\,m^{-2}\,sr^{-1}}. \tag{2.163}$$

For a uniformly emitting circular effective source of diameter d_s and a small emission semi-angle α,

$$\boxed{B \simeq \frac{4I}{\pi^2 d_s^2 \alpha^2}.} \tag{2.164}$$

Derivation.

The source area is

$$A_s = \pi \left(\frac{d_s}{2}\right)^2 = \frac{\pi d_s^2}{4}. \tag{2.165}$$

The solid angle of a cone with semi-angle α is

$$\Omega = 2\pi(1 - \cos\alpha). \tag{2.166}$$

For $\alpha \ll 1$,

$$\cos\alpha \simeq 1 - \frac{\alpha^2}{2}, \tag{2.167}$$

so

$$\Omega \simeq \pi\alpha^2. \tag{2.168}$$

For a uniform source,

$$B \simeq \frac{I}{A_s\Omega} = \frac{4I}{\pi^2 d_s^2 \alpha^2}, \tag{2.169}$$

which gives Eq. (2.164).

Brightness itself changes during acceleration because the longitudinal momentum increases while the transverse angular spread decreases. It is therefore convenient to introduce a *reduced brightness*

$$\boxed{B_r \triangleq \frac{B}{\widehat{\Phi}},} \tag{2.170}$$

where $\widehat{\Phi}$ is the relativistic acceleration potential defined in Eq. (2.77). The units of B_r are

$$[B_r] = \mathrm{A\,m^{-2}\,sr^{-1}\,V^{-1}}. \tag{2.171}$$

The reason for using $\widehat{\Phi}$ rather than the uncorrected voltage is that the relativistic momentum satisfies Eq. (2.78),

$$p^2 = 2m_0 e\widehat{\Phi}. \tag{2.172}$$

Why acceleration increases geometrical brightness.

Consider ideal acceleration parallel to the optic axis and neglect transverse forces. The transverse momentum $p_\perp$ is then unchanged, whereas the total momentum p increases. For a paraxial electron,

$$\alpha \simeq \frac{p_\perp}{p}. \tag{2.173}$$

Consequently,

$$\alpha^2 \propto \frac{1}{p^2}. \tag{2.174}$$

Since the corresponding small solid angle satisfies $\Omega \simeq \pi\alpha^2$,

$$\Omega \propto \frac{1}{p^2}. \tag{2.175}$$

If the current and transverse area are unchanged during this ideal acceleration,

$$B \propto p^2. \tag{2.176}$$

Using $p^2 = 2m_0 e\widehat{\Phi}$ from Eq. (2.78),

$$B \propto \widehat{\Phi}. \tag{2.177}$$

Hence the ratio $B/\widehat{\Phi}$ remains constant during ideal axial acceleration, which motivates the definition in Eq. (2.170).

The more general statement that ideal conservative electron optics cannot increase phase-space density is a form of the brightness or Liouville theorem. Its natural place is the geometrical electron-optics treatment, where transverse ray phase space will be introduced explicitly.

2.5.3 Energy spread and temporal coherence

A real source emits electrons with a distribution of initial kinetic energies K_0. According to Eq. (2.97), ideal acceleration through a fixed potential adds eV to every electron:

$$K = K_0 + eV. \tag{2.178}$$

Thus an intrinsic spread in K_0 survives acceleration as an absolute energy spread, although it becomes a progressively smaller fraction of the total kinetic energy.

Additional fluctuations in the accelerating voltage also modify the final electron energy. To first order,

$$\delta K = \delta K_0 + e\,\delta V. \tag{2.179}$$

If the two contributions are statistically independent and characterized by variances $\sigma_{K_0}^2$ and σ_V^2, respectively, then

$$\boxed{\sigma_K^2 = \sigma_{K_0}^2 + e^2\sigma_V^2.} \tag{2.180}$$

Derivation.

From Eq. (2.179),

$$\sigma_K^2 = \left\langle (\delta K_0 + e\delta V)^2 \right\rangle \tag{2.181}$$

$$= \sigma_{K_0}^2 + e^2\sigma_V^2 + 2e\left\langle \delta K_0 \delta V \right\rangle . \tag{2.182}$$

For statistically independent emission-energy and voltage fluctuations,

$$\langle \delta K_0 \delta V \rangle = 0, \tag{2.183}$$

which gives Eq. (2.180).

Energy spread is commonly specified by the full width at half maximum of the measured energy distribution. For a Gaussian distribution, the conversion between FWHM and standard deviation has already been derived in Eq. (3.773).

The relation between this energy distribution and temporal coherence is developed rigorously in Sec. 3.11.4. In particular, different energies should not simply be declared "incoherent." The mutual coherence between energy components is described by Eq. (3.759), and the incoherent-energy limit leads to the intensity average Eq. (3.764).

The physical chain connecting the source to image formation is therefore

$$\boxed{\begin{array}{c}\text{emission-energy distribution and voltage fluctuations}\\ \Downarrow \\ \text{distribution of electron energies}\\ \Downarrow \\ \text{distribution of wavelengths and focal conditions}\\ \Downarrow \\ \text{temporal partial coherence and chromatic effects.}\end{array}} \tag{2.184}$$

2.5.4 Finite source size and spatial coherence

The second important source property is its finite transverse extent. Different points of the effective source correspond, after propagation through the condenser system, to slightly different illumination directions at the specimen.

The rigorous wave-optical treatment is given in Sec. 3.11.4, where the source is described as an ensemble of illumination tilts $\boldsymbol{\theta} = (\theta_x, \theta_y)$ and

the observed intensity is obtained by averaging the coherently propagated intensities over that distribution.

Thus a small effective source is desirable not simply because it produces a geometrically narrow beam. It reduces the distribution of illumination directions and thereby improves spatial coherence. This is one of the major advantages of field-emission sources over thermionic emitters.

As emphasized in Sec. 3.11.2, however, finite source size and spatial incoherence are not mathematically identical concepts. The rigorous quantity is the mutual coherence function of the field, Eq. (3.732). Source size is a physical mechanism that usually reduces that coherence.

2.5.5 Source-current stability

A final property of practical importance is stability. The mean emission current should remain constant over the time scale of an exposure and, ideally, over the much longer time scale of an automated acquisition.

Thermionic emitters are comparatively insensitive to minor changes in surface state, but their emission depends very strongly on temperature through Eq. (2.2). Field emitters are sensitive to both the extraction field and the work function. In particular, Eq. (2.30) shows that small changes of local field or surface condition can produce substantial changes in a cold field-emission current.

The practical design of an electron gun therefore includes not only the emitter itself but also extremely stable high-voltage supplies, vacuum systems, extraction electronics, and feedback mechanisms. For cryo-EM this stability is particularly important because a single dataset may consist of many thousands of movies recorded over periods of many hours or days.

2.6 Summary

The electron source establishes the initial conditions for all subsequent electron optics. Three groups of quantities should be distinguished.

First, the emission mechanism determines the microscopic origin of the beam. Thermionic emission is described by Eq. (2.2); an extraction field lowers the surface barrier according to Eq. (2.17); and sufficiently strong fields permit quantum tunnelling, leading in the simplest approximation to the Fowler–Nordheim law Eq. (2.30).

Second, the accelerating electrostatic field determines the relativistic kinematics. The central quantity is the acceleration potential $\widehat{\Phi}$ introduced in Eq. (2.77), from which momentum, velocity, kinetic energy, and

Lorentz factor follow through Eq. (2.78). For an electron accelerated from rest through V,

$$K = eV, \tag{2.185}$$

$$\gamma = 1 + \frac{eV}{m_0 c^2}, \tag{2.186}$$

$$v = c\sqrt{1 - \left(1 + \frac{eV}{m_0 c^2}\right)^{-2}}, \tag{2.187}$$

$$p = \frac{1}{c}\sqrt{eV(2m_0 c^2 + eV)}, \tag{2.188}$$

$$\lambda = \frac{hc}{\sqrt{eV(2m_0 c^2 + eV)}}. \tag{2.189}$$

Third, the finite source determines the statistical quality of the illumination. The effective source size and angular distribution are connected to spatial partial coherence, whereas the energy distribution is connected to temporal partial coherence. These relations are treated formally in Sec. 3.11.2, with energy spread developed in Sec. 3.11.4 and finite-source effects in Sec. 3.11.4.

The present chapter therefore supplies both the particle kinematics and the statistical source parameters needed by the next chapter. The bridge to wave optics is completed in Sec. 3.4.1, where the relativistic electron wave is shown to satisfy the Helmholtz equation Eq. (3.159) with

$$k = \frac{p}{\hbar} = \frac{2\pi}{\lambda}, \tag{2.190}$$

as derived in Eq. (3.163).

Chapter 3

Wave diffraction theory

This chapter develops the mathematical framework required to describe the propagation and diffraction of waves. The treatment is intentionally general. At this stage, we are not concerned with the particular physical nature of the wave, but with the mathematical consequences of wave propagation itself. The same framework applies, under the appropriate assumptions, to electromagnetic radiation, acoustic waves, X-rays, and electron matter waves.

This separation between general wave theory and electron-specific physics will be useful throughout the book. In the present chapter, we will derive how a wave propagates through free space, how it is diffracted by apertures, how lenses modify its phase, and how these operations can be expressed in the Fourier domain. In Chapter 4, we will determine which additional ingredients are required when the propagating wave describes an electron moving through the electromagnetic fields of an electron microscope.

Wave diffraction theory extends geometrical optics. Geometrical optics describes propagation in terms of rays and is recovered when the wavelength is sufficiently small compared with the characteristic spatial scales of the system. Wave optics retains the phase of the propagating field and therefore accounts for interference and diffraction. These effects are essential for understanding image formation in cryo-electron microscopy.

A recurrent mathematical tool throughout this chapter, and indeed throughout the rest of the book, will be the Fourier transform. We begin by fixing its definition and our notation.

3.1 Fourier transform convention

Let

$$f : \mathbb{R}^n \longrightarrow \mathbb{C} \tag{3.1}$$

be an n-dimensional function of the spatial variable $\mathbf{r} \in \mathbb{R}^n$. Throughout this book, we define its Fourier transform as

$$\boxed{\hat{f}(\mathbf{R}) = \mathcal{F}\{f\}(\mathbf{R}) = \int_{\mathbb{R}^n} f(\mathbf{r}) \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}\right)\, d\mathbf{r}.} \tag{3.2}$$

The corresponding inverse Fourier transform is

$$\boxed{f(\mathbf{r}) = \mathcal{F}^{-1}\{\hat{f}\}(\mathbf{r}) = \int_{\mathbb{R}^n} \hat{f}(\mathbf{R}) \exp\left(j2\pi\mathbf{R}\cdot\mathbf{r}\right)\, d\mathbf{R}.} \tag{3.3}$$

Here, $j = \sqrt{-1}$, $\mathbf{R} \in \mathbb{R}^n$ is the spatial-frequency vector, and $\mathbf{R}\cdot\mathbf{r}$ denotes the Euclidean inner product.

The variables $\mathbf{r}$ and $\mathbf{R}$ belong to reciprocal spaces. If the components of $\mathbf{r}$ are measured in units of length, then the components of $\mathbf{R}$ have units of inverse length. In particular, if

$$\mathbf{r} = (r_1, \ldots, r_n)^T, \qquad \mathbf{R} = (R_1, \ldots, R_n)^T, \tag{3.4}$$

then

$$\mathbf{R}\cdot\mathbf{r} = \sum_{i=1}^{n} R_i r_i \tag{3.5}$$

is dimensionless, as required for the argument of the complex exponential.

With the convention in Eq. (3.2), the components of $\mathbf{R}$ represent spatial frequencies measured in cycles per unit length. This convention should be distinguished from the angular wavevector, which will be denoted by $\mathbf{k}$ and measured in radians per unit length. The two are related by

$$\boxed{\mathbf{k} = 2\pi\mathbf{R}.} \tag{3.6}$$

For a monochromatic wave of wavelength λ, the magnitude of the wavevector is

$$k = \|\mathbf{k}\| = \frac{2\pi}{\lambda}, \tag{3.7}$$

whereas the corresponding scalar spatial frequency is

$$R = \|\mathbf{R}\| = \frac{1}{\lambda}. \tag{3.8}$$

This distinction between spatial frequency and angular wavenumber will be maintained throughout the book.

The Fourier transform is particularly natural in wave theory because its basis functions,

$$\exp\left(j2\pi\mathbf{R}\cdot\mathbf{r}\right), \tag{3.9}$$

are complex plane waves. Equation (3.3) therefore states that an arbitrary sufficiently well-behaved spatial field can be represented as a continuous superposition of plane-wave components, each characterized by a spatial frequency $\mathbf{R}$ and a complex coefficient $\hat{f}(\mathbf{R})$.

The magnitude $|\hat{f}(\mathbf{R})|$ determines the amplitude with which that Fourier component participates in the decomposition, while $\arg \hat{f}(\mathbf{R})$ determines its phase. This interpretation will become especially important when we introduce the angular-spectrum description of wave propagation: each Fourier component will be associated with a plane wave propagating in a particular direction, and propagation will act independently on these components.

Although many physical quantities considered later will be real-valued, there is no requirement that f itself be real. In wave optics, the quantities being propagated will generally be complex-valued fields, with their complex phase carrying essential physical information.

The Fourier transform generally does not have the same physical units as the function being transformed. This follows directly from the integration measure in Eq. (3.2). If the spatial coordinates $\mathbf{r} \in \mathbb{R}^n$ are measured in metres, then

$$d\mathbf{r} = dr_1 \cdots dr_n \tag{3.10}$$

has units m^n. The complex exponential is dimensionless because $\mathbf{R}\cdot\mathbf{r}$ is dimensionless. Consequently,

$$\boxed{[\hat{f}] = [f]\,\mathrm{m}^n.} \tag{3.11}$$

For example, a one-dimensional Fourier transform acquires one factor of length, a two-dimensional Fourier transform acquires a factor of area, and a three-dimensional Fourier transform acquires a factor of volume.

The inverse transform provides a useful consistency check. Since the components of $\mathbf{R}$ have units m^{-1},

$$[d\mathbf{R}] = \mathrm{m}^{-n}. \tag{3.12}$$

Therefore,

$$[\hat{f}]\,[d\mathbf{R}] = [f]\,\mathrm{m}^n\,\mathrm{m}^{-n} = [f], \tag{3.13}$$

as required for the inverse transform to recover $f(\mathbf{r})$.

This dimensional factor is a property of the continuous Fourier transform and should not be confused with the discrete Fourier transform commonly used in numerical computations, where sampling intervals may be absorbed into the normalization convention.

The principal quantities introduced in this section are

Symbol	Meaning	Units
$\mathbf{r}$	Spatial coordinate	m
$\mathbf{R}$	Spatial-frequency vector	$\mathrm{cycles\,m^{-1}}$
$\mathbf{k}$	Angular wavevector	$\mathrm{rad\,m^{-1}}$
R	Magnitude of the spatial frequency	$\mathrm{cycles\,m^{-1}}$
k	Angular wavenumber	$\mathrm{rad\,m^{-1}}$
λ	Wavelength	m
$f(\mathbf{r})$	Function in real space	problem dependent
$\hat{f}(\mathbf{R})$	Fourier transform of f	units of f times m^n

Strictly speaking, both cycles and radians are dimensionless, so the SI dimensions of both $\mathbf{R}$ and $\mathbf{k}$ are m^{-1}. We retain "cycles" and "rad" in the unit notation to emphasize the distinction between ordinary spatial frequency and angular spatial frequency:

$$R = \frac{1}{\lambda}, \qquad k = 2\pi R = \frac{2\pi}{\lambda}. \tag{3.14}$$

3.2 Wave propagation in homogeneous, isotropic, and non-dispersive media

We begin with the propagation of electromagnetic waves in a simple medium. The purpose of this section is not specifically to describe electrons. Rather, it is to derive the scalar wave equation that underlies classical wave optics. In Section 3.4, we will show that the spatial wave function of a free electron satisfies an equation of the same mathematical form. This equivalence will allow the diffraction theory developed in this chapter to be applied later to electron microscopy.

Let $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$ denote the electric and magnetic fields, respectively. We consider a linear medium without free electric charges or free

currents. Maxwell's equations in SI units are then

$$\begin{aligned} \nabla \times \mathbf{E} &= -\mu \frac{\partial \mathbf{H}}{\partial t}, \\ \nabla \times \mathbf{H} &= \epsilon \frac{\partial \mathbf{E}}{\partial t}, \\ \nabla \cdot (\epsilon \mathbf{E}) &= 0, \\ \nabla \cdot (\mu \mathbf{H}) &= 0. \end{aligned} \tag{3.15}$$

Here, ϵ and μ denote the electric permittivity and magnetic permeability of the medium. The differential operator

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)^T \tag{3.16}$$

is the spatial gradient, while $\nabla\cdot$ and $\nabla\times$ denote the divergence and curl operators.

We now impose three assumptions on the propagation medium:

- **Homogeneity:** ϵ and μ do not depend on position.
- **Isotropy:** the constitutive properties of the medium do not depend on the direction of the fields, so ϵ and μ are scalars rather than tensors.
- **Non-dispersiveness:** ϵ and μ do not depend on temporal frequency.

Under these assumptions, both ϵ and μ are constants. Maxwell's equations then imply that the electric and magnetic fields satisfy

$$\boxed{\nabla^2 \mathbf{E} - \mu\epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mathbf{0}, \qquad \nabla^2 \mathbf{H} - \mu\epsilon \frac{\partial^2 \mathbf{H}}{\partial t^2} = \mathbf{0}.} \tag{3.17}$$

Defining

$$\boxed{v = \frac{1}{\sqrt{\mu\epsilon}},} \tag{3.18}$$

these equations take the familiar form

$$\boxed{\nabla^2 \mathbf{E} - \frac{1}{v^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mathbf{0}, \qquad \nabla^2 \mathbf{H} - \frac{1}{v^2} \frac{\partial^2 \mathbf{H}}{\partial t^2} = \mathbf{0}.} \tag{3.19}$$

Because the differential operator acts independently on each Cartesian component, every component of $\mathbf{E}$ and $\mathbf{H}$ satisfies the same scalar equation. We may therefore introduce a generic scalar field $U(\mathbf{r}, t)$ satisfying

$$\boxed{\nabla^2 U(\mathbf{r}, t) - \frac{1}{v^2}\frac{\partial^2 U(\mathbf{r}, t)}{\partial t^2} = 0.} \tag{3.20}$$

This is the scalar wave equation that will form the starting point for the diffraction theory developed throughout this chapter.

Derivation of the vector wave equation.

Because the medium is homogeneous, ϵ and μ are independent of position. Consequently,

$$\nabla \cdot (\epsilon \mathbf{E}) = \epsilon \nabla \cdot \mathbf{E}. \tag{3.21}$$

Since $\epsilon \neq 0$, the third Maxwell equation in Eq. (3.15) therefore implies

$$\nabla \cdot \mathbf{E} = 0. \tag{3.22}$$

Similarly,

$$\nabla \cdot \mathbf{H} = 0. \tag{3.23}$$

We now apply the curl operator to Faraday's law,

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}. \tag{3.24}$$

Taking the curl of both sides gives

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \nabla \times \frac{\partial \mathbf{H}}{\partial t}. \tag{3.25}$$

For any sufficiently differentiable vector field $\mathbf{U}$,

$$\nabla \times (\nabla \times \mathbf{U}) = \nabla(\nabla \cdot \mathbf{U}) - \nabla^2 \mathbf{U}, \tag{3.26}$$

where

$$\nabla^2 \mathbf{U} = \left(\nabla^2 U_x, \nabla^2 U_y, \nabla^2 U_z\right)^T \tag{3.27}$$

and the scalar Laplacian is

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{3.28}$$

Applying Eq. (3.26) to the left-hand side of Eq. (3.25), and using Eq. (3.22), gives

$$-\nabla^2 \mathbf{E} = -\mu \nabla \times \frac{\partial \mathbf{H}}{\partial t}. \tag{3.29}$$

Since spatial and temporal derivatives commute for sufficiently smooth fields,

$$\nabla \times \frac{\partial \mathbf{H}}{\partial t} = \frac{\partial}{\partial t}(\nabla \times \mathbf{H}). \tag{3.30}$$

Therefore,

$$-\nabla^2 \mathbf{E} = -\mu \frac{\partial}{\partial t}(\nabla \times \mathbf{H}). \tag{3.31}$$

Ampère–Maxwell's law gives

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}. \tag{3.32}$$

Substituting this expression,

$$-\nabla^2 \mathbf{E} = -\mu \frac{\partial}{\partial t}\left(\epsilon \frac{\partial \mathbf{E}}{\partial t}\right) \tag{3.33}$$

$$= -\mu\epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}, \tag{3.34}$$

where the second equality follows because ϵ is independent of time. Multiplying by -1 yields

$$\nabla^2 \mathbf{E} - \mu\epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} = \mathbf{0}. \tag{3.35}$$

The derivation for $\mathbf{H}$ is completely analogous. Starting from Ampère–Maxwell's law, taking its curl, and using Faraday's law gives

$$\nabla^2 \mathbf{H} - \mu\epsilon \frac{\partial^2 \mathbf{H}}{\partial t^2} = \mathbf{0}. \tag{3.36}$$

Thus both fields propagate according to the same vector wave equation.

3.2.1 Wave velocity and refractive index

In vacuum,

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}, \tag{3.37}$$

where μ_0 and ϵ_0 are the vacuum permeability and permittivity. Comparing this expression with Eq. (3.18), the refractive index of a homogeneous medium may be defined as

$$\boxed{n = \frac{c}{v} = \sqrt{\frac{\mu\epsilon}{\mu_0\epsilon_0}}.} \tag{3.38}$$

For a non-magnetic medium, $\mu \simeq \mu_0$, and therefore

$$\boxed{n \simeq \sqrt{\frac{\epsilon}{\epsilon_0}}, \qquad v = \frac{c}{n}.} \tag{3.39}$$

The quantity n is dimensionless. In a non-dispersive medium it is independent of frequency, so all frequency components propagate with the same phase velocity v. When n depends on frequency, different frequency components acquire different propagation velocities and the medium is said to be dispersive.

3.2.2 From vector to scalar wave theory

Equation (3.19) remains a vector equation: an electromagnetic wave possesses both magnitude and polarization. Nevertheless, in a homogeneous and isotropic medium its Cartesian components obey identical, uncoupled differential equations. For example, the electric-field equation is equivalent to

$$\nabla^2 E_x - \frac{1}{v^2}\frac{\partial^2 E_x}{\partial t^2} = 0, \tag{3.40}$$

$$\nabla^2 E_y - \frac{1}{v^2}\frac{\partial^2 E_y}{\partial t^2} = 0, \tag{3.41}$$

$$\nabla^2 E_z - \frac{1}{v^2}\frac{\partial^2 E_z}{\partial t^2} = 0. \tag{3.42}$$

The same holds for the three components of $\mathbf{H}$.

Therefore, whenever polarization effects are not relevant to the problem under consideration, we may study a single representative component and write

$$\nabla^2 U - \frac{1}{v^2}\frac{\partial^2 U}{\partial t^2} = 0. \tag{3.43}$$

This scalarization does not imply that the electromagnetic field itself is physically scalar. It means that, under the assumptions made above, the propagation of each field component is governed by the same scalar differential equation. The scalar theory retains the amplitude and phase phenomena required to describe diffraction and interference while avoiding the additional complexity associated with polarization.

The assumptions made in this section are important. If the material is inhomogeneous, then ϵ or μ depend on position and can no longer be taken outside spatial derivatives. Additional terms involving their gradients appear, and the simple scalar wave equation Eq. (3.20) is no longer obtained directly. Likewise, in an anisotropic medium, ϵ and μ may become tensors and different field components can become coupled. At interfaces and sharp boundaries, vector boundary conditions may also become important.

For the purposes of this chapter, however, Eq. (3.20) provides the fundamental mathematical model of wave propagation in a homogeneous region. We emphasize again that this derivation has been made for electromagnetic waves. An electron is not an electromagnetic wave, and its wave function does not obey Maxwell's equations. The relevance of the present derivation to cryo-electron microscopy lies instead in the fact that the stationary quantum-mechanical equation governing a freely propagating electron will later reduce to the same Helmholtz equation obtained from Eq. (3.20). Once this mathematical equivalence has been established, the machinery of scalar diffraction theory developed below can be transferred to electron waves.

3.3 Solutions of the wave equation

In the previous section we showed that, under appropriate assumptions, each component of an electromagnetic field satisfies the scalar wave equation

$$\nabla^2 u(\mathbf{r}, t) - \frac{1}{v^2} \frac{\partial^2 u(\mathbf{r}, t)}{\partial t^2} = 0. \tag{3.44}$$

Here we have denoted the real, time-dependent scalar field by $u(\mathbf{r}, t)$. Later, $U(\mathbf{r})$ will denote its complex time-independent phasor.

A particularly important family of solutions consists of travelling plane waves. Consider a field of the form

$$u(\mathbf{r}, t) = g\left(\omega t - \mathbf{k} \cdot \mathbf{r}\right), \tag{3.45}$$

where g is any twice-differentiable scalar function, ω is the angular temporal frequency, and $\mathbf{k}$ is the wavevector. This expression describes a waveform

whose surfaces of constant argument satisfy

$$\omega t - \mathbf{k} \cdot \mathbf{r} = \text{constant}. \tag{3.46}$$

These surfaces are planes perpendicular to $\mathbf{k}$ and move in the direction of $\mathbf{k}$.

Equation (3.45) satisfies the scalar wave equation provided

$$\boxed{\omega = vk, \qquad k = \|\mathbf{k}\|.} \tag{3.47}$$

This is the dispersion relation for a homogeneous, isotropic, non-dispersive medium.

It is important to emphasize that Eq. (3.45) is not the most general solution of the three-dimensional wave equation. Rather, it describes a single travelling plane-wave component. More general solutions can be constructed as superpositions of such components. This observation will become the basis of the angular-spectrum representation developed later.

Derivation of the plane-wave dispersion relation.

Let

$$\xi = \omega t - \mathbf{k} \cdot \mathbf{r}, \tag{3.48}$$

so that

$$u(\mathbf{r}, t) = g(\xi). \tag{3.49}$$

The first temporal derivative is

$$\frac{\partial u}{\partial t} = g'(\xi)\frac{\partial \xi}{\partial t} = \omega g'(\xi), \tag{3.50}$$

and therefore

$$\frac{\partial^2 u}{\partial t^2} = \omega^2 g''(\xi). \tag{3.51}$$

For the spatial derivatives,

$$\frac{\partial \xi}{\partial x_i} = -k_i, \tag{3.52}$$

and hence

$$\frac{\partial u}{\partial x_i} = -k_i g'(\xi). \tag{3.53}$$

Differentiating once more,

$$\frac{\partial^2 u}{\partial x_i^2} = k_i^2 g''(\xi). \tag{3.54}$$

Summing over all spatial coordinates gives

$$\nabla^2 u = \sum_i k_i^2 g''(\xi) \tag{3.55}$$

$$= \|\mathbf{k}\|^2 g''(\xi) = k^2 g''(\xi). \tag{3.56}$$

Substituting Eqs. (3.51) and (3.56) into Eq. (3.44) yields

$$\left(k^2 - \frac{\omega^2}{v^2}\right) g''(\xi) = 0. \tag{3.57}$$

For a non-trivial waveform, g'' is not identically zero, so

$$k^2 = \frac{\omega^2}{v^2}. \tag{3.58}$$

Taking the positive root for a positive frequency gives

$$\omega = vk. \tag{3.59}$$

The ordinary temporal frequency ν and angular temporal frequency ω are related by

$$\omega = 2\pi\nu. \tag{3.60}$$

Similarly, the wavelength λ and angular wavenumber k satisfy

$$k = \frac{2\pi}{\lambda}. \tag{3.61}$$

Substitution of these definitions into Eq. (3.47) gives

$$2\pi\nu = v\frac{2\pi}{\lambda}, \tag{3.62}$$

and therefore

$$\boxed{v = \lambda\nu.} \tag{3.63}$$

Thus, for a monochromatic wave,

$$\boxed{k = \frac{\omega}{v} = \frac{2\pi\nu}{v} = \frac{2\pi}{\lambda}.} \tag{3.64}$$

When a monochromatic wave crosses a stationary interface between two media, its temporal frequency remains unchanged. Its propagation velocity, however, may change. Consequently, its wavelength changes according to Eq. (3.63). If the refractive index of the medium is $n = c/v$, then

$$\lambda = \frac{v}{\nu} = \frac{c}{n\nu}. \tag{3.65}$$

Thus a change in refractive index modifies the spatial oscillation of the wave while preserving its temporal frequency.

3.3.1 Monochromatic waves and the phasor representation

Of particular importance in diffraction theory are monochromatic fields, for which the temporal dependence consists of a single frequency ν, or equivalently a single angular frequency $\omega = 2\pi\nu$.

A general real monochromatic scalar field may be written as

$$u(\mathbf{r}, t) = A(\mathbf{r}) \cos\left(\varphi(\mathbf{r}) - \omega t\right), \tag{3.66}$$

where $A(\mathbf{r}) \geq 0$ is the spatially varying amplitude and $\varphi(\mathbf{r})$ is the spatial phase.

Using Euler's identity,

$$e^{j\theta} = \cos\theta + j\sin\theta, \tag{3.67}$$

we may represent the same physical field as

$$\boxed{u(\mathbf{r}, t) = \mathrm{Re}\left\{U(\mathbf{r})e^{-j\omega t}\right\},} \tag{3.68}$$

where

$$\boxed{U(\mathbf{r}) = A(\mathbf{r})e^{j\varphi(\mathbf{r})}} \tag{3.69}$$

is the complex spatial amplitude, or phasor.

The phasor contains all spatial information about the monochromatic field:

$$A(\mathbf{r}) = |U(\mathbf{r})|, \qquad \varphi(\mathbf{r}) = \arg U(\mathbf{r}). \tag{3.70}$$

The rapid and completely known temporal factor $e^{-j\omega t}$ can therefore be separated from the spatial dependence. This is the main advantage of the phasor representation.

For a plane wave, the amplitude is constant and the spatial phase is linear,

$$\varphi(\mathbf{r}) = \mathbf{k} \cdot \mathbf{r} + \varphi_0, \tag{3.71}$$

so that

$$U(\mathbf{r}) = A_0 \exp\left[j\left(\mathbf{k}\cdot\mathbf{r} + \varphi_0\right)\right]. \tag{3.72}$$

The corresponding physical field is

$$u(\mathbf{r}, t) = A_0 \cos\left(\mathbf{k}\cdot\mathbf{r} - \omega t + \varphi_0\right). \tag{3.73}$$

3.3.2 The Helmholtz equation

For a monochromatic field, the temporal dependence can be eliminated entirely from the wave equation. Substitution of the phasor representation Eq. (3.68) into the scalar wave equation gives

$$\boxed{\left(\nabla^2 + k^2\right) U(\mathbf{r}) = 0,} \tag{3.74}$$

where

$$k = \frac{\omega}{v} = \frac{2\pi}{\lambda}. \tag{3.75}$$

Equation (3.74) is the Helmholtz equation. It is one of the central equations of this chapter. Whereas the wave equation describes the evolution of a field in both space and time, the Helmholtz equation describes the spatial structure of a single-frequency component.

Consequently, once monochromaticity is assumed, the wave-propagation problem becomes entirely spatial. The temporal oscillation $e^{-j\omega t}$ is known and can be restored at any time if required. Most of the diffraction theory developed in the remainder of this chapter will therefore be formulated directly in terms of the complex phasor $U(\mathbf{r})$.

Derivation of the Helmholtz equation.

Starting from

$$u(\mathbf{r}, t) = \mathrm{Re}\left\{U(\mathbf{r})e^{-j\omega t}\right\}, \tag{3.76}$$

it is sufficient, because the wave equation is linear with real coefficients, to work with the associated complex field

$$u_c(\mathbf{r}, t) = U(\mathbf{r})e^{-j\omega t}. \tag{3.77}$$

Since the temporal exponential is independent of position,

$$\nabla^2 u_c = e^{-j\omega t}\nabla^2 U. \tag{3.78}$$

The first temporal derivative is

$$\frac{\partial u_c}{\partial t} = -j\omega U(\mathbf{r})e^{-j\omega t}, \tag{3.79}$$

and the second derivative is

$$\frac{\partial^2 u_c}{\partial t^2} = (-j\omega)^2 U(\mathbf{r})e^{-j\omega t} = -\omega^2 U(\mathbf{r})e^{-j\omega t}. \tag{3.80}$$

Substituting these expressions into the scalar wave equation gives

$$e^{-j\omega t}\nabla^2 U + \frac{\omega^2}{v^2} U e^{-j\omega t} = 0. \tag{3.81}$$

Since $e^{-j\omega t}$ is never zero,

$$\nabla^2 U + \frac{\omega^2}{v^2} U = 0. \tag{3.82}$$

Defining

$$k = \frac{\omega}{v} \tag{3.83}$$

gives

$$\left(\nabla^2 + k^2\right) U = 0, \tag{3.84}$$

which is Eq. (3.74).

3.3.3 Integral representation of a Helmholtz field

The Helmholtz equation is a local differential equation: it specifies how the field and its spatial derivatives are related at every point. An equivalent and extremely useful viewpoint is to express the field at an interior point in terms of quantities defined on a closed surface surrounding that point. This boundary representation will provide the mathematical connection between the Helmholtz equation and the diffraction integrals derived in the following sections.

Let V be a volume bounded by a closed surface S, and let $\mathbf{r}_o \in V$ be an observation point. We denote by $\mathbf{r}_s \in S$ a generic point on the boundary and by $\mathbf{n}(\mathbf{r}_s)$ the outward unit normal to S at that point, as illustrated in Fig. 3.1. The derivative of a field along the outward normal direction is defined as

$$\frac{\partial U(\mathbf{r}_s)}{\partial n} = \mathbf{n}(\mathbf{r}_s) \cdot \nabla_{\mathbf{r}_s} U(\mathbf{r}_s). \tag{3.85}$$

The central mathematical object in this representation is the free-space Green function of the Helmholtz operator,

$$\boxed{G(\mathbf{r}_s, \mathbf{r}_o) = \frac{\exp\left(jk\|\mathbf{r}_s - \mathbf{r}_o\|\right)}{\|\mathbf{r}_s - \mathbf{r}_o\|}.} \tag{3.86}$$

Apart from normalization, this function has the form of an outgoing spherical wave generated at $\mathbf{r}_s$ and observed at $\mathbf{r}_o$. It is not a solution of the homogeneous Helmholtz equation at the source point itself. Instead, it satisfies

$$\boxed{\left(\nabla^2_{\mathbf{r}_s} + k^2\right) G(\mathbf{r}_s, \mathbf{r}_o) = -4\pi\delta(\mathbf{r}_s - \mathbf{r}_o).} \tag{3.87}$$

Away from $\mathbf{r}_s = \mathbf{r}_o$, the delta function vanishes and G satisfies the ordinary homogeneous Helmholtz equation.

Why the Green function contains a point source.

Let

$$R = \|\mathbf{r}_s - \mathbf{r}_o\| \tag{3.88}$$

and consider the outgoing spherical wave

$$G(R) = \frac{e^{jkR}}{R}. \tag{3.89}$$

For $R > 0$, a radially symmetric function satisfies

$$\nabla^2 G = \frac{1}{R^2}\frac{d}{dR}\left(R^2 \frac{dG}{dR}\right). \tag{3.90}$$

Since

$$\frac{dG}{dR} = e^{jkR}\left(\frac{jk}{R} - \frac{1}{R^2}\right), \tag{3.91}$$

we have

$$R^2 \frac{dG}{dR} = e^{jkR}(jkR - 1). \tag{3.92}$$

Differentiating once more,

$$\frac{d}{dR}\left[e^{jkR}(jkR - 1)\right] = -k^2 R e^{jkR}, \tag{3.93}$$

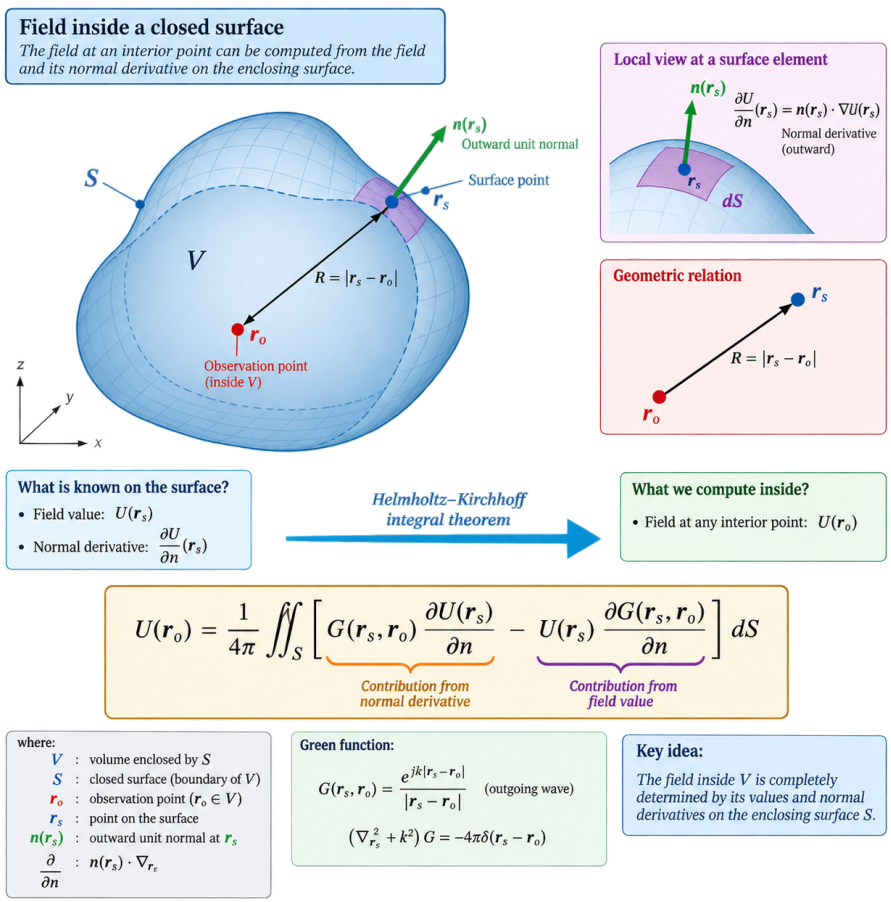


Figure 3.1: Geometrical interpretation of the Helmholtz–Kirchhoff integral theorem. A volume V is enclosed by a closed surface S. The field at an interior observation point $\mathbf{r}_o$ can be expressed in terms of the field value $U(\mathbf{r}_s)$ and its outward normal derivative $\partial U(\mathbf{r}_s)/\partial n$ over the enclosing surface. The Green function $G(\mathbf{r}_s, \mathbf{r}_o)$ describes propagation from each surface point $\mathbf{r}_s$ to the observation point.

and therefore

$$\nabla^2 G = -k^2 G, \qquad R > 0. \tag{3.94}$$

Thus the Green function satisfies the homogeneous Helmholtz equation everywhere except at the source point.

At $R = 0$, however, G is singular. Expanding the exponential,

$$G(R) = \frac{1}{R} + jk + O(R), \tag{3.95}$$

so its singular part is exactly $1/R$. In three dimensions,

$$\nabla^2 \frac{1}{R} = -4\pi\delta(\mathbf{R}). \tag{3.96}$$

To see this, integrate the Laplacian over a sphere V_a of radius a centered at the origin. By the divergence theorem,

$$\iiint_{V_a} \nabla^2 \frac{1}{R}\, dV = \iint_{\partial V_a} \nabla \frac{1}{R} \cdot \mathbf{n}\, dS \tag{3.97}$$

$$= \iint_{\partial V_a} \left(-\frac{1}{a^2}\right) dS \tag{3.98}$$

$$= -\frac{1}{a^2}(4\pi a^2) = -4\pi. \tag{3.99}$$

This result is independent of a, even in the limit $a \to 0$. Therefore the singularity behaves distributionally as a point source of strength -4π, proving Eq. (3.96).

Consequently,

$$\boxed{\left(\nabla^2_{\mathbf{r}_s} + k^2\right) G(\mathbf{r}_s, \mathbf{r}_o) = -4\pi\delta(\mathbf{r}_s - \mathbf{r}_o).} \tag{3.100}$$

If U satisfies the Helmholtz equation throughout V, Green's second identity allows its value at the interior point $\mathbf{r}_o$ to be written as

$$\boxed{U(\mathbf{r}_o) = \frac{1}{4\pi} \iint_S \left[G(\mathbf{r}_s, \mathbf{r}_o) \frac{\partial U(\mathbf{r}_s)}{\partial n} - U(\mathbf{r}_s) \frac{\partial G(\mathbf{r}_s, \mathbf{r}_o)}{\partial n} \right] dS.} \tag{3.101}$$

This expression is the Helmholtz–Kirchhoff integral theorem. It replaces the local differential description of the Helmholtz equation by a boundary description: once the field generated by the physical problem is known on a closed surface, its value at every interior point can be reconstructed from the field and its normal derivative over that surface.

The appearance of both U and $\partial U/\partial n$ is closely related to the fact that the Helmholtz equation is a second-order differential equation. The field value

$$U(\mathbf{r}_s) \tag{3.102}$$

specifies the complex amplitude and phase of the wave at the boundary, whereas

$$\frac{\partial U(\mathbf{r}_s)}{\partial n} \tag{3.103}$$

describes how that complex field changes when moving through the boundary in the direction normal to the surface. The two quantities therefore contain complementary information about the local state of the wave at S.

The two terms in Eq. (3.101) also admit a useful equivalent-source interpretation. Consider first

$$G(\mathbf{r}_s, \mathbf{r}_o)\frac{\partial U(\mathbf{r}_s)}{\partial n}. \tag{3.104}$$

Since G has the form of the spherical field produced by a point source, this term may be interpreted as the field produced by an equivalent point source located at $\mathbf{r}_s$, whose strength is proportional to $\partial U/\partial n$. Integration over S therefore corresponds to a continuous sheet of such point sources. In boundary-integral terminology, this is known as a *single-layer potential*, or equivalently as a distribution of monopole-like sources over the boundary.

The second contribution,

$$-U(\mathbf{r}_s)\frac{\partial G(\mathbf{r}_s, \mathbf{r}_o)}{\partial n}, \tag{3.105}$$

has a different interpretation. The normal derivative of the Green function can be written as

$$\frac{\partial G}{\partial n} = \lim_{\Delta n \to 0} \frac{G(\mathbf{r}_s + \Delta n\, \mathbf{n}, \mathbf{r}_o) - G(\mathbf{r}_s, \mathbf{r}_o)}{\Delta n}. \tag{3.106}$$

It therefore represents the limiting difference between two nearby point sources displaced along the normal direction. Such a pair behaves as a dipole. Consequently, the second term in Eq. (3.101) can be regarded as a continuous distribution of equivalent dipoles over S, whose strengths are proportional to the boundary value $U(\mathbf{r}_s)$. This contribution is known as a *double-layer potential*.

It is important not to interpret these monopoles and dipoles as actual physical sources located on S. The enclosing surface is an arbitrary mathematical construction and may lie entirely within a region containing no

physical sources. The Helmholtz–Kirchhoff theorem states instead that, as far as the field inside V is concerned, the effect of the physical sources and propagation history outside the surface can be represented exactly by an equivalent distribution of monopole- and dipole-like sources on its boundary.

This equivalent-source viewpoint provides a precise mathematical interpretation of Huygens' principle. Rather than following the wave all the way from its original physical source, we may replace the exterior problem by secondary sources distributed over a surface through which the wave has already propagated. Their superposition reconstructs the field at subsequent observation points.

The Helmholtz–Kirchhoff theorem is therefore the natural bridge between the Helmholtz differential equation and diffraction theory. In the following sections, suitable choices of the integration surface together with appropriate boundary conditions will transform Eq. (3.101) into practical diffraction and propagation integrals, culminating in the Huygens–Fresnel principle.

Derivation of the Helmholtz–Kirchhoff integral theorem.

For two sufficiently regular scalar functions U and G, Green's second identity states that

$$\iiint_V \left(U\nabla^2 G - G\nabla^2 U\right)\, dV = \iint_S \left(U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right)\, dS. \tag{3.107}$$

The field U satisfies the homogeneous Helmholtz equation,

$$\nabla^2 U = -k^2 U, \tag{3.108}$$

whereas the Green function satisfies

$$\nabla^2 G = -k^2 G - 4\pi\delta(\mathbf{r}_s - \mathbf{r}_o). \tag{3.109}$$

Substituting both expressions into the volume integral on the left-hand

side of Eq. (3.107) gives

$$\iiint_V \left(U\nabla^2 G - G\nabla^2 U\right)\, dV = \iiint_V \left[U\left(-k^2 G - 4\pi\delta(\mathbf{r}_s - \mathbf{r}_o)\right) - G(-k^2 U)\right]\, dV \tag{3.110}$$

$$= -4\pi \iiint_V U(\mathbf{r}_s)\delta(\mathbf{r}_s - \mathbf{r}_o)\, dV. \tag{3.111}$$

The two terms proportional to $k^2 UG$ cancel exactly. Using the sifting property of the Dirac delta,

$$\iiint_V U(\mathbf{r}_s)\delta(\mathbf{r}_s - \mathbf{r}_o)\, dV = U(\mathbf{r}_o), \tag{3.112}$$

provided $\mathbf{r}_o$ lies inside V. Therefore,

$$-4\pi U(\mathbf{r}_o) = \iint_S \left(U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right)\, dS. \tag{3.113}$$

Multiplying by $-1/(4\pi)$ gives

$$U(\mathbf{r}_o) = \frac{1}{4\pi}\iint_S \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n}\right)\, dS, \tag{3.114}$$

which is Eq. (3.101).

The principal results of this section that will be used in the subsequent development are therefore

$$\omega = 2\pi\nu, \tag{3.115}$$

$$k = \frac{2\pi}{\lambda}, \tag{3.116}$$

$$\omega = vk, \tag{3.117}$$

$$v = \lambda\nu, \tag{3.118}$$

$$u(\mathbf{r}, t) = \mathrm{Re}\left\{U(\mathbf{r})e^{-j\omega t}\right\}, \tag{3.119}$$

$$(\nabla^2 + k^2)U = 0. \tag{3.120}$$

The last equation is particularly important: from this point onward, the propagation of a monochromatic wave can be studied as a purely spatial problem. The Helmholtz–Kirchhoff theorem then allows this local differential description to be replaced by an integral description in which the field inside a region is determined by the field on its boundary.

3.4 Connecting classical optics to quantum mechanics

The diffraction theory developed so far originated from Maxwell's equations and therefore describes classical electromagnetic waves. An electron, however, is not an electromagnetic wave: it is a quantum particle with mass, charge, and spin. Before applying the scalar diffraction theory developed above to electron microscopy, we must therefore establish why the same mathematical equations can describe the spatial propagation of an electron wave.

The appropriate relativistic description of a free electron is provided by the Dirac equation,

$$\boxed{j\hbar\frac{\partial \boldsymbol{\Psi}(\mathbf{r},t)}{\partial t} = \left(-j\hbar c\,\boldsymbol{\alpha}\cdot\nabla + \beta m_0 c^2\right)\boldsymbol{\Psi}(\mathbf{r},t).} \tag{3.121}$$

Here, $\boldsymbol{\Psi}(\mathbf{r},t)$ is a four-component spinor, m_0 is the electron rest mass, $\hbar = h/(2\pi)$ is the reduced Planck constant, and

$$\boldsymbol{\alpha} = (\alpha_x, \alpha_y, \alpha_z) \tag{3.122}$$

together with β are 4×4 matrices known as the Dirac matrices.

The four components of $\boldsymbol{\Psi}$ arise from the relativistic description of a spin-1/2 particle. They contain the two possible spin states together with the positive- and negative-energy sectors associated with electrons and positrons. Thus, unlike the scalar field U used in the previous sections, the relativistic electron wave function possesses an internal spinor structure.

The operator appearing on the right-hand side of Eq. (3.121) occurs so frequently that it is given its own symbol,

$$\boxed{H_D = -j\hbar c\,\boldsymbol{\alpha}\cdot\nabla + \beta m_0 c^2.} \tag{3.123}$$

It is called the *Dirac Hamiltonian*. With this notation, the Dirac equation becomes simply

$$\boxed{j\hbar\frac{\partial \boldsymbol{\Psi}}{\partial t} = H_D\boldsymbol{\Psi}.} \tag{3.124}$$

The word *Hamiltonian* is used throughout quantum mechanics for the operator associated with the total energy of a system. More importantly for our purposes, the Hamiltonian determines how the quantum state evolves in time. Equation (3.124) states that the temporal rate of change of the

electron wave function is determined by the action of H_D on that wave function.

It is important to distinguish an operator from an ordinary number. The Hamiltonian H_D contains the spatial differential operator ∇. Consequently, writing

$$H_D \mathbf{\Psi} \tag{3.125}$$

means that the derivatives and matrices contained in H_D act on the spinor $\mathbf{\Psi}$. Likewise,

$$H_D^2 \mathbf{\Psi} \equiv H_D \left(H_D \mathbf{\Psi} \right) \tag{3.126}$$

means that the same operator is applied twice.

The Dirac matrices are constructed so that they satisfy the anticommutation relations

$$\alpha_i \alpha_j + \alpha_j \alpha_i = 2\delta_{ij} I, \tag{3.127}$$

$$\alpha_i \beta + \beta \alpha_i = 0, \tag{3.128}$$

$$\beta^2 = I, \tag{3.129}$$

where I is the 4×4 identity matrix and δ_{ij} is the Kronecker delta.

These particular algebraic properties have an important consequence. Although the Dirac equation is a first-order matrix differential equation, every component of a free Dirac spinor also satisfies the second-order scalar equation

$$\boxed{\frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} - \nabla^2 \psi + \frac{m_0^2 c^2}{\hbar^2} \psi = 0.} \tag{3.130}$$

This has the form of the free Klein–Gordon equation. The derivation below shows explicitly how this equation follows from the Dirac equation.

From the Dirac equation to a second-order scalar equation.

We start from the compact form of the free Dirac equation,

$$j\hbar \frac{\partial \mathbf{\Psi}}{\partial t} = H_D \mathbf{\Psi}. \tag{3.131}$$

Our objective is to obtain an equation containing a second derivative with respect to time. We therefore apply the operator $j\hbar\, \partial/\partial t$ to both sides:

$$j\hbar \frac{\partial}{\partial t} \left(j\hbar \frac{\partial \mathbf{\Psi}}{\partial t} \right) = j\hbar \frac{\partial}{\partial t} \left(H_D \mathbf{\Psi} \right). \tag{3.132}$$

The left-hand side is immediately

$$j^2\hbar^2\frac{\partial^2\mathbf{\Psi}}{\partial t^2} = -\hbar^2\frac{\partial^2\mathbf{\Psi}}{\partial t^2}. \tag{3.133}$$

The right-hand side requires a little more care. In general, if an operator H itself depends on time, the product rule gives

$$\frac{\partial}{\partial t}\left(H\mathbf{\Psi}\right) = \frac{\partial H}{\partial t}\mathbf{\Psi} + H\frac{\partial\mathbf{\Psi}}{\partial t}. \tag{3.134}$$

For the free Dirac Hamiltonian,

$$H_D = -j\hbar c\,\boldsymbol{\alpha}\cdot\nabla + \beta m_0 c^2, \tag{3.135}$$

none of its ingredients depends explicitly on time: m_0, c, and the Dirac matrices are constants, while ∇ differentiates only with respect to the spatial coordinates. Therefore,

$$\frac{\partial H_D}{\partial t} = 0. \tag{3.136}$$

Equation (3.134) then reduces to

$$\frac{\partial}{\partial t}\left(H_D\mathbf{\Psi}\right) = H_D\frac{\partial\mathbf{\Psi}}{\partial t}. \tag{3.137}$$

Consequently, the right-hand side of Eq. (3.132) becomes

$$j\hbar H_D\frac{\partial\mathbf{\Psi}}{\partial t} = H_D\left(j\hbar\frac{\partial\mathbf{\Psi}}{\partial t}\right). \tag{3.138}$$

We may now use the original Dirac equation, Eq. (3.131), once again:

$$j\hbar\frac{\partial\mathbf{\Psi}}{\partial t} = H_D\mathbf{\Psi}. \tag{3.139}$$

Therefore,

$$H_D\left(j\hbar\frac{\partial\mathbf{\Psi}}{\partial t}\right) = H_D\left(H_D\mathbf{\Psi}\right) = H_D^2\mathbf{\Psi}. \tag{3.140}$$

Combining both sides gives

$$\boxed{-\hbar^2 \frac{\partial^2 \mathbf{\Psi}}{\partial t^2} = H_D^2 \mathbf{\Psi}.} \tag{3.141}$$

We must therefore determine the square of the Dirac Hamiltonian,

$$H_D^2 = \left(-j\hbar c\, \boldsymbol{\alpha} \cdot \nabla + \beta m_0 c^2\right)^2. \tag{3.142}$$

Expanding the square gives

$$\begin{aligned} H_D^2 = &-\hbar^2 c^2 (\boldsymbol{\alpha} \cdot \nabla)^2 \\ &- j\hbar m_0 c^3 \left[(\boldsymbol{\alpha} \cdot \nabla)\beta + \beta(\boldsymbol{\alpha} \cdot \nabla)\right] + \beta^2 m_0^2 c^4. \end{aligned} \tag{3.143}$$

Because the Dirac matrices are independent of position, the mixed term can be written as

$$\sum_i (\alpha_i \beta + \beta \alpha_i) \frac{\partial}{\partial x_i}. \tag{3.144}$$

According to Eq. (3.128),

$$\alpha_i \beta + \beta \alpha_i = 0, \tag{3.145}$$

so the complete mixed term vanishes.

We now consider the first term,

$$(\boldsymbol{\alpha} \cdot \nabla)^2 = \sum_{i,j} \alpha_i \alpha_j \frac{\partial^2}{\partial x_i \partial x_j}. \tag{3.146}$$

Since ordinary spatial derivatives commute,

$$\frac{\partial^2}{\partial x_i \partial x_j} = \frac{\partial^2}{\partial x_j \partial x_i}, \tag{3.147}$$

only the symmetric part of the matrix product contributes. Thus,

$$(\boldsymbol{\alpha}\cdot\nabla)^2 = \frac{1}{2}\sum_{i,j}(\alpha_i\alpha_j+\alpha_j\alpha_i)\frac{\partial^2}{\partial x_i\partial x_j} \tag{3.148}$$

$$= \sum_{i,j}\delta_{ij}I\frac{\partial^2}{\partial x_i\partial x_j} \tag{3.149}$$

$$= I\nabla^2. \tag{3.150}$$

Using also $\beta^2 = I$, we obtain

$$\boxed{H_D^2 = \left(-\hbar^2c^2\nabla^2 + m_0^2c^4\right)I.} \tag{3.151}$$

Substitution into Eq. (3.141) gives

$$-\hbar^2\frac{\partial^2\boldsymbol{\Psi}}{\partial t^2} = \left(-\hbar^2c^2\nabla^2 + m_0^2c^4\right)\boldsymbol{\Psi}. \tag{3.152}$$

The operator on the right-hand side is now proportional to the identity matrix. Consequently, it acts in exactly the same way on each of the four components of $\boldsymbol{\Psi}$. If ψ denotes any one of these components, then

$$-\hbar^2\frac{\partial^2\psi}{\partial t^2} = -\hbar^2c^2\nabla^2\psi + m_0^2c^4\psi. \tag{3.153}$$

Dividing by $\hbar^2c^2$ and rearranging gives

$$\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi + \frac{m_0^2c^2}{\hbar^2}\psi = 0, \tag{3.154}$$

which is Eq. (3.130).

The fact that the second-order equation is scalar does not mean that the spin structure of the Dirac theory has disappeared physically. Rather, it means that every component of a free Dirac spinor must individually satisfy the same second-order differential equation. The original Dirac equation still imposes relationships between those components.

Consequently, every solution of the Dirac equation satisfies Eq. (3.130) component by component, but the converse is not true: four arbitrary solutions of the Klein–Gordon equation cannot in general be assembled into a valid Dirac spinor. Squaring the Dirac equation has produced a necessary

second-order condition while discarding some of the information contained in the original first-order equation.

For our present purpose, however, this second-order equation is precisely what we need. It allows us to isolate the spatial propagation of a free electron and compare it directly with the Helmholtz equation obtained previously from classical wave optics.

3.4.1 The stationary free electron

Consider a free electron with a well-defined total relativistic energy E. A state of definite energy has a temporal dependence of the form

$$\boxed{\psi(\mathbf{r},t) = \psi_E(\mathbf{r}) \exp\left(-j\frac{E}{\hbar}t\right),} \tag{3.155}$$

where $\psi_E(\mathbf{r})$ contains the spatial dependence of the state.

The origin of this exponential time dependence can be seen directly from the quantum evolution equation. If a state has a definite energy E, it is an eigenstate of the Hamiltonian,

$$H_D \mathbf{\Psi}_E = E\mathbf{\Psi}_E. \tag{3.156}$$

The Dirac equation then implies

$$j\hbar\frac{\partial \mathbf{\Psi}_E}{\partial t} = E\mathbf{\Psi}_E. \tag{3.157}$$

At every spatial position this is an ordinary first-order differential equation in time, whose solution is

$$\mathbf{\Psi}_E(\mathbf{r},t) = \mathbf{\Psi}_E(\mathbf{r},0)e^{-jEt/\hbar}. \tag{3.158}$$

Thus a state with a precisely defined energy changes in time only through a complex phase factor.

Substituting Eq. (3.155) into Eq. (3.130) gives

$$\boxed{\left(\nabla^2 + k^2\right)\psi_E(\mathbf{r}) = 0,} \tag{3.159}$$

where

$$\boxed{k^2 = \frac{E^2 - m_0^2c^4}{\hbar^2c^2}.} \tag{3.160}$$

Using the relativistic energy–momentum relation

$$E^2 = p^2c^2 + m_0^2c^4, \tag{3.161}$$

we obtain

$$k^2 = \frac{p^2}{\hbar^2}, \tag{3.162}$$

and therefore

$$\boxed{k = \frac{p}{\hbar} = \frac{2\pi}{\lambda}.} \tag{3.163}$$

The last equality follows directly from the de Broglie relation $\lambda = h/p$.

We have therefore arrived at the central connection of this section. A monochromatic classical scalar wave satisfies

$$(\nabla^2 + k^2)U(\mathbf{r}) = 0, \tag{3.164}$$

while the spatial part of a stationary free-electron state satisfies

$$(\nabla^2 + k^2)\psi_E(\mathbf{r}) = 0. \tag{3.165}$$

The two quantities have completely different physical meanings: $U(\mathbf{r})$ is a classical field amplitude, whereas $\psi_E(\mathbf{r})$ is a quantum-mechanical probability amplitude. Nevertheless, their propagation through a homogeneous, field-free region is governed by exactly the same mathematical operator.

This mathematical equivalence is the reason why the machinery of scalar wave optics can be transferred to electron microscopy. Plane-wave decompositions, Green functions, diffraction integrals, Fresnel propagation, Fourier propagation, and phase transformations can all be applied to electron waves, provided that the relevant wavenumber is the electron wavenumber

$$k = \frac{p}{\hbar} = \frac{2\pi}{\lambda}. \tag{3.166}$$

No physical identification between an electron and an electromagnetic wave is being made: what is shared is the mathematical equation governing their spatial propagation.

Derivation of the stationary electron Helmholtz equation.

Starting from

$$\psi(\mathbf{r}, t) = \psi_E(\mathbf{r})e^{-jEt/\hbar}, \tag{3.167}$$

the second temporal derivative is

$$\frac{\partial^2 \psi}{\partial t^2} = -\frac{E^2}{\hbar^2}\psi_E(\mathbf{r})e^{-jEt/\hbar}. \tag{3.168}$$

Since the temporal exponential does not depend on position,

$$\nabla^2 \psi = e^{-jEt/\hbar}\nabla^2 \psi_E. \tag{3.169}$$

Substitution into Eq. (3.130) gives

$$-\frac{E^2}{\hbar^2 c^2}\psi_E e^{-jEt/\hbar} - (\nabla^2 \psi_E)e^{-jEt/\hbar} + \frac{m_0^2 c^2}{\hbar^2}\psi_E e^{-jEt/\hbar} = 0. \tag{3.170}$$

Dividing by the non-zero temporal factor and rearranging,

$$\nabla^2 \psi_E + \frac{E^2 - m_0^2 c^4}{\hbar^2 c^2}\psi_E = 0. \tag{3.171}$$

Using

$$E^2 - m_0^2 c^4 = p^2 c^2 \tag{3.172}$$

gives

$$\nabla^2 \psi_E + \frac{p^2}{\hbar^2}\psi_E = 0. \tag{3.173}$$

Defining

$$k = \frac{p}{\hbar} \tag{3.174}$$

yields

$$(\nabla^2 + k^2)\psi_E = 0. \tag{3.175}$$

Finally, since $p = h/\lambda$ and $h = 2\pi\hbar$,

$$k = \frac{h}{\hbar\lambda} = \frac{2\pi}{\lambda}. \tag{3.176}$$

3.4.2 The non-relativistic limit

Although the relativistic treatment developed above is the appropriate one for the high-energy electrons used in cryo-electron microscopy, it is useful to see how the familiar Schrödinger equation emerges as the non-relativistic limit of the same theory.

We start from the second-order equation satisfied by each component of

a free Dirac spinor,

$$\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi + \frac{m_0^2c^2}{\hbar^2}\psi = 0. \tag{3.177}$$

A positive-energy electron has a total energy that can be written as

$$E = m_0c^2 + E_k, \tag{3.178}$$

where m_0c^2 is its rest energy and E_k its kinetic energy. The corresponding temporal phase can therefore be separated as

$$e^{-jEt/\hbar} = e^{-jm_0c^2t/\hbar}e^{-jE_kt/\hbar}. \tag{3.179}$$

This motivates writing the wave function in the form

$$\boxed{\psi(\mathbf{r},t) = \varphi(\mathbf{r},t)e^{-jm_0c^2t/\hbar},} \tag{3.180}$$

where the rapidly oscillating phase associated with the rest energy has been factored out. This factorization is exact; no approximation has yet been introduced.

Substitution of Eq. (3.180) into Eq. (3.177) gives

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} - \frac{2jm_0}{\hbar}\frac{\partial\varphi}{\partial t} - \nabla^2\varphi = 0. \tag{3.181}$$

The advantage of the factorization is now apparent: the large term associated directly with the rest energy has disappeared.

In the non-relativistic regime,

$$\boxed{E_k \ll m_0c^2.} \tag{3.182}$$

As established in Section 2.4, an electron accelerated through a potential difference V acquires a kinetic energy

$$E_k = eV. \tag{3.183}$$

For the accelerating voltages commonly used in cryo-electron microscopy, this kinetic energy is not negligible compared with the electron rest energy,

$$m_0c^2 \simeq 511 \text{ keV}. \tag{3.184}$$

For example, at $V = 300$ kV,

$$\frac{E_k}{m_0c^2} = \frac{300 \text{ keV}}{511 \text{ keV}} \simeq 0.59, \tag{3.185}$$

which is clearly not $\ll 1$. This immediately shows why a purely non-relativistic treatment is inadequate for modern cryo-EM electrons. Nevertheless, considering this limit is useful because it shows how the familiar Schrödinger equation emerges from the relativistic theory.

After removing the rest-energy oscillation, the remaining function φ varies on the slower time scale associated with E_k. In the formal non-relativistic limit of Eq. (3.182), the first term in Eq. (3.181), containing $\partial^2\varphi/\partial t^2$, is much smaller than the term containing $\partial\varphi/\partial t$ and may therefore be neglected. We then obtain

$$-\frac{2jm_0}{\hbar}\frac{\partial\varphi}{\partial t} - \nabla^2\varphi \simeq 0, \tag{3.186}$$

or, equivalently,

$$\boxed{j\hbar\frac{\partial\varphi}{\partial t} = -\frac{\hbar^2}{2m_0}\nabla^2\varphi.} \tag{3.187}$$

This is the time-dependent Schrödinger equation for a free particle.

Thus, the Schrödinger equation is not an independent description unrelated to relativistic quantum mechanics. It emerges from the relativistic equation when the kinetic energy is small compared with the electron rest energy and the rapid rest-energy phase is removed.

This limit is conceptually useful, but it is not the approximation on which our treatment of cryo-electron microscopy will be based. Since $m_0c^2 \simeq$ 511 keV, electrons accelerated through 100–300 kV are sufficiently energetic that relativistic corrections are important. The direct relativistic derivation of the electron Helmholtz equation in Section 3.4.1 is therefore the appropriate connection to scalar diffraction theory.

Derivation of the non-relativistic limit.

Starting from

$$\psi(\mathbf{r},t) = \varphi(\mathbf{r},t)e^{-jm_0c^2t/\hbar}, \tag{3.188}$$

the first temporal derivative is

$$\frac{\partial\psi}{\partial t} = e^{-jm_0c^2t/\hbar}\left(\frac{\partial\varphi}{\partial t} - j\frac{m_0c^2}{\hbar}\varphi\right), \tag{3.189}$$

and the second derivative is

$$\frac{\partial^2\psi}{\partial t^2} = e^{-jm_0c^2t/\hbar}\left[\frac{\partial^2\varphi}{\partial t^2} - 2j\frac{m_0c^2}{\hbar}\frac{\partial\varphi}{\partial t} - \frac{m_0^2c^4}{\hbar^2}\varphi\right]. \tag{3.190}$$

Since the exponential depends only on time,

$$\nabla^2\psi = e^{-jm_0c^2t/\hbar}\nabla^2\varphi. \tag{3.191}$$

Substituting these expressions into Eq. (3.177), the terms proportional to $m_0^2c^2/\hbar^2$ cancel exactly, leaving

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} - \frac{2jm_0}{\hbar}\frac{\partial\varphi}{\partial t} - \nabla^2\varphi = 0. \tag{3.192}$$

We now estimate the relative magnitude of the two temporal terms. If φ varies on the characteristic time scale associated with the kinetic energy E_k, then

$$\left|\frac{\partial\varphi}{\partial t}\right| \sim \frac{E_k}{\hbar}|\varphi| \tag{3.193}$$

and

$$\left|\frac{\partial^2\varphi}{\partial t^2}\right| \sim \frac{E_k^2}{\hbar^2}|\varphi|. \tag{3.194}$$

The ratio between the first and second temporal terms in Eq. (3.181) is therefore of order

$$\frac{|c^{-2}\partial_t^2\varphi|}{|(2m_0/\hbar)\partial_t\varphi|} \sim \frac{E_k}{2m_0c^2}. \tag{3.195}$$

Thus, when

$$E_k \ll m_0c^2, \tag{3.196}$$

the second-order temporal derivative is negligible. Equation (3.181) consequently reduces to

$$-\frac{2jm_0}{\hbar}\frac{\partial\varphi}{\partial t} - \nabla^2\varphi \simeq 0. \tag{3.197}$$

Multiplying by $-\hbar^2/(2m_0)$ gives

$$j\hbar\frac{\partial\varphi}{\partial t} = -\frac{\hbar^2}{2m_0}\nabla^2\varphi, \tag{3.198}$$

which is the free-particle Schrödinger equation.

3.4.3 Why a stationary description is useful in electron microscopy

The use of a stationary wave function may initially appear surprising. A physical electron is emitted by the source, traverses the microscope, interacts with the specimen, and is eventually detected at a particular location. An individual electron may therefore be described more generally as a time-dependent wave packet rather than as an infinitely extended stationary wave.

Nevertheless, stationary scattering states are the natural description for image-formation theory when the illuminating electrons have a narrowly defined energy and the specimen and microscope potentials can be regarded as time independent during the passage of an electron. Under these conditions, each energy component evolves only through a global temporal phase,

$$\psi(\mathbf{r}, t) = \psi_E(\mathbf{r}) e^{-jEt/\hbar}. \tag{3.199}$$

Consequently, its probability density is independent of time:

$$|\psi(\mathbf{r}, t)|^2 = |\psi_E(\mathbf{r})|^2. \tag{3.200}$$

Likewise, the probability current associated with a stationary scattering state is time independent. The detector therefore does not need to resolve the extremely rapid factor $e^{-jEt/\hbar}$; it records discrete electron arrival events whose spatial distribution is governed by the stationary scattering solution.

A cryo-EM exposure consists of many such electron detection events. At the beam currents relevant to structural imaging, electron–electron interactions within the imaging process are normally neglected, so each arrival can be treated as an independent realization of the same single-electron scattering experiment, apart from source-energy spread, illumination coherence, specimen motion, radiation damage, and other effects that will be introduced later.

The stationary formulation therefore does not assert that an individual electron is literally an infinite monochromatic wave extending through all space. Rather, it isolates the energy-resolved spatial scattering problem that determines the probabilities and fluxes measured experimentally. A wave-packet description may always be recovered by superposing stationary solutions with different energies.

This distinction will become especially important when partial temporal coherence is introduced. A perfectly monochromatic beam corresponds to a single energy and hence a single stationary solution, whereas a realistic source possesses a distribution of electron energies. The observed image

must then account for the corresponding ensemble of stationary energy components.

3.5 Wave diffraction by an aperture

We now apply the Helmholtz–Kirchhoff integral theorem to one of the central problems of wave optics: the diffraction of a wave by an aperture in an otherwise opaque planar screen.

The geometry is shown in Fig. 3.2. A source located at $\mathbf{r}_i$ illuminates an aperture Σ contained in a planar screen. A generic point of the aperture is denoted by $\mathbf{r}_s$, and the field is to be calculated at an observation point $\mathbf{r}_o$ on the opposite side of the screen. We define the distances

$$r_{is} = \|\mathbf{r}_s - \mathbf{r}_i\|, \qquad r_{os} = \|\mathbf{r}_o - \mathbf{r}_s\|. \tag{3.201}$$

Let $\hat{\mathbf{z}}$ denote the unit normal to the aperture plane pointing from the source side toward the observation side. We define θ_i as the angle between the incident propagation direction at $\mathbf{r}_s$ and $\hat{\mathbf{z}}$, and θ_o as the angle between the direction from $\mathbf{r}_s$ to $\mathbf{r}_o$ and $\hat{\mathbf{z}}$.

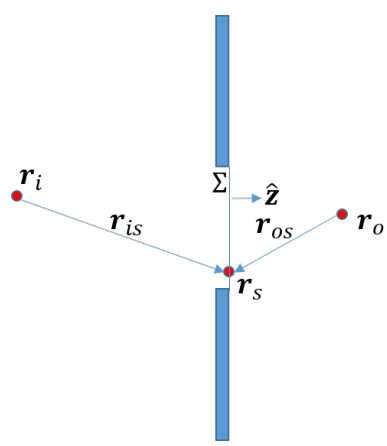


Figure 3.2: Diffraction by an aperture Σ in an opaque planar screen. A spherical wave originating at $\mathbf{r}_i$ illuminates a generic aperture point $\mathbf{r}_s$, and the diffracted field is evaluated at $\mathbf{r}_o$. The distances from the source to the aperture and from the aperture to the observation point are denoted by r_{is} and r_{os}, respectively. The unit vector $\hat{\mathbf{z}}$ is normal to the screen and points toward the observation half-space.

To use the Helmholtz–Kirchhoff theorem, the observation region is enclosed by a surface consisting of the screen plane together with a large hemisphere extending into the observation half-space. The Sommerfeld radiation condition,

$$\lim_{r\to\infty} r\left(\frac{\partial U}{\partial r} - jkU\right) = 0, \tag{3.202}$$

selects waves propagating outward rather than waves arriving from infinity. Under this condition, the contribution from the distant hemisphere vanishes as its radius tends to infinity, leaving only the contribution from the plane containing the aperture.

The difficulty is that the exact field and its normal derivative on that plane are not known once the screen is introduced. Kirchhoff proposed the following approximation:

1. Inside the aperture Σ, both the field and its normal derivative are assumed to be equal to those that would exist if the screen were absent.

2. Immediately behind the opaque part of the screen, both the field and its normal derivative are assumed to vanish.

These assumptions are known as the *Kirchhoff boundary conditions*. They provide a practical way of converting the general Helmholtz–Kirchhoff representation into a diffraction integral.

3.5.1 Diffraction of a spherical wave

Consider first a monochromatic spherical wave of amplitude A emitted by a point source at $\mathbf{r}_i$. In the absence of the screen, the field at the aperture would be

$$U_i(\mathbf{r}_s) = A\frac{e^{jkr_{is}}}{r_{is}}. \tag{3.203}$$

The Green function that propagates a contribution from an aperture point $\mathbf{r}_s$ to the observation point $\mathbf{r}_o$ is

$$G(\mathbf{r}_s, \mathbf{r}_o) = \frac{e^{jkr_{os}}}{r_{os}}. \tag{3.204}$$

When both r_{is} and r_{os} are large compared with the wavelength,

$$kr_{is} \gg 1, \qquad kr_{os} \gg 1, \tag{3.205}$$

the terms proportional to $1/r^2$ in the normal derivatives of the spherical waves are negligible compared with those proportional to k/r. Applying

the Kirchhoff boundary conditions to the Helmholtz–Kirchhoff integral then gives

$$\boxed{U(\mathbf{r}_o) = \frac{A}{j\lambda} \iint_\Sigma \frac{e^{jk(r_{is}+r_{os})}}{r_{is} r_{os}} K_K(\theta_i, \theta_o)\, dS,} \tag{3.206}$$

where

$$\boxed{K_K(\theta_i, \theta_o) = \frac{1}{2}\left(\cos\theta_i + \cos\theta_o\right)} \tag{3.207}$$

is the *Kirchhoff obliquity factor*.

The two factors $e^{jkr_{is}}/r_{is}$ and $e^{jkr_{os}}/r_{os}$ have a simple interpretation. The first describes propagation from the original source to the aperture point, while the second describes propagation from that aperture point to the observation point. The obliquity factor accounts for the dependence of the contribution on the directions of incidence and observation.

From the Helmholtz–Kirchhoff theorem to the aperture integral.

The Helmholtz–Kirchhoff theorem gives

$$U(\mathbf{r}_o) = \frac{1}{4\pi} \iint_S \left[G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n} \right] dS. \tag{3.208}$$

After applying the Sommerfeld radiation condition, the distant part of the closed integration surface does not contribute. Kirchhoff's boundary conditions furthermore make the contribution from the opaque part of the screen vanish, so only the aperture Σ remains.

For the incident spherical wave,

$$U_i = A \frac{e^{jkr_{is}}}{r_{is}}, \tag{3.209}$$

we have

$$\frac{\partial U_i}{\partial r_{is}} = \left(jk - \frac{1}{r_{is}} \right) U_i. \tag{3.210}$$

Similarly,

$$\frac{\partial G}{\partial r_{os}} = \left(jk - \frac{1}{r_{os}} \right) G. \tag{3.211}$$

The normal derivatives introduce the projections of the corresponding propagation directions onto the normal to the integration surface.

At the aperture, the outward normal of the closed observation volume points opposite to the forward direction $\hat{\mathbf{z}}$. Taking this sign into account gives, when $kr_{is} \gg 1$ and $kr_{os} \gg 1$,

$$\frac{\partial U_i}{\partial n} \simeq -jk\cos\theta_i\, U_i \tag{3.212}$$

and

$$\frac{\partial G}{\partial n} \simeq jk\cos\theta_o\, G. \tag{3.213}$$

Therefore,

$$G\frac{\partial U_i}{\partial n} - U_i\frac{\partial G}{\partial n} \simeq -jkU_iG\left(\cos\theta_i + \cos\theta_o\right). \tag{3.214}$$

Substitution into Eq. (3.208) gives

$$U(\mathbf{r}_o) \simeq -\frac{jk}{4\pi}\iint_\Sigma U_i(\mathbf{r}_s)G(\mathbf{r}_s,\mathbf{r}_o)\left(\cos\theta_i + \cos\theta_o\right)\, dS. \tag{3.215}$$

Since

$$k = \frac{2\pi}{\lambda} \tag{3.216}$$

and $1/j = -j$, we have

$$-\frac{jk}{4\pi} = \frac{1}{2j\lambda}. \tag{3.217}$$

Thus,

$$U(\mathbf{r}_o) = \frac{1}{j\lambda}\iint_\Sigma U_i(\mathbf{r}_s)G(\mathbf{r}_s,\mathbf{r}_o)\frac{\cos\theta_i + \cos\theta_o}{2}\, dS, \tag{3.218}$$

which yields Eq. (3.206).

3.5.2 The Huygens–Fresnel interpretation

Equation (3.206) has a particularly important physical interpretation. It may be written as

$$U(\mathbf{r}_o) = \iint_\Sigma U_{\text{sec}}(\mathbf{r}_s;\mathbf{r}_o)\frac{e^{jkr_{os}}}{r_{os}}\, dS, \tag{3.219}$$

where

$$U_{\text{sec}}(\mathbf{r}_s;\mathbf{r}_o) = \frac{1}{j\lambda}U_i(\mathbf{r}_s)K_K(\theta_i,\theta_o). \tag{3.220}$$

Thus, every point of the aperture can be regarded as an *equivalent secondary source* emitting a spherical wave toward the observation point. The complex amplitude of each secondary wave is determined by the field illuminating that point of the aperture together with the corresponding angular factor.

The field at $\mathbf{r}_o$ is obtained by coherently adding all these secondary spherical waves. Because the contributions are complex, this superposition includes both constructive and destructive interference. Diffraction therefore arises naturally from the different phases accumulated along the paths passing through different points of the aperture.

This is the mathematical content of the Huygens–Fresnel principle: a wavefront may be replaced by a continuous distribution of secondary wavelets, and the propagated field is obtained from their coherent superposition.

3.5.3 Plane-wave illumination

A particularly important limiting case occurs when the source is very far from the aperture. The incident spherical wave then becomes locally indistinguishable from a plane wave.

For normal incidence,

$$\theta_i = 0, \qquad \cos\theta_i = 1, \tag{3.221}$$

and the Kirchhoff obliquity factor becomes

$$\boxed{K_K(\theta_o) = \frac{1}{2}\left(1 + \cos\theta_o\right).} \tag{3.222}$$

Therefore, if the complex amplitude over the aperture is $U(\mathbf{r}_s)$, the Kirchhoff diffraction formula takes the form

$$U(\mathbf{r}_o) = \frac{1}{j\lambda}\iint_\Sigma U(\mathbf{r}_s)\frac{e^{jkr_{os}}}{r_{os}}\frac{1+\cos\theta_o}{2}\,dS. \tag{3.223}$$

For observation directions close to the optical axis, $\theta_o \ll 1$, and hence

$$\cos\theta_o \simeq 1. \tag{3.224}$$

The obliquity factor is then approximately unity,

$$K_K(\theta_o) \simeq 1, \tag{3.225}$$

which leads to the particularly simple Huygens–Fresnel expression

$$\boxed{U(\mathbf{r}_o) \simeq \frac{1}{j\lambda}\iint_\Sigma U(\mathbf{r}_s)\frac{e^{jkr_{os}}}{r_{os}}\,dS.} \tag{3.226}$$

3.5.4 Propagation as a linear integral operator

The previous expressions reveal a more general principle. Suppose that the complex field immediately across the aperture is an arbitrary function $U(\mathbf{r}_s)$. Because the Helmholtz equation is linear, the contribution produced by every infinitesimal aperture element may be calculated independently and the resulting fields added coherently.

We may therefore write the propagated field in the general form

$$\boxed{U(\mathbf{r}_o) = \iint_\Sigma U(\mathbf{r}_s) h_\lambda(\mathbf{r}_s, \mathbf{r}_o)\, dS,} \tag{3.227}$$

where $h_\lambda(\mathbf{r}_s, \mathbf{r}_o)$ is the propagation kernel, or impulse response, connecting the aperture point $\mathbf{r}_s$ with the observation point $\mathbf{r}_o$.

Within the Kirchhoff formulation for normally incident illumination,

$$h_\lambda^{(K)}(\mathbf{r}_s, \mathbf{r}_o) = \frac{1}{j\lambda} \frac{e^{jkr_{os}}}{r_{os}} \frac{1 + \cos\theta_o}{2}. \tag{3.228}$$

In the small-angle limit,

$$\boxed{h_\lambda(\mathbf{r}_s, \mathbf{r}_o) \simeq \frac{1}{j\lambda} \frac{e^{jkr_{os}}}{r_{os}}.} \tag{3.229}$$

Equation (3.227) makes explicit that free-space propagation is a linear operation on the complex wave amplitude. This observation will become especially important when the propagation problem is reformulated in Fourier space.

3.5.5 Limitations of the Kirchhoff boundary conditions

Kirchhoff's construction provides an intuitive and remarkably successful description of diffraction, but its boundary assumptions are not mathematically self-consistent. On the opaque part of the screen, it assumes simultaneously that

$$U = 0 \qquad \text{and} \qquad \frac{\partial U}{\partial n} = 0. \tag{3.230}$$

For a solution of the Helmholtz equation, specifying both the field and its normal derivative independently over a finite open portion of a boundary generally overdetermines the problem. In particular, requiring both to vanish is incompatible with a non-trivial wave field under the assumptions used in the Kirchhoff construction.

This difficulty is avoided by the Rayleigh–Sommerfeld formulations of scalar diffraction. Instead of imposing both quantities simultaneously, one chooses a Green function constructed so that one of the two boundary terms vanishes on the screen plane. This leads to two closely related Rayleigh–Sommerfeld diffraction formulas.

For example, in the regime $kr_{os} \gg 1$, one commonly encounters the first Rayleigh–Sommerfeld form,

$$U(\mathbf{r}_o) \simeq \frac{1}{j\lambda} \iint_\Sigma U(\mathbf{r}_s) \frac{e^{jkr_{os}}}{r_{os}} \cos\theta_o \, dS. \tag{3.231}$$

Its obliquity factor is therefore

$$K_{\mathrm{RS1}}(\theta_o) = \cos\theta_o, \tag{3.232}$$

whereas the Kirchhoff result for normally incident illumination gives

$$K_K(\theta_o) = \frac{1 + \cos\theta_o}{2}. \tag{3.233}$$

The second Rayleigh–Sommerfeld formulation leads to another closely related expression, and the Kirchhoff result may be interpreted as lying between these two formulations.

For small diffraction angles,

$$\theta_o \ll 1, \tag{3.234}$$

we have

$$\cos\theta_o \simeq \frac{1 + \cos\theta_o}{2} \simeq 1. \tag{3.235}$$

Thus, the Kirchhoff and Rayleigh–Sommerfeld formulations converge to the same paraxial propagation law.

This is the regime of primary interest in electron microscopy. Electron wavelengths are of the order of picometres, whereas apertures and propagation distances in the microscope are enormously larger. Moreover, the useful electron trajectories are generally confined to small angles about the optical axis. Under these conditions, the differences between the scalar diffraction formulations become negligible for the purposes of the propagation theory developed below.

3.5.6 Extension to non-monochromatic waves

The derivation above assumed a single wavelength. A general time-dependent wave, however, can be decomposed into temporal-frequency components by a Fourier transform. Each monochromatic component has its own wavelength and therefore its own propagation kernel.

Because the wave equation is linear, every temporal-frequency component may be propagated independently. If

$$u(\mathbf{r}, t) = \int_{-\infty}^{\infty} U(\mathbf{r}, \nu) e^{-j2\pi\nu t}\, d\nu, \tag{3.236}$$

then, for every frequency ν,

$$U(\mathbf{r}_o, \nu) = \iint_{\Sigma} U(\mathbf{r}_s, \nu) h_\nu(\mathbf{r}_s, \mathbf{r}_o)\, dS. \tag{3.237}$$

The complete time-dependent field is recovered by superposing all frequency components.

This frequency-by-frequency description is also the natural starting point for the treatment of temporal coherence. A perfectly monochromatic wave contains a single frequency, whereas a realistic source contains a distribution of energies or frequencies whose propagated contributions must ultimately be combined according to the measurement process.

3.6 The angular spectrum of a wave

The diffraction integrals derived in the previous section describe propagation by regarding every point of a wavefront as an equivalent secondary source of spherical waves. There is a complementary and equally powerful representation: instead of decomposing a field into spherical waves originating at different spatial points, we decompose it into plane waves propagating in different directions.

This is the *angular-spectrum representation*. The relation between the Huygens–Fresnel and angular-spectrum representations is illustrated in Fig. 3.3. The same field that can be regarded as a superposition of spherical secondary waves may equivalently be decomposed into a spectrum of plane waves propagating in different directions.

Consider a monochromatic scalar field $U(x, y, z)$ satisfying the Helmholtz equation

$$\left(\nabla^2 + k^2\right) U(x, y, z) = 0, \qquad k = \frac{2\pi}{\lambda}. \tag{3.238}$$

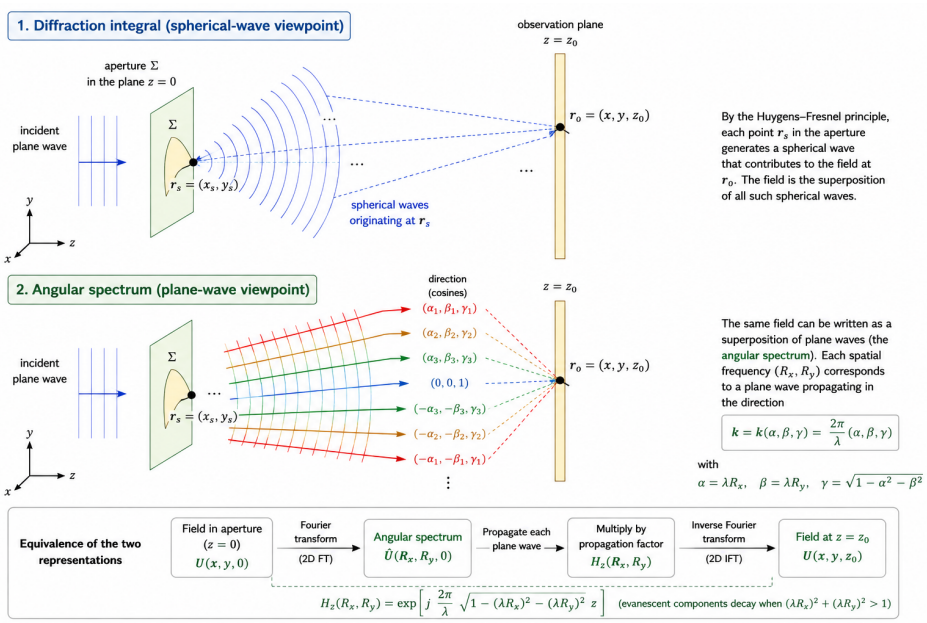


Figure 3.3: Two equivalent representations of wave propagation from an aperture. In the Huygens–Fresnel viewpoint (top), the field transmitted through the aperture is represented as a superposition of spherical secondary waves originating from points $\mathbf{r}_s$ in the aperture plane. In the angular-spectrum viewpoint (bottom), the same field is decomposed by a two-dimensional Fourier transform into a superposition of plane waves propagating in different directions. Each transverse spatial frequency (R_x, R_y) corresponds to a plane wave whose direction cosines satisfy $\alpha = \lambda R_x$, $\beta = \lambda R_y$, and $\gamma = \sqrt{1 - \alpha^2 - \beta^2}$. Propagation to a subsequent plane is performed independently for each plane-wave component by multiplication with the corresponding propagation factor, after which the propagated field is recovered by an inverse Fourier transform. The spherical-wave and plane-wave descriptions are therefore two mathematically equivalent representations of the same propagated field.

Suppose that the field is known over the plane $z = 0$,

$$U(x, y, 0). \tag{3.239}$$

Our objective is to calculate the field over a parallel plane located at an arbitrary distance z,

$$U(x, y, z). \tag{3.240}$$

The key observation is that the two-dimensional Fourier basis functions used to represent $U(x, y, 0)$ are precisely the transverse parts of plane waves. Consequently, the two-dimensional Fourier transform of the field does more than provide an alternative mathematical representation: each Fourier component identifies a plane wave propagating in a particular direction.

Propagation can therefore be performed in three steps:

1. decompose the field at $z = 0$ into plane waves;
2. propagate each plane-wave component independently;
3. superpose the propagated components to reconstruct the field at the desired plane.

Because every spatial-frequency component evolves independently, this representation will also reveal that free-space propagation between parallel planes is a linear, shift-invariant operation.

3.6.1 Plane waves and spatial frequency

Consider first a monochromatic plane wave,

$$u(\mathbf{r}, t) = \exp\left[j\left(\mathbf{k} \cdot \mathbf{r} - \omega t\right)\right], \tag{3.241}$$

where

$$\mathbf{k} = \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \tag{3.242}$$

is the angular wavevector. Since the magnitude of the wavevector is

$$\|\mathbf{k}\| = k = \frac{2\pi}{\lambda}, \tag{3.243}$$

its components satisfy

$$\boxed{k_x^2 + k_y^2 + k_z^2 = \left(\frac{2\pi}{\lambda}\right)^2.} \tag{3.244}$$

It is often convenient to describe the propagation direction using the direction cosines

$$\alpha = \frac{k_x}{k}, \qquad \beta = \frac{k_y}{k}, \qquad \gamma = \frac{k_z}{k}. \tag{3.245}$$

These quantities are dimensionless and satisfy

$$\alpha^2 + \beta^2 + \gamma^2 = 1. \tag{3.246}$$

For propagation toward increasing z,

$$\gamma = \sqrt{1 - \alpha^2 - \beta^2}. \tag{3.247}$$

The corresponding time-independent phasor is

$$U(\mathbf{r}) = \exp\left[j(k_x x + k_y y + k_z z)\right]. \tag{3.248}$$

At the plane $z = 0$, Eq. (3.248) reduces to

$$U(x, y, 0) = \exp\left[j(k_x x + k_y y)\right]. \tag{3.249}$$

Using

$$k_x = 2\pi R_x, \qquad k_y = 2\pi R_y, \tag{3.250}$$

we may write

$$U(x, y, 0) = \exp\left[j2\pi(R_x x + R_y y)\right]. \tag{3.251}$$

Thus, the Fourier basis function associated with the transverse spatial frequency (R_x, R_y) is the restriction to $z = 0$ of a plane wave whose transverse wavevector components are

$$k_x = 2\pi R_x, \qquad k_y = 2\pi R_y. \tag{3.252}$$

Since

$$k_x^2 + k_y^2 + k_z^2 = k^2, \tag{3.253}$$

the longitudinal component must satisfy

$$\boxed{k_z = 2\pi\sqrt{\frac{1}{\lambda^2} - R_x^2 - R_y^2}.} \tag{3.254}$$

For a forward-propagating wave, the branch of the square root is chosen with non-negative real part and non-negative imaginary part.

Equivalently, the direction cosines of the corresponding plane wave are

$$\boxed{\alpha = \lambda R_x, \qquad \beta = \lambda R_y, \qquad \gamma = \sqrt{1 - (\lambda R_x)^2 - (\lambda R_y)^2}.} \tag{3.255}$$

This equation gives the geometrical meaning of spatial frequency: the transverse Fourier coordinates (R_x, R_y) specify the direction of propagation of a plane-wave component.

3.6.2 The angular-spectrum decomposition

An arbitrary monochromatic field over the plane $z = 0$ can be decomposed using the two-dimensional Fourier transform,

$$\hat{U}(R_x, R_y, 0) = \iint_{-\infty}^{\infty} U(x, y, 0) \exp\left[-j2\pi(R_x x + R_y y)\right] \, dx \, dy. \tag{3.256}$$

The inverse transform gives

$$\boxed{U(x, y, 0) = \iint_{-\infty}^{\infty} \hat{U}(R_x, R_y, 0) \exp\left[j2\pi(R_x x + R_y y)\right] \, dR_x \, dR_y.} \tag{3.257}$$

According to the interpretation developed above, each exponential in Eq. (3.257) corresponds to a plane-wave component with transverse spatial frequency (R_x, R_y) and propagation direction given by Eq. (3.255). Its complex amplitude is

$$\hat{U}(R_x, R_y, 0). \tag{3.258}$$

For this reason, $\hat{U}(R_x, R_y, 0)$ is called the *angular spectrum* of the field. Its magnitude specifies how strongly each plane-wave direction is represented, while its phase specifies the phase of that component relative to the others.

3.6.3 Propagation of the angular spectrum

At a subsequent plane z, the field may likewise be represented as

$$U(x, y, z) = \iint_{-\infty}^{\infty} \hat{U}(R_x, R_y, z) \exp\left[j2\pi(R_x x + R_y y)\right] \, dR_x \, dR_y. \tag{3.259}$$

The question is therefore reduced to determining how each coefficient $\hat{U}(R_x, R_y, z)$ evolves with z.

Substitution of Eq. (3.259) into the Helmholtz equation gives, independently for every pair (R_x, R_y),

$$\frac{\partial^2 \hat{U}}{\partial z^2} + \left[k^2 - 4\pi^2(R_x^2 + R_y^2)\right] \hat{U} = 0. \tag{3.260}$$

Introducing k_z from Eq. (3.254), this becomes

$$\frac{\partial^2 \hat{U}}{\partial z^2} + k_z^2 \hat{U} = 0. \tag{3.261}$$

The two possible solutions are proportional to

$$e^{+jk_z z} \qquad \text{and} \qquad e^{-jk_z z}, \tag{3.262}$$

corresponding to propagation in opposite longitudinal directions. If only forward propagation toward increasing z is retained, then

$$\boxed{\hat{U}(R_x, R_y, z) = \hat{U}(R_x, R_y, 0)e^{jk_z z}.} \tag{3.263}$$

Using Eq. (3.254),

$$\boxed{\hat{U}(R_x, R_y, z) = \hat{U}(R_x, R_y, 0)\exp\left[j2\pi z\sqrt{\frac{1}{\lambda^2} - R_x^2 - R_y^2}\right].} \tag{3.264}$$

This is the exact angular-spectrum propagation law for a monochromatic scalar field in a homogeneous medium.

Derivation of the angular-spectrum propagation equation.

Starting from

$$U(x, y, z) = \iint \hat{U}(R_x, R_y, z)e^{j2\pi(R_x x + R_y y)}\, dR_x\, dR_y, \tag{3.265}$$

the second derivative with respect to x is

$$\frac{\partial^2 U}{\partial x^2} = \iint -4\pi^2 R_x^2 \hat{U}(R_x, R_y, z)e^{j2\pi(R_x x + R_y y)}\, dR_x\, dR_y. \tag{3.266}$$

Similarly,

$$\frac{\partial^2 U}{\partial y^2} = \iint -4\pi^2 R_y^2 \hat{U}(R_x, R_y, z)e^{j2\pi(R_x x + R_y y)}\, dR_x\, dR_y. \tag{3.267}$$

Since the Fourier exponential does not depend on z,

$$\frac{\partial^2 U}{\partial z^2} = \iint \frac{\partial^2 \hat{U}(R_x, R_y, z)}{\partial z^2} e^{j2\pi(R_x x + R_y y)}\, dR_x\, dR_y. \tag{3.268}$$

Substitution into

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} + k^2\right) U = 0 \tag{3.269}$$

gives

$$\iint \left[\frac{\partial^2 \hat{U}}{\partial z^2} + \left(k^2 - 4\pi^2 R_x^2 - 4\pi^2 R_y^2 \right) \hat{U} \right] e^{j2\pi(R_x x + R_y y)} \, dR_x \, dR_y = 0. \tag{3.270}$$

Since the Fourier basis functions are linearly independent, the quantity inside square brackets must vanish for every (R_x, R_y):

$$\frac{\partial^2 \hat{U}}{\partial z^2} + \left[k^2 - 4\pi^2 (R_x^2 + R_y^2) \right] \hat{U} = 0. \tag{3.271}$$

Defining

$$k_z^2 = k^2 - 4\pi^2 (R_x^2 + R_y^2) \tag{3.272}$$

gives

$$\frac{\partial^2 \hat{U}}{\partial z^2} + k_z^2 \hat{U} = 0, \tag{3.273}$$

whose general solution is

$$\hat{U}(R_x, R_y, z) = C_+(R_x, R_y) e^{jk_z z} + C_-(R_x, R_y) e^{-jk_z z}. \tag{3.274}$$

Retaining only propagation toward increasing z and imposing the field at $z = 0$ gives

$$C_+(R_x, R_y) = \hat{U}(R_x, R_y, 0), \qquad C_-(R_x, R_y) = 0, \tag{3.275}$$

which yields Eq. (3.264).

3.6.4 Propagating and evanescent components

Equation (3.264) reveals two qualitatively different regimes. Define the transverse spatial-frequency magnitude

$$R_\perp = \sqrt{R_x^2 + R_y^2}. \tag{3.276}$$

If

$$R_\perp < \frac{1}{\lambda}, \tag{3.277}$$

then k_z is real:

$$k_z = 2\pi \sqrt{\frac{1}{\lambda^2} - R_\perp^2}. \tag{3.278}$$

Propagation therefore multiplies the corresponding Fourier component by a pure phase factor,

$$e^{jk_z z}, \tag{3.279}$$

whose magnitude is unity. These are the ordinary *propagating components* of the angular spectrum.

Different transverse spatial frequencies correspond to different propagation directions and hence accumulate different longitudinal phases between the two planes. It is precisely this frequency-dependent phase accumulation that produces diffraction: after propagation, the plane-wave components interfere differently from how they did at the initial plane.

At the limiting value

$$R_\perp = \frac{1}{\lambda}, \tag{3.280}$$

we have $k_z = 0$, corresponding formally to propagation parallel to the plane $z = 0$.

If instead

$$R_\perp > \frac{1}{\lambda}, \tag{3.281}$$

then k_z becomes purely imaginary. Writing

$$k_z = j\kappa, \qquad \kappa = 2\pi\sqrt{R_\perp^2 - \frac{1}{\lambda^2}}, \tag{3.282}$$

the forward-propagation factor becomes

$$e^{jk_z z} = e^{-\kappa z}. \tag{3.283}$$

These components therefore do not propagate indefinitely through free space. Their amplitudes decay exponentially with distance from the plane where they are generated, and they are called *evanescent waves*.

3.6.5 Spatial frequency and the wavelength scale

The distinction between propagating and evanescent components establishes an important physical scale. Consider a one-dimensional periodic structure of period d. Its fundamental spatial frequency is

$$R = \frac{1}{d}. \tag{3.284}$$

For this component to propagate through free space, it must satisfy

$$R \leq \frac{1}{\lambda}, \tag{3.285}$$

or equivalently

$$d \geq \lambda. \tag{3.286}$$

Spatial variations with periods appreciably smaller than the wavelength are therefore represented by transverse spatial frequencies that are evanescent rather than freely propagating.

This result should not be confused with the practical resolution limit of an imaging system. A real optical system collects only a restricted range of propagation angles. If the largest accepted angle is $\theta_{\max}$, then

$$R_{\perp,\max} = \frac{\sin\theta_{\max}}{\lambda} = \frac{\mathrm{NA}}{\lambda} \tag{3.287}$$

for propagation in vacuum, where $\mathrm{NA} = \sin\theta_{\max}$ is the numerical aperture. The accessible spatial-frequency range is therefore normally smaller than the absolute free-space limit $1/\lambda$.

Thus, the wavelength establishes the largest transverse spatial frequency that can propagate in the far field, while the numerical aperture and the rest of the imaging system determine how much of that range is actually recorded.

For high-energy electrons the de Broglie wavelength is of the order of picometres. Consequently, the wavelength itself is not the principal limitation on the spatial resolution attained in modern cryo-electron microscopy. Aberrations, coherence, specimen scattering, radiation damage, motion, detector response, sampling, and the reconstruction procedure usually impose much stronger practical limitations.

3.6.6 Free-space propagation as a linear filter

Equation (3.263) may be written as

$$\boxed{\hat{U}(R_x, R_y, z) = \hat{U}(R_x, R_y, 0)\hat{h}_z(R_x, R_y),} \tag{3.288}$$

where

$$\boxed{\hat{h}_z(R_x, R_y) = e^{jk_z z}} \tag{3.289}$$

is the exact free-space transfer function.

Explicitly,

$$\hat{h}_z(R_x, R_y) = \begin{cases} \exp\left[j2\pi z\sqrt{\dfrac{1}{\lambda^2} - R_x^2 - R_y^2}\right], & R_x^2 + R_y^2 \leq \frac{1}{\lambda^2}, \\ \exp\left[-2\pi z\sqrt{R_x^2 + R_y^2 - \dfrac{1}{\lambda^2}}\right], & R_x^2 + R_y^2 > \frac{1}{\lambda^2}. \end{cases} \tag{3.290}$$

If evanescent components are neglected, as is normally justified when the propagation distance is many wavelengths, this becomes

$$\boxed{\hat{h}_z(R_x, R_y) = \begin{cases} \exp\left[j2\pi z\sqrt{\frac{1}{\lambda^2} - R_x^2 - R_y^2}\right], & R_x^2 + R_y^2 \leq \frac{1}{\lambda^2}, \\ 0, & R_x^2 + R_y^2 > \frac{1}{\lambda^2}. \end{cases}} \tag{3.291}$$

Since multiplication in Fourier space corresponds to convolution in real space,

$$\boxed{U(x, y, z) = U(x, y, 0) * h_z(x, y),} \tag{3.292}$$

where

$$h_z(x, y) = \mathcal{F}^{-1}\left\{\hat{h}_z(R_x, R_y)\right\}. \tag{3.293}$$

Free-space propagation between two parallel planes is therefore a linear, shift-invariant system.

The real-space function $h_z(x, y)$ has a direct physical interpretation. It is the *free-space propagation kernel*, or *impulse response*: it represents the field at (x, y, z) produced by a point excitation at the origin of the input plane $z = 0$. Thus, the convolution

$$U(x, y, z) = U(x, y, 0) * h_z(x, y) \tag{3.294}$$

expresses the propagated field as a coherent superposition of translated copies of the field generated by a single point of the input plane.

The exact scalar impulse response is closely related to the outgoing spherical Green function. Defining

$$r = \sqrt{x^2 + y^2 + z^2}, \tag{3.295}$$

it can be written as

$$\boxed{h_z(x, y) = -\frac{1}{2\pi}\frac{\partial}{\partial z}\left(\frac{e^{jkr}}{r}\right).} \tag{3.296}$$

Evaluating the derivative gives

$$\boxed{h_z(x, y) = \frac{1}{j\lambda}\frac{z}{r}\frac{e^{jkr}}{r}\left(1 - \frac{1}{jkr}\right).} \tag{3.297}$$

For propagation distances much larger than the wavelength, $kr \gg 1$, the last factor approaches unity and

$$h_z(x,y) \simeq \frac{1}{j\lambda}\cos\theta\,\frac{e^{jkr}}{r}, \qquad \cos\theta = \frac{z}{r}. \tag{3.298}$$

This is precisely the spherical-wave propagation factor encountered in the Rayleigh–Sommerfeld diffraction integral. The real-space spherical-wave description and the Fourier-space angular-spectrum description are therefore two representations of the same free-space propagation operator.

From the angular-spectrum transfer function to the real-space propagation kernel.

We want to determine the real-space impulse response corresponding to the angular-spectrum transfer function

$$\hat{h}_z(R_x,R_y) = e^{jk_z z}, \qquad z > 0, \tag{3.299}$$

where

$$k_z = \sqrt{k^2 - 4\pi^2(R_x^2 + R_y^2)}. \tag{3.300}$$

For propagating components, k_z is chosen to be positive and real. For evanescent components, the branch is chosen such that $\mathrm{Im}(k_z) > 0$, ensuring that $e^{jk_z z}$ decays for increasing z.

By definition, the corresponding real-space impulse response is

$$h_z(x,y) = \mathcal{F}_{x,y}^{-1}\left\{e^{jk_z z}\right\}. \tag{3.301}$$

Evaluating this inverse Fourier transform directly is possible, but not particularly transparent. A more instructive route starts from the outgoing spherical Green function

$$G(x,y,z) = \frac{e^{jkr}}{r}, \qquad r = \sqrt{x^2+y^2+z^2}. \tag{3.302}$$

As shown previously, this function satisfies

$$\left(\nabla^2 + k^2\right) G(x,y,z) = -4\pi\delta(x)\delta(y)\delta(z). \tag{3.303}$$

Our first objective is to determine the two-dimensional Fourier transform of G with respect to x and y. We therefore define

$$\hat{G}(R_x, R_y, z) = \iint_{-\infty}^{\infty} G(x, y, z) e^{-j2\pi(R_x x + R_y y)}\, dx\, dy. \tag{3.304}$$

Notice that this is only a *transverse* Fourier transform. The coordinates x and y are transformed into R_x and R_y, but z remains a real-space coordinate.

We now Fourier transform every term of Eq. (3.303). Consider first the second derivative with respect to x. The Fourier transform of the first derivative is

$$\mathcal{F}_{x,y}\left\{\frac{\partial G}{\partial x}\right\} = \iint \frac{\partial G}{\partial x} e^{-j2\pi(R_x x + R_y y)}\, dx\, dy \tag{3.305}$$

$$= \iint G\,(j2\pi R_x)\, e^{-j2\pi(R_x x + R_y y)}\, dx\, dy \tag{3.306}$$

$$= j2\pi R_x \hat{G}, \tag{3.307}$$

where integration by parts has been used and the boundary term has been assumed to vanish at infinity. Applying the same property a second time gives

$$\boxed{\mathcal{F}_{x,y}\left\{\frac{\partial^2 G}{\partial x^2}\right\} = -4\pi^2 R_x^2 \hat{G}.} \tag{3.308}$$

Similarly,

$$\boxed{\mathcal{F}_{x,y}\left\{\frac{\partial^2 G}{\partial y^2}\right\} = -4\pi^2 R_y^2 \hat{G}.} \tag{3.309}$$

The z derivative behaves differently because z has not been Fourier transformed. Since the Fourier exponential in Eq. (3.304) is independent of z,

$$\mathcal{F}_{x,y}\left\{\frac{\partial^2 G}{\partial z^2}\right\} = \iint \frac{\partial^2 G}{\partial z^2} e^{-j2\pi(R_x x + R_y y)}\, dx\, dy \tag{3.310}$$

$$= \frac{\partial^2}{\partial z^2} \iint G e^{-j2\pi(R_x x + R_y y)}\, dx\, dy \tag{3.311}$$

$$= \boxed{\frac{\partial^2 \hat{G}}{\partial z^2}}. \tag{3.312}$$

Likewise, since k is constant,

$$\mathcal{F}_{x,y}\{k^2 G\} = k^2 \hat{G}. \tag{3.313}$$

Finally, the source term transforms as

$$\mathcal{F}_{x,y}\left\{\delta(x)\delta(y)\delta(z)\right\} = \delta(z) \iint \delta(x)\delta(y) e^{-j2\pi(R_x x + R_y y)}\, dx\, dy \tag{3.314}$$

$$= \delta(z), \tag{3.315}$$

because the delta functions select $x = y = 0$, where the Fourier exponential is equal to unity.

Consequently, the transverse Fourier transform of Eq. (3.303) is

$$-4\pi^2 R_x^2 \hat{G} - 4\pi^2 R_y^2 \hat{G} + \frac{\partial^2 \hat{G}}{\partial z^2} + k^2 \hat{G} = -4\pi\delta(z). \tag{3.316}$$

Collecting the terms that multiply $\hat{G}$ gives

$$\frac{\partial^2 \hat{G}}{\partial z^2} + \left[k^2 - 4\pi^2 (R_x^2 + R_y^2)\right] \hat{G} = -4\pi\delta(z). \tag{3.317}$$

Using the definition of k_z in Eq. (3.300), we obtain

$$\boxed{\frac{\partial^2 \hat{G}}{\partial z^2} + k_z^2 \hat{G} = -4\pi\delta(z).} \tag{3.318}$$

For $z \neq 0$, the delta function vanishes, so

$$\frac{\partial^2 \hat{G}}{\partial z^2} + k_z^2 \hat{G} = 0. \tag{3.319}$$

The source is located at $z = 0$, and the outgoing-wave condition requires waves to propagate away from this plane on both sides. The appropriate solution is therefore

$$\hat{G}(R_x, R_y, z) = C(R_x, R_y) e^{jk_z|z|}, \tag{3.320}$$

where the constant $C(R_x, R_y)$ remains to be determined.

The delta function in Eq. (3.318) determines this constant. We integrate that equation across an infinitesimally thin interval containing $z = 0$:

$$\int_{-\varepsilon}^{+\varepsilon} \left(\frac{\partial^2 \hat{G}}{\partial z^2} + k_z^2 \hat{G} \right) dz = -4\pi \int_{-\varepsilon}^{+\varepsilon} \delta(z)\, dz. \tag{3.321}$$

The right-hand side is

$$-4\pi. \tag{3.322}$$

As $\varepsilon \to 0$, the integral containing $\hat{G}$ vanishes because $\hat{G}$ remains finite at $z = 0$. The integral of the second derivative, however, gives the jump of the first derivative:

$$\left. \frac{\partial \hat{G}}{\partial z} \right|_{0^+} - \left. \frac{\partial \hat{G}}{\partial z} \right|_{0^-} = -4\pi. \tag{3.323}$$

From Eq. (3.320),

$$\left. \frac{\partial \hat{G}}{\partial z} \right|_{0^+} = jk_z C \tag{3.324}$$

whereas, because $|z| = -z$ for $z < 0$,

$$\left. \frac{\partial \hat{G}}{\partial z} \right|_{0^-} = -jk_z C. \tag{3.325}$$

Hence,

$$jk_z C - (-jk_z C) = 2jk_z C = -4\pi, \tag{3.326}$$

which gives

$$C = \frac{2\pi j}{k_z}. \tag{3.327}$$

We have therefore obtained

$$\boxed{\hat{G}(R_x, R_y, z) = \frac{2\pi j}{k_z} e^{jk_z|z|}.} \tag{3.328}$$

This relation is the Fourier-space form of the *Weyl representation* of a spherical wave. It shows explicitly that a spherical wave can be represented as a continuous superposition of plane waves with different transverse spatial frequencies.

We now restrict ourselves to propagation toward positive z. For $z > 0$, Eq. (3.328) becomes

$$\hat{G}(R_x, R_y, z) = \frac{2\pi j}{k_z} e^{jk_z z}. \tag{3.329}$$

Differentiating with respect to z,

$$\frac{\partial \hat{G}}{\partial z} = \frac{2\pi j}{k_z}(jk_z)e^{jk_z z} \tag{3.330}$$

$$= -2\pi e^{jk_z z}. \tag{3.331}$$

Therefore,

$$\boxed{e^{jk_z z} = -\frac{1}{2\pi}\frac{\partial \hat{G}}{\partial z}.} \tag{3.332}$$

But $e^{jk_z z}$ is precisely the angular-spectrum transfer function $\hat{h}_z$. Thus,

$$\hat{h}_z(R_x, R_y) = -\frac{1}{2\pi}\frac{\partial \hat{G}(R_x, R_y, z)}{\partial z}. \tag{3.333}$$

Taking the inverse two-dimensional Fourier transform of both sides gives

$$h_z(x, y) = -\frac{1}{2\pi}\mathcal{F}^{-1}_{x,y}\left\{\frac{\partial \hat{G}}{\partial z}\right\}. \tag{3.334}$$

Because the Fourier transform acts only on x and y, differentiation with respect to z commutes with the inverse transform:

$$\mathcal{F}^{-1}_{x,y}\left\{\frac{\partial \hat{G}}{\partial z}\right\} = \frac{\partial}{\partial z}\mathcal{F}^{-1}_{x,y}\left\{\hat{G}\right\}. \tag{3.335}$$

Since

$$\mathcal{F}^{-1}_{x,y}\{\hat{G}\} = G(x, y, z), \tag{3.336}$$

we finally obtain

$$\boxed{h_z(x, y) = -\frac{1}{2\pi}\frac{\partial}{\partial z}\left(\frac{e^{jkr}}{r}\right), \qquad r = \sqrt{x^2 + y^2 + z^2}.} \tag{3.337}$$

The derivative can now be evaluated explicitly. Since

$$\frac{\partial r}{\partial z} = \frac{z}{r}, \tag{3.338}$$

the chain rule gives

$$\frac{\partial}{\partial z}\left(\frac{e^{jkr}}{r}\right) = \frac{\partial r}{\partial z}\frac{d}{dr}\left(\frac{e^{jkr}}{r}\right) \tag{3.339}$$

$$= \frac{z}{r}e^{jkr}\left(\frac{jk}{r} - \frac{1}{r^2}\right). \tag{3.340}$$

Substitution into Eq. (3.337) yields

$$h_z(x,y) = \frac{ze^{jkr}}{2\pi}\left(\frac{1}{r^3} - \frac{jk}{r^2}\right). \tag{3.341}$$

Using

$$k = \frac{2\pi}{\lambda}, \tag{3.342}$$

this can be rearranged into the more physically suggestive form

$$\boxed{h_z(x,y) = \frac{1}{j\lambda}\frac{z}{r}\frac{e^{jkr}}{r}\left(1 - \frac{1}{jkr}\right).} \tag{3.343}$$

The three factors in this expression have direct interpretations:

$$\frac{e^{jkr}}{r} \tag{3.344}$$

is the outgoing spherical-wave Green function,

$$\frac{z}{r} = \cos\theta \tag{3.345}$$

is the obliquity factor, and

$$1 - \frac{1}{jkr} \tag{3.346}$$

is a near-field correction. When $kr \gg 1$, this last factor approaches unity and the kernel reduces to

$$h_z(x,y) \simeq \frac{1}{j\lambda}\cos\theta\frac{e^{jkr}}{r}, \tag{3.347}$$

which is precisely the propagation kernel appearing in the first Rayleigh–Sommerfeld diffraction formula.

The appearance of the normal derivative of the spherical Green function is not accidental. In Fourier space, the Green function contains the factor $1/k_z$, whereas differentiation with respect to z contributes a factor jk_z. The two factors cancel, leaving exactly the free-space angular-spectrum propagation factor $e^{jk_z z}$. Thus, the spherical-wave Rayleigh–Sommerfeld kernel and the plane-wave angular-spectrum propagator are two representations of precisely the same propagation operator.

3.6.7 Apertures and phase-modifying elements

The angular-spectrum formalism also provides a simple way to describe elements that modify the wave over a plane. Suppose that an optical element is located at $z = 0$. Its complex amplitude transmittance is defined as

$$\boxed{t_A(x,y) = \frac{U_t(x,y,0)}{U_i(x,y,0)},} \tag{3.348}$$

where U_i and U_t denote the incident and transmitted fields, respectively. Thus,

$$\boxed{U_t(x,y,0) = t_A(x,y)U_i(x,y,0).} \tag{3.349}$$

A general complex transmittance can be written as

$$t_A(x,y) = a(x,y)e^{j\phi(x,y)}, \tag{3.350}$$

where $a(x,y)$ modifies the amplitude and $\phi(x,y)$ modifies the phase.

For an ideal aperture Σ,

$$t_A(x,y) = \begin{cases} 1, & (x,y) \in \Sigma, \\ 0, & (x,y) \notin \Sigma. \end{cases} \tag{3.351}$$

An ideal phase-modifying element, on the other hand, has

$$|t_A(x,y)| = 1 \tag{3.352}$$

and modifies only the phase,

$$t_A(x, y) = e^{j\phi(x,y)}. \tag{3.353}$$

Because multiplication in real space becomes convolution in Fourier space, Eq. (3.349) implies

$$\boxed{\hat{U}_t(R_x, R_y, 0) = \hat{U}_i(R_x, R_y, 0) * \hat{t}_A(R_x, R_y).} \tag{3.354}$$

Thus an aperture or phase-modifying element changes the angular spectrum of the incident wave by convolving it with the Fourier transform of its transmittance.

This result provides another view of diffraction by an aperture. Consider, for example, an incident plane wave at normal incidence. Its transverse angular spectrum is concentrated at

$$R_x = R_y = 0. \tag{3.355}$$

After multiplication by the aperture function, the transmitted angular spectrum becomes proportional to the Fourier transform of the aperture,

$$\hat{U}_t(R_x, R_y, 0) \propto \hat{t}_A(R_x, R_y). \tag{3.356}$$

A spatially extended aperture has a relatively narrow Fourier transform and therefore produces a relatively narrow range of propagation directions. Conversely, reducing the dimensions of the aperture broadens its Fourier transform and hence broadens the angular distribution of the transmitted wave.

In the limiting idealization of an infinitesimally small point aperture,

$$t_A(x, y) \propto \delta(x, y), \tag{3.357}$$

whose Fourier transform is constant:

$$\hat{t}_A(R_x, R_y) = \text{constant}. \tag{3.358}$$

The aperture would therefore generate an equally weighted spectrum of transverse spatial frequencies, subject ultimately to the distinction between propagating and evanescent components. This is the Fourier-space expression of the familiar fact that a very small aperture diffracts a wave over a very wide angular range.

3.6.8 Relation to the diffraction-integral formulation

The angular-spectrum method and the diffraction-integral method are not competing theories. They are two representations of the same solution of the Helmholtz equation.

In the diffraction-integral viewpoint, the field is propagated by superposing spherical secondary waves originating from points of the input plane. In the angular-spectrum viewpoint, the same field is decomposed into plane waves, each of which acquires an independently calculable propagation factor.

For propagation between parallel planes, the angular-spectrum solution is equivalent to the first Rayleigh–Sommerfeld solution of scalar diffraction theory (Goodman, 2017, Sec. 3.10). The choice between the two representations is therefore primarily one of convenience. The angular-spectrum formulation is especially useful for Fourier optics because propagation appears simply as multiplication by a transfer function.

3.7 Fresnel and Fraunhofer approximations

The propagation laws derived in the previous sections are exact within scalar wave theory. The angular-spectrum representation, for example, propagates every transverse spatial frequency according to

$$\hat{U}(R_x, R_y, z) = \hat{U}(R_x, R_y, 0) \exp\left[j2\pi z \sqrt{\frac{1}{\lambda^2} - R_x^2 - R_y^2} \right]. \tag{3.359}$$

Although exact, this expression can often be simplified considerably in an imaging system because the relevant waves propagate at small angles with respect to the optical axis.

This leads to two successive approximations. The first is the *Fresnel*, or *paraxial*, approximation. It assumes that the angular spectrum is concentrated close to the optical axis, so that the transverse components of the wavevector are much smaller than its total magnitude. The second is the *Fraunhofer*, or *far-field*, approximation. It introduces an additional condition on the propagation distance and reduces the propagated field, apart from simple phase and scale factors, to the Fourier transform of the input field.

The hierarchy is therefore

$$\boxed{\text{exact propagation} \quad \longrightarrow \quad \text{Fresnel propagation} \quad \longrightarrow \quad \text{Fraunhofer propagation.}} \tag{3.360}$$

3.7.1 The Fresnel or paraxial approximation

Define the magnitude of the transverse spatial frequency as

$$R_\perp = \sqrt{R_x^2 + R_y^2}. \tag{3.361}$$

To relate this quantity to the propagation angle, recall that the wavevector of a plane-wave component is

$$\mathbf{k} = (k_x, k_y, k_z)^T, \qquad k = \|\mathbf{k}\| = \frac{2\pi}{\lambda}. \tag{3.362}$$

The magnitude of its component perpendicular to the optical axis is

$$k_\perp = \sqrt{k_x^2 + k_y^2}. \tag{3.363}$$

If θ denotes the angle between $\mathbf{k}$ and the positive z-axis, the geometry of the wavevector gives

$$\sin\theta = \frac{k_\perp}{k}. \tag{3.364}$$

Our Fourier-transform convention relates transverse angular wavevector and spatial frequency through

$$k_x = 2\pi R_x, \qquad k_y = 2\pi R_y. \tag{3.365}$$

Consequently,

$$k_\perp = 2\pi\sqrt{R_x^2 + R_y^2} = 2\pi R_\perp. \tag{3.366}$$

Using $k = 2\pi/\lambda$ in Eq. (3.364) therefore gives

$$\boxed{\sin\theta = \lambda R_\perp.} \tag{3.367}$$

Thus, increasing transverse spatial frequency corresponds directly to increasing propagation angle.

The paraxial approximation assumes that the relevant plane-wave components make only small angles with the optical axis,

$$\theta \ll 1. \tag{3.368}$$

For small angles, $\sin\theta \simeq \theta \ll 1$. From Eq. (3.367), this is equivalent to

$$\boxed{\lambda R_\perp \ll 1.} \tag{3.369}$$

Using the Taylor expansion

$$\sqrt{1-u} = 1 - \frac{u}{2} - \frac{u^2}{8} - \cdots, \tag{3.370}$$

and retaining only the first two terms, we obtain

$$\sqrt{1-\lambda^2 R_\perp^2} \simeq 1 - \frac{\lambda^2 R_\perp^2}{2}. \tag{3.371}$$

The exact propagation transfer function therefore becomes

$$\hat{h}_z(R_x, R_y) \simeq \exp\left[j\frac{2\pi z}{\lambda}\left(1 - \frac{\lambda^2}{2}(R_x^2 + R_y^2)\right)\right] \tag{3.372}$$

$$= \boxed{e^{jkz}\exp\left[-j\pi\lambda z(R_x^2 + R_y^2)\right].} \tag{3.373}$$

Thus, within the paraxial approximation, free-space propagation corresponds in Fourier space to multiplication by a quadratic phase function.

The inverse Fourier transform of Eq. (3.373) gives the corresponding real-space impulse response,

$$\boxed{h_z(x,y) = \frac{e^{jkz}}{j\lambda z}\exp\left[j\frac{\pi}{\lambda z}(x^2+y^2)\right].} \tag{3.374}$$

This is the *Fresnel propagation kernel*. Its magnitude is constant,

$$|h_z(x,y)| = \frac{1}{\lambda z}, \tag{3.375}$$

whereas its phase varies quadratically with the transverse distance from the optical axis. For this reason the Fresnel kernel is also called a *quadratic-phase function* or a two-dimensional *chirp*.

Propagation between two parallel planes is therefore given by the convolution

$$\boxed{U(x,y,z) = U(x,y,0) * h_z(x,y).} \tag{3.376}$$

Explicitly,

$$\boxed{U(x,y,z) = \frac{e^{jkz}}{j\lambda z}\iint_{-\infty}^{\infty} U(x_s,y_s,0)\exp\left[j\frac{\pi}{\lambda z}\left((x-x_s)^2 + (y-y_s)^2\right)\right]dx_s\,dy_s.} \tag{3.377}$$

This is the Fresnel diffraction integral.

The geometrical meaning of the Fresnel approximation can be seen directly from the distance between a point

$$\mathbf{r}_s = (x_s, y_s, 0)^T \tag{3.378}$$

on the input plane and an observation point

$$\mathbf{r}_o = (x, y, z)^T. \tag{3.379}$$

Their separation is

$$r_{os} = \sqrt{z^2 + (x - x_s)^2 + (y - y_s)^2} \tag{3.380}$$

$$= z\sqrt{1 + \frac{(x - x_s)^2 + (y - y_s)^2}{z^2}}. \tag{3.381}$$

For paraxial propagation,

$$(x - x_s)^2 + (y - y_s)^2 \ll z^2, \tag{3.382}$$

so that

$$\boxed{r_{os} \simeq z + \frac{(x - x_s)^2 + (y - y_s)^2}{2z}.} \tag{3.383}$$

Consequently,

$$e^{jkr_{os}} \simeq e^{jkz} \exp\left[j\frac{k}{2z}\left((x - x_s)^2 + (y - y_s)^2\right)\right], \tag{3.384}$$

which is precisely the quadratic phase appearing in Eq. (3.377).

The Fresnel approximation may therefore be interpreted geometrically as approximating the spherical secondary wavefronts by paraboloidal wavefronts. Unlike a plane-wave approximation, this retains the curvature of the wavefront to second order in the transverse coordinates.

Derivation of the Fresnel propagation kernel.

Starting from the exact angular-spectrum transfer function,

$$\hat{h}_z(R_x, R_y) = \exp\left[j\frac{2\pi z}{\lambda}\sqrt{1 - \lambda^2(R_x^2 + R_y^2)}\right], \tag{3.385}$$

the paraxial condition

$$\lambda^2(R_x^2 + R_y^2) \ll 1 \tag{3.386}$$

allows us to use

$$\sqrt{1-u} \simeq 1 - \frac{u}{2}. \tag{3.387}$$

Therefore,

$$\hat{h}_z(R_x, R_y) \simeq \exp\left[j\frac{2\pi z}{\lambda}\left(1 - \frac{\lambda^2}{2}(R_x^2 + R_y^2)\right)\right] \tag{3.388}$$

$$= e^{jkz} \exp\left[-j\pi\lambda z(R_x^2 + R_y^2)\right]. \tag{3.389}$$

We now use the one-dimensional Fourier-transform identity

$$\mathcal{F}^{-1}\left\{e^{-j\pi a R^2}\right\} = \frac{1}{\sqrt{ja}} \exp\left(j\frac{\pi x^2}{a}\right), \qquad a > 0. \tag{3.390}$$

Since the two-dimensional transfer function separates into independent x and y factors,

$$h_z(x, y) = e^{jkz}\mathcal{F}^{-1}\left\{e^{-j\pi\lambda z R_x^2}\right\}\mathcal{F}^{-1}\left\{e^{-j\pi\lambda z R_y^2}\right\} \tag{3.391}$$

$$= e^{jkz}\frac{1}{\sqrt{j\lambda z}} \exp\left(j\frac{\pi x^2}{\lambda z}\right)\frac{1}{\sqrt{j\lambda z}} \exp\left(j\frac{\pi y^2}{\lambda z}\right). \tag{3.392}$$

Since

$$\frac{1}{\sqrt{j\lambda z}\sqrt{j\lambda z}} = \frac{1}{j\lambda z}, \tag{3.393}$$

we obtain

$$\boxed{h_z(x, y) = \frac{e^{jkz}}{j\lambda z} \exp\left[j\frac{\pi}{\lambda z}(x^2 + y^2)\right].} \tag{3.394}$$

3.7.2 The Fraunhofer or far-field approximation

The Fresnel diffraction integral can be rearranged to reveal another useful limit. Expanding

$$(x - x_s)^2 + (y - y_s)^2 = x^2 + y^2 + x_s^2 + y_s^2 - 2(xx_s + yy_s), \tag{3.395}$$

Eq. (3.377) becomes

$$U(x,y,z) = \frac{e^{jkz}}{j\lambda z} \exp\left[j\frac{\pi}{\lambda z}(x^2+y^2)\right] \times \iint U(x_s,y_s,0) \exp\left[j\frac{\pi}{\lambda z}(x_s^2+y_s^2)\right] \times \exp\left[-j2\pi\left(\frac{x}{\lambda z}x_s + \frac{y}{\lambda z}y_s\right)\right] dx_s\, dy_s. \tag{3.396}$$

This expression already resembles a Fourier transform. The only additional factor inside the integral is the quadratic phase

$$\exp\left[j\frac{\pi}{\lambda z}(x_s^2+y_s^2)\right]. \tag{3.397}$$

If the input field is confined to a region of characteristic radius a, then throughout that region

$$x_s^2 + y_s^2 \leq a^2. \tag{3.398}$$

If

$$\boxed{\frac{\pi a^2}{\lambda z} \ll 1,} \tag{3.399}$$

the quadratic phase varies negligibly across the input plane and may be approximated by unity. Equation (3.396) then reduces to

$$\boxed{U(x,y,z) \simeq \frac{e^{jkz}}{j\lambda z} \exp\left[j\frac{\pi}{\lambda z}(x^2+y^2)\right] \hat{U}\left(\frac{x}{\lambda z}, \frac{y}{\lambda z}, 0\right).} \tag{3.400}$$

This is the Fraunhofer diffraction formula. Apart from a constant amplitude factor and a quadratic phase factor, the field in the far-field plane is the two-dimensional Fourier transform of the field in the input plane, evaluated at

$$\boxed{R_x = \frac{x}{\lambda z}, \qquad R_y = \frac{y}{\lambda z}.} \tag{3.401}$$

Thus, transverse position in the Fraunhofer pattern directly represents transverse spatial frequency in the input field.

3.7.3 The Fresnel number

The importance of the quadratic phase across an object or aperture is commonly characterized by the dimensionless *Fresnel number*

$$\boxed{F = \frac{a^2}{\lambda z},} \tag{3.402}$$

where a is a characteristic transverse size.

The regimes may be summarized qualitatively as

$$\begin{aligned} F \gtrsim 1 \quad &: \quad \text{pronounced Fresnel or near-field diffraction,} \\ F \ll 1 \quad &: \quad \text{Fraunhofer or far-field diffraction.} \end{aligned} \tag{3.403}$$

More precisely, the phase criterion used in Eq. (3.399) is

$$\pi F \ll 1. \tag{3.404}$$

The precise numerical boundary between the two regimes is therefore not fundamental; it depends on how much phase error is considered acceptable.

For electron waves, the very small value of λ can make the free-space Fraunhofer distance surprisingly large. For example, consider a circular aperture of radius

$$a = 10\ \mu\text{m} \tag{3.405}$$

and an electron wavelength of approximately

$$\lambda = 2\ \text{pm}. \tag{3.406}$$

Then

$$\frac{a^2}{\lambda} = \frac{(10^{-5}\ \text{m})^2}{2 \times 10^{-12}\ \text{m}} \simeq 50\ \text{m}. \tag{3.407}$$

Thus, achieving $F \ll 1$ by free-space propagation alone would require distances of many tens or even hundreds of metres.

This does not imply that Fraunhofer diffraction is inaccessible in an electron microscope. As will be shown later, a lens can transform angular information into spatial position and form a Fourier, or diffraction, plane at a finite distance. The back focal plane of an electron lens therefore plays a role analogous to the far-field diffraction plane without requiring the corresponding free-space propagation distance.

The transition between the Fresnel and Fraunhofer regimes is illustrated in Fig. 3.4. At relatively short propagation distances, the field still retains a recognizable geometrical relation to the aperture, although interference produces oscillations near and within its projected boundary. As z increases, these Fresnel features evolve progressively into the far-field diffraction pattern.

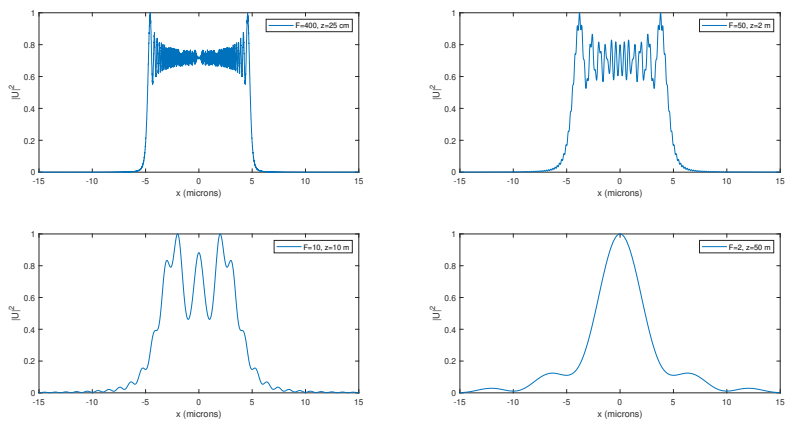


Figure 3.4: Evolution of the diffraction pattern produced by a uniformly illuminated aperture as the propagation distance increases. At large Fresnel numbers the field lies in the Fresnel regime and retains features related to the geometrical projection of the aperture. As the Fresnel number decreases, the pattern approaches the Fraunhofer regime, in which the field is proportional, apart from scale and phase factors, to the Fourier transform of the aperture function.

3.7.4 Fraunhofer diffraction by a circular aperture

A particularly important example is the Fraunhofer diffraction pattern produced by a circular aperture. Let the aperture have radius a and amplitude transmittance

$$t_A(x_s, y_s) = \mathrm{circ}\left(\frac{\rho_s}{a}\right), \qquad \rho_s = \sqrt{x_s^2 + y_s^2}, \tag{3.408}$$

where

$$\mathrm{circ}(q) = \begin{cases} 1, & q \leq 1, \\ 0, & q > 1. \end{cases} \tag{3.409}$$

For uniform normally incident illumination of unit amplitude, the field immediately after the aperture is simply

$$U(x_s, y_s, 0) = t_A(x_s, y_s). \tag{3.410}$$

Because the aperture is rotationally symmetric, its two-dimensional Fourier transform depends only on the radial spatial frequency

$$R = \sqrt{R_x^2 + R_y^2}. \tag{3.411}$$

The result is

$$\boxed{\hat{t}_A(R) = \pi a^2 \frac{2J_1(2\pi a R)}{2\pi a R},} \tag{3.412}$$

where J_1 is the Bessel function of the first kind of order one. The appearance of a Bessel function is a direct consequence of the rotational symmetry of the aperture.

Fourier transform of a circular aperture.

The two-dimensional Fourier transform of the aperture is

$$\hat{t}_A(R_x, R_y) = \iint_{\mathbb{R}^2} t_A(x_s, y_s) e^{-j2\pi(R_x x_s + R_y y_s)}\, dx_s\, dy_s. \tag{3.413}$$

Since t_A is rotationally symmetric, it is convenient to use polar coordinates in both real and Fourier space:

$$x_s = \rho_s \cos\phi, \qquad y_s = \rho_s \sin\phi \tag{3.414}$$

and

$$R_x = R\cos\phi_R, \qquad R_y = R\sin\phi_R. \tag{3.415}$$

The scalar product in the Fourier exponential becomes

$$R_x x_s + R_y y_s = R\rho_s \left(\cos\phi_R \cos\phi + \sin\phi_R \sin\phi\right) \tag{3.416}$$
$$= R\rho_s \cos(\phi - \phi_R). \tag{3.417}$$

Since

$$dx_s\, dy_s = \rho_s\, d\rho_s\, d\phi, \tag{3.418}$$

and the aperture is non-zero only for $0 \le \rho_s \le a$, we obtain

$$\hat{t}_A(R) = \int_0^a \int_0^{2\pi} e^{-j2\pi R\rho_s \cos(\phi-\phi_R)} \rho_s\, d\phi\, d\rho_s. \tag{3.419}$$

The angular integral is related to the Bessel function J_0 through

$$J_0(q) = \frac{1}{2\pi} \int_0^{2\pi} e^{-jq\cos\phi}\, d\phi. \tag{3.420}$$

A shift of the angular variable does not change the integral, and therefore

$$\int_0^{2\pi} e^{-j2\pi R\rho_s \cos(\phi-\phi_R)}\, d\phi = 2\pi J_0(2\pi R\rho_s). \tag{3.421}$$

Hence,

$$\hat{t}_A(R) = 2\pi \int_0^a \rho_s J_0(2\pi R\rho_s)\, d\rho_s. \tag{3.422}$$

We now use the identity

$$\frac{d}{dq}\left[qJ_1(q)\right] = qJ_0(q). \tag{3.423}$$

Introducing

$$q = 2\pi R\rho_s, \qquad d\rho_s = \frac{dq}{2\pi R}, \tag{3.424}$$

gives

$$\hat{t}_A(R) = \frac{1}{2\pi R^2} \int_0^{2\pi Ra} qJ_0(q)\, dq \tag{3.425}$$
$$= \frac{1}{2\pi R^2} \left[qJ_1(q)\right]_0^{2\pi Ra} \tag{3.426}$$
$$= \frac{a}{R} J_1(2\pi aR). \tag{3.427}$$

Rearranging,

$$\boxed{\hat{t}_A(R) = \pi a^2 \frac{2J_1(2\pi a R)}{2\pi a R}.} \tag{3.428}$$

At $R = 0$, this expression has the apparent indeterminate form 0/0. Using

$$J_1(q) \simeq \frac{q}{2} \qquad (q \to 0), \tag{3.429}$$

we obtain

$$\hat{t}_A(0) = \pi a^2, \tag{3.430}$$

which is exactly the area of the aperture, as expected for the zero-frequency Fourier component.

In the Fraunhofer regime, transverse position in the observation plane is related to transverse spatial frequency by

$$R_x = \frac{x}{\lambda z}, \qquad R_y = \frac{y}{\lambda z}. \tag{3.431}$$

For a rotationally symmetric pattern,

$$R = \sqrt{R_x^2 + R_y^2} = \frac{\rho_o}{\lambda z}, \qquad \rho_o = \sqrt{x^2 + y^2}. \tag{3.432}$$

Consequently, apart from an overall amplitude factor and a phase factor, the field in the observation plane is

$$U(\rho_o, z) \propto \frac{2J_1\left(\dfrac{2\pi a \rho_o}{\lambda z}\right)}{\dfrac{2\pi a \rho_o}{\lambda z}}. \tag{3.433}$$

The corresponding intensity is

$$\boxed{I(\rho_o) \propto \left[\frac{2J_1\left(\dfrac{2\pi a \rho_o}{\lambda z}\right)}{\dfrac{2\pi a \rho_o}{\lambda z}}\right]^2.} \tag{3.434}$$

This radially symmetric intensity distribution is the familiar *Airy pattern*.

Fraunhofer diffraction pattern of a circular aperture.

In the Fraunhofer regime, the propagated field is

$$U(x,y,z) \simeq \frac{e^{jkz}}{j\lambda z} \exp\left[j\frac{\pi}{\lambda z}(x^2+y^2)\right] \hat{U}\left(\frac{x}{\lambda z}, \frac{y}{\lambda z}, 0\right). \tag{3.435}$$

For a uniformly illuminated circular aperture,

$$U(x_s, y_s, 0) = t_A(x_s, y_s), \tag{3.436}$$

so Eq. (3.412) gives

$$\hat{U}(R) = \pi a^2 \frac{2J_1(2\pi a R)}{2\pi a R}. \tag{3.437}$$

Using the Fraunhofer relation

$$R = \frac{\rho_o}{\lambda z}, \tag{3.438}$$

we obtain

$$\hat{U}\left(\frac{\rho_o}{\lambda z}\right) = \pi a^2 \frac{2J_1\left(\frac{2\pi a \rho_o}{\lambda z}\right)}{\frac{2\pi a \rho_o}{\lambda z}}. \tag{3.439}$$

Substitution into Eq. (3.435) gives

$$\boxed{U(\rho_o, z) = \frac{\pi a^2 e^{jkz}}{j\lambda z} \exp\left(j\frac{\pi \rho_o^2}{\lambda z}\right) \frac{2J_1\left(\frac{2\pi a \rho_o}{\lambda z}\right)}{\frac{2\pi a \rho_o}{\lambda z}}.} \tag{3.440}$$

The exponential

$$\exp\left(j\frac{\pi \rho_o^2}{\lambda z}\right) \tag{3.441}$$

has unit magnitude, so it modifies the phase of the field but not its intensity. Consequently,

$$I(\rho_o) = |U(\rho_o, z)|^2 \propto \left[\frac{2J_1\left(\frac{2\pi a \rho_o}{\lambda z}\right)}{\frac{2\pi a \rho_o}{\lambda z}} \right]^2, \tag{3.442}$$

which is Eq. (3.434).

The central maximum of the Airy pattern is surrounded by alternating dark and bright rings. Its first dark ring occurs at

$$\boxed{\rho_{\text{zero}} \simeq 0.61 \frac{\lambda z}{a} = 1.22 \frac{\lambda z}{D},} \tag{3.443}$$

where

$$D = 2a \tag{3.444}$$

is the aperture diameter. In angular coordinates, for small angles,

$$\boxed{\theta_{\text{zero}} \simeq 1.22 \frac{\lambda}{D}.} \tag{3.445}$$

This angular scale provides the basis of the classical Rayleigh resolution criterion. Two mutually incoherent point sources producing identical Airy patterns are conventionally said to be *just resolved* when the central maximum of one pattern coincides with the first minimum of the other. The corresponding angular separation is therefore

$$\boxed{\Delta\theta_{\text{R}} \simeq 1.22 \frac{\lambda}{D}.} \tag{3.446}$$

The Rayleigh criterion should not be interpreted as a fundamental distinction between resolvable and unresolvable objects. It is a conventional criterion arising from a particular aperture geometry, the corresponding point-spread function, and a specific definition of visual separability.

Position of the first dark ring of the Airy pattern.

The zeros of the Airy intensity occur whenever the numerator of

Eq. (3.434) vanishes:

$$J_1\left(\frac{2\pi a\rho_o}{\lambda z}\right) = 0. \tag{3.447}$$

The first positive zero of $J_1(q)$ occurs at

$$q_1 \simeq 3.8317. \tag{3.448}$$

Therefore,

$$\frac{2\pi a\rho_{\text{zero}}}{\lambda z} = 3.8317. \tag{3.449}$$

Solving for ρ_{zero},

$$\rho_{\text{zero}} = \frac{3.8317}{2\pi}\frac{\lambda z}{a} \tag{3.450}$$

$$\simeq 0.6098\frac{\lambda z}{a}. \tag{3.451}$$

Hence,

$$\boxed{\rho_{\text{zero}} \simeq 0.61\frac{\lambda z}{a}.} \tag{3.452}$$

Since $D = 2a$,

$$\boxed{\rho_{\text{zero}} \simeq 1.22\frac{\lambda z}{D}.} \tag{3.453}$$

For small diffraction angles,

$$\tan\theta \simeq \theta \simeq \frac{\rho_o}{z}. \tag{3.454}$$

Therefore,

$$\boxed{\theta_{\text{zero}} \simeq 1.22\frac{\lambda}{D}.} \tag{3.455}$$

From the Airy pattern to the Rayleigh criterion.

Consider two mutually incoherent point sources imaged through the same circular aperture. Each source produces an Airy intensity distribution.

Rayleigh proposed that the two sources be considered *just resolved*

when the central maximum of one Airy pattern coincides with the first dark ring of the other. If their angular separation is $\Delta\theta_{\mathrm{R}}$, this definition gives

$$\Delta\theta_{\mathrm{R}} = \theta_{\mathrm{zero}}. \tag{3.456}$$

Using the angular position of the first Airy minimum,

$$\theta_{\mathrm{zero}} \simeq 1.22\frac{\lambda}{D}, \tag{3.457}$$

we obtain

$$\boxed{\Delta\theta_{\mathrm{R}} \simeq 1.22\frac{\lambda}{D}.} \tag{3.458}$$

The factor 1.22 is therefore not a universal constant of resolution. It arises from the first zero of the Bessel function associated with a circular aperture, together with Rayleigh's chosen definition of when two overlapping diffraction patterns should be regarded as distinguishable.

3.7.5 From optical resolution to resolution in cryo-EM

The Airy pattern and the Rayleigh criterion are useful because they illustrate how a finite angular acceptance limits the spatial information transmitted by a wave-optical system. These concepts remain relevant in electron microscopy. Electron lenses form diffraction planes, finite objective and probe-forming apertures restrict the angular spectrum, and the resulting point-spread function is influenced by diffraction.

However, the practical resolution of a modern electron microscope is not usually determined by wavelength and aperture alone. Lens aberrations, defocus, spatial and temporal coherence, specimen scattering, radiation damage, specimen motion, detector response, and sampling all influence the transfer of spatial information. In modern instruments these effects are generally much more restrictive than the fundamental free-space limit associated with the electron wavelength itself.

There is a further distinction in cryo-electron microscopy. The "resolution" quoted for a three-dimensional reconstruction usually does not refer directly to the point resolution of the microscope. Instead, it quantifies the spatial frequencies for which the reconstructed structure is considered sufficiently reliable according to a specified criterion.

Several criteria can therefore be defined. One possibility is to determine the spatial frequency at which a signal-to-noise measure falls below a specified threshold. Another, widely used in single-particle cryo-EM, is

the spatial frequency at which the Fourier Shell Correlation between two independently reconstructed half-maps crosses a selected threshold. Additional measures can assess local resolution, directional resolution, or the compatibility of reconstructed density with expected structural features.

These different criteria answer different questions and need not yield the same numerical value. Moreover, the quality of a reconstruction may vary with position and direction because of anisotropic angular sampling, preferred orientations, conformational heterogeneity, local flexibility, or spatial variations in signal-to-noise ratio. A reconstruction may therefore possess a local and directional distribution of resolvable spatial frequencies rather than a single universal resolution value.

Resolution is consequently best understood as a quantity defined relative to a clearly specified criterion and experimental context, rather than as an absolute scalar property of either the microscope or the reconstructed object. A broader discussion of these issues can be found in Sorzano et al. (2017).

3.8 Phase transformation produced by a thin lens

A lens is said to be *thin* when its thickness is sufficiently small that the transverse displacement of a ray while traversing the lens can be neglected. A point of the incident field at transverse coordinate

$$\mathbf{r} = (x, y)^T \tag{3.459}$$

may then be regarded as leaving the lens at the same transverse coordinate. The action of the lens is consequently described locally by a complex amplitude transmittance $t_l(\mathbf{r})$,

$$U^+(\mathbf{r}) = t_l(\mathbf{r})U^-(\mathbf{r}), \tag{3.460}$$

where U^- and U^+ denote the complex fields immediately before and after the lens, respectively.

We first consider an ordinary refractive lens surrounded by vacuum or air, with refractive index n and two spherical surfaces of signed radii R_1 and R_2, as illustrated in Fig. 3.5. The wave propagates from left to right. We adopt the usual sign convention in which $R_i > 0$ when the centre of curvature of the corresponding surface lies to the right of its vertex and $R_i < 0$ when it lies to the left. Thus, for a standard biconvex converging lens,

$$R_1 > 0, \qquad R_2 < 0. \tag{3.461}$$

Let Δ_0 be the thickness of the lens on the optical axis and let $\Delta(\mathbf{r})$ be its thickness at transverse position $\mathbf{r}$. If absorption and reflections are neglected, the principal effect of the lens is to modify the phase of the transmitted field. Including the finite aperture of the lens through a pupil function $P(\mathbf{r})$, its amplitude transmittance is

$$t_l(\mathbf{r}) = P(\mathbf{r}) \exp\left[jkn\Delta(\mathbf{r})\right] \exp\left[jk\left(\Delta_0 - \Delta(\mathbf{r})\right)\right], \tag{3.462}$$

where $k = 2\pi/\lambda$ is the wavenumber in the surrounding medium. The first exponential accounts for propagation through the lens material and the second for propagation through the remaining free-space distance within the reference slab of total thickness Δ_0.

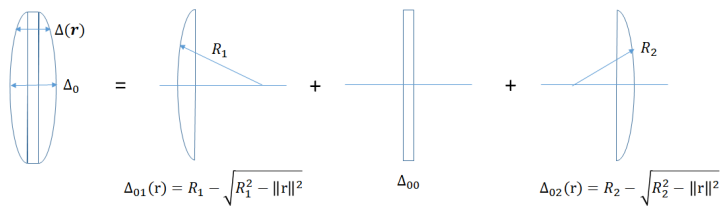


Figure 3.5: Geometry of a thin refractive lens. The two surfaces are spherical caps with signed radii of curvature R_1 and R_2. The axial thickness is Δ_0, whereas $\Delta(\mathbf{r})$ denotes the amount of lens material traversed by a ray at transverse coordinate $\mathbf{r}$. Under the thin-lens approximation, the transverse displacement of the ray while crossing the lens is neglected.

For a spherical surface of signed radius R, the displacement of the surface from its vertex at radial distance

$$\rho = \|\mathbf{r}\| \tag{3.463}$$

is

$$s_R(\rho) = R\left(1 - \sqrt{1 - \frac{\rho^2}{R^2}}\right). \tag{3.464}$$

The front surface therefore lies at $s_{R_1}(\rho)$ relative to its vertex, whereas the rear surface lies at $\Delta_0 + s_{R_2}(\rho)$. Hence, the local thickness of the lens is

$$\boxed{\Delta(\rho) = \Delta_0 - R_1\left(1 - \sqrt{1 - \frac{\rho^2}{R_1^2}}\right) + R_2\left(1 - \sqrt{1 - \frac{\rho^2}{R_2^2}}\right).} \tag{3.465}$$

In the paraxial regime,

$$\rho \ll |R_1|, \qquad \rho \ll |R_2|, \tag{3.466}$$

and the spherical surfaces can be approximated locally by paraboloids. Using

$$\sqrt{1-u} \simeq 1 - \frac{u}{2}, \qquad |u| \ll 1, \tag{3.467}$$

the lens thickness becomes

$$\boxed{\Delta(\rho) \simeq \Delta_0 - \frac{\rho^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right).} \tag{3.468}$$

Paraxial thickness of a spherical thin lens.

For a sphere of radius $|R|$, the surface coordinates satisfy

$$\rho^2 + (z-R)^2 = R^2, \tag{3.469}$$

where the vertex has been chosen at $z = 0$. Solving for the branch passing through the vertex gives

$$z = R\left(1 - \sqrt{1 - \frac{\rho^2}{R^2}}\right) = s_R(\rho). \tag{3.470}$$

This expression remains valid with the adopted signed-radius convention.

The front surface is therefore located at

$$z_1(\rho) = s_{R_1}(\rho), \tag{3.471}$$

whereas the rear surface is located at

$$z_2(\rho) = \Delta_0 + s_{R_2}(\rho). \tag{3.472}$$

The thickness traversed through the lens material is their difference,

$$\Delta(\rho) = z_2(\rho) - z_1(\rho) \tag{3.473}$$

$$= \Delta_0 + s_{R_2}(\rho) - s_{R_1}(\rho), \tag{3.474}$$

which gives Eq. (3.465).

For $\rho \ll |R|$,

$$\sqrt{1 - \frac{\rho^2}{R^2}} \simeq 1 - \frac{\rho^2}{2R^2}, \tag{3.475}$$

and therefore

$$s_R(\rho) \simeq \frac{\rho^2}{2R}. \tag{3.476}$$

Substituting this approximation for the two surfaces gives

$$\Delta(\rho) \simeq \Delta_0 - \frac{\rho^2}{2R_1} + \frac{\rho^2}{2R_2} \tag{3.477}$$

$$= \boxed{\Delta_0 - \frac{\rho^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right).} \tag{3.478}$$

Equation (3.462) can be rewritten as

$$t_l(\mathbf{r}) = P(\mathbf{r})e^{jk\Delta_0}e^{jk(n-1)\Delta(\mathbf{r})}. \tag{3.479}$$

Substituting the paraxial thickness Eq. (3.468) gives

$$t_l(\mathbf{r}) = P(\mathbf{r})e^{jkn\Delta_0}\exp\left[-jk\frac{(n-1)\rho^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\right]. \tag{3.480}$$

The factor $e^{jkn\Delta_0}$ is independent of transverse position and therefore represents a global phase. For a single optical path it has no effect on the measured intensity or on the spatial structure of the wave and will henceforth be omitted.

A positive thin lens must transform an incident plane wave into a wave that converges toward a point on the optical axis at distance f. In the paraxial approximation, such a converging spherical wave has the transverse phase

$$\exp\left(-jk\frac{\rho^2}{2f}\right). \tag{3.481}$$

Comparing Eqs. (3.480) and (3.481) gives

$$\boxed{\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right),} \tag{3.482}$$

which is the thin-lens Lensmaker equation for a lens surrounded by air or vacuum.

The paraxial amplitude transmittance of the lens can consequently be written in the compact form

$$\boxed{t_l(\mathbf{r}) = P(\mathbf{r})\exp\left(-jk\frac{\|\mathbf{r}\|^2}{2f}\right).} \tag{3.483}$$

Thus, apart from its finite pupil, the defining action of an ideal thin lens is to multiply the incident field by a quadratic phase.

Why the focusing phase is quadratic.

Consider a point on the lens plane at transverse distance ρ from the optical axis and a desired focal point located a distance f behind the lens. Their separation is

$$r_f = \sqrt{f^2 + \rho^2} = f\sqrt{1 + \frac{\rho^2}{f^2}}. \tag{3.484}$$

For $\rho \ll f$,

$$r_f \simeq f + \frac{\rho^2}{2f}. \tag{3.485}$$

With the convention $e^{-j\omega t}$, a wave converging toward the focal point has, at the lens plane, the spatial dependence

$$e^{-jkr_f}. \tag{3.486}$$

Therefore,

$$e^{-jkr_f} \simeq e^{-jkf} \exp\left(-jk\frac{\rho^2}{2f}\right). \tag{3.487}$$

The factor e^{-jkf} is spatially constant. Hence the transverse phase required to produce a paraxially converging spherical wave is

$$\boxed{\exp\left(-jk\frac{\rho^2}{2f}\right).} \tag{3.488}$$

Equating this required phase with the quadratic phase produced by the spherical refractive lens,

$$\exp\left[-jk\frac{(n-1)\rho^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\right], \tag{3.489}$$

immediately gives

$$\boxed{\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).} \tag{3.490}$$

Thus, the Lensmaker equation may be understood wave-optically as the condition that the phase accumulated through the lens have exactly the quadratic curvature required for focusing.

For a uniformly illuminated lens, the incident field is independent of $\mathbf{r}$ and Eq. (3.483) shows directly that a positive lens, $f > 0$, produces the quadratic phase of a converging spherical wave. The wave converges to the *back focal point*, located on the optical axis a distance f behind the lens. The plane through this point and perpendicular to the optical axis is the *back focal plane*.

The reverse situation follows from the same phase relation. A point source placed at the *front focal point*, a distance f in front of a positive lens, produces at the lens a paraxially diverging spherical wave with phase

$$\exp\left(+jk\frac{\rho^2}{2f}\right). \tag{3.491}$$

Multiplication by the lens phase

$$\exp\left(-jk\frac{\rho^2}{2f}\right) \tag{3.492}$$

cancels this curvature, leaving a plane wave. These two reciprocal configurations are illustrated in Fig. 3.6.

For $f < 0$, the quadratic phase has the opposite sign. An incident plane wave is then transformed into a diverging wave that appears to originate from a virtual focal point located a distance $|f|$ in front of the lens.

Focusing and collimation by phase cancellation.

For an incident plane wave of constant amplitude U_0,

$$U^-(\mathbf{r}) = U_0. \tag{3.493}$$

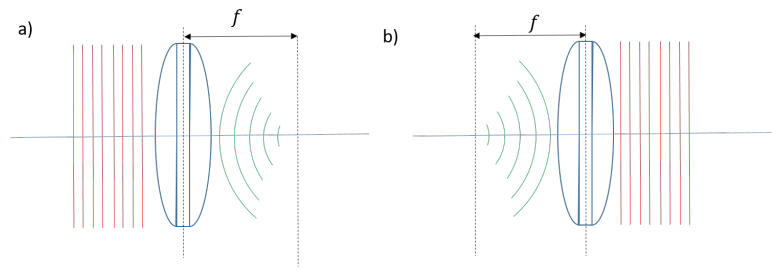


Figure 3.6: Phase transformations produced by a positive thin lens in the paraxial approximation. (a) A normally incident plane wave acquires a quadratic phase and becomes a converging spherical wave whose centre is the back focal point, a distance f behind the lens. (b) Conversely, a diverging spherical wave produced by a point source at the front focal point has its quadratic phase cancelled by the lens and emerges as a collimated plane wave.

Immediately after an ideal positive lens,

$$U^+(\mathbf{r}) = U_0 P(\mathbf{r}) \exp\left(-jk\frac{\rho^2}{2f}\right), \tag{3.494}$$

which is the paraxial phase of a spherical wave converging toward the back focal point.

Conversely, an on-axis point source located at the front focal point produces, at the lens plane,

$$U^-(\mathbf{r}) \propto \exp\left(+jk\frac{\rho^2}{2f}\right). \tag{3.495}$$

After multiplication by the lens transmittance,

$$U^+(\mathbf{r}) \propto P(\mathbf{r}) \exp\left(+jk\frac{\rho^2}{2f}\right) \exp\left(-jk\frac{\rho^2}{2f}\right) \tag{3.496}$$

$$= P(\mathbf{r}). \tag{3.497}$$

Within the pupil, the phase is therefore constant: the emerging wave is collimated.

3.8.1 The back focal plane as a Fourier plane

An especially important consequence of the quadratic lens phase is that propagation over one focal length after the lens converts transverse position in the back focal plane into spatial frequency in the field incident on the lens.

Let $\mathbf{r}$ denote a transverse coordinate in the lens plane and $\mathbf{r}_o$ a transverse coordinate in the back focal plane. Under the Fresnel approximation, the field at a distance f behind the lens is

$$\boxed{U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \exp\left(j\frac{k}{2f}\|\mathbf{r}_o\|^2\right) \widehat{PU^-}\left(\frac{\mathbf{r}_o}{\lambda f}\right).} \tag{3.498}$$

Thus, apart from a constant amplitude factor and a quadratic phase factor, the field in the back focal plane is the two-dimensional Fourier transform of the field incident on the lens, truncated by the pupil.

The spatial-frequency coordinate represented at position $\mathbf{r}_o = (x_o, y_o)^T$ is

$$\boxed{R_x = \frac{x_o}{\lambda f}, \qquad R_y = \frac{y_o}{\lambda f}.} \tag{3.499}$$

A lens therefore brings the angular spectrum to a finite plane: plane-wave components propagating at different angles are focused to different positions in the back focal plane.

Fourier-transform property of a thin lens.

Immediately after the lens,

$$U^+(\mathbf{r}) = P(\mathbf{r})U^-(\mathbf{r}) \exp\left(-jk\frac{\|\mathbf{r}\|^2}{2f}\right). \tag{3.500}$$

We now propagate this field by a distance f using the Fresnel diffraction integral:

$$U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \int_{\mathbb{R}^2} U^+(\mathbf{r}) \exp\left[j\frac{k}{2f}\|\mathbf{r}_o - \mathbf{r}\|^2\right] d\mathbf{r}. \tag{3.501}$$

Substituting Eq. (3.500),

$$U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \int P(\mathbf{r}) U^-(\mathbf{r}) \exp\left(-j\frac{k}{2f}\|\mathbf{r}\|^2\right) \times \exp\left[j\frac{k}{2f}\|\mathbf{r}_o - \mathbf{r}\|^2\right] d\mathbf{r}. \tag{3.502}$$

Expanding the squared distance,

$$\|\mathbf{r}_o - \mathbf{r}\|^2 = \|\mathbf{r}_o\|^2 + \|\mathbf{r}\|^2 - 2\mathbf{r}_o \cdot \mathbf{r}, \tag{3.503}$$

we find

$$U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \exp\left(j\frac{k}{2f}\|\mathbf{r}_o\|^2\right) \times \int P(\mathbf{r}) U^-(\mathbf{r}) \exp\left(-j\frac{k}{f}\mathbf{r}_o \cdot \mathbf{r}\right) d\mathbf{r}. \tag{3.504}$$

The quadratic phase in $\mathbf{r}$ introduced during propagation has cancelled exactly with the opposite quadratic phase introduced by the lens.

Using

$$k = \frac{2\pi}{\lambda}, \tag{3.505}$$

the remaining exponential becomes

$$\exp\left[-j2\pi\left(\frac{\mathbf{r}_o}{\lambda f}\right) \cdot \mathbf{r}\right]. \tag{3.506}$$

According to our Fourier-transform convention, the integral is therefore the Fourier transform of $P(\mathbf{r})U^-(\mathbf{r})$ evaluated at

$$\mathbf{R} = \frac{\mathbf{r}_o}{\lambda f}. \tag{3.507}$$

Hence,

$$\boxed{U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \exp\left(j\frac{k}{2f}\|\mathbf{r}_o\|^2\right) \widehat{PU^-}\left(\frac{\mathbf{r}_o}{\lambda f}\right).} \tag{3.508}$$

If the pupil does not truncate the incident field, $P = 1$ over its relevant support, the back focal plane contains, apart from scale and phase factors, the Fourier transform of the incident field itself.

3.8.2 Limits of the ideal thin-lens approximation

The ideal quadratic phase in Eq. (3.483) is a paraxial approximation. Retaining higher-order terms in the expansion of the spherical surfaces shows that the phase of a real spherical lens is not exactly quadratic away from the optical axis. These higher-order terms produce departures from the ideal converging wavefront and contribute to lens aberrations.

For example,

$$\sqrt{1-u} = 1 - \frac{u}{2} - \frac{u^2}{8} - \cdots , \tag{3.509}$$

so the sag of a spherical surface contains terms proportional to ρ^4, ρ^6, and higher powers. The first correction to the quadratic thin-lens phase is therefore fourth order in the aperture coordinate. Such higher-order phase errors are associated with spherical aberration: rays or wave components passing through different radial regions of the lens do not acquire exactly the phase required to converge to a single geometrical point.

First non-paraxial correction to the lens phase.

Expanding the spherical sag one order further gives

$$s_R(\rho) = R\left(1 - \sqrt{1 - \frac{\rho^2}{R^2}}\right) \tag{3.510}$$

$$\simeq \frac{\rho^2}{2R} + \frac{\rho^4}{8R^3} + \cdots . \tag{3.511}$$

Consequently,

$$\Delta(\rho) \simeq \Delta_0 - \frac{\rho^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) - \frac{\rho^4}{8}\left(\frac{1}{R_1^3} - \frac{1}{R_2^3}\right) + \cdots . \tag{3.512}$$

After removing the constant phase, the phase introduced by the lens is

therefore

$$\phi_l(\rho) \simeq -k\frac{\rho^2}{2f} - k(n-1)\frac{\rho^4}{8}\left(\frac{1}{R_1^3} - \frac{1}{R_2^3}\right) + \cdots . \tag{3.513}$$

The first term is the desired quadratic focusing phase. The following fourth-order term is one indication of the departure of a spherical lens from ideal paraxial behavior and contributes to spherical aberration.

3.8.3 Relation to electron lenses

The geometrical derivation above applies directly to a refractive optical lens and leads to the Lensmaker equation Eq. (3.482). An electron microscope lens is physically very different: electrons propagate predominantly in vacuum and are focused by electrostatic or, more commonly, magnetic fields rather than by refraction at material interfaces. Consequently, the focal length of an electron lens is not given by the refractive Lensmaker equation.

Nevertheless, the paraxial wave-optical description has the same mathematical structure. To lowest order, an ideal electron lens also acts on the electron wavefunction through an effective quadratic phase of the form

$$\boxed{t_l(\mathbf{r}) \propto P(\mathbf{r}) \exp\left(-jk\frac{\|\mathbf{r}\|^2}{2f}\right),} \tag{3.514}$$

where f is now determined by the electromagnetic field distribution of the electron lens and by the electron energy. This common quadratic-phase form is what allows the Fourier-optical formalism developed here to be transferred directly to electron microscopy. The departures from this ideal phase—including defocus, spherical aberration, astigmatism, coma, and higher-order aberrations—will be treated later.

3.9 Impulse response of a thin-lens imaging system

We now combine the two operations derived in the preceding sections: Fresnel propagation (Sec. 3.7) and the quadratic phase transformation produced by a thin lens (Sec. 3.8). Our objective is to determine how a point in an

object plane is represented in an image plane. This defines the *impulse response* of the imaging system.

Consider the geometry shown in Fig. 3.7. The object plane is located a distance z_1 in front of the lens, and the observation plane is located a distance z_2 behind it. Transverse coordinates in the three planes will be denoted by

$$\mathbf{r}_0 = (x_0, y_0)^T, \qquad \mathbf{r}_1 = (x_1, y_1)^T, \qquad \mathbf{r}_2 = (x_2, y_2)^T, \tag{3.515}$$

respectively.

To determine the impulse response, we consider a point object located at $\mathbf{r}_0$,

$$U_0(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_0). \tag{3.516}$$

The field produced in the observation plane by this point object will be denoted by

$$h(\mathbf{r}_2, \mathbf{r}_0). \tag{3.517}$$

Once this function is known, the response to an arbitrary object follows from linear superposition.

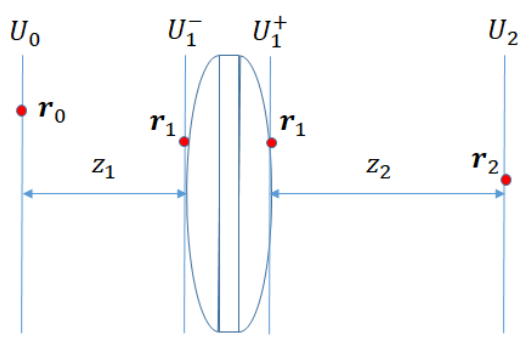


Figure 3.7: Geometry used to calculate the impulse response of a thin-lens imaging system. A point source at transverse position $\mathbf{r}_0$ in the object plane produces a spherical wave that propagates a distance z_1 to the lens. The lens transforms the incident field $U_1^-(\mathbf{r}_1)$ into $U_1^+(\mathbf{r}_1)$, which subsequently propagates a distance z_2 to the observation plane, where the field $U_2(\mathbf{r}_2)$ is formed.

Under the Fresnel approximation, propagation over a distance z is described by the kernel

$$h_z(\mathbf{r}) = \frac{e^{jkz}}{j\lambda z} \exp\left(j\frac{k}{2z}\|\mathbf{r}\|^2\right). \tag{3.518}$$

Therefore, the field immediately before the lens is

$$U_1^-(\mathbf{r}_1) = \frac{e^{jkz_1}}{j\lambda z_1} \exp\left[j\frac{k}{2z_1}\|\mathbf{r}_1 - \mathbf{r}_0\|^2\right]. \tag{3.519}$$

The amplitude transmittance of an ideal thin lens is (Eq. 3.483)

$$t_l(\mathbf{r}_1) = P(\mathbf{r}_1)\exp\left(-j\frac{k}{2f}\|\mathbf{r}_1\|^2\right), \tag{3.520}$$

so the field immediately after the lens becomes

$$U_1^+(\mathbf{r}_1) = t_l(\mathbf{r}_1)U_1^-(\mathbf{r}_1). \tag{3.521}$$

A second Fresnel propagation over the distance z_2 gives the field in the observation plane,

$$U_2(\mathbf{r}_2) = \frac{e^{jkz_2}}{j\lambda z_2}\int_{\mathbb{R}^2} U_1^+(\mathbf{r}_1)\exp\left[j\frac{k}{2z_2}\|\mathbf{r}_2 - \mathbf{r}_1\|^2\right] d\mathbf{r}_1. \tag{3.522}$$

Impulse response of the lens system.

Substituting Eq. (3.519) and the thin-lens transmittance into Eq. (3.522) gives

$$\begin{aligned} U_2(\mathbf{r}_2) = \frac{e^{jk(z_1+z_2)}}{(j\lambda)^2 z_1 z_2}\int P(\mathbf{r}_1)\exp\left(-j\frac{k}{2f}\|\mathbf{r}_1\|^2\right) \\ \times \exp\left[j\frac{k}{2z_1}\|\mathbf{r}_1 - \mathbf{r}_0\|^2\right]\exp\left[j\frac{k}{2z_2}\|\mathbf{r}_2 - \mathbf{r}_1\|^2\right] d\mathbf{r}_1. \end{aligned} \tag{3.523}$$

We expand the two squared distances,

$$\|\mathbf{r}_1 - \mathbf{r}_0\|^2 = \|\mathbf{r}_1\|^2 + \|\mathbf{r}_0\|^2 - 2\mathbf{r}_0\cdot\mathbf{r}_1 \tag{3.524}$$

and

$$\|\mathbf{r}_2 - \mathbf{r}_1\|^2 = \|\mathbf{r}_2\|^2 + \|\mathbf{r}_1\|^2 - 2\mathbf{r}_2\cdot\mathbf{r}_1. \tag{3.525}$$

Collecting all terms independent of $\mathbf{r}_1$ outside the integral, and using

$j^2 = -1$, gives

$$U_2(\mathbf{r}_2) = -\frac{e^{jk(z_1+z_2)}}{\lambda^2 z_1 z_2} \exp\left[j\frac{k}{2}\left(\frac{\|\mathbf{r}_0\|^2}{z_1} + \frac{\|\mathbf{r}_2\|^2}{z_2}\right)\right] \times \int P(\mathbf{r}_1) \exp\left[j\frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{f}\right)\|\mathbf{r}_1\|^2\right] \times \exp\left[-jk\left(\frac{\mathbf{r}_0}{z_1} + \frac{\mathbf{r}_2}{z_2}\right)\cdot \mathbf{r}_1\right] d\mathbf{r}_1. \tag{3.526}$$

Since the input was a point object, this field is the impulse response $h(\mathbf{r}_2, \mathbf{r}_0)$ of the complete optical system.

Equation (3.526) contains two qualitatively different phase factors inside the pupil integral. The term linear in $\mathbf{r}_1$ determines where the image of the point is located. The quadratic term,

$$\exp\left[j\frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{f}\right)\|\mathbf{r}_1\|^2\right], \tag{3.527}$$

measures the mismatch between the curvature required to form an image at the chosen observation plane and the curvature actually supplied by propagation and by the lens.

A geometrically focused image is obtained when this residual quadratic phase vanishes. Therefore,

$$\boxed{\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}.} \tag{3.528}$$

This is the thin-lens imaging equation. In the wave-optical description it appears naturally as a *phase-curvature matching condition.*

Wave-optical origin of the thin-lens imaging equation.

The wave emitted by a point object acquires, after propagation over z_1, the quadratic phase

$$\exp\left(+j\frac{k}{2z_1}\|\mathbf{r}_1\|^2\right). \tag{3.529}$$

The lens contributes the opposite curvature

$$\exp\left(-j\frac{k}{2f}\|\mathbf{r}_1\|^2\right), \tag{3.530}$$

and propagation from the lens to the observation plane contributes

$$\exp\left(+j\frac{k}{2z_2}\|\mathbf{r}_1\|^2\right) \tag{3.531}$$

to the pupil integral.

Multiplying these three factors gives

$$\exp\left[j\frac{k}{2}\left(\frac{1}{z_1}+\frac{1}{z_2}-\frac{1}{f}\right)\|\mathbf{r}_1\|^2\right]. \tag{3.532}$$

For every point across the pupil to contribute with the phase curvature appropriate to the same image point, the coefficient of $\|\mathbf{r}_1\|^2$ must vanish. Hence,

$$\boxed{\frac{1}{z_1}+\frac{1}{z_2}=\frac{1}{f}.} \tag{3.533}$$

Thus, the familiar lens equation from geometrical optics is recovered here as the condition for cancellation of the quadratic wavefront curvature.

When Eq. (3.528) is satisfied, the quadratic phase inside the integral disappears. Define the transverse magnification

$$\boxed{M=-\frac{z_2}{z_1}.} \tag{3.534}$$

The minus sign represents the inversion produced by a single positive lens. From this definition,

$$\frac{\mathbf{r}_0}{z_1}+\frac{\mathbf{r}_2}{z_2}=\frac{\mathbf{r}_2-M\mathbf{r}_0}{z_2}. \tag{3.535}$$

The focused point response therefore becomes

$$h(\mathbf{r}_2,\mathbf{r}_0)=C(\mathbf{r}_2,\mathbf{r}_0)\int P(\mathbf{r}_1)\exp\left[-j2\pi\frac{\mathbf{r}_2-M\mathbf{r}_0}{\lambda z_2}\cdot\mathbf{r}_1\right]d\mathbf{r}_1, \tag{3.536}$$

where

$$C(\mathbf{r}_2,\mathbf{r}_0)=-\frac{e^{jk(z_1+z_2)}}{\lambda^2 z_1 z_2}\exp\left[j\frac{k}{2}\left(\frac{\|\mathbf{r}_0\|^2}{z_1}+\frac{\|\mathbf{r}_2\|^2}{z_2}\right)\right]. \tag{3.537}$$

Using our Fourier-transform convention, Eq. (3.536) can be written as

$$\boxed{h(\mathbf{r}_2, \mathbf{r}_0) = C(\mathbf{r}_2, \mathbf{r}_0)\, \hat{P}\left(\frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2}\right).} \tag{3.538}$$

Thus, the coherent amplitude produced by a point object is, apart from known phase and scale factors, the Fourier transform of the pupil centred at the geometrical image position

$$\boxed{\mathbf{r}_2 = M\mathbf{r}_0.} \tag{3.539}$$

Why the pupil determines the point response.

The Fourier transform of the pupil is defined as

$$\hat{P}(\mathbf{R}) = \int P(\mathbf{r}_1) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_1}\, d\mathbf{r}_1. \tag{3.540}$$

Comparing this expression with Eq. (3.536), the corresponding spatial frequency is

$$\mathbf{R} = \frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2}. \tag{3.541}$$

Therefore,

$$\int P(\mathbf{r}_1) \exp\left[-j2\pi \frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2} \cdot \mathbf{r}_1\right] d\mathbf{r}_1 = \hat{P}\left(\frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2}\right). \tag{3.542}$$

The maximum of the point response occurs around the zero-frequency value of $\hat{P}$, which corresponds to

$$\mathbf{r}_2 - M\mathbf{r}_0 = 0. \tag{3.543}$$

Hence,

$$\boxed{\mathbf{r}_2 = M\mathbf{r}_0.} \tag{3.544}$$

The geometrical image location therefore emerges directly from the wave-optical integral.

The connection with geometrical optics is now transparent. An ideal geometrical lens maps a point object into a point image. Wave optics approaches this limit when the pupil admits an arbitrarily broad range of transverse spatial frequencies.

Formally, consider an infinite pupil,

$$P(\mathbf{r}_1) = 1. \tag{3.545}$$

Its Fourier transform is

$$\hat{P}(\mathbf{R}) = \delta(\mathbf{R}). \tag{3.546}$$

Consequently, the point response becomes concentrated at

$$\mathbf{r}_2 = M\mathbf{r}_0. \tag{3.547}$$

Ignoring phase factors that do not affect the intensity distribution of an individual point, the geometrical impulse response can be written as

$$\boxed{h_g(\mathbf{r}_2, \mathbf{r}_0) \propto \delta(\mathbf{r}_2 - M\mathbf{r}_0).} \tag{3.548}$$

A physical pupil has finite support, and its Fourier transform therefore has finite width. The image of a point is consequently not a mathematical point but a spatially extended diffraction pattern. The finite aperture of the microscope is thus directly responsible for the finite width of the coherent point-spread function.

For a general coherent object field $U_0(\mathbf{r}_0)$, linearity gives

$$U_2(\mathbf{r}_2) = \int U_0(\mathbf{r}_0) h(\mathbf{r}_2, \mathbf{r}_0)\, d\mathbf{r}_0. \tag{3.549}$$

Equation (3.538) contains known quadratic phase factors depending separately on object and image coordinates. It is therefore useful to define reduced fields in which these deterministic phase curvatures have been removed:

$$\widetilde{U}_0(\mathbf{r}_0) = U_0(\mathbf{r}_0) \exp\left(j\frac{k}{2z_1}\|\mathbf{r}_0\|^2\right) \tag{3.550}$$

and

$$\widetilde{U}_2(\mathbf{r}_2) = U_2(\mathbf{r}_2) \exp\left(-j\frac{k}{2z_2}\|\mathbf{r}_2\|^2\right). \tag{3.551}$$

We also express object coordinates in image-plane units by defining

$$\mathbf{r}' = M\mathbf{r}_0. \tag{3.552}$$

Apart from an overall constant, the imaging equation then takes the shift-invariant form

$$\boxed{\widetilde{U}_2(\mathbf{r}_2) = \widetilde{U}_2^g(\mathbf{r}_2) * h_c(\mathbf{r}_2),} \tag{3.553}$$

where $\widetilde{U}_2^g$ is the geometrically magnified and inverted object field, including the corresponding deterministic phase factor, and the coherent point-spread function is

$$\boxed{h_c(\mathbf{r}) \propto \hat{P}\left(\frac{\mathbf{r}}{\lambda z_2}\right).} \tag{3.554}$$

Thus, within the paraxial and isoplanatic approximations, coherent imaging has the familiar convolution structure

$$\boxed{\text{geometrical image} \quad * \quad \text{diffraction point-spread function.}} \tag{3.555}$$

Convolution form of coherent imaging.

Under the imaging condition, Eq. (3.538) gives

$$\begin{aligned} U_2(\mathbf{r}_2) = C_0 \exp\left(j\frac{k}{2z_2}\|\mathbf{r}_2\|^2\right) \\ \times \int U_0(\mathbf{r}_0) \exp\left(j\frac{k}{2z_1}\|\mathbf{r}_0\|^2\right) \hat{P}\left(\frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2}\right) d\mathbf{r}_0, \end{aligned} \tag{3.556}$$

where C_0 contains only spatially constant factors.

Removing the output quadratic phase gives

$$\widetilde{U}_2(\mathbf{r}_2) = C_0 \int \widetilde{U}_0(\mathbf{r}_0) \hat{P}\left(\frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2}\right) d\mathbf{r}_0. \tag{3.557}$$

Now introduce the image-plane coordinate

$$\mathbf{r}' = M\mathbf{r}_0. \tag{3.558}$$

Because the problem is two-dimensional,

$$d\mathbf{r}_0 = \frac{1}{|M|^2} d\mathbf{r}'. \tag{3.559}$$

Therefore,

$$\widetilde{U}_2(\mathbf{r}_2) = \frac{C_0}{|M|^2} \int \widetilde{U}_0\left(\frac{\mathbf{r}'}{M}\right) \hat{P}\left(\frac{\mathbf{r}_2 - \mathbf{r}'}{\lambda z_2}\right) d\mathbf{r}'. \tag{3.560}$$

Defining the geometrically transformed field

$$\widetilde{U}_2^g(\mathbf{r}') = \frac{C_0}{|M|^2}\widetilde{U}_0\left(\frac{\mathbf{r}'}{M}\right) \tag{3.561}$$

and

$$h_c(\mathbf{r}) = \hat{P}\left(\frac{\mathbf{r}}{\lambda z_2}\right), \tag{3.562}$$

we obtain

$$\boxed{\widetilde{U}_2 = \widetilde{U}_2^g * h_c.} \tag{3.563}$$

Thus, after accounting for magnification and the known quadratic phase curvatures, the coherent imaging system is shift invariant and its point-spread function is the scaled Fourier transform of the pupil.

For a circular pupil of radius a,

$$P(\mathbf{r}_1) = \mathrm{circ}\left(\frac{\|\mathbf{r}_1\|}{a}\right), \tag{3.564}$$

the Fourier transform of the pupil was derived previously in Sec. 3.7.4. Consequently, the coherent amplitude point-spread function has the corresponding jinc form, and its intensity is the Airy pattern discussed there.

No new diffraction phenomenon is involved here. The result obtained previously for Fraunhofer diffraction by a circular aperture now acquires a different interpretation: it is the *point-spread function of a focused imaging system with a circular pupil*.

The characteristic width scales as

$$\frac{\lambda z_2}{a}, \tag{3.565}$$

or, equivalently, in angular coordinates as λ/a. Thus, increasing the pupil size admits a broader angular spectrum and produces a narrower image of a point.

3.9.1 Defocus as a quadratic pupil phase

Suppose now that the object and observation planes do not satisfy the imaging condition exactly. Define the curvature mismatch

$$\Delta = \frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{f}. \tag{3.566}$$

Equation (3.526) then contains the residual quadratic phase

$$\exp\left(j\frac{k\Delta}{2}\|\mathbf{r}_1\|^2\right). \tag{3.567}$$

Rather than destroying the Fourier description of the imaging system, this term can be incorporated into a *complex pupil function*,

$$\boxed{P_\Delta(\mathbf{r}_1) = P(\mathbf{r}_1)\exp\left(j\frac{k\Delta}{2}\|\mathbf{r}_1\|^2\right).} \tag{3.568}$$

The point response is consequently determined by the Fourier transform of P_Δ rather than by the Fourier transform of P alone.

Thus,

$$\boxed{h_\Delta \propto \mathcal{F}\{P_\Delta\}.} \tag{3.569}$$

Defocus is therefore naturally described in wave optics as a *quadratic phase aberration across the pupil.*

Why defocus appears as a pupil phase.

The general pupil-dependent part of Eq. (3.526) is

$$\int P(\mathbf{r}_1)\exp\left(j\frac{k\Delta}{2}\|\mathbf{r}_1\|^2\right) \times \exp\left[-j2\pi\frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2}\cdot\mathbf{r}_1\right] d\mathbf{r}_1. \tag{3.570}$$

Defining

$$P_\Delta(\mathbf{r}_1) = P(\mathbf{r}_1)\exp\left(j\frac{k\Delta}{2}\|\mathbf{r}_1\|^2\right), \tag{3.571}$$

the integral becomes

$$\int P_\Delta(\mathbf{r}_1)e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_1}\,d\mathbf{r}_1 = \hat{P}_\Delta(\mathbf{R}), \tag{3.572}$$

with

$$\mathbf{R} = \frac{\mathbf{r}_2 - M\mathbf{r}_0}{\lambda z_2}. \tag{3.573}$$

Hence,

$$\boxed{h_\Delta \propto \hat{P}_\Delta.} \tag{3.574}$$

> Defocus therefore does not invalidate the Fourier-optical description. It changes the pupil from a purely amplitude-limiting function into a complex function carrying an additional quadratic phase.

This formulation is especially important in transmission electron microscopy. In conventional photography one normally attempts to satisfy the imaging condition as accurately as possible, because defocus spreads the image of a point. In cryo-electron microscopy, by contrast, controlled defocus is deliberately introduced because a weak-phase biological specimen would otherwise produce very little intensity contrast.

The relevant quantity is consequently not merely the geometrical width of a point image. What matters is how the phase modifications introduced by defocus, spherical aberration, and the finite objective aperture alter the transmission of the specimen's spatial frequencies. These effects are naturally collected into a complex pupil, or equivalently into the Contrast Transfer Function of the microscope.

The result derived in this section therefore provides the bridge from elementary diffraction theory to image formation in cryo-EM:

$$\boxed{\begin{array}{c}\text{pupil amplitude} + \text{lens and aberration phases}\\ \Downarrow \\ \text{complex pupil} \\ \xleftrightarrow{\mathcal{F}} \\ \text{coherent point-spread function.}\end{array}} \tag{3.575}$$

The explicit aberration function and the corresponding electron-microscope Contrast Transfer Function will be derived in the following sections.

3.10 Complex optical configurations

A transmission electron microscope is not formed by a single propagation distance followed by a single lens. Instead, the electron wave passes through a sequence of free-space propagation regions, electromagnetic lenses, apertures, and intermediate image and diffraction planes. A schematic example is shown in Fig. 3.8.

The propagation of the electron wave through such a column can be analyzed using the same mathematical operations introduced in the preceding sections. The apparent complexity of the microscope arises primarily from the repeated composition of a small number of elementary operations.

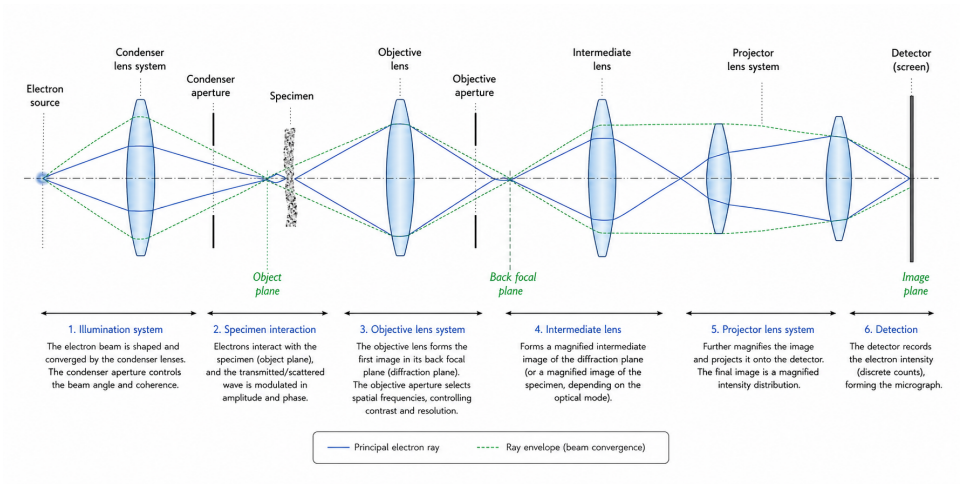


Figure 3.8: Schematic organization of a transmission electron microscope. The condenser system prepares the illumination incident on the specimen. The objective lens forms both a primary image and, in its back focal plane, a diffraction representation of the specimen exit wave. Intermediate and projector lenses subsequently select and magnify either an image plane or a diffraction plane before the wave reaches the detector. The exact number and arrangement of lenses depend on the microscope design.

3.10.1 Elementary paraxial operators

We first introduce a compact notation for the elementary operations that will repeatedly appear below.

Quadratic phase. Define

$$\boxed{\mathcal{Q}[s]\{U(\mathbf{r})\} = U(\mathbf{r}) \exp\left(j\frac{k}{2s}\|\mathbf{r}\|^2\right),} \tag{3.576}$$

where s has dimensions of length. Its inverse is

$$\mathcal{Q}[s]^{-1} = \mathcal{Q}[-s]. \tag{3.577}$$

An ideal thin lens of focal length f, apart from its pupil, therefore acts as

$$\boxed{\mathcal{L}[f] = \mathcal{Q}[-f].} \tag{3.578}$$

Pupil or aperture. If $P(\mathbf{r})$ is the pupil function, define

$$\boxed{\mathcal{P}[P]\{U(\mathbf{r})\} = P(\mathbf{r})U(\mathbf{r}).} \tag{3.579}$$

A real lens is therefore represented, in the ideal paraxial model, by

$$\mathcal{P}[P]\mathcal{Q}[-f]. \tag{3.580}$$

Free-space propagation. Under the Fresnel approximation, propagation over a distance d is described by

$$\boxed{\begin{aligned}\mathcal{R}[d]\{U\}(\mathbf{r}) = \frac{e^{jkd}}{j\lambda d}\int_{\mathbb{R}^2} U(\mathbf{r}_s) \\ \times \exp\left[j\frac{k}{2d}\|\mathbf{r}-\mathbf{r}_s\|^2\right] d\mathbf{r}_s.\end{aligned}} \tag{3.581}$$

Successive propagation distances add,

$$\boxed{\mathcal{R}[d_2]\mathcal{R}[d_1] = \mathcal{R}[d_1+d_2].} \tag{3.582}$$

Transverse magnification. For later convenience, define

$$\boxed{\mathcal{V}[M]\{U(\mathbf{r})\} = \frac{1}{|M|} U\left(\frac{\mathbf{r}}{M}\right),} \tag{3.583}$$

where M is dimensionless. With this normalization,

$$\int |\mathcal{V}[M]U|^2 \, d\mathbf{r} = \int |U|^2 \, d\mathbf{r}. \tag{3.584}$$

Successive magnifications satisfy

$$\mathcal{V}[M_2]\mathcal{V}[M_1] = \mathcal{V}[M_1 M_2]. \tag{3.585}$$

The Fourier transform operator will be denoted by $\mathcal{F}$. With the convention adopted in this book,

$$\mathcal{F}^2 = \mathcal{V}[-1], \tag{3.586}$$

because applying the Fourier transform twice reverses the transverse coordinates. Scaling and Fourier transformation satisfy

$$\boxed{\mathcal{F}\mathcal{V}[M] = \mathcal{V}[1/M]\mathcal{F}.} \tag{3.587}$$

Quadratic phases also combine particularly simply:

$$\boxed{\mathcal{Q}[s_2]\mathcal{Q}[s_1] = \mathcal{Q}\left[\left(\frac{1}{s_1} + \frac{1}{s_2}\right)^{-1}\right].} \tag{3.588}$$

These identities are sufficient for most of the manipulations that follow.

3.10.2 A lens converts angular spectrum into position

We encountered an apparently surprising consequence of the very short electron wavelength in Sec. 3.7. For an aperture or object of characteristic radius a, the Fresnel number is

$$F = \frac{a^2}{\lambda z}. \tag{3.589}$$

The characteristic propagation distance at which F becomes of order unity is therefore

$$z_F = \frac{a^2}{\lambda}. \tag{3.590}$$

For example, for

$$a = 10 \ \mu\text{m}, \qquad \lambda = 2 \text{ pm}, \tag{3.591}$$

we obtain

$$z_F = \frac{(10^{-5}\,\mathrm{m})^2}{2 \times 10^{-12}\,\mathrm{m}} \simeq 50\ \mathrm{m}. \tag{3.592}$$

A strict Fraunhofer condition $F \ll 1$ would require an even larger distance.

Clearly, an electron microscope cannot wait tens or hundreds of metres for the wave to evolve naturally into its far-field diffraction pattern. A lens provides a remarkable alternative. Its quadratic phase exactly compensates the quadratic phase that otherwise distinguishes Fresnel from Fraunhofer propagation.

Let $U(\mathbf{r})$ be the field immediately before an ideal thin lens of focal length f. After the lens and a further propagation distance f, the field is

$$\boxed{\mathcal{R}[f]\mathcal{Q}[-f]\{U\}(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \exp\left(j\frac{k}{2f}\|\mathbf{r}_o\|^2\right) \hat{U}\left(\frac{\mathbf{r}_o}{\lambda f}\right).} \tag{3.593}$$

Thus, apart from a quadratic phase and an amplitude factor, the field one focal length behind the lens is the Fourier transform of the field at the lens.

Why a lens produces a Fourier plane after one focal length.

Immediately after the lens,

$$U^+(\mathbf{r}) = U(\mathbf{r}) \exp\left(-j\frac{k}{2f}\|\mathbf{r}\|^2\right). \tag{3.594}$$

Propagation by a distance f gives

$$\begin{aligned} U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \int U(\mathbf{r}) \exp\left(-j\frac{k}{2f}\|\mathbf{r}\|^2\right) \\ \times \exp\left[j\frac{k}{2f}\|\mathbf{r}_o - \mathbf{r}\|^2\right] d\mathbf{r}. \end{aligned} \tag{3.595}$$

Expanding

$$\|\mathbf{r}_o - \mathbf{r}\|^2 = \|\mathbf{r}_o\|^2 + \|\mathbf{r}\|^2 - 2\mathbf{r}_o \cdot \mathbf{r}, \tag{3.596}$$

the two quadratic terms in $\mathbf{r}$ cancel:

$$\begin{aligned} U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f} \exp\left(j\frac{k}{2f}\|\mathbf{r}_o\|^2\right) \\ \times \int U(\mathbf{r}) \exp\left(-j\frac{k}{f}\mathbf{r}_o \cdot \mathbf{r}\right) d\mathbf{r}. \end{aligned} \tag{3.597}$$

Since

$$k = \frac{2\pi}{\lambda}, \tag{3.598}$$

the remaining exponential is

$$\exp\left[-j2\pi\left(\frac{\mathbf{r}_o}{\lambda f}\right)\cdot\mathbf{r}\right], \tag{3.599}$$

which is exactly the Fourier-transform kernel. Hence,

$$\boxed{U_f(\mathbf{r}_o) = \frac{e^{jkf}}{j\lambda f}\exp\left(j\frac{k}{2f}\|\mathbf{r}_o\|^2\right)\hat{U}\left(\frac{\mathbf{r}_o}{\lambda f}\right).} \tag{3.600}$$

The lens has therefore converted transverse spatial frequency into transverse position at a finite propagation distance.

This result explains what is sometimes described informally as a lens "bringing the far field closer." The wave has not physically propagated through the free-space distance required for the Fraunhofer regime. Instead, the lens has imposed the quadratic phase needed to cancel the Fresnel curvature, so that propagation over only one focal length produces the same Fourier mapping that would otherwise emerge asymptotically in the far field.

An especially important configuration occurs when the input field itself lies in the front focal plane of the lens. Propagation from the front focal plane to the lens, followed by the lens and propagation to the back focal plane, is represented by

$$\boxed{\mathcal{R}[f]\mathcal{Q}[-f]\mathcal{R}[f].} \tag{3.601}$$

In this configuration the additional quadratic phases cancel as well, and, apart from a spatially constant factor,

$$\boxed{U_{\mathrm{BFP}}(\mathbf{r}) \propto \hat{U}_{\mathrm{FFP}}\left(\frac{\mathbf{r}}{\lambda f}\right).} \tag{3.602}$$

Thus, conjugate front and back focal planes form a genuine Fourier pair.

Fourier relation between the front and back focal planes.

Starting with a field $U_0(\mathbf{r}_0)$ in the front focal plane, the field imme-

diately before the lens is

$$U_1^- = \mathcal{R}[f]\{U_0\}. \tag{3.603}$$

After multiplication by the lens phase and propagation over another focal length,

$$U_2 = \mathcal{R}[f]\mathcal{Q}[-f]\mathcal{R}[f]\{U_0\}. \tag{3.604}$$

Substituting the two Fresnel integrals and performing the intermediate Gaussian integral gives

$$U_2(\mathbf{r}_2) = C_f \int U_0(\mathbf{r}_0) \exp\left(-j\frac{k}{f}\mathbf{r}_0 \cdot \mathbf{r}_2\right) d\mathbf{r}_0, \tag{3.605}$$

where C_f is independent of $\mathbf{r}_0$ and $\mathbf{r}_2$. Using $k = 2\pi/\lambda$,

$$U_2(\mathbf{r}_2) = C_f \int U_0(\mathbf{r}_0) \exp\left[-j2\pi\left(\frac{\mathbf{r}_2}{\lambda f}\right) \cdot \mathbf{r}_0\right] d\mathbf{r}_0. \tag{3.606}$$

Therefore,

$$\boxed{U_2(\mathbf{r}_2) = C_f \hat{U}_0\left(\frac{\mathbf{r}_2}{\lambda f}\right).} \tag{3.607}$$

Unlike Eq. (3.593), no transverse quadratic phase remains when the field is propagated from the front focal plane all the way to the back focal plane.

The physical coordinate in the back focal plane has a simple relation to both spatial frequency and scattering angle. From Eq. (3.602),

$$\boxed{\mathbf{r}_{\mathrm{BFP}} = \lambda f\, \mathbf{R}.} \tag{3.608}$$

Since, as derived previously in Eq. (3.367),

$$\sin\theta = \lambda R, \tag{3.609}$$

we also have

$$\boxed{r_{\mathrm{BFP}} = f \sin\theta \simeq f\theta} \tag{3.610}$$

for paraxial scattering angles.

Thus, the back focal plane is simultaneously a map of propagation direction and of transverse spatial frequency.

For example, consider an idealized objective lens with

$$f_{\mathrm{obj}} = 2 \text{ mm} \tag{3.611}$$

and electrons with

$$\lambda \simeq 2 \text{ pm}. \tag{3.612}$$

A spatial frequency

$$R = 0.5 \text{ Å}^{-1}, \tag{3.613}$$

corresponding to a real-space period of 2 Å, has

$$\lambda R = (2 \times 10^{-12})(5 \times 10^{9}) = 10^{-2}. \tag{3.614}$$

The corresponding scattering angle is therefore approximately

$$\theta \simeq 10 \text{ mrad}, \tag{3.615}$$

and its location in the back focal plane is

$$r_{\text{BFP}} \simeq f_{\text{obj}}\theta = (2 \text{ mm})(10 \text{ mrad}) = 20\ \mu\text{m}. \tag{3.616}$$

Thus, atomic-scale spatial frequencies that would require tens of metres to approach their free-space Fraunhofer pattern are mapped by the objective lens into physically accessible positions only a few tens of micrometres from the optical axis in a back focal plane situated only millimetres from the lens.

This interpretation also immediately explains the role of an objective aperture placed in, or close to, the back focal plane. If the aperture has radius a_{obj}, the maximum transmitted transverse spatial frequency is approximately

$$\boxed{R_{\max} \simeq \frac{a_{\text{obj}}}{\lambda f_{\text{obj}}}.} \tag{3.617}$$

Equivalently, the maximum accepted scattering angle is

$$\boxed{\theta_{\max} \simeq \frac{a_{\text{obj}}}{f_{\text{obj}}}.} \tag{3.618}$$

Thus, an aperture located in a Fourier plane performs a direct spatial-frequency selection: restricting physical radius in the back focal plane restricts the angular spectrum transmitted by the microscope.

This is fundamentally different from placing an aperture in an image plane, where the aperture primarily restricts spatial position rather than spatial frequency.

3.10.3 ABCD representation of a paraxial lens cascade

The wave operators introduced above describe the evolution of the complex field itself. For a long paraxial optical cascade, however, it is often useful to complement the wave description with a geometrical one that tracks the position and direction of a ray. This leads to the *ABCD matrix formalism*, also known as the ray-transfer matrix formalism.

We begin with a single transverse Cartesian direction, say x. At a given longitudinal position z, a paraxial ray is described by

$$\boxed{\mathbf{q}_x = \begin{pmatrix} x \\ \theta_x \end{pmatrix},} \tag{3.619}$$

where x is the transverse displacement of the ray from the optical axis and θ_x is the signed angle made by the ray with the optical axis in the x–z plane. The geometrical meaning of these quantities and their relation to the ABCD formalism are summarized in Fig. 3.9.

Under the paraxial approximation,

$$|\theta_x| \ll 1, \tag{3.620}$$

and therefore

$$\tan\theta_x \simeq \theta_x \simeq \frac{dx}{dz}. \tag{3.621}$$

The corresponding quantities in the second transverse direction are

$$\mathbf{q}_y = \begin{pmatrix} y \\ \theta_y \end{pmatrix}. \tag{3.622}$$

For an ideal rotationally symmetric optical system, the x and y directions are uncoupled and the same ABCD matrix acts independently on both. It is therefore sufficient to derive the formalism for one transverse coordinate. More general systems exhibiting astigmatism or coupling between the transverse directions require separate matrices, or a four-dimensional ray-transfer matrix.

The propagation angles also have a direct wave-optical interpretation. In Sec. 3.6, transverse spatial frequencies were related to the direction of the corresponding plane-wave components. In particular, Eq. (3.367) gives

$$\sin\theta = \lambda R_\perp \tag{3.623}$$

for the magnitude of the transverse propagation angle.

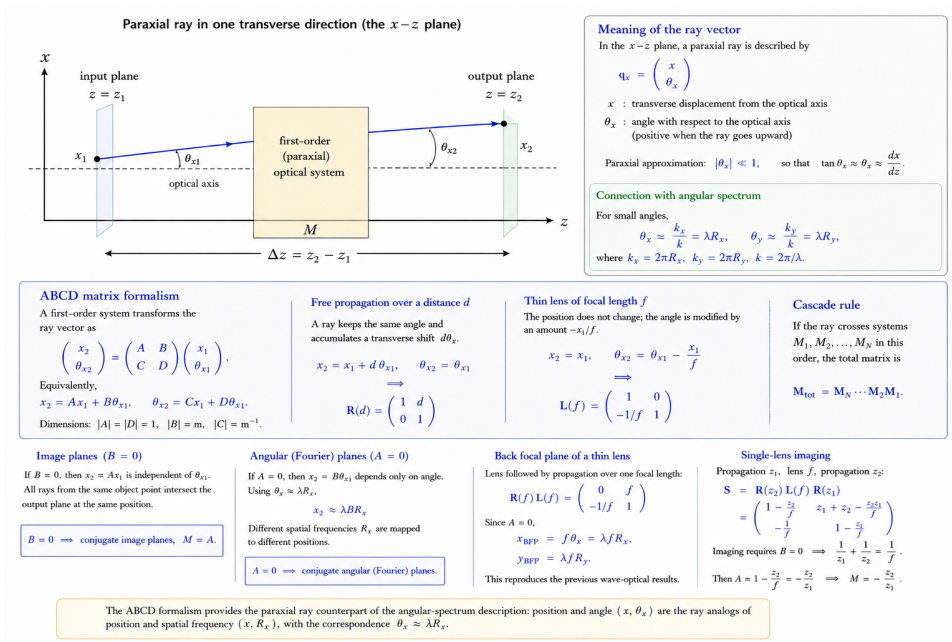


Figure 3.9: Geometrical interpretation of the paraxial ray vector and the ABCD formalism. The transverse coordinate x measures the displacement from the optical axis, while θ_x is the signed propagation angle in the x–z plane. The lower panels summarize the elementary ray-transfer matrices for free propagation and a thin lens, as well as the conditions defining conjugate image planes ($B = 0$) and angular or Fourier planes ($A = 0$).

Componentwise, in the paraxial regime,

$$\theta_x \simeq \frac{k_x}{k} = \lambda R_x, \qquad \theta_y \simeq \frac{k_y}{k} = \lambda R_y, \tag{3.624}$$

where we have used

$$k_x = 2\pi R_x, \qquad k_y = 2\pi R_y, \qquad k = \frac{2\pi}{\lambda}, \tag{3.625}$$

as established in Sec. 3.6. Thus, the ray angle used in geometrical optics is the paraxial counterpart of transverse spatial frequency in the angular-spectrum description.

A first-order paraxial optical system transforms an input ray (x_1, θ_{x1}) into an output ray (x_2, θ_{x2}) according to

$$\boxed{\begin{pmatrix} x_2 \\ \theta_{x2} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_1 \\ \theta_{x1} \end{pmatrix}.} \tag{3.626}$$

Equivalently,

$$x_2 = Ax_1 + B\theta_{x1}, \qquad \theta_{x2} = Cx_1 + D\theta_{x1}. \tag{3.627}$$

Since x has dimensions of length whereas θ_x is dimensionless, the matrix elements have dimensions

$$[A] = [D] = 1, \qquad [B] = \mathrm{m}, \qquad [C] = \mathrm{m}^{-1}. \tag{3.628}$$

If an electron successively traverses systems represented by $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N$, the total ray-transfer matrix is

$$\boxed{\mathbf{M}_{\mathrm{tot}} = \mathbf{M}_N \cdots \mathbf{M}_2 \mathbf{M}_1.} \tag{3.629}$$

Thus, matrices appear from right to left in the order in which the corresponding transformations act on the ray.

The two elementary matrices needed for the systems considered here correspond to free propagation and to an ideal thin lens.

For free propagation over a distance d, the angle does not change. Under the paraxial approximation, the transverse displacement accumulated over this distance is $d\theta_x$. Thus,

$$x_2 = x_1 + d\theta_{x1}, \qquad \theta_{x2} = \theta_{x1}, \tag{3.630}$$

and the propagation matrix is

$$\boxed{\mathbf{R}(d) = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}.} \tag{3.631}$$

For a thin lens, the transverse position does not change while the ray crosses the lens,

$$x_{\text{out}} = x_{\text{in}}. \tag{3.632}$$

What changes is the slope of the ray. In the paraxial approximation, a thin lens of focal length f changes the slope according to

$$x'_{\text{out}} = x'_{\text{in}} - \frac{x_{\text{in}}}{f}. \tag{3.633}$$

Equivalently, defining the change of slope as

$$\Delta x' \triangleq x'_{\text{out}} - x'_{\text{in}}, \tag{3.634}$$

we obtain

$$\boxed{\Delta x' = -\frac{x}{f}.} \tag{3.635}$$

Here $x = x_{\text{in}} = x_{\text{out}}$, since the transverse position is unchanged across an ideal infinitesimally thin lens.

Equations (3.632) and (3.633) can be written together as

$$\begin{pmatrix} x_{\text{out}} \\ x'_{\text{out}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} x_{\text{in}} \\ x'_{\text{in}} \end{pmatrix}. \tag{3.636}$$

The corresponding ABCD matrix is therefore

$$\boxed{\mathbf{L}(f) = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}.} \tag{3.637}$$

The same relation applies independently to the other transverse coordinate,

$$\Delta y' = -\frac{y}{f}. \tag{3.638}$$

Thus an ideal rotationally symmetric thin lens changes the transverse slope by an amount proportional to the distance from the optical axis, with the minus sign indicating a deflection toward the axis for a converging lens.

Ray-transfer matrices for propagation and a thin lens.

During free propagation, a ray travels in a straight line. If its initial transverse position is x_1 and its angle with the optical axis is θ_{x1}, then after travelling a distance d,

$$x_2 = x_1 + d \tan \theta_{x1}. \tag{3.639}$$

Using the paraxial approximation Eq. (3.621),

$$\tan \theta_{x1} \simeq \theta_{x1}, \tag{3.640}$$

so

$$x_2 = x_1 + d\theta_{x1}. \tag{3.641}$$

The propagation direction is unchanged,

$$\theta_{x2} = \theta_{x1}. \tag{3.642}$$

These two equations immediately give

$$\begin{pmatrix} x_2 \\ \theta_{x2} \end{pmatrix} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \theta_{x1} \end{pmatrix}. \tag{3.643}$$

For a thin lens, the transverse displacement while traversing the lens is neglected, as assumed in Sec. 3.8, so

$$x_2 = x_1. \tag{3.644}$$

Consider first a ray parallel to the optical axis, $\theta_{x1} = 0$, entering a positive lens at height x_1. The lens directs the ray toward the back focal point. Its outgoing slope is therefore

$$\tan \theta_{x2} = -\frac{x_1}{f}. \tag{3.645}$$

In the paraxial regime,

$$\theta_{x2} \simeq -\frac{x_1}{f}. \tag{3.646}$$

For a ray that already has an incident angle θ_{x1}, first-order linearity gives

$$\theta_{x2} = \theta_{x1} - \frac{x_1}{f}. \tag{3.647}$$

Hence,

$$\begin{pmatrix} x_2 \\ \theta_{x2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \theta_{x1} \end{pmatrix}. \tag{3.648}$$

Two particular conditions on the ABCD matrix have an immediate physical interpretation.

Consider first

$$\boxed{B = 0.} \tag{3.649}$$

Equation (3.627) then gives

$$x_2 = Ax_1. \tag{3.650}$$

The output position is independent of the angle θ_{x1}. Thus, all paraxial rays leaving the same object point x_1, regardless of their initial direction, intersect the output plane at the same transverse position x_2. The input and output planes are therefore conjugate image planes, and

$$\boxed{M = A} \tag{3.651}$$

is their transverse magnification.

Consider instead

$$\boxed{A = 0.} \tag{3.652}$$

Then

$$x_2 = B\theta_{x1}. \tag{3.653}$$

The output position is independent of the initial transverse position and depends only on propagation direction. Using Eq. (3.624),

$$x_2 \simeq \lambda B R_x. \tag{3.654}$$

The output plane is therefore an angular, or Fourier, plane: plane-wave components with different transverse spatial frequencies are mapped to different physical positions.

As an important example, consider an ideal thin lens followed by propagation over one focal length. Its matrix is

$$\mathbf{R}(f)\mathbf{L}(f) = \begin{pmatrix} 1 & f \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \tag{3.655}$$

$$= \boxed{\begin{pmatrix} 0 & f \\ -1/f & 1 \end{pmatrix}.} \tag{3.656}$$

Since $A = 0$, Eq. (3.653) gives

$$\boxed{x_{\mathrm{BFP}} = f\theta_x.} \tag{3.657}$$

Using Eq. (3.624),

$$\boxed{x_{\mathrm{BFP}} = \lambda f R_x.} \tag{3.658}$$

The corresponding two-dimensional relations are

$$x_{\mathrm{BFP}} = \lambda f R_x, \qquad y_{\mathrm{BFP}} = \lambda f R_y. \tag{3.659}$$

These results are precisely the ray-optical counterparts of Eqs. (3.610) and (3.608), obtained previously from the wave-optical analysis of the back focal plane. Thus, geometrical ray optics and Fourier wave optics lead to the same angular-to-position mapping.

As a second example, consider propagation over a distance z_1 from an object plane to the lens, passage through a thin lens of focal length f, and subsequent propagation over a distance z_2 to an observation plane. The complete ray-transfer matrix is

$$\mathbf{S} = \mathbf{R}(z_2)\mathbf{L}(f)\mathbf{R}(z_1). \tag{3.660}$$

Multiplying the matrices gives

$$\boxed{\mathbf{S} = \begin{pmatrix} 1 - \dfrac{z_2}{f} & z_1 + z_2 - \dfrac{z_1 z_2}{f} \\ -\dfrac{1}{f} & 1 - \dfrac{z_1}{f} \end{pmatrix}.} \tag{3.661}$$

For the observation plane to be conjugate to the object plane, the output position must be independent of the initial ray angle. As discussed above, this requires

$$B = 0. \tag{3.662}$$

Using Eq. (3.661),

$$z_1 + z_2 - \frac{z_1 z_2}{f} = 0. \tag{3.663}$$

Dividing by $z_1 z_2$ gives

$$\boxed{\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f},} \tag{3.664}$$

which is exactly the thin-lens imaging condition previously derived from wavefront-curvature cancellation in Eq. (3.528).

When this condition is satisfied,

$$A = 1 - \frac{z_2}{f} \tag{3.665}$$

$$= 1 - z_2 \left(\frac{1}{z_1} + \frac{1}{z_2} \right) \tag{3.666}$$

$$= -\frac{z_2}{z_1}. \tag{3.667}$$

Since A is the transverse magnification when $B = 0$,

$$\boxed{M = -\frac{z_2}{z_1}.} \tag{3.668}$$

The negative sign indicates the inversion of the image produced by a single positive lens.

The two fundamental plane-to-plane transformations can therefore be recognized directly from the ABCD matrix:

$$\boxed{\begin{array}{lll} B = 0 & \Longrightarrow & \text{conjugate image planes,} \\ A = 0 & \Longrightarrow & \text{conjugate angular/Fourier planes.} \end{array}} \tag{3.669}$$

In the first case, transverse position at the input determines transverse position at the output. In the second, propagation angle—and therefore, through Eq. (3.624), transverse spatial frequency—determines transverse position at the output.

This distinction will be particularly useful for understanding the electron-microscope column in Fig. 3.8: successive lens systems can be adjusted either to reimage a real-space plane or to reimage a diffraction plane, and a long cascade can be analyzed simply by multiplying the corresponding ABCD matrices.

3.10.4 Alternating image and Fourier planes in a focal relay

The distinction between image planes and Fourier planes becomes especially transparent in a sequence of identical lenses arranged as a focal relay. Consider first a field in the front focal plane of a lens of focal length f. Propagation over a distance f to the lens, passage through the lens, and propagation over another distance f to its back focal plane is represented by

$$\boxed{\mathbf{M}_F = \mathbf{R}(f)\mathbf{L}(f)\mathbf{R}(f).} \tag{3.670}$$

Using the elementary matrices derived in Eqs. (3.631) and (3.637),

$$\mathbf{M}_F = \begin{pmatrix} 1 & f \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & f \\ 0 & 1 \end{pmatrix} \tag{3.671}$$

$$= \boxed{\begin{pmatrix} 0 & f \\ -1/f & 0 \end{pmatrix}.} \tag{3.672}$$

Since $A = 0$, the output plane is an angular or Fourier plane according to Eq. (3.669). The ray coordinates transform as

$$\boxed{x_2 = f\theta_{x1}, \qquad \theta_{x2} = -\frac{x_1}{f}.} \tag{3.673}$$

Using the correspondence between angle and transverse spatial frequency derived in Eq. (3.624),

$$x_2 \simeq \lambda f R_x. \tag{3.674}$$

This is the same position–spatial-frequency mapping obtained previously for the back focal plane in Eq. (3.608).

Suppose that a second identical lens is arranged so that this Fourier plane is its front focal plane. After a second focal relay, the total ray-transfer matrix is

$$\mathbf{M}_F^2 = \begin{pmatrix} 0 & f \\ -1/f & 0 \end{pmatrix}^2 = \boxed{\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.} \tag{3.675}$$

Now $B = 0$, so according to Eq. (3.669) the output is an image plane. In particular,

$$x_3 = -x_1, \qquad \theta_{x3} = -\theta_{x1}. \tag{3.676}$$

Thus, two successive Fourier-transform stages produce an inverted image.

Applying further identical stages gives

$$\mathbf{M}_F^3 = \begin{pmatrix} 0 & -f \\ 1/f & 0 \end{pmatrix}, \tag{3.677}$$

$$\mathbf{M}_F^4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.678}$$

The sequence therefore repeats every four Fourier-transform stages:

$$\boxed{\begin{array}{lcl} \mathbf{M}_F^0 & : & \text{image,} \\ \mathbf{M}_F^1 & : & \text{Fourier plane,} \\ \mathbf{M}_F^2 & : & \text{inverted image,} \\ \mathbf{M}_F^3 & : & \text{inverted Fourier plane,} \\ \mathbf{M}_F^4 & : & \text{original image.} \end{array}} \tag{3.679}$$

The same result follows directly from wave optics. As derived previously for propagation between the front and back focal planes, Eq. (3.602), one focal relay performs, apart from known scale and phase factors, a Fourier transform:

$$U_1 \propto \mathcal{F}\{U_0\}. \tag{3.680}$$

A second identical relay therefore gives

$$U_2 \propto \mathcal{F}^2\{U_0\}. \tag{3.681}$$

Using the Fourier-transform property introduced earlier,

$$\mathcal{F}^2 = \mathcal{V}[-1], \tag{3.682}$$

we obtain

$$U_2(\mathbf{r}) \propto U_0(-\mathbf{r}), \tag{3.683}$$

which is an inverted image. Similarly,

$$\mathcal{F}^3 = \mathcal{F}\mathcal{V}[-1], \qquad \mathcal{F}^4 = \mathcal{I}, \tag{3.684}$$

apart from the scaling and phase conventions associated with the physical focal transforms.

Geometrically, two successive Fourier stages constitute the familiar $4f$ imaging system. For identical lenses, the sequence along the optical axis is

$$\boxed{\begin{array}{lcl} \text{object} \xrightarrow{f} \text{lens 1} \xrightarrow{f} & & \text{Fourier plane} \\ & & \downarrow \\ & \xrightarrow{f} & \text{lens 2} \xrightarrow{f} \text{image.} \end{array}} \tag{3.685}$$

The Fourier plane lies midway between the two lenses, while the image plane lies a total distance $4f$ from the original object plane. The resulting transverse magnification is $M = -1$ for two identical lenses.

More generally, if the two lenses have focal lengths f_1 and f_2, the object is placed in the front focal plane of the first lens and the output plane in the back focal plane of the second. The two lenses are separated by $f_1 + f_2$. The intermediate common focal plane is a Fourier plane, while the final plane is an image plane with transverse magnification

$$\boxed{M = -\frac{f_2}{f_1}.} \tag{3.686}$$

This alternating sequence provides a useful idealized model for the electron-microscope column. The objective lens first makes the angular spectrum of the specimen exit wave physically accessible in its back focal plane, as discussed in Sec. 3.10.2. Subsequent intermediate and projector lenses can then reimage either this diffraction plane or a real-space image plane onto later planes of the microscope.

Thus, the succession of lenses in a TEM can be viewed, at the paraxial level, as a sequence of transformations between two complementary representations:

$$\boxed{\text{real space} \xleftrightarrow{\text{lens relay}} \text{spatial-frequency space} \xleftrightarrow{\text{lens relay}} \text{real space.}} \tag{3.687}$$

This is why image and diffraction planes repeatedly appear throughout the electron-optical column.

3.10.5 Interpretation of the main lens systems of a TEM

The principal lens systems of a transmission electron microscope can now be interpreted using these ideas.

Condenser system. The condenser lenses act primarily on the illumination before it reaches the specimen. By changing their excitation, the electron source or a source crossover can be demagnified and reimaged at selected planes, while the convergence angle and illuminated specimen area can be controlled. The condenser aperture restricts the angular spectrum admitted by the illumination system. In the language introduced above, the condenser system is therefore a sequence of $\mathcal{R}$ and $\mathcal{Q}$ operations whose parameters are chosen to create the desired conjugate planes, together with pupil operations that restrict the illumination angles.

Objective lens. The objective lens is the critical first imaging lens after the specimen. In the idealized thin-lens description, the specimen exit wave lies close to the front focal plane of the objective. The back focal plane therefore contains, to a good approximation, a scaled Fourier transform of the specimen exit wave:

$$U_{\mathrm{BFP}}(\mathbf{r}) \propto \hat{U}_{\mathrm{exit}}\left(\frac{\mathbf{r}}{\lambda f_{\mathrm{obj}}}\right). \tag{3.688}$$

The objective lens simultaneously forms a primary real-space image at a different downstream plane satisfying the imaging condition.

Consequently, one physical lens naturally produces two particularly important planes: an *image plane*, containing a magnified real-space representation of the specimen, and a *back focal plane*, containing its angular or spatial-frequency representation.

Objective aperture. An aperture placed in the objective back focal plane directly restricts the spatial frequencies and scattering angles transmitted by the microscope. In conventional TEM this provides a means of selecting particular portions of the diffraction pattern. In high-resolution applications, however, such angular restriction must be considered carefully because it can remove high-spatial-frequency information.

Intermediate lens. The intermediate lens determines which plane produced by the objective system is subsequently magnified. By changing its excitation, the microscope can be configured so that the downstream projector system is conjugate either to the objective image plane or to the objective back focal plane.

In the first configuration, the detector eventually receives an enlarged real-space image. In the second, it receives an enlarged diffraction pattern. Thus, switching between *image mode* and *diffraction mode* can be understood as changing which earlier plane satisfies the $B = 0$ imaging condition with respect to the detector.

Projector system. The projector lenses provide the final stages of magnification. If the successive imaging stages have transverse magnifications $M_1, M_2, \ldots, M_N$, the total magnification is

$$\boxed{M_{\mathrm{tot}} = \prod_{i=1}^{N} M_i.} \tag{3.689}$$

This cascade is what converts the very small spatial scales formed by the objective system into a pattern large enough to be sampled by the detector.

3.10.6 Why an electron microscope contains both image and Fourier planes

The preceding analysis resolves an apparent paradox. Because the electron wavelength is extremely small, the natural free-space distance required to approach Fraunhofer diffraction can be enormous. For the illustrative values

$$a = 10\ \mu\text{m}, \qquad \lambda = 2\ \text{pm}, \tag{3.690}$$

the characteristic distance a^2/λ is approximately 50 m.

An objective lens with a millimetre-scale focal length does not somehow accelerate free-space propagation. Instead, it supplies a quadratic phase that cancels the Fresnel wavefront curvature. After only one focal length of propagation, the field therefore exhibits the same angular-to-position mapping that characterizes the far field:

$$\boxed{\mathbf{r}_{\text{BFP}} = \lambda f_{\text{obj}} \mathbf{R} \simeq f_{\text{obj}} \boldsymbol{\theta}.} \tag{3.691}$$

For an illustrative focal length

$$f_{\text{obj}} = 2\ \text{mm}, \tag{3.692}$$

the Fourier plane is produced on a millimetre rather than a tens-of-metres length scale. Compared with the characteristic free-space distance of 50 m, the relevant longitudinal scale has been reduced by a factor of approximately

$$\frac{50\ \text{m}}{2 \times 10^{-3}\ \text{m}} \simeq 2.5 \times 10^4. \tag{3.693}$$

The optical column then repeatedly reimages either real-space planes or Fourier planes, depending on the excitation of the intermediate and projector lenses. The complexity of a TEM can therefore be understood as a controlled alternation between two fundamental operations:

$$\boxed{\begin{aligned} B = 0 \quad &\Longrightarrow \quad \text{imaging of transverse position,} \\ A = 0 \quad &\Longrightarrow \quad \text{mapping of angle/spatial frequency to position.} \end{aligned}} \tag{3.694}$$

From the wave-optical point of view, the essential structure of the microscope can thus be summarized as

$$\boxed{\begin{gathered} \text{propagation} \longrightarrow \text{quadratic lens phase} \longrightarrow \\ \text{image or Fourier plane} \longrightarrow \text{reimaging and magnification.} \end{gathered}} \tag{3.695}$$

Apertures and lens aberrations modify this ideal sequence by restricting or phase-modulating the angular spectrum. These effects will become central when we derive the Contrast Transfer Function of the electron microscope.

3.11 Detection and partial coherence

The preceding sections describe how a complex wave amplitude propagates through free space, apertures, and lenses. For example, Eq. (3.292) expresses free-space propagation as a linear operation on the complex field, while Eq. (3.549) gives the corresponding linear superposition for a lens-based imaging system.

A detector, however, does not directly measure this complex amplitude. Its output is related to the flux of energy or particles arriving at the detector. This distinction becomes particularly important when the source is not perfectly coherent, because different realizations of the incident wave must then be combined statistically rather than simply added as complex amplitudes.

We therefore proceed in two steps. We first establish the relation between the electron wave function and the measurable detector signal. We then generalize the coherent propagation equations to statistically fluctuating electron waves and introduce temporal and spatial partial coherence.

3.11.1 From wave amplitude to detected intensity

It is useful first to recall the corresponding result for classical electromagnetic waves. The instantaneous energy-flux density is given by the Poynting vector,

$$\mathbf{S}(\mathbf{r},t) = \mathbf{E}(\mathbf{r},t) \times \mathbf{H}(\mathbf{r},t). \tag{3.696}$$

For monochromatic fields written as

$$\mathbf{E}(\mathbf{r},t) = \mathrm{Re}\left\{\mathbf{E}(\mathbf{r})e^{-j\omega t}\right\}, \tag{3.697}$$

and analogously for $\mathbf{H}$, averaging over many periods gives

$$\boxed{\langle \mathbf{S}(\mathbf{r})\rangle_t = \frac{1}{2}\,\mathrm{Re}\left\{\mathbf{E}(\mathbf{r}) \times \mathbf{H}^*(\mathbf{r})\right\}.} \tag{3.698}$$

For a monochromatic plane wave in a homogeneous isotropic medium,

$$\boxed{\langle \mathbf{S}(\mathbf{r})\rangle_t = \frac{|\mathbf{E}(\mathbf{r})|^2}{2\eta}\frac{\mathbf{k}}{k},} \tag{3.699}$$

where

$$\eta = \sqrt{\frac{\mu}{\epsilon}} \tag{3.700}$$

is the wave impedance. Thus, apart from a proportionality constant, classical optical intensity is proportional to the squared magnitude of the complex field.

Time-averaged Poynting vector.

Write the real fields as

$$\mathbf{E}(\mathbf{r},t) = \frac{1}{2}\left[\mathbf{E}(\mathbf{r})e^{-j\omega t} + \mathbf{E}^*(\mathbf{r})e^{j\omega t}\right], \tag{3.701}$$

$$\mathbf{H}(\mathbf{r},t) = \frac{1}{2}\left[\mathbf{H}(\mathbf{r})e^{-j\omega t} + \mathbf{H}^*(\mathbf{r})e^{j\omega t}\right]. \tag{3.702}$$

Their cross product contains four terms,

$$\begin{aligned}\mathbf{E}(t)\times\mathbf{H}(t) = \frac{1}{4}\Big[&\mathbf{E}\times\mathbf{H}\,e^{-j2\omega t} + \mathbf{E}\times\mathbf{H}^* \\ &+ \mathbf{E}^*\times\mathbf{H} + \mathbf{E}^*\times\mathbf{H}^* e^{j2\omega t}\Big].\end{aligned} \tag{3.703}$$

The terms oscillating at 2ω vanish upon averaging over many periods. The remaining two terms are complex conjugates, so

$$\langle\mathbf{S}\rangle_t = \frac{1}{4}\left(\mathbf{E}\times\mathbf{H}^* + \mathbf{E}^*\times\mathbf{H}\right) = \frac{1}{2}\,\mathrm{Re}\left\{\mathbf{E}\times\mathbf{H}^*\right\}. \tag{3.704}$$

For a plane wave,

$$\mathbf{H} = \frac{1}{\eta}\frac{\mathbf{k}}{k}\times\mathbf{E}, \tag{3.705}$$

and, because $\mathbf{E}\perp\mathbf{k}$,

$$\mathbf{E}\times\left(\frac{\mathbf{k}}{k}\times\mathbf{E}^*\right) = |\mathbf{E}|^2\frac{\mathbf{k}}{k}. \tag{3.706}$$

Therefore,

$$\langle\mathbf{S}\rangle_t = \frac{|\mathbf{E}|^2}{2\eta}\frac{\mathbf{k}}{k}, \tag{3.707}$$

which is Eq. (3.699).

An electron is not an electromagnetic wave, so the Poynting vector is not

the appropriate quantity for electron detection. The corresponding quantity is the *probability current* associated with the electron wave function.

In Sec. 3.4.1, the relativistic electron was described by the Dirac equation, Eq. (3.121). The corresponding wave function $\mathbf{\Psi}(\mathbf{r}, t)$ is a four-component Dirac spinor. Its probability density is

$$\boxed{\rho(\mathbf{r}, t) = \mathbf{\Psi}^\dagger(\mathbf{r}, t)\mathbf{\Psi}(\mathbf{r}, t),} \tag{3.708}$$

where $\mathbf{\Psi}^\dagger$ denotes the Hermitian conjugate of the spinor.

The corresponding probability-current density is

$$\boxed{\mathbf{J}(\mathbf{r}, t) = c\,\mathbf{\Psi}^\dagger(\mathbf{r}, t)\boldsymbol{\alpha}\mathbf{\Psi}(\mathbf{r}, t),} \tag{3.709}$$

where

$$\boldsymbol{\alpha} = (\alpha_x, \alpha_y, \alpha_z) \tag{3.710}$$

is the vector formed by the three 4×4 Dirac matrices already introduced in Eq. (3.121). In the standard Dirac representation,

$$\boxed{\alpha_i = \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix}, \qquad i \in \{x, y, z\},} \tag{3.711}$$

where σ_x, σ_y, and σ_z are the Pauli matrices,

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -j \\ j & 0 \end{pmatrix}, \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.712}$$

Equation (3.709) is shorthand for the three Cartesian components

$$\boxed{J_i(\mathbf{r}, t) = c\,\mathbf{\Psi}^\dagger(\mathbf{r}, t)\alpha_i\mathbf{\Psi}(\mathbf{r}, t), \qquad i \in \{x, y, z\}.} \tag{3.713}$$

Thus, each quantity $\mathbf{\Psi}^\dagger\alpha_i\mathbf{\Psi}$ is a scalar, and the three resulting scalars form the vector $\mathbf{J}$.

The physical interpretation is particularly transparent because $c\alpha_i$ is the velocity operator in the Dirac theory. The probability current therefore has the general structure of a probability density multiplied by a velocity, generalized to the relativistic four-component electron wave function.

The probability density and probability current satisfy the continuity equation

$$\boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0,} \tag{3.714}$$

which expresses local conservation of probability: a decrease of probability within a given volume must be accompanied by a net outward probability flux through its boundary.

Probability conservation for the Dirac equation.

The free Dirac equation derived in Eq. (3.121) is

$$j\hbar\frac{\partial\mathbf{\Psi}}{\partial t} = -j\hbar c\,\boldsymbol{\alpha}\cdot\nabla\mathbf{\Psi} + \beta m_0 c^2\mathbf{\Psi}. \tag{3.715}$$

Dividing by $j\hbar$ gives

$$\frac{\partial\mathbf{\Psi}}{\partial t} = -c\,\boldsymbol{\alpha}\cdot\nabla\mathbf{\Psi} - j\frac{m_0c^2}{\hbar}\beta\mathbf{\Psi}. \tag{3.716}$$

Since the Dirac matrices are Hermitian, the Hermitian conjugate equation is

$$\frac{\partial\mathbf{\Psi}^\dagger}{\partial t} = -c\,(\nabla\mathbf{\Psi}^\dagger)\cdot\boldsymbol{\alpha} + j\frac{m_0c^2}{\hbar}\mathbf{\Psi}^\dagger\beta. \tag{3.717}$$

Differentiating the probability density,

$$\frac{\partial}{\partial t}\left(\mathbf{\Psi}^\dagger\mathbf{\Psi}\right) = \frac{\partial\mathbf{\Psi}^\dagger}{\partial t}\mathbf{\Psi} + \mathbf{\Psi}^\dagger\frac{\partial\mathbf{\Psi}}{\partial t}. \tag{3.718}$$

Substituting the two evolution equations, the terms containing the rest energy cancel exactly, leaving

$$\frac{\partial\rho}{\partial t} = -c\left[(\nabla\mathbf{\Psi}^\dagger)\cdot\boldsymbol{\alpha}\mathbf{\Psi} + \mathbf{\Psi}^\dagger\boldsymbol{\alpha}\cdot\nabla\mathbf{\Psi}\right] \tag{3.719}$$

$$= -\nabla\cdot\left(c\mathbf{\Psi}^\dagger\boldsymbol{\alpha}\mathbf{\Psi}\right). \tag{3.720}$$

Defining

$$\mathbf{J} = c\mathbf{\Psi}^\dagger\boldsymbol{\alpha}\mathbf{\Psi} \tag{3.721}$$

therefore gives

$$\boxed{\frac{\partial\rho}{\partial t} + \nabla\cdot\mathbf{J} = 0.} \tag{3.722}$$

For a narrow, paraxial electron beam with essentially fixed energy and velocity v, the longitudinal probability flux is, to an excellent approximation,

$$J_z(\mathbf{r}) \propto v|\psi_E(\mathbf{r})|^2, \tag{3.723}$$

where ψ_E is the scalar spatial wave function satisfying the Helmholtz equation derived in Eq. (3.159). Since v is essentially constant across a monochro-

matic image, relative electron intensity can be represented by

$$\boxed{I(\mathbf{r}) \propto |\psi_E(\mathbf{r})|^2.} \tag{3.724}$$

Notice that this result does *not* arise because a detector averages away the temporal oscillation of a stationary electron state. For the definite-energy state derived in Eq. (3.155),

$$\psi(\mathbf{r}, t) = \psi_E(\mathbf{r}) e^{-jEt/\hbar}, \tag{3.725}$$

the temporal phase cancels exactly:

$$|\psi(\mathbf{r}, t)|^2 = |\psi_E(\mathbf{r})|^2. \tag{3.726}$$

The probability density of a stationary state is therefore time independent before any detector averaging is invoked.

If A denotes the area of one detector pixel and $\mathbf{n}$ its normal, the expected number of electrons crossing that pixel during an exposure time T is, with an appropriate normalization of the electron flux,

$$\boxed{\mu_A = T \int_A \mathbf{J}(\mathbf{r}) \cdot \mathbf{n}\, dA.} \tag{3.727}$$

Individual detections are discrete events. An ideal detector therefore samples the incident electron flux statistically, while a real detector also introduces finite detection efficiency, charge spreading, readout noise, coincidence losses, and other imperfections.

These effects are commonly summarized in electron microscopy by the Detective Quantum Efficiency,

$$\mathrm{DQE}(\mathbf{R}) = \frac{\mathrm{SNR}^2_{\mathrm{out}}(\mathbf{R})}{\mathrm{SNR}^2_{\mathrm{in}}(\mathbf{R})}, \tag{3.728}$$

which is generally spatial-frequency dependent. Detector physics will be considered in greater detail later; for the present wave-optical discussion, the essential result is that the measured signal is determined by the incident electron probability flux.

3.11.2 Statistical description of partial coherence

So far, the electron wave has been treated as a deterministic complex amplitude. A real source, however, does not reproduce exactly the same wave

function for every emitted electron. Energy, emission position, illumination angle, and phase may fluctuate between emission events.

The detector consequently samples an *ensemble* of wave fields. Throughout this section, brackets

$$\langle \cdot \rangle \tag{3.729}$$

will denote an ensemble average over realizations of the electron beam. This should not be confused with the temporal averaging used to obtain Eq. (3.698) for a classical electromagnetic wave.

Consider a general linear propagation system with impulse response $h(\mathbf{r}, \mathbf{r}_s)$. For one realization of the incident electron wave,

$$\psi_o(\mathbf{r}) = \int h(\mathbf{r}, \mathbf{r}_s)\psi_s(\mathbf{r}_s)\, d\mathbf{r}_s. \tag{3.730}$$

This is the same linear-superposition structure used previously in Eq. (3.549).

The detector measures the ensemble-averaged intensity,

$$I_o(\mathbf{r}) \propto \left\langle |\psi_o(\mathbf{r})|^2 \right\rangle. \tag{3.731}$$

To evaluate this quantity, we introduce the *mutual coherence function*

$$\boxed{\Gamma_s(\mathbf{r}_1, \mathbf{r}_2) = \langle \psi_s(\mathbf{r}_1)\psi_s^*(\mathbf{r}_2) \rangle.} \tag{3.732}$$

In quantum-mechanical language, this function is closely related to the coordinate representation of the one-electron density matrix.

The diagonal values contain the ordinary intensity,

$$\boxed{\Gamma_s(\mathbf{r}, \mathbf{r}) = \left\langle |\psi_s(\mathbf{r})|^2 \right\rangle \propto I_s(\mathbf{r}),} \tag{3.733}$$

whereas the off-diagonal values describe the degree to which the fields at two different positions preserve a reproducible relative phase.

A convenient normalized quantity is the complex degree of coherence,

$$\boxed{\mu(\mathbf{r}_1, \mathbf{r}_2) = \frac{\Gamma_s(\mathbf{r}_1, \mathbf{r}_2)}{\sqrt{\Gamma_s(\mathbf{r}_1, \mathbf{r}_1)\Gamma_s(\mathbf{r}_2, \mathbf{r}_2)}}.} \tag{3.734}$$

Its magnitude satisfies

$$0 \le |\mu| \le 1. \tag{3.735}$$

Propagation of a partially coherent field.

For one realization, Eq. (3.730) gives

$$\psi_o(\mathbf{r}) = \int h(\mathbf{r}, \mathbf{r}_1)\psi_s(\mathbf{r}_1)\, d\mathbf{r}_1. \tag{3.736}$$

Its complex conjugate is

$$\psi_o^*(\mathbf{r}) = \int h^*(\mathbf{r}, \mathbf{r}_2)\psi_s^*(\mathbf{r}_2)\, d\mathbf{r}_2. \tag{3.737}$$

Multiplying,

$$|\psi_o(\mathbf{r})|^2 = \iint \psi_s(\mathbf{r}_1)\psi_s^*(\mathbf{r}_2) \times h(\mathbf{r}, \mathbf{r}_1)h^*(\mathbf{r}, \mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2. \tag{3.738}$$

Taking the ensemble average and using Eq. (3.732),

$$\boxed{I_o(\mathbf{r}) \propto \iint \Gamma_s(\mathbf{r}_1, \mathbf{r}_2)h(\mathbf{r}, \mathbf{r}_1)h^*(\mathbf{r}, \mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2.} \tag{3.739}$$

Thus, the propagation of intensity through a partially coherent system depends not only on the intensity at each source point, but also on the correlations between every pair of source points.

3.11.3 Coherent and incoherent limits

Equation (3.739) contains both coherent and incoherent imaging as limiting cases.

For perfectly coherent illumination, every realization differs at most by a global complex factor. The mutual coherence therefore factorizes as

$$\Gamma_s(\mathbf{r}_1, \mathbf{r}_2) = \psi_s(\mathbf{r}_1)\psi_s^*(\mathbf{r}_2). \tag{3.740}$$

Substitution into Eq. (3.739) gives

$$\boxed{I_o(\mathbf{r}) \propto \left| \int h(\mathbf{r}, \mathbf{r}_s)\psi_s(\mathbf{r}_s)\, d\mathbf{r}_s \right|^2.} \tag{3.741}$$

Thus, for coherent illumination, complex amplitudes are first propagated and added, and only afterwards is the squared magnitude calculated.

Perfectly coherent limit.

Substituting Eq. (3.740) into Eq. (3.739),

$$I_o(\mathbf{r}) \propto \iint \psi_s(\mathbf{r}_1)\psi_s^*(\mathbf{r}_2)h(\mathbf{r},\mathbf{r}_1)h^*(\mathbf{r},\mathbf{r}_2)\,d\mathbf{r}_1\,d\mathbf{r}_2 \tag{3.742}$$

$$= \left[\int \psi_s(\mathbf{r}_1)h(\mathbf{r},\mathbf{r}_1)\,d\mathbf{r}_1\right] \times \left[\int \psi_s(\mathbf{r}_2)h(\mathbf{r},\mathbf{r}_2)\,d\mathbf{r}_2\right]^*, \tag{3.743}$$

and hence

$$I_o(\mathbf{r}) \propto \left|\int \psi_s(\mathbf{r}_s)h(\mathbf{r},\mathbf{r}_s)\,d\mathbf{r}_s\right|^2. \tag{3.744}$$

At the opposite extreme, a perfectly spatially incoherent field has no correlation between distinct source points. Its mutual coherence is idealized as

$$\boxed{\Gamma_s(\mathbf{r}_1,\mathbf{r}_2) = I_s(\mathbf{r}_1)\delta(\mathbf{r}_1-\mathbf{r}_2).} \tag{3.745}$$

Equation (3.739) then becomes

$$\boxed{I_o(\mathbf{r}) \propto \int I_s(\mathbf{r}_s)|h(\mathbf{r},\mathbf{r}_s)|^2\,d\mathbf{r}_s.} \tag{3.746}$$

Thus, mutually incoherent source points contribute intensities rather than complex amplitudes.

Perfectly incoherent limit.

Substituting Eq. (3.745) into Eq. (3.739) gives

$$I_o(\mathbf{r}) \propto \iint I_s(\mathbf{r}_1)\delta(\mathbf{r}_1-\mathbf{r}_2) \times h(\mathbf{r},\mathbf{r}_1)h^*(\mathbf{r},\mathbf{r}_2)\,d\mathbf{r}_1\,d\mathbf{r}_2. \tag{3.747}$$

Using the sifting property of the Dirac delta,

$$\int \delta(\mathbf{r}_1-\mathbf{r}_2)h^*(\mathbf{r},\mathbf{r}_2)\,d\mathbf{r}_2 = h^*(\mathbf{r},\mathbf{r}_1), \tag{3.748}$$

so

$$I_o(\mathbf{r}) \propto \int I_s(\mathbf{r}_1)|h(\mathbf{r},\mathbf{r}_1)|^2\, d\mathbf{r}_1. \tag{3.749}$$

For a shift-invariant system,

$$h(\mathbf{r},\mathbf{r}_s) = h(\mathbf{r}-\mathbf{r}_s), \tag{3.750}$$

as occurs for the free-space propagation derived in Eq. (3.292), these two limits reduce to

$$\boxed{I_o(\mathbf{r}) \propto |\psi_s * h|^2} \qquad \text{coherent illumination,} \tag{3.751}$$

and

$$\boxed{I_o(\mathbf{r}) \propto I_s * |h|^2} \qquad \text{incoherent illumination.} \tag{3.752}$$

It is important to interpret this distinction correctly. The detector always responds to electron flux. What changes between coherent and incoherent illumination is the statistical rule by which contributions from different parts of the incident wave combine before reaching the detector.

3.11.4 Temporal and spatial partial coherence in electron microscopy

A real electron beam generally lies between the two ideal limits described above. Two different physical mechanisms are particularly important in electron microscopy.

Temporal partial coherence arises primarily because the source has a finite energy spread. Different energies have different wavelengths and, more importantly, are focused slightly differently by the electron lenses.

Spatial partial coherence arises because the effective electron source has a finite size. Different source points illuminate the specimen with slightly different propagation directions. Since different source points are generally mutually incoherent, the observed image is an average over these illumination directions.

The two mechanisms are physically different but mathematically parallel:

$$\boxed{\begin{array}{rcl} \text{temporal coherence} & \longleftrightarrow & \text{distribution over energy } E, \\ \text{spatial coherence} & \longleftrightarrow & \text{distribution over illumination angle } \boldsymbol{\theta}. \end{array}} \tag{3.753}$$

In both cases, each member of the ensemble propagates coherently through the microscope, and the detected signal results from the statistical combination of those coherent realizations.

Energy spread and temporal coherence

The monochromatic electron considered in Sec. 3.4.1 has a definite energy E and therefore the time dependence given by Eq. (3.155). A real electron source, however, produces a distribution of energies around a mean value E_0.

Energy spread should not itself be identified with incoherence. A coherent superposition of different energies is possible:

$$\boxed{\Psi(\mathbf{r},t) = \int a(E)\psi_E(\mathbf{r})e^{-jEt/\hbar}\,dE.} \tag{3.754}$$

If the coefficients $a(E)$ possess reproducible relative phases, different energy components interfere.

Interference between different energy components.

From Eq. (3.754),

$$|\Psi(\mathbf{r},t)|^2 = \iint a(E)a^*(E')\psi_E(\mathbf{r})\psi^*_{E'}(\mathbf{r}) \times e^{-j(E-E')t/\hbar}\,dE\,dE'. \tag{3.755}$$

Terms with $E = E'$ are diagonal intensity contributions, whereas terms with $E \neq E'$ are interference terms.

For two energies,

$$\Psi = a_1\psi_1 e^{-jE_1t/\hbar} + a_2\psi_2 e^{-jE_2t/\hbar}, \tag{3.756}$$

we obtain

$$\begin{aligned}|\Psi|^2 = {} & |a_1|^2|\psi_1|^2 + |a_2|^2|\psi_2|^2 \\ & + a_1a_2^*\psi_1\psi_2^* e^{-j(E_1-E_2)t/\hbar} \\ & + a_1^*a_2\psi_1^*\psi_2 e^{j(E_1-E_2)t/\hbar}.\end{aligned} \tag{3.757}$$

The cross terms oscillate at

$$\Delta\omega = \frac{E_1 - E_2}{\hbar}. \tag{3.758}$$

They are present whenever the relative phase between the two components is coherent. Thus, different energies do not automatically add incoherently.

For a statistical electron source, define the mutual coherence between energy components as

$$\boxed{W(E, E') = \langle a(E)a^*(E')\rangle\,.} \tag{3.759}$$

This is the energy-domain counterpart of the spatial mutual coherence defined in Eq. (3.732). Its diagonal values describe the energy distribution, whereas its off-diagonal values describe coherence between different energies.

For an incoherent mixture in energy,

$$\boxed{W(E, E') = p(E)\delta(E - E'),} \tag{3.760}$$

where

$$\int p(E)\,dE = 1. \tag{3.761}$$

Energy averaging for an incoherent source.

Averaging Eq. (3.755) gives

$$I(\mathbf{r}, t) \propto \iint W(E, E')\psi_E(\mathbf{r})\psi^*_{E'}(\mathbf{r})e^{-j(E-E')t/\hbar}\,dE\,dE'. \tag{3.762}$$

Using Eq. (3.760),

$$I(\mathbf{r}) \propto \iint p(E)\delta(E - E')\psi_E(\mathbf{r})\psi^*_{E'}(\mathbf{r})\,dE\,dE' \tag{3.763}$$

$$= \boxed{\int p(E)|\psi_E(\mathbf{r})|^2\,dE.} \tag{3.764}$$

Thus, an energy-incoherent electron source produces an incoherent average over the intensities corresponding to the individual energies.

Each individual energy nevertheless propagates coherently through the microscope. If $\mathcal{L}_E$ denotes the complete wave-propagation operator at energy E,

$$\psi^{\text{out}}_E = \mathcal{L}_E\psi^{\text{in}}_E, \tag{3.765}$$

then

$$\boxed{I(\mathbf{r}) \propto \int p(E) \left|\mathcal{L}_E \psi_E^{\text{in}}(\mathbf{r})\right|^2 dE.} \tag{3.766}$$

The operator $\mathcal{L}_E$ depends on energy for at least two reasons. First, the electron wavelength depends on energy, as follows from Eq. (3.160) and Eq. (3.163). Second, the focal strength of an electron lens is energy dependent. Electrons with different energies therefore experience slightly different effective focal positions.

This latter effect is *chromatic aberration.* If C_c is the chromatic-aberration coefficient, the magnitude of the focal displacement produced by a small energy deviation δE is conventionally written, to first order, as

$$\boxed{|\delta f| \simeq C_c \frac{|\delta E|}{E_0}.} \tag{3.767}$$

The sign depends on the adopted defocus convention.

Thus, an energy distribution produces a distribution of effective defocus values. In Sec. 3.9.1, defocus was shown to introduce a quadratic pupil phase, Eq. (3.568). Energy spread therefore causes the microscope to sample a family of slightly different complex pupils rather than one unique pupil.

When image formation can be linearized so that each specimen spatial frequency is multiplied by a complex transfer factor, this statistical average naturally produces a temporal-coherence envelope.

Generic temporal-coherence envelope.

Let the linear transfer factor for spatial frequency $\mathbf{R}$ and energy E contain the phase

$$H(\mathbf{R}, E) = e^{j\chi(\mathbf{R},E)}. \tag{3.768}$$

For a narrow distribution centered at E_0,

$$\chi(\mathbf{R}, E) \simeq \chi(\mathbf{R}, E_0) + \left.\frac{\partial \chi(\mathbf{R}, E)}{\partial E}\right|_{E_0} (E - E_0). \tag{3.769}$$

The ensemble-averaged transfer factor is therefore

$$\langle H(\mathbf{R}, E)\rangle_E = \int p(E) e^{j\chi(\mathbf{R},E)}\, dE \tag{3.770}$$

$$\simeq e^{j\chi(\mathbf{R},E_0)} \int p(E) \exp\left[j \left.\frac{\partial \chi}{\partial E}\right|_{E_0} (E - E_0)\right] dE. \tag{3.771}$$

The second factor is the characteristic function of the energy distribution.

For a Gaussian energy distribution,

$$p(E) = \frac{1}{\sqrt{2\pi}\sigma_E} \exp\left[-\frac{(E - E_0)^2}{2\sigma_E^2}\right], \tag{3.772}$$

where σ_E is the standard deviation of the electron energy distribution. In electron-microscopy practice, the energy spread is often reported instead as the full width at half maximum (FWHM). For a Gaussian distribution,

$$\boxed{\Delta E_{\mathrm{FWHM}} = 2\sqrt{2\ln 2}\,\sigma_E \simeq 2.355\,\sigma_E,} \tag{3.773}$$

or equivalently,

$$\sigma_E = \frac{\Delta E_{\mathrm{FWHM}}}{2\sqrt{2\ln 2}}. \tag{3.774}$$

The Gaussian integral gives

$$\boxed{\langle H(\mathbf{R}, E)\rangle_E \simeq H(\mathbf{R}, E_0) E_t(\mathbf{R}),} \tag{3.775}$$

with

$$\boxed{E_t(\mathbf{R}) = \exp\left[-\frac{\sigma_E^2}{2} \left(\left.\frac{\partial \chi(\mathbf{R}, E)}{\partial E}\right|_{E_0} \right)^2 \right].} \tag{3.776}$$

Using the FWHM energy spread from Eq. (3.773), the same envelope can be written as

$$\boxed{E_t(\mathbf{R}) = \exp\left[-\frac{\Delta E_{\mathrm{FWHM}}^2}{16\ln 2} \left(\left.\frac{\partial \chi(\mathbf{R}, E)}{\partial E}\right|_{E_0} \right)^2 \right].} \tag{3.777}$$

This form is often more convenient when the source energy spread is taken directly from microscope or electron-gun specifications.

Thus, whenever the transfer phase becomes increasingly sensitive to energy, the corresponding spatial frequencies are progressively attenuated by energy averaging.

Equation (3.776) is a generic result. The exact temporal-coherence envelope of a cryo-EM image depends on the specific aberration phase of the objective lens and on all contributions to focal spread. It will be derived after the electron-microscope Contrast Transfer Function has been established.

Finite source size and spatial coherence

Temporal partial coherence concerns correlations between different energies. Spatial partial coherence instead concerns correlations between different transverse points or propagation directions.

An ideal point source produces, after suitable condenser optics, a single well-defined illumination wavefront. A finite source can be regarded as an ensemble of source points, each producing a slightly different incident direction at the specimen. If these source points are mutually incoherent, their corresponding illumination states contribute incoherently to the detected image.

Let $\boldsymbol{\theta} = (\theta_x, \theta_y)$ denote a small illumination tilt. In the paraxial regime, its magnitude is related to transverse spatial frequency through the angular-spectrum relation derived in Eq. (3.367),

$$\sin\theta = \lambda R_\perp \simeq \theta. \tag{3.778}$$

Thus, varying the source position is equivalent, at the specimen, to sampling a distribution of incident transverse wavevectors.

If $p(\boldsymbol{\theta})$ is the normalized angular distribution of the illumination and $\mathcal{L}_{\boldsymbol{\theta}}$ is the coherent imaging operator for a particular illumination direction, the detected intensity is

$$\boxed{I(\mathbf{r}) \propto \int p(\boldsymbol{\theta}) \left|\mathcal{L}_{\boldsymbol{\theta}} \psi^{\text{in}}_{\boldsymbol{\theta}}(\mathbf{r})\right|^2 d^2\boldsymbol{\theta}.} \tag{3.779}$$

This equation is the spatial analogue of the energy average in Eq. (3.766).

Different illumination directions sample slightly different regions of the objective-lens aberration function. At sufficiently high spatial frequency, the corresponding phases differ appreciably and their statistical average reduces

the transferred contrast. In the spatial-frequency description this effect appears as a *spatial-coherence envelope*.

The statistical structure of image formation can now be summarized as

$$\boxed{\begin{array}{c}\text{source statistics}\\ \Downarrow\\ \text{ensemble of coherent electron waves}\\ \Downarrow\\ \text{coherent propagation of each realization}\\ \Downarrow\\ \text{electron probability flux at the detector}\\ \Downarrow\\ \text{statistical average of detected intensities.}\end{array}} \tag{3.780}$$

For a perfectly coherent source, the entire ensemble reduces to a single complex wave and Eq. (3.741) applies. For a partially coherent source, the general result is Eq. (3.739). Energy spread and finite source size are two important physical mechanisms that reduce coherence, leading respectively to temporal- and spatial-coherence envelopes.

This distinction will become especially useful when the Contrast Transfer Function is derived: the coherent aberration phase determines which spatial frequencies are transferred and with what phase, while temporal and spatial partial coherence determine how strongly those oscillations survive averaging over the real electron source.

3.12 Summary

The principal objective of this chapter has been to establish a rigorous wave-optical framework that can subsequently be transferred to electron microscopy. Classical electromagnetic waves and electrons are physically very different objects: the former are described by Maxwell's equations, whereas the latter are relativistic quantum particles described by the Dirac equation. Nevertheless, under the conditions relevant to propagation through a field-free region, their spatial evolution is governed by the same mathematical equation.

This common mathematical structure is the central result of the chapter.

For a monochromatic scalar wave, Sec. 3.3.2 derived the Helmholtz equation

$$\boxed{\left(\nabla^2 + k^2\right) U(\mathbf{r}) = 0, \qquad k = \frac{2\pi}{\lambda}.} \tag{3.781}$$

For a stationary relativistic electron, starting from the Dirac equation in Eq. (3.121), Sec. 3.4.1 showed that each spatial component satisfies

$$\boxed{\left(\nabla^2 + k^2\right) \psi_E(\mathbf{r}) = 0, \qquad k = \frac{p}{\hbar} = \frac{2\pi}{\lambda}.} \tag{3.782}$$

The complete derivation is given in Eqs. (3.159)–(3.163).

Equations (3.781) and (3.782) do not imply that an electron is an electromagnetic wave. The quantities U and ψ_E have different physical meanings. What they establish is that their spatial propagation in a homogeneous, field-free region is governed by the same Helmholtz operator. This is why the mathematical machinery of scalar diffraction and Fourier optics can be carried over to electron waves.

The Fourier-optical description of propagation

Throughout the chapter we have used the Fourier-transform convention introduced in Eq. (3.2),

$$\boxed{\hat{U}(\mathbf{R}) = \int U(\mathbf{r}) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}}\, d\mathbf{r},} \tag{3.783}$$

where $\mathbf{R}$ is spatial frequency in cycles per unit length and the angular wavevector is related to it through Eq. (3.6),

$$\boxed{\mathbf{k} = 2\pi \mathbf{R}.} \tag{3.784}$$

The angular-spectrum representation developed in Sec. 3.6 gives this Fourier transform a direct physical interpretation. Each transverse spatial frequency (R_x, R_y) corresponds to a plane-wave component propagating in a particular direction. The exact propagation law between two parallel planes is

$$\boxed{\hat{U}(R_x, R_y, z) = \hat{U}(R_x, R_y, 0) \exp\left[j2\pi z \sqrt{\frac{1}{\lambda^2} - R_x^2 - R_y^2}\right],} \tag{3.785}$$

as derived in Eq. (3.264). Propagation therefore acts independently on every spatial-frequency component.

The connection between spatial frequency and propagation angle follows from Eq. (3.367),

$$\boxed{\sin\theta = \lambda R_\perp, \qquad R_\perp = \sqrt{R_x^2 + R_y^2}.} \tag{3.786}$$

This relation will remain fundamental in electron optics: spatial frequency, scattering angle, and position in a diffraction plane are different representations of the same information.

For paraxial propagation,

$$\lambda R_\perp \ll 1, \tag{3.787}$$

as established in Eq. (3.369). Expanding the exact propagation law then gives the Fresnel transfer function derived in Eq. (3.373),

$$\boxed{\hat{h}_z^{\text{Fresnel}}(R_x, R_y) = e^{jkz} \exp\left[-j\pi\lambda z\left(R_x^2 + R_y^2\right)\right].} \tag{3.788}$$

Equivalently, propagation in real space is convolution with the quadratic-phase kernel of Eq. (3.374).

The Fraunhofer regime derived in Sec. 3.7.2 introduces the additional far-field condition and reduces propagation, apart from known phase and scale factors, to a Fourier transform of the input field, as shown in Eq. (3.400). The corresponding Fresnel number is

$$\boxed{F = \frac{a^2}{\lambda z},} \tag{3.789}$$

from Eq. (3.402). The far-field regime requires $F \ll 1$.

For electrons, the picometre-scale wavelength makes the natural free-space distance a^2/λ very large. As discussed in Sec. 3.10.2, for $a = 10\ \mu$m and $\lambda = 2$ pm this characteristic distance is approximately 50 m. Electron microscopy avoids this impractical propagation distance by using lenses to produce Fourier planes at finite distances.

Lenses, image planes, and Fourier planes

Within the paraxial approximation, an ideal thin lens modifies the wave primarily through a quadratic phase. As derived in Sec. 3.8, its complex amplitude transmittance is

$$\boxed{t_l(\mathbf{r}) = P(\mathbf{r}) \exp\left(-j\frac{k}{2f}\|\mathbf{r}\|^2\right),} \tag{3.790}$$

which is Eq. (3.483). Here $P(\mathbf{r})$ represents the finite pupil and f is the focal length.

The importance of this quadratic phase is that it can cancel the quadratic phase generated by Fresnel propagation. Consequently, one focal length behind a lens the transverse angular spectrum is converted into position. The result derived in Eq. (3.593) is

$$\boxed{U_f(\mathbf{r}) \propto e^{jk\|\mathbf{r}\|^2/(2f)}\,\hat{U}\left(\frac{\mathbf{r}}{\lambda f}\right).} \tag{3.791}$$

In particular, propagation between conjugate front and back focal planes produces the Fourier relation derived in Eq. (3.602). Therefore,

$$\boxed{\mathbf{r}_{\mathrm{BFP}} = \lambda f\,\mathbf{R} \simeq f\,\boldsymbol{\theta},} \tag{3.792}$$

combining Eqs. (3.608) and (3.610).

This equation explains the physical role of the back focal plane in an electron microscope. A diffraction pattern that would require a very long distance to emerge naturally in free space is made accessible within a millimetre-scale electron-optical column by the quadratic phase introduced by the objective lens.

A lens also forms real-space images. For an object plane at distance z_1 and an image plane at distance z_2, cancellation of the residual quadratic phase gives the imaging condition derived in Eq. (3.528),

$$\boxed{\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}, \qquad M = -\frac{z_2}{z_1}.} \tag{3.793}$$

The ABCD formalism developed in Sec. 3.10.3 generalizes these results to arbitrary paraxial lens cascades. If

$$\begin{pmatrix} x_2 \\ \theta_{x2} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_1 \\ \theta_{x1} \end{pmatrix}, \tag{3.794}$$

then the two especially important conditions derived in Eq. (3.669) are

$$\boxed{\begin{array}{rcl} B = 0 & \Longrightarrow & \text{conjugate image planes,} \\ A = 0 & \Longrightarrow & \text{conjugate angular/Fourier planes.} \end{array}} \tag{3.795}$$

This provides a compact interpretation of the complete TEM column: successive lenses repeatedly form and reimage either real-space image planes or spatial-frequency diffraction planes.

What the detector measures

Although wave propagation can be described by the same Helmholtz mathematics for photons and electrons, detection has a different physical meaning. For electromagnetic radiation, measurable intensity is related to the time-averaged Poynting vector derived in Eq. (3.698). For an electron, the corresponding conserved quantity is the probability current derived in Sec. 3.11.

For the relativistic electron,

$$\rho = \mathbf{\Psi}^\dagger \mathbf{\Psi}, \qquad \mathbf{J} = c\, \mathbf{\Psi}^\dagger \boldsymbol{\alpha} \mathbf{\Psi}, \tag{3.796}$$

as given in Eqs. (3.708) and (3.709), and these quantities satisfy the continuity equation

$$\boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0.} \tag{3.797}$$

For the paraxial, essentially monoenergetic beam used in image-formation theory, the relative detected intensity is therefore represented by Eq. (3.724),

$$\boxed{I(\mathbf{r}) \propto |\psi_E(\mathbf{r})|^2.} \tag{3.798}$$

For a stationary state, this quantity is time independent because the factor $e^{-jEt/\hbar}$ cancels exactly in the squared modulus, as already shown in Eq. (3.155). No assumption of ergodicity or detector averaging is required to establish this result.

A real electron source is not perfectly coherent. The appropriate general description is statistical and is expressed through the mutual coherence function introduced in Eq. (3.732),

$$\boxed{\Gamma(\mathbf{r}_1, \mathbf{r}_2) = \langle \psi(\mathbf{r}_1) \psi^*(\mathbf{r}_2) \rangle\,.} \tag{3.799}$$

The observed intensity after a linear propagation system therefore depends on both the propagation kernel and the correlations between different parts of the incident wave, as shown in the general result Eq. (3.739).

Two important sources of partial coherence were identified in Sec. 3.11.2:

- *temporal partial coherence*, principally associated with the finite energy spread of the electron source and the corresponding chromatic focal spread;
- *spatial partial coherence*, principally associated with the finite effective source size and the resulting distribution of illumination directions.

For an energy-incoherent distribution, Eq. (3.764) gives

$$\boxed{I(\mathbf{r}) \propto \int p(E)|\psi_E(\mathbf{r})|^2\, dE.} \tag{3.800}$$

Similarly, finite spatial coherence leads to an average over illumination directions, as expressed in Eq. (3.779).

When the imaging problem is expressed through a linear transfer phase, these statistical averages lead to envelope functions. For example, for a Gaussian energy distribution, the generic temporal-coherence envelope derived in Eq. (3.776) is

$$\boxed{E_t(\mathbf{R}) = \exp\left[-\frac{\sigma_E^2}{2}\left(\frac{\partial\chi(\mathbf{R},E)}{\partial E}\bigg|_{E_0}\right)^2\right].} \tag{3.801}$$

The specific temporal- and spatial-coherence envelopes relevant to the electron-microscope Contrast Transfer Function will be developed only after the aberration function of the objective lens has been established.

Domain of validity of the framework

The wave-optical framework established in this chapter rests on a number of assumptions whose limits should remain explicit.

- **Stationary, energy-resolved propagation.** Image formation is formulated primarily in terms of stationary scattering states,

  $$\psi(\mathbf{r},t) = \psi_E(\mathbf{r})e^{-jEt/\hbar}, \tag{3.802}$$

  as discussed in Sec. 3.4.1. This does not imply that a physical electron is literally an infinitely extended monochromatic wave. It isolates the energy-resolved spatial problem; wave packets can be constructed by superposing such stationary states. The experimentally recorded image consists of many individual electron detection events, normally treated as independent realizations of this single-electron scattering problem.

- **Scalar approximation.** Classical electromagnetic fields are intrinsically vectorial and the relativistic electron is intrinsically a four-component spinor. Nevertheless, under the conditions established in Secs. 3.2 and 3.4.1, the relevant spatial components obey scalar

Helmholtz equations. Polarization, spin coupling, and other vector effects are neglected when they do not materially affect the image-formation problem.

- **Paraxial approximation.** Much of the lens theory assumes $\lambda R_\perp \ll 1$, equivalently $\theta \ll 1$, as derived in Eq. (3.369). This converts exact spherical propagation into quadratic-phase Fresnel propagation and gives the ideal lens its simple Fourier-transform properties. Terms beyond the paraxial approximation become important at larger angles. Real electron lenses additionally deviate from the ideal quadratic phase because of spherical aberration, astigmatism, coma, and higher-order aberrations.

- **Homogeneous field-free propagation between optical elements.** The Helmholtz propagation laws derived in Sec. 3.6 apply directly to homogeneous, field-free regions. Electron lenses are not pieces of refractive material: their focusing action arises from electromagnetic fields. In this chapter they have been represented by an effective paraxial phase transformation, Eq. (3.483). The physical origin of that phase and the aberrations of real electron lenses will be developed in the electron-optics chapter.

- **Linear wave propagation.** The propagation operators used throughout this chapter are linear in the complex amplitude. Partial coherence does not destroy this linearity: each realization of the source propagates coherently and linearly, after which the appropriate statistical average is taken, as established in Sec. 3.11.2.

- **The specimen has not yet been included.** The present chapter establishes how an electron wave propagates and is transformed by an idealized optical system. The crucial remaining problem is how the specimen modifies the incident electron wave. Electron–specimen interaction introduces the electrostatic potential, scattering, phase and amplitude contrast, and ultimately the exit wave that enters the electron-optical system. These effects are developed in the subsequent chapters.

What is carried forward

The central conceptual achievement of this chapter can be summarized in one chain:

$$\boxed{\begin{array}{c}\text{relativistic electron}\\ \Downarrow\\ \text{stationary spatial Helmholtz equation}\\ \Downarrow\\ \text{Fourier and angular-spectrum propagation}\\ \Downarrow\\ \text{Fresnel propagation and quadratic phases}\\ \Downarrow\\ \text{lenses, image planes, and Fourier planes}\\ \Downarrow\\ \text{probability flux and partially coherent detection.}\end{array}} \tag{3.803}$$

The importance of this chain is that most of the mathematical language required for electron microscopy is already familiar from conventional wave optics. The electron wavelength is different, the physical mechanism of focusing is different, the specimen interaction is different, and the detector counts particles rather than measuring an electromagnetic energy flux. Nevertheless, propagation, diffraction, interference, Fourier planes, pupils, point-spread functions, and coherence can all be organized within the same wave-optical framework.

The next step is therefore not to replace wave optics, but to specialize it. We must determine how electromagnetic lens fields transform a relativistic electron wave, how departures from the ideal quadratic lens phase generate aberrations, and how the electrostatic potential of the specimen produces the exit wave that is ultimately transformed into the recorded cryo-EM image.

Chapter 4

Electron optics

4.1 Electron trajectories in a uniform axial magnetic field

Before considering the spatially varying magnetic fields of a real electron lens, it is useful to study the simpler problem of an electron moving in a uniform magnetic field. This problem can be solved exactly and provides a clear physical interpretation of the bending produced by the magnetic part of the Lorentz force. It also establishes the characteristic length scale over which the transverse velocity of an electron rotates in a magnetic field.

It is important, however, not to interpret the uniform-field model as a complete model of a magnetic lens. A real round lens has a magnetic field that varies strongly along the optical axis and possesses transverse components away from the axis. These spatial variations will be essential when we derive focusing and image rotation later.

Consider a region in which

$$\mathbf{B} = B_0\hat{\mathbf{z}}, \qquad B_0 > 0, \tag{4.1}$$

and in which there is no electric field,

$$\mathbf{E} = -\nabla\Phi = \mathbf{0}. \tag{4.2}$$

The electrostatic potential Φ is therefore *constant*, but it need not be zero. In particular, for an electron that has already been accelerated through the microscope high voltage, its value relative to the cathode is approximately the accelerating potential.

Because the electrostatic potential is constant, Eq. (2.53) shows that the kinetic energy is constant. Equivalently, the acceleration potential defined

in Eq. (2.77) is constant throughout the region,

$$\widehat{\Phi} = \text{constant}. \tag{4.3}$$

Consequently, the momentum magnitude p, the velocity magnitude v, and the relativistic factor γ are also constant, according to Eq. (2.78).

4.1.1 Exact helical motion

The Lorentz equation derived in Eq. (2.48) becomes

$$\frac{d\mathbf{p}}{dt} = -e\mathbf{v} \times \mathbf{B}. \tag{4.4}$$

Since

$$\mathbf{p} = \gamma m_0 \mathbf{v} \tag{4.5}$$

and γ is constant in the present problem, the magnetic field changes only the direction of the momentum.

Let

$$\mathbf{p} = p_x \hat{\mathbf{x}} + p_y \hat{\mathbf{y}} + p_z \hat{\mathbf{z}}. \tag{4.6}$$

Substitution of Eq. (4.1) into Eq. (4.4) gives

$$\boxed{\begin{aligned} \frac{dp_x}{dt} &= -\omega_c p_y, \\ \frac{dp_y}{dt} &= \omega_c p_x, \\ \frac{dp_z}{dt} &= 0, \end{aligned}} \tag{4.7}$$

where

$$\boxed{\omega_c \triangleq \frac{eB_0}{\gamma m_0}} \tag{4.8}$$

is the relativistic cyclotron angular frequency.

Thus the momentum component parallel to the magnetic field remains constant, whereas the transverse momentum rotates uniformly in the x–y plane.

Derivation.

For the field in Eq. (4.1),

$$\mathbf{v} \times \mathbf{B} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ v_x & v_y & v_z \\ 0 & 0 & B_0 \end{vmatrix} \tag{4.9}$$

$$= B_0 v_y \hat{\mathbf{x}} - B_0 v_x \hat{\mathbf{y}}. \tag{4.10}$$

Therefore Eq. (4.4) gives

$$\frac{dp_x}{dt} = -eB_0 v_y, \tag{4.11}$$

$$\frac{dp_y}{dt} = eB_0 v_x, \tag{4.12}$$

$$\frac{dp_z}{dt} = 0. \tag{4.13}$$

Using

$$v_x = \frac{p_x}{\gamma m_0}, \qquad v_y = \frac{p_y}{\gamma m_0}, \tag{4.14}$$

we obtain

$$\frac{dp_x}{dt} = -\frac{eB_0}{\gamma m_0} p_y, \tag{4.15}$$

$$\frac{dp_y}{dt} = \frac{eB_0}{\gamma m_0} p_x, \tag{4.16}$$

which proves Eq. (4.7) and motivates the definition in Eq. (4.8).

Suppose that at $t = 0$ the transverse velocity is directed along the x-axis,

$$v_x(0) = v_\perp, \qquad v_y(0) = 0, \tag{4.17}$$

while the longitudinal velocity is $v_\parallel > 0$. The solution of Eq. (4.7) is then

$$\boxed{\begin{aligned} v_x(t) &= v_\perp \cos(\omega_c t), \\ v_y(t) &= v_\perp \sin(\omega_c t), \\ v_z(t) &= v_\parallel. \end{aligned}} \tag{4.18}$$

Integrating once with respect to time gives

$$\boxed{\begin{aligned} x(t) &= x_{\text{in}} + r_c \sin(\omega_c t), \\ y(t) &= y_{\text{in}} + r_c \left[1 - \cos(\omega_c t)\right], \\ z(t) &= v_\parallel t, \end{aligned}} \tag{4.19}$$

where

$$\boxed{r_c \triangleq \frac{v_\perp}{\omega_c} = \frac{p_\perp}{eB_0}} \tag{4.20}$$

is the cyclotron radius.

The geometry of this motion is summarized in Fig. 4.1. The magnetic field does not change the magnitude of the electron velocity, but continuously rotates its transverse component $v_\perp$. The component $v_\parallel$ parallel to the field remains unchanged. Their combination produces a helix of radius r_c and pitch P_c around a guiding center parallel to the optical axis.

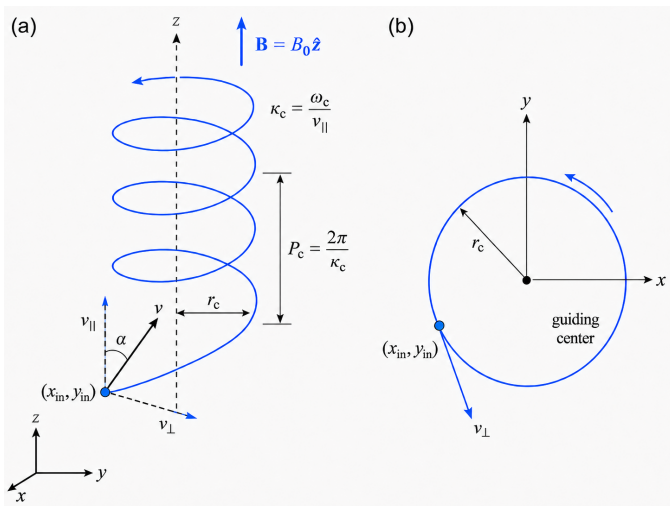


Figure 4.1: Helical motion of an electron in a uniform axial magnetic field. (a) Three-dimensional trajectory in a field $\mathbf{B} = B_0\hat{\mathbf{z}}$. The electron velocity is decomposed into a component $v_\parallel$ parallel to the magnetic field and a transverse component $v_\perp$. The latter produces circular motion with cyclotron radius r_c, while the longitudinal motion advances the electron along the optical axis. The resulting trajectory is a helix whose spatial cyclotron wavenumber is $\kappa_c = \omega_c/v_\parallel$ and whose pitch is $P_c = 2\pi/\kappa_c$. (b) Projection of the same trajectory onto the transverse x–y plane. The electron executes circular motion around the guiding center, with $v_\perp$ tangent to the orbit and r_c equal to the distance from the guiding center to the electron trajectory.

Derivation.

Differentiating the first equation in Eq. (4.7) once more gives

$$\frac{d^2 p_x}{dt^2} = -\omega_c \frac{dp_y}{dt} = -\omega_c^2 p_x. \tag{4.21}$$

Hence p_x, and therefore v_x, performs harmonic motion. The initial conditions give

$$v_x(t) = v_\perp \cos(\omega_c t). \tag{4.22}$$

The second equation in Eq. (4.7) then gives

$$v_y(t) = v_\perp \sin(\omega_c t). \tag{4.23}$$

Since $dp_z/dt = 0$, $v_z = v_\parallel$ is constant.

Integrating these three velocity components and imposing

$$x(0) = x_{\text{in}}, \qquad y(0) = y_{\text{in}}, \qquad z(0) = 0, \tag{4.24}$$

gives Eq. (4.19).

Finally,

$$r_c = \frac{v_\perp}{\omega_c} = \frac{\gamma m_0 v_\perp}{eB_0} = \frac{p_\perp}{eB_0}, \tag{4.25}$$

which proves Eq. (4.20).

Equation (4.19) describes a helix whose axis is parallel to the magnetic field. Its projection onto the transverse plane is a circle centred at

$$(x_{\text{in}}, y_{\text{in}} + r_c)\,. \tag{4.26}$$

Reversing either the sign of the charge or the direction of the magnetic field reverses the sense of rotation.

4.1.2 Representation with the optical coordinate z

For electron optics, the geometrical trajectory is usually more useful than its time parameterization. Since $v_\parallel$ is constant in the uniform field,

$$t = \frac{z}{v_\parallel}. \tag{4.27}$$

We therefore define the *spatial cyclotron wavenumber*

$$\boxed{\kappa_c \triangleq \frac{\omega_c}{v_\parallel} = \frac{eB_0}{p_\parallel}.} \tag{4.28}$$

Its units are $\mathrm{rad\,m^{-1}}$: it measures the cyclotron phase accumulated per unit distance along the optical axis.

Let α be the angle between the incident trajectory and the optical axis. Then

$$\frac{v_\perp}{v_\parallel} = \frac{p_\perp}{p_\parallel} = \tan\alpha. \tag{4.29}$$

Substitution of $t = z/v_\parallel$ into Eq. (4.19) gives the exact geometrical trajectory

$$\boxed{\begin{aligned} x(z) &= x_{\mathrm{in}} + r_c \sin(\kappa_c z), \\ y(z) &= y_{\mathrm{in}} + r_c \left[1 - \cos(\kappa_c z)\right], \end{aligned}} \tag{4.30}$$

with

$$\boxed{r_c = \frac{\tan\alpha}{\kappa_c} = \frac{p_\perp}{eB_0}.} \tag{4.31}$$

The longitudinal distance required for one complete cyclotron revolution is the pitch of the helix,

$$\boxed{P_c = \frac{2\pi}{\kappa_c} = \frac{2\pi p_\parallel}{eB_0}.} \tag{4.32}$$

Notice that no paraxial approximation has yet been made in Eqs. (4.30)–(4.32). The only requirement for using z as the trajectory parameter is $v_\parallel > 0$.

4.1.3 Paraxial form

In conventional electron optics, trajectories remain close to the optical axis, so that

$$|\alpha| \ll 1. \tag{4.33}$$

Therefore,

$$\tan\alpha \simeq \alpha, \qquad p_\parallel = p\cos\alpha \simeq p. \tag{4.34}$$

The spatial cyclotron wavenumber then becomes

$$\boxed{\kappa_c \simeq \frac{eB_0}{p}.} \tag{4.35}$$

Using the momentum–acceleration-potential relation from Eq. (2.78),

$$p = \sqrt{2m_0 e\widehat{\Phi}}, \tag{4.36}$$

and the definition of η in Eq. (2.76), we obtain

$$\boxed{\kappa_c \simeq \eta \frac{B_0}{\sqrt{\widehat{\Phi}}}.} \tag{4.37}$$

Derivation.

From Eq. (4.35),

$$\kappa_c \simeq \frac{eB_0}{\sqrt{2m_0 e\widehat{\Phi}}} \tag{4.38}$$

$$= \sqrt{\frac{e}{2m_0}} \frac{B_0}{\sqrt{\widehat{\Phi}}}. \tag{4.39}$$

Since

$$\eta = \sqrt{\frac{e}{2m_0}} \tag{4.40}$$

by Eq. (2.76), we obtain Eq. (4.37).

The same result can be written directly in terms of the electron wavelength. From the de Broglie relation, Eq. (2.101),

$$p = \frac{h}{\lambda}, \tag{4.41}$$

so that

$$\boxed{\kappa_c \simeq \frac{eB_0\lambda}{h}.} \tag{4.42}$$

This expression provides a direct connection between the particle description used here and the wave description developed in Chapter 3.

For a paraxial ray entering in the x–z plane,

$$x'(0) \simeq \alpha, \qquad y'(0) = 0, \tag{4.43}$$

and Eq. (4.31) reduces to

$$\boxed{r_c \simeq \frac{\alpha}{\kappa_c}.} \tag{4.44}$$

The angle α is the same small propagation angle that appears in the angular-spectrum description of Sec. 3.6; its relation to transverse spatial frequency was established in Eq. (3.367).

4.1.4 Numerical scale for a 300 kV electron

Consider, as an illustrative example, an electron accelerated through $V = 300$ kV. The initial emission energy is negligible on this scale, so that $\Phi_0 \simeq 0$, while the potential in the region after acceleration is

$$\Phi \simeq 3.0 \times 10^5 \text{ V}. \tag{4.45}$$

This distinction is important: Φ_0 denotes the small initial kinetic energy expressed as a voltage, whereas V is the accelerating potential difference.

Using the acceleration potential defined in Eq. (2.77),

$$\widehat{\Phi} = V(1 + \epsilon V) \tag{4.46}$$

$$= (3.0 \times 10^5)\left[1 + (0.9785 \times 10^{-6})(3.0 \times 10^5)\right] \tag{4.47}$$

$$\simeq 3.88 \times 10^5 \text{ V}. \tag{4.48}$$

For an illustrative axial magnetic field

$$B_0 = 2 \text{ T}, \tag{4.49}$$

Eq. (4.37) gives

$$\kappa_c \simeq (2.965 \times 10^5)\frac{2}{\sqrt{3.88 \times 10^5}} \tag{4.50}$$

$$\simeq 9.52 \times 10^2 \text{ rad}\,\text{m}^{-1}. \tag{4.51}$$

For an entrance angle of

$$\alpha = 1 \text{ mrad}, \tag{4.52}$$

the cyclotron radius is, from Eq. (4.44),

$$r_c \simeq \frac{10^{-3}}{9.52 \times 10^2} \tag{4.53}$$

$$\simeq 1.05 \times 10^{-6} \text{ m} \tag{4.54}$$

$$\simeq 1.05\ \mu\text{m}. \tag{4.55}$$

The pitch of the helix is

$$P_c = \frac{2\pi}{\kappa_c} \tag{4.56}$$

$$\simeq 6.60 \times 10^{-3} \text{ m} \tag{4.57}$$

$$\simeq 6.6 \text{ mm}. \tag{4.58}$$

These values illustrate that, despite the very high longitudinal velocity of a 300 kV electron, a magnetic field of a few tesla can produce an appreciable rotation of its transverse momentum over millimetre distances.

4.1.5 What the uniform-field model does and does not describe

The helical trajectory derived above provides an exactly solvable example of the action of a magnetic field on an electron. It shows that the magnetic Lorentz force does no work, but continuously changes the direction of the transverse momentum, producing cyclotron motion with spatial wavenumber κ_c given by Eq. (4.28).

A uniform axial magnetic field, however, is not yet a magnetic *lens*. In particular, the field

$$\mathbf{B} = B_0\hat{\mathbf{z}} \tag{4.59}$$

has no transverse component and no spatial variation. It bends individual electron trajectories into helices, but it does not contain the fringe fields that are responsible for the focusing action of a real round magnetic lens.

A realistic magnetic lens has an axial field whose magnitude changes with position,

$$B_z = B_0(z) \qquad \text{on the optical axis.} \tag{4.60}$$

As we shall see below, Maxwell's equations then require the existence of a radial magnetic-field component away from the axis. It is the combined action of the axial and radial components that produces two effects: the transverse coordinates rotate about the optical axis and, simultaneously, the electron bundle is focused.

The first of these effects is conventionally called *Larmor rotation*. Rather than introducing it as an additional physical principle, we shall derive it directly from the Lorentz equation. We will find that the local spatial rate of Larmor rotation is

$$\boxed{\frac{d\theta_L}{dz} = \frac{\kappa(z)}{2},} \tag{4.61}$$

where

$$\kappa(z) = \frac{eB_0(z)}{p} \tag{4.62}$$

is the local generalization of the cyclotron spatial wavenumber introduced in Eq. (4.28). Thus Larmor rotation occurs at *one half* of the local cyclotron rate. This factor of two will emerge naturally from the equations below.

4.2 Paraxial trajectories in a rotationally symmetric magnetic lens

Consider now a static magnetic lens that is rotationally symmetric about the optical axis. Cylindrical coordinates (r, ϕ, z) are particularly appropriate. In the current-free region traversed by the electron,

$$\nabla \cdot \mathbf{B} = 0, \qquad \nabla \times \mathbf{B} = \mathbf{0}, \tag{4.63}$$

and rotational symmetry implies

$$B_\phi = 0, \qquad \frac{\partial}{\partial \phi} = 0. \tag{4.64}$$

Let

$$B_0(z) \triangleq B_z(0, z) \tag{4.65}$$

denote the magnetic field on the optical axis. Close to the axis, the magnetic field is

$$\boxed{\begin{aligned} B_z(r,z) &= B_0(z) - \frac{r^2}{4}\frac{d^2 B_0}{dz^2} + O(r^4), \\ B_r(r,z) &= -\frac{r}{2}\frac{dB_0}{dz} + O(r^3). \end{aligned}} \tag{4.66}$$

Equation (4.66) already contains the essential physics of a magnetic lens. Wherever the axial field changes with z, there must be a radial field proportional to $-rB_0'(z)/2$.

Near-axis magnetic field of a round lens.

Rotational symmetry allows us to expand the field near $r = 0$ as

$$B_z(r,z) = B_0(z) + b_2(z)r^2 + O(r^4), \tag{4.67}$$

$$B_r(r,z) = a_1(z)r + a_3(z)r^3 + O(r^5). \tag{4.68}$$

Only even powers occur in B_z and odd powers in B_r because the field must remain regular and symmetric about the axis.

The divergence-free condition in cylindrical coordinates is

$$\frac{1}{r}\frac{\partial}{\partial r}(rB_r) + \frac{\partial B_z}{\partial z} = 0. \tag{4.69}$$

Substitution of the expansions gives, to zeroth order in r,

$$2a_1(z) + B_0'(z) = 0, \tag{4.70}$$

and hence

$$a_1(z) = -\frac{1}{2}B_0'(z). \tag{4.71}$$

Therefore,

$$B_r(r,z) = -\frac{r}{2}B_0'(z) + O(r^3). \tag{4.72}$$

The azimuthal component of $\nabla \times \mathbf{B} = \mathbf{0}$ is

$$\frac{\partial B_r}{\partial z} - \frac{\partial B_z}{\partial r} = 0. \tag{4.73}$$

At first order in r,

$$a_1'(z)r - 2b_2(z)r = 0, \tag{4.74}$$

so that

$$b_2(z) = \frac{1}{2}a_1'(z) = -\frac{1}{4}B_0''(z). \tag{4.75}$$

This gives Eq. (4.66).

4.2.1 Paraxial trajectory equations

We now consider an electron whose kinetic energy is constant throughout the magnetic lens. This follows from Eq. (2.53), since a static magnetic field performs no work. The momentum magnitude p is therefore constant.

In Cartesian coordinates, the radial field in Eq. (4.66) has the components

$$\boxed{B_x \simeq -\frac{x}{2}B_0'(z), \qquad B_y \simeq -\frac{y}{2}B_0'(z), \qquad B_z \simeq B_0(z).} \tag{4.76}$$

Cartesian components of the paraxial magnetic field.

The magnetic field of a rotationally symmetric lens has no azimuthal component, so near the optical axis

$$\mathbf{B} = B_r\hat{\mathbf{r}} + B_z\hat{\mathbf{z}}. \tag{4.77}$$

The cylindrical radial unit vector is

$$\hat{\mathbf{r}} = \cos\phi\,\hat{\mathbf{x}} + \sin\phi\,\hat{\mathbf{y}}. \tag{4.78}$$

Since

$$x = r\cos\phi, \qquad y = r\sin\phi, \tag{4.79}$$

we may write

$$\hat{\mathbf{r}} = \frac{x}{r}\hat{\mathbf{x}} + \frac{y}{r}\hat{\mathbf{y}}. \tag{4.80}$$

Therefore,

$$B_x = B_r\frac{x}{r}, \qquad B_y = B_r\frac{y}{r}. \tag{4.81}$$

Using the near-axis result Eq. (4.66),

$$B_r = -\frac{r}{2}B_0'(z) + O(r^3), \tag{4.82}$$

we obtain

$$B_x = -\frac{x}{2}B_0'(z) + O(r^3), \tag{4.83}$$

$$B_y = -\frac{y}{2}B_0'(z) + O(r^3). \tag{4.84}$$

The longitudinal component is

$$B_z(r,z) = B_0(z) - \frac{r^2}{4}B_0''(z) + O(r^4). \tag{4.85}$$

In the first-order paraxial approximation the terms proportional to $r^2 = x^2 + y^2$ are neglected, so

$$B_z(r,z) \simeq B_0(z). \tag{4.86}$$

Hence,

$$\boxed{B_x \simeq -\frac{x}{2}B_0'(z), \qquad B_y \simeq -\frac{y}{2}B_0'(z), \qquad B_z \simeq B_0(z).} \tag{4.87}$$

For electron-optical imaging we are principally interested in trajectories that remain close to the optical axis and propagate predominantly in the positive z direction. Before applying the paraxial approximation, it is useful to make its geometrical meaning explicit.

The electron trajectory is parameterized as

$$\mathbf{r}(z) = x(z)\hat{\mathbf{x}} + y(z)\hat{\mathbf{y}} + z\hat{\mathbf{z}}. \tag{4.88}$$

Differentiating with respect to z gives

$$\mathbf{r}'(z) = x'(z)\hat{\mathbf{x}} + y'(z)\hat{\mathbf{y}} + \hat{\mathbf{z}}. \tag{4.89}$$

Its magnitude is

$$\varrho \triangleq \|\mathbf{r}'\| = \sqrt{1 + x'^2 + y'^2}. \tag{4.90}$$

Therefore the unit vector tangent to the electron trajectory is

$$\boxed{\mathbf{t} = \frac{x'\hat{\mathbf{x}} + y'\hat{\mathbf{y}} + \hat{\mathbf{z}}}{\sqrt{1 + x'^2 + y'^2}}.} \tag{4.91}$$

Let α denote the angle between the local electron trajectory and the optical axis. Since

$$\cos\alpha = \mathbf{t} \cdot \hat{\mathbf{z}}, \tag{4.92}$$

Eq. (4.91) gives

$$\cos\alpha = \frac{1}{\sqrt{1 + x'^2 + y'^2}}. \tag{4.93}$$

Consequently,

$$\boxed{\tan\alpha = \sqrt{x'^2 + y'^2}.} \tag{4.94}$$

Relation between trajectory slopes and propagation angle.

From Eq. (4.91),

$$\cos^2\alpha = \frac{1}{1 + x'^2 + y'^2}. \tag{4.95}$$

Hence,

$$\tan^2\alpha = \frac{1 - \cos^2\alpha}{\cos^2\alpha} \tag{4.96}$$

$$= x'^2 + y'^2. \tag{4.97}$$

Since α denotes the non-negative angle from the optical axis,

$$\tan\alpha = \sqrt{x'^2 + y'^2}, \tag{4.98}$$

which proves Eq. (4.94).

There is also a direct dynamical interpretation of x' and y'. Since

$$\mathbf{v} = \frac{d\mathbf{r}}{dt}, \tag{4.99}$$

the chain rule gives

$$\frac{dx}{dz} = \frac{dx/dt}{dz/dt} = \frac{v_x}{v_z}, \qquad \frac{dy}{dz} = \frac{v_y}{v_z}. \tag{4.100}$$

Therefore,

$$\boxed{x' = \frac{v_x}{v_z} = \frac{p_x}{p_z}, \qquad y' = \frac{v_y}{v_z} = \frac{p_y}{p_z}.} \tag{4.101}$$

The second equalities follow from $\mathbf{p} = \gamma m_0 \mathbf{v}$: the common factor γm_0 cancels between numerator and denominator.

The paraxial approximation means that the electron propagates at a small angle with respect to the optical axis,

$$|\alpha| \ll 1. \tag{4.102}$$

From Eq. (4.94), this is equivalent to

$$\boxed{x'^2 + y'^2 \ll 1,} \tag{4.103}$$

or, equivalently,

$$|x'| \ll 1, \qquad |y'| \ll 1. \tag{4.104}$$

From Eq. (4.123), the same approximation can be written in momentum space as

$$|p_x| \ll p_z, \qquad |p_y| \ll p_z. \tag{4.105}$$

To first order in the transverse slopes,

$$\tan\alpha \simeq \alpha, \tag{4.106}$$

so that

$$\alpha \simeq \sqrt{x'^2 + y'^2}. \tag{4.107}$$

Moreover,

$$\varrho = \sqrt{1 + x'^2 + y'^2} = 1 + \frac{x'^2 + y'^2}{2} + O(\alpha^4). \tag{4.108}$$

Thus, to first paraxial order,

$$\boxed{\varrho \simeq 1.} \tag{4.109}$$

There is also a direct relation between the trajectory slopes and the components of the electron momentum. Since

$$\mathbf{p} = \gamma m_0 \mathbf{v}, \tag{4.110}$$

the momentum and velocity vectors are parallel. Therefore their unit vectors are identical:

$$\frac{\mathbf{p}}{p} = \frac{\mathbf{v}}{v} = \mathbf{t}, \tag{4.111}$$

where $p = \|\mathbf{p}\|$ and $\mathbf{t}$ is the unit tangent to the trajectory given in Eq. (4.91).

Using Eq. (4.91),

$$\frac{\mathbf{p}}{p} = \frac{x'\hat{\mathbf{x}} + y'\hat{\mathbf{y}} + \hat{\mathbf{z}}}{\sqrt{1 + x'^2 + y'^2}}. \tag{4.112}$$

Consequently, the momentum components are exactly

$$\boxed{\begin{aligned} p_x &= \frac{px'}{\sqrt{1 + x'^2 + y'^2}}, \\ p_y &= \frac{py'}{\sqrt{1 + x'^2 + y'^2}}, \\ p_z &= \frac{p}{\sqrt{1 + x'^2 + y'^2}}. \end{aligned}} \tag{4.113}$$

Relation between trajectory slopes and momentum components.

The velocity components satisfy

$$v_x = \frac{dx}{dt}, \qquad v_y = \frac{dy}{dt}, \qquad v_z = \frac{dz}{dt}. \tag{4.114}$$

Since x and y are regarded as functions of z, the chain rule gives

$$v_x = \frac{dx}{dz}\frac{dz}{dt} = x' v_z, \tag{4.115}$$

$$v_y = \frac{dy}{dz}\frac{dz}{dt} = y' v_z. \tag{4.116}$$

Hence,

$$\mathbf{v} = v_z \left(x'\hat{\mathbf{x}} + y'\hat{\mathbf{y}} + \hat{\mathbf{z}} \right). \tag{4.117}$$

Its magnitude is therefore

$$v = v_z\sqrt{1 + x'^2 + y'^2}. \tag{4.118}$$

Thus,

$$v_z = \frac{v}{\sqrt{1 + x'^2 + y'^2}}, \tag{4.119}$$

and consequently

$$v_x = \frac{vx'}{\sqrt{1 + x'^2 + y'^2}}, \tag{4.120}$$

$$v_y = \frac{vy'}{\sqrt{1 + x'^2 + y'^2}}. \tag{4.121}$$

Multiplying each velocity component by the common relativistic factor γm_0, and using

$$p = \gamma m_0 v, \tag{4.122}$$

gives Eq. (4.113).

An immediate consequence of Eq. (4.113) is

$$\boxed{x' = \frac{p_x}{p_z}, \qquad y' = \frac{p_y}{p_z}.} \tag{4.123}$$

These relations are exact; no paraxial approximation has yet been made.

We now use the paraxial condition Eq. (4.103),

$$x'^2 + y'^2 \ll 1. \tag{4.124}$$

It implies

$$\sqrt{1 + x'^2 + y'^2} \simeq 1, \tag{4.125}$$

and therefore

$$\boxed{p_z \simeq p, \qquad p_x \simeq px', \qquad p_y \simeq py'.} \tag{4.126}$$

Thus, in the paraxial regime, the total electron momentum is almost entirely longitudinal, while the small transverse momentum components are the total momentum multiplied by the corresponding trajectory slopes.

These relations are the approximations that will be used below to linearize the Lorentz trajectory equations.

For paraxial trajectories, the transverse equations of motion become

$$\boxed{\begin{aligned} x'' &= -\kappa(z)y' - \frac{1}{2}\kappa'(z)y, \\ y'' &= \kappa(z)x' + \frac{1}{2}\kappa'(z)x, \end{aligned}} \tag{4.127}$$

where

$$\boxed{\kappa(z) \triangleq \frac{eB_0(z)}{p}.} \tag{4.128}$$

For constant B_0, $\kappa'(z) = 0$ and Eq. (4.127) reduces exactly to the uniform-field equations derived previously.

Paraxial equations in a round magnetic lens.

For an electron moving in a purely magnetic field, the Lorentz equation, Eq. (2.48), is

$$\frac{d\mathbf{p}}{dt} = -e\,\mathbf{v} \times \mathbf{B}. \tag{4.129}$$

Because the magnetic force is always perpendicular to the velocity,

$$\mathbf{v} \cdot (\mathbf{v} \times \mathbf{B}) = 0, \tag{4.130}$$

the magnetic field performs no work. Hence the kinetic energy is constant, as already established from Eq. (2.53). Since the momentum magnitude is uniquely determined by the kinetic energy through Eq. (2.72), its magnitude

$$p = \|\mathbf{p}\| \tag{4.131}$$

is also constant throughout the magnetic lens.

We parameterize the trajectory by the optical coordinate z,

$$\mathbf{r}(z) = x(z)\hat{\mathbf{x}} + y(z)\hat{\mathbf{y}} + z\hat{\mathbf{z}}. \tag{4.132}$$

As derived in Eq. (4.113), the exact momentum components are

$$\begin{aligned} p_x &= \frac{px'}{\sqrt{1+x'^2+y'^2}}, \\ p_y &= \frac{py'}{\sqrt{1+x'^2+y'^2}}, \\ p_z &= \frac{p}{\sqrt{1+x'^2+y'^2}}. \end{aligned} \tag{4.133}$$

For paraxial trajectories, Eq. (4.103) states that

$$x'^2 + y'^2 \ll 1. \tag{4.134}$$

Therefore,

$$\sqrt{1+x'^2+y'^2} \simeq 1, \tag{4.135}$$

and Eq. (4.126) gives

$$\boxed{p_x \simeq px', \qquad p_y \simeq py', \qquad p_z \simeq p.} \tag{4.136}$$

We next relate differentiation with respect to time to differentiation with respect to z. Let F be any quantity evaluated along the trajectory. Since the trajectory is parameterized by z, we may write

$$F(t) = F(z(t)). \tag{4.137}$$

The chain rule then gives

$$\frac{dF}{dt} = \frac{dF}{dz}\frac{dz}{dt}. \tag{4.138}$$

But

$$\frac{dz}{dt} = v_z, \tag{4.139}$$

and hence

$$\boxed{\frac{d}{dt} = v_z \frac{d}{dz}.} \tag{4.140}$$

This relation is exact, not paraxial, provided that $z(t)$ varies monotonically so that z can be used as a trajectory parameter.

The same parameterization gives the exact relations

$$v_x = \frac{dx}{dt} = \frac{dx}{dz}\frac{dz}{dt} = x' v_z, \tag{4.141}$$

and

$$v_y = y' v_z. \tag{4.142}$$

Thus,

$$\boxed{v_x = x' v_z, \qquad v_y = y' v_z.} \tag{4.143}$$

Again, these relations are exact.

We can now evaluate the transverse components of the Lorentz equation. The x component of Eq. (4.129) is

$$\frac{dp_x}{dt} = -e\left(v_y B_z - v_z B_y\right). \tag{4.144}$$

Using Eq. (4.143),

$$\frac{dp_x}{dt} = -e v_z\left(y' B_z - B_y\right). \tag{4.145}$$

On the left-hand side, Eq. (4.140) gives

$$\frac{dp_x}{dt} = v_z \frac{dp_x}{dz}. \tag{4.146}$$

Using the paraxial expression $p_x \simeq p x'$ from Eq. (4.136), together with the fact that p is constant,

$$\frac{dp_x}{dz} \simeq p x''. \tag{4.147}$$

Consequently,

$$\frac{dp_x}{dt} \simeq p v_z x''. \tag{4.148}$$

Equating Eqs. (4.145) and (4.148), and cancelling the common factor v_z, gives

$$x'' = -\frac{e}{p}\left(y' B_z - B_y\right). \tag{4.149}$$

Close to the optical axis, the round-lens field has the Cartesian components derived in Eq. (4.76),

$$B_x \simeq -\frac{x}{2}B_0'(z), \qquad B_y \simeq -\frac{y}{2}B_0'(z), \qquad B_z \simeq B_0(z). \tag{4.150}$$

Substitution into Eq. (4.149) gives

$$x'' = -\frac{e}{p}\left[y'B_0(z) + \frac{y}{2}B_0'(z)\right] \tag{4.151}$$

$$= -\frac{eB_0(z)}{p}y' - \frac{eB_0'(z)}{2p}y. \tag{4.152}$$

The y component of the Lorentz equation is

$$\frac{dp_y}{dt} = -e\left(v_z B_x - v_x B_z\right). \tag{4.153}$$

Using $v_x = x'v_z$ from Eq. (4.143),

$$\frac{dp_y}{dt} = -ev_z\left(B_x - x'B_z\right) = ev_z\left(x'B_z - B_x\right). \tag{4.154}$$

On the other hand,

$$\frac{dp_y}{dt} = v_z\frac{dp_y}{dz} \simeq pv_z y'', \tag{4.155}$$

where Eq. (4.136) and the constancy of p have again been used. Cancelling v_z gives

$$y'' = \frac{e}{p}\left(x'B_z - B_x\right). \tag{4.156}$$

Substitution of Eq. (4.76) yields

$$y'' = \frac{e}{p}\left[x'B_0(z) + \frac{x}{2}B_0'(z)\right] \tag{4.157}$$

$$= \frac{eB_0(z)}{p}x' + \frac{eB_0'(z)}{2p}x. \tag{4.158}$$

Finally, introducing the local spatial cyclotron wavenumber

$$\boxed{\kappa(z) \triangleq \frac{eB_0(z)}{p},} \tag{4.159}$$

and noting that p is constant, so that

$$\kappa'(z) = \frac{eB_0'(z)}{p}, \tag{4.160}$$

Eqs. (4.152) and (4.158) become

$$\boxed{\begin{aligned} x'' &= -\kappa(z)y' - \frac{1}{2}\kappa'(z)y, \\ y'' &= \kappa(z)x' + \frac{1}{2}\kappa'(z)x. \end{aligned}} \tag{4.161}$$

These are the linear paraxial trajectory equations for a rotationally symmetric magnetic lens.

4.2.2 Larmor rotation emerges from the trajectory equations

The coupled equations Eq. (4.127) appear more complicated than the uniform-field equations, but they possess a remarkable simplification.

Introduce the complex transverse coordinate

$$u(z) \triangleq x(z) + jy(z). \tag{4.162}$$

Equation (4.127) then becomes

$$\boxed{u'' - j\kappa(z)u' - \frac{j}{2}\kappa'(z)u = 0.} \tag{4.163}$$

We now write the trajectory as

$$\boxed{u(z) = w(z)\exp\left[j\theta_L(z)\right],} \tag{4.164}$$

where the rotation angle is defined by

$$\boxed{\theta_L'(z) = \frac{\kappa(z)}{2} = \frac{eB_0(z)}{2p}.} \tag{4.165}$$

Thus

$$\boxed{\theta_L(z) - \theta_L(z_0) = \frac{e}{2p}\int_{z_0}^{z} B_0(z')\,dz'.} \tag{4.166}$$

The complex factor $\exp[j\theta_L(z)]$ represents an actual rotation of the transverse coordinate system. Hence the electron bundle, viewed in the laboratory frame, rotates through the angle θ_L as it passes through the lens. This rotation is what is meant by *Larmor rotation*.

The remaining coordinate $w(z)$ satisfies

$$\boxed{w'' + \frac{\kappa^2(z)}{4} w = 0.} \tag{4.167}$$

This equation is of central importance. Once the trivial rotation of the coordinate system has been removed, a round magnetic lens acts as an ordinary isotropic focusing system. The focusing strength is proportional to $B_0^2(z)$.

The Larmor transformation.

Combining the two equations in Eq. (4.127),

$$u'' = x'' + jy'' \tag{4.168}$$

$$= -\kappa y' - \frac{\kappa'}{2} y + j\left(\kappa x' + \frac{\kappa'}{2} x\right) \tag{4.169}$$

$$= j\kappa u' + \frac{j}{2}\kappa' u, \tag{4.170}$$

which gives Eq. (4.163).

Now let

$$u = we^{j\theta_L}. \tag{4.171}$$

Then

$$u' = e^{j\theta_L}\left(w' + j\theta_L' w\right), \tag{4.172}$$

and

$$u'' = e^{j\theta_L}\left[w'' + 2j\theta_L' w' + \left(j\theta_L'' - \theta_L'^2\right) w\right]. \tag{4.173}$$

Substitution into Eq. (4.163) gives

$$0 = w'' + j(2\theta_L' - \kappa)w' \tag{4.174}$$

$$+ \left[j\theta_L'' - \theta_L'^2 + \kappa\theta_L' - \frac{j}{2}\kappa'\right] w. \tag{4.175}$$

Choosing

$$\theta_L' = \frac{\kappa}{2} \tag{4.176}$$

eliminates the first derivative. Since

$$\theta_L'' = \frac{\kappa'}{2}, \tag{4.177}$$

the imaginary terms also cancel, leaving

$$w'' + \frac{\kappa^2}{4} w = 0, \tag{4.178}$$

which proves Eq. (4.167).

This derivation explains both effects of the magnetic lens without introducing any additional phenomenology:

$$\boxed{\begin{array}{lll} B_0(z) & \longrightarrow & \theta_L'(z) = \dfrac{eB_0(z)}{2p} \\ & & \text{rotation,} \\ B_0^2(z) & \longrightarrow & w'' + \dfrac{e^2 B_0^2(z)}{4p^2} w = 0 \\ & & \text{focusing.} \end{array}} \tag{4.179}$$

4.2.3 Conservation of axial angular momentum

The factor of 1/2 that appears in the Larmor rotation can also be understood from a conservation law associated with the rotational symmetry of the magnetic lens. We derive this conservation law explicitly before using it.

Consider a static magnetic field that is rotationally symmetric about the optical axis. In cylindrical coordinates,

$$\mathbf{B} = B_r(r,z)\hat{\mathbf{r}} + B_z(r,z)\hat{\mathbf{z}}, \tag{4.180}$$

with

$$B_\phi = 0, \qquad \frac{\partial}{\partial \phi} = 0. \tag{4.181}$$

The mechanical angular momentum of the electron about the optical axis is the z component of $\mathbf{r} \times \mathbf{p}$,

$$\boxed{L_z^{\text{mech}} = \gamma m_0 r^2 \dot{\phi}.} \tag{4.182}$$

Unlike the corresponding quantity for a free particle, L_z^{mech} is not generally conserved in a magnetic lens, because the magnetic Lorentz force can exert a torque about the optical axis.

Mechanical angular momentum about the optical axis.

Let the electron position be described in cylindrical coordinates (r, ϕ, z). To avoid confusing the radial coordinate r with the position vector, we denote the latter by $\mathbf{R}$:

$$\mathbf{R} = r\hat{\mathbf{r}} + z\hat{\mathbf{z}}. \tag{4.183}$$

The velocity in cylindrical coordinates is

$$\boxed{\mathbf{v} = \dot{r}\,\hat{\mathbf{r}} + r\dot{\phi}\,\hat{\boldsymbol{\phi}} + \dot{z}\,\hat{\mathbf{z}}.} \tag{4.184}$$

To see the origin of the factor r in the azimuthal component, note that an angular displacement $d\phi$ at radius r corresponds to an arc length

$$ds_\phi = r\,d\phi. \tag{4.185}$$

Therefore,

$$v_\phi = \frac{ds_\phi}{dt} = r\dot{\phi}. \tag{4.186}$$

The relativistic mechanical momentum is consequently

$$\mathbf{p} = \gamma m_0 \mathbf{v} = \gamma m_0 \left(\dot{r}\,\hat{\mathbf{r}} + r\dot{\phi}\,\hat{\boldsymbol{\phi}} + \dot{z}\,\hat{\mathbf{z}}\right). \tag{4.187}$$

The mechanical angular momentum is defined by

$$\mathbf{L}^{\text{mech}} = \mathbf{R} \times \mathbf{p}. \tag{4.188}$$

We only need its component along the optical axis. Since

$$\hat{\mathbf{r}} \times \hat{\boldsymbol{\phi}} = \hat{\mathbf{z}}, \tag{4.189}$$

the only contribution to the z component comes from the product of the radial part of $\mathbf{R}$ and the azimuthal part of $\mathbf{p}$:

$$L_z^{\text{mech}} = rp_\phi \tag{4.190}$$

$$= r\left(\gamma m_0 r\dot{\phi}\right). \tag{4.191}$$

Hence,

$$\boxed{L_z^{\text{mech}} = \gamma m_0 r^2 \dot{\phi},} \tag{4.192}$$

which proves Eq. (4.182).

To describe this torque conveniently, define the magnetic flux $\Psi(r, z)$ through a circular disc of radius r, centred on the optical axis and lying in the plane $z =$ constant:

$$\boxed{\Psi(r, z) \triangleq 2\pi \int_0^r r' B_z(r', z)\, dr'.} \tag{4.193}$$

The axial and radial magnetic-field components can then be expressed as

$$\boxed{B_z(r, z) = \frac{1}{2\pi r}\frac{\partial \Psi}{\partial r}, \qquad B_r(r, z) = -\frac{1}{2\pi r}\frac{\partial \Psi}{\partial z}.} \tag{4.194}$$

Magnetic-field components from the flux function.

Differentiating Eq. (4.193) with respect to r gives directly

$$\frac{\partial \Psi}{\partial r} = 2\pi r B_z(r, z), \tag{4.195}$$

and hence

$$B_z = \frac{1}{2\pi r}\frac{\partial \Psi}{\partial r}. \tag{4.196}$$

To obtain the radial component, we use

$$\nabla \cdot \mathbf{B} = 0. \tag{4.197}$$

For a rotationally symmetric field this becomes

$$\frac{1}{r}\frac{\partial}{\partial r}(r B_r) + \frac{\partial B_z}{\partial z} = 0. \tag{4.198}$$

Differentiating Eq. (4.193) with respect to z,

$$\frac{\partial \Psi}{\partial z} = 2\pi \int_0^r r' \frac{\partial B_z(r', z)}{\partial z}\, dr' \tag{4.199}$$

$$= -2\pi \int_0^r \frac{\partial}{\partial r'}\left[r' B_r(r', z)\right]\, dr'. \tag{4.200}$$

Therefore,

$$\frac{\partial \Psi}{\partial z} = -2\pi r B_r(r, z), \tag{4.201}$$

where regularity of the magnetic field on the optical axis implies $rB_r \to 0$ as $r \to 0$. Hence,

$$B_r = -\frac{1}{2\pi r}\frac{\partial \Psi}{\partial z}. \tag{4.202}$$

This proves Eq. (4.194).

We can now derive the conserved quantity. For an electron in a purely magnetic field, Eq. (2.48) gives

$$\frac{d\mathbf{p}}{dt} = -e\,\mathbf{v} \times \mathbf{B}. \tag{4.203}$$

Taking the torque about the optical axis yields

$$\boxed{\frac{dL_z^{\text{mech}}}{dt} = -er\left(\dot{z}B_r - \dot{r}B_z\right).} \tag{4.204}$$

Conserved axial angular momentum.

The time derivative of the mechanical angular momentum is

$$\frac{d}{dt}\left(\mathbf{r} \times \mathbf{p}\right) = \mathbf{r} \times \frac{d\mathbf{p}}{dt}, \tag{4.205}$$

because

$$\frac{d\mathbf{r}}{dt} \times \mathbf{p} = \mathbf{v} \times \gamma m_0 \mathbf{v} = \mathbf{0}. \tag{4.206}$$

Taking the z component,

$$\frac{dL_z^{\text{mech}}}{dt} = rF_\phi, \tag{4.207}$$

where F_ϕ is the azimuthal component of the Lorentz force.

The velocity is

$$\mathbf{v} = \dot{r}\,\hat{\mathbf{r}} + r\dot{\phi}\,\hat{\boldsymbol{\phi}} + \dot{z}\,\hat{\mathbf{z}}, \tag{4.208}$$

whereas the rotationally symmetric magnetic field is

$$\mathbf{B} = B_r\hat{\mathbf{r}} + B_z\hat{\mathbf{z}}. \tag{4.209}$$

The azimuthal component of their cross product is

$$(\mathbf{v} \times \mathbf{B})_\phi = \dot{z} B_r - \dot{r} B_z. \tag{4.210}$$

Since the electron charge is $-e$,

$$F_\phi = -e\left(\dot{z} B_r - \dot{r} B_z\right), \tag{4.211}$$

which gives Eq. (4.204).

Now evaluate the time derivative of the magnetic flux along the electron trajectory. Since

$$\Psi = \Psi(r(t), z(t)), \tag{4.212}$$

the chain rule gives

$$\frac{d\Psi}{dt} = \frac{\partial \Psi}{\partial r}\dot{r} + \frac{\partial \Psi}{\partial z}\dot{z}. \tag{4.213}$$

Using Eq. (4.194),

$$\frac{d\Psi}{dt} = 2\pi r B_z \dot{r} - 2\pi r B_r \dot{z} \tag{4.214}$$

$$= -2\pi r\left(\dot{z} B_r - \dot{r} B_z\right). \tag{4.215}$$

Consequently, Eq. (4.204) can be written as

$$\frac{dL_z^{\text{mech}}}{dt} = \frac{e}{2\pi}\frac{d\Psi}{dt}. \tag{4.216}$$

Rearranging,

$$\frac{d}{dt}\left[L_z^{\text{mech}} - \frac{e}{2\pi}\Psi\right] = 0. \tag{4.217}$$

Therefore the quantity

$$\boxed{\mathcal{L}_z = \gamma m_0 r^2 \dot{\phi} - \frac{e}{2\pi}\Psi(r, z)} \tag{4.218}$$

is conserved along the electron trajectory.

The meaning of Eq. (4.218) is important. The mechanical angular momentum L_z^{mech} is not itself conserved: as the electron enters or leaves a magnetic field, the fringe field exerts an azimuthal torque and changes it. What is conserved is the combination of the mechanical angular momentum and a term determined by the magnetic flux enclosed between the trajectory

and the optical axis.

For a particle of arbitrary charge Q, the corresponding conservation law is

$$\boxed{\mathcal{L}_z = \gamma m_0 r^2 \dot{\phi} + \frac{Q}{2\pi}\Psi(r,z) = \text{constant}.} \tag{4.219}$$

Equation (4.218) follows by setting $Q = -e$.

4.2.4 Why is the Larmor frequency one half of the cyclotron frequency?

The factor 1/2 appearing in the Larmor rotation rate, Eq. (4.165), deserves a physical explanation. It is not an arbitrary feature of the transformation introduced in Sec. 4.2.2; it follows directly from the conservation of generalized axial angular momentum derived in Sec. 4.2.3.

For an electron in a static rotationally symmetric magnetic field, we showed in Eq. (4.218) that

$$\boxed{\mathcal{L}_z = \gamma m_0 r^2 \dot{\phi} - \frac{e}{2\pi}\Psi(r,z) = \text{constant},} \tag{4.220}$$

where

$$\Psi(r,z) = 2\pi \int_0^r r' B_z(r',z)\, dr' \tag{4.221}$$

is the magnetic flux through the coaxial disc bounded by the electron trajectory, as defined in Eq. (4.193).

The first term in Eq. (4.220),

$$L_z^{\text{mech}} = \gamma m_0 r^2 \dot{\phi}, \tag{4.222}$$

is the mechanical angular momentum about the optical axis. As discussed in Sec. 4.2.3, it is *not* conserved by itself. The fringe field exerts an azimuthal torque on the electron, changing the mechanical angular momentum, while the combination with the magnetic-flux term remains constant.

Close to the optical axis, the paraxial field expansion Eq. (4.66) gives

$$B_z(r,z) = B_0(z) + O(r^2). \tag{4.223}$$

The magnetic flux enclosed by a trajectory of small radius r is therefore

$$\Psi(r,z) = 2\pi \int_0^r r' \left[B_0(z) + O(r'^2)\right] dr' \tag{4.224}$$

$$= \pi r^2 B_0(z) + O(r^4). \tag{4.225}$$

Thus, to the order required by the paraxial theory,

$$\boxed{\Psi(r,z) \simeq \pi r^2 B_0(z).} \tag{4.226}$$

Now consider a trajectory that intersects the optical axis at some point. At that point $r = 0$, and therefore both terms in Eq. (4.220) vanish:

$$L_z^{\text{mech}} = 0, \qquad \Psi = 0. \tag{4.227}$$

Hence the conserved value is

$$\mathcal{L}_z = 0 \tag{4.228}$$

along the entire trajectory.

Substitution of Eq. (4.226) into Eq. (4.220) therefore gives

$$\gamma m_0 r^2 \dot{\phi} - \frac{e}{2} r^2 B_0(z) = 0. \tag{4.229}$$

For $r \neq 0$, the common factor r^2 cancels and we obtain

$$\boxed{\dot{\phi} = \frac{eB_0(z)}{2\gamma m_0}.} \tag{4.230}$$

The relativistic cyclotron angular frequency in a uniform magnetic field was defined in Eq. (4.8) as

$$\omega_c = \frac{eB_0}{\gamma m_0}. \tag{4.231}$$

Consequently,

$$\boxed{\dot{\phi} = \omega_L = \frac{\omega_c}{2}.} \tag{4.232}$$

The quantity ω_L is called the *Larmor angular frequency*. Thus the factor $1/2$ appearing in the Larmor rotation is a direct consequence of conservation of generalized axial angular momentum.

It is useful to express the same result per unit distance travelled along the optical axis. Since

$$\dot{\phi} = \frac{d\phi}{dt} = \frac{d\phi}{dz}\frac{dz}{dt} = v_z \frac{d\phi}{dz}, \tag{4.233}$$

where we have used the exact change of parameter derived in Eq. (4.140), we have

$$\frac{d\phi}{dz} = \frac{\dot{\phi}}{v_z}. \tag{4.234}$$

Using

$$p_z = \gamma m_0 v_z, \tag{4.235}$$

Eq. (4.230) gives the exact relation

$$\boxed{\frac{d\phi}{dz} = \frac{eB_0(z)}{2p_z}.} \tag{4.236}$$

For a paraxial trajectory, Eq. (4.126) gives $p_z \simeq p$, so

$$\frac{d\phi}{dz} \simeq \frac{eB_0(z)}{2p}. \tag{4.237}$$

Introducing the local spatial cyclotron wavenumber $\kappa(z)$ from Eq. (4.159),

$$\kappa(z) = \frac{eB_0(z)}{p}, \tag{4.238}$$

we finally obtain

$$\boxed{\frac{d\phi}{dz} \simeq \frac{\kappa(z)}{2},} \tag{4.239}$$

in agreement with the Larmor rotation rate derived independently in Eq. (4.165) from the coupled paraxial trajectory equations.

Cyclotron rotation and Larmor rotation are not the same motion

The factor of two becomes less surprising once the two geometrical situations are distinguished.

In the uniform-field problem of Sec. 4.1, an individual electron with a transverse velocity component executes a circular orbit around its own *guiding centre*. Its transverse velocity completes one revolution at the cyclotron rate

$$\omega_c = \frac{eB_0}{\gamma m_0}. \tag{4.240}$$

This is the motion represented in Fig. 4.1.

A round magnetic lens involves a different geometrical question. We follow the position of each paraxial trajectory with respect to the common *optical axis*. As the electron enters the fringe field, the radial magnetic-field component derived in Eq. (4.66) generates mechanical axial angular momentum. Conservation of the generalized angular momentum, Eq. (4.218), requires the trajectory to rotate about the optical axis at the Larmor rate $\omega_c/2$.

At the same time, the transverse coordinates measured in the rotating Larmor frame satisfy the focusing equation

$$w'' + \frac{\kappa^2(z)}{4} w = 0, \tag{4.241}$$

derived in Eq. (4.167). Thus the action of a round magnetic lens naturally separates into

$$\boxed{\begin{aligned} \frac{d\theta_L}{dz} &= \frac{\kappa(z)}{2} && \Longrightarrow \quad \text{rotation}, \\ w'' + \frac{\kappa^2(z)}{4} w &= 0 && \Longrightarrow \quad \text{focusing}. \end{aligned}} \tag{4.242}$$

These are not independent effects imposed on the electron by two different mechanisms. Both arise from the same Lorentz force acting in the spatially varying magnetic field of the round lens. The Larmor transformation simply separates the rotational part of that motion from the radial focusing part.

4.2.5 Focal length in the thin-lens approximation

Equation (4.167) also allows us to connect the magnetic-field distribution with the familiar focal length of a thin lens.

If the electron does not move appreciably in the transverse direction while crossing the magnetic field, we may approximate $w(z)$ as constant inside the lens. Integration of Eq. (4.167) then gives

$$\Delta w' \simeq -\frac{w}{4} \int_{-\infty}^{\infty} \kappa^2(z)\, dz. \tag{4.243}$$

Integration of Eq. (4.167) across the lens gives, in the thin-lens approximation,

$$\Delta w' \simeq -\frac{w}{4} \int_{-\infty}^{\infty} \kappa^2(z)\, dz. \tag{4.244}$$

The thin-lens slope relation derived in Eq. (3.635) is

$$\Delta x' = -\frac{x}{f}. \tag{4.245}$$

Because the round magnetic lens is isotropic in the transverse plane, the same relation holds for the Larmor-frame transverse coordinate w:

$$\Delta w' = -\frac{w}{f}. \tag{4.246}$$

Comparing the two expressions for $\Delta w'$ gives

$$\boxed{\frac{1}{f} = \frac{1}{4}\int_{-\infty}^{\infty} \kappa^2(z)\,dz = \frac{e^2}{4p^2}\int_{-\infty}^{\infty} B_0^2(z)\,dz.} \tag{4.247}$$

Using $p^2 = 2m_0 e\widehat{\Phi}$ from Eq. (2.78), this may also be written

$$\boxed{\frac{1}{f} = \frac{e}{8m_0\widehat{\Phi}}\int_{-\infty}^{\infty} B_0^2(z)\,dz.} \tag{4.248}$$

Thus a magnetic lens is stronger when the magnetic field is larger or extends over a longer distance, whereas electrons of larger momentum are more difficult to deflect and therefore experience a longer focal length.

4.2.6 An analytically solvable model of a round magnetic lens

The preceding results are valid for an arbitrary paraxial on-axis magnetic field $B_0(z)$. In a real microscope this field can be calculated from the pole-piece geometry or measured experimentally, and Eq. (4.167) can then be integrated numerically.

It is nevertheless useful to consider a simple localized field for which the trajectory can be obtained analytically. This provides considerably more physical insight than the uniform-field model of Sec. 4.1, while retaining a closed-form solution.

We choose the on-axis field

$$\boxed{B_0(z) = \frac{B_m}{1 + (z/a)^2},} \tag{4.249}$$

where B_m is the maximum field, reached at $z = 0$, and a determines the longitudinal extent of the lens. In particular,

$$B_0(\pm a) = \frac{B_m}{2}, \tag{4.250}$$

so that the full width at half maximum of the axial field is $2a$.

This is not intended to reproduce exactly the field of a particular pole-piece geometry. It captures, however, three essential features of a real magnetic lens: the field is localized, it is symmetric about the centre of the lens, and it varies with z, thereby necessarily producing transverse fringe fields.

The corresponding fringe field

From the general near-axis result Eq. (4.66),

$$B_r(r,z) = -\frac{r}{2}\frac{dB_0}{dz} + O(r^3). \tag{4.251}$$

Differentiating Eq. (4.249),

$$\frac{dB_0}{dz} = -\frac{2B_m z/a^2}{\left[1+(z/a)^2\right]^2}, \tag{4.252}$$

and therefore

$$\boxed{B_r(r,z) = \frac{B_m r z/a^2}{\left[1+(z/a)^2\right]^2} + O(r^3).} \tag{4.253}$$

The radial component changes sign at the centre of the lens. On the entrance side, $z < 0$, it points in one radial direction, whereas on the exit side, $z > 0$, it points in the opposite direction.

This observation helps explain why the uniform-field problem of Sec. 4.1 cannot by itself describe focusing. Consider an electron entering the lens parallel to the optical axis but at a finite distance r from it. In a strictly uniform axial field, $\mathbf{v} \parallel \mathbf{B}$, and hence

$$\mathbf{v} \times \mathbf{B} = \mathbf{0}. \tag{4.254}$$

Such an electron would not be deflected at all.

In a real lens, by contrast, the entrance fringe field contains the radial component of Eq. (4.253). The longitudinal electron velocity then experiences an azimuthal Lorentz force,

$$-\,e\left(v_z\hat{\mathbf{z}} \times B_r\hat{\mathbf{r}}\right), \tag{4.255}$$

which generates an azimuthal velocity. Once this azimuthal motion exists, its interaction with the strong axial component B_z produces a radial Lorentz force. The combined action appears mathematically as the focusing term in Eq. (4.167).

Thus a magnetic lens focuses through a two-stage mechanism: the fringe field generates rotation, and the rotating trajectory interacts with the axial field to produce a net restoring action toward the optical axis.

Larmor rotation through the localized field

For the field in Eq. (4.249), define the maximum spatial cyclotron wavenumber

$$\boxed{\kappa_m \triangleq \frac{eB_m}{p}.} \tag{4.256}$$

The local quantity introduced in Eq. (4.159) is then

$$\kappa(z) = \frac{\kappa_m}{1 + (z/a)^2}. \tag{4.257}$$

The local Larmor rotation rate, Eq. (4.165), becomes

$$\frac{d\theta_L}{dz} = \frac{\kappa_m}{2} \frac{1}{1 + (z/a)^2}. \tag{4.258}$$

Taking the accumulated rotation to vanish far before the lens,

$$\theta_L(-\infty) = 0, \tag{4.259}$$

gives

$$\boxed{\theta_L(z) = \frac{a\kappa_m}{2} \left[\arctan\left(\frac{z}{a}\right) + \frac{\pi}{2}\right].} \tag{4.260}$$

Larmor angle for the localized lens.

From Eq. (4.166),

$$\theta_L(z) = \frac{1}{2} \int_{-\infty}^{z} \kappa(z')\, dz'. \tag{4.261}$$

Using Eq. (4.257),

$$\theta_L(z) = \frac{\kappa_m}{2} \int_{-\infty}^{z} \frac{dz'}{1 + (z'/a)^2}. \tag{4.262}$$

Introducing

$$s = \frac{z'}{a}, \qquad dz' = a\, ds, \tag{4.263}$$

gives

$$\theta_L(z) = \frac{a\kappa_m}{2}\int_{-\infty}^{z/a}\frac{ds}{1+s^2} \tag{4.264}$$

$$= \frac{a\kappa_m}{2}\left[\arctan s\right]_{-\infty}^{z/a} \tag{4.265}$$

$$= \frac{a\kappa_m}{2}\left[\arctan\left(\frac{z}{a}\right)+\frac{\pi}{2}\right], \tag{4.266}$$

which proves Eq. (4.260).

After the electron has passed completely through the lens,

$$\boxed{\Delta\theta_L = \theta_L(+\infty)-\theta_L(-\infty) = \frac{\pi a\kappa_m}{2} = \frac{\pi a e B_m}{2p}.} \tag{4.267}$$

The image rotation generated by this model is therefore proportional to the integrated axial magnetic field. In fact,

$$\Delta\theta_L = \frac{e}{2p}\int_{-\infty}^{\infty} B_0(z)\,dz, \tag{4.268}$$

as expected from the general result Eq. (4.166).

Exact paraxial trajectory in the rotating frame

The more interesting question is whether the focusing equation can also be solved analytically. In the Larmor frame, Eq. (4.167) becomes

$$\boxed{w'' + \frac{\kappa_m^2}{4\left[1+(z/a)^2\right]^2}w = 0.} \tag{4.269}$$

Although Eq. (4.269) does not immediately resemble a familiar differential equation, it can be transformed into a harmonic oscillator.

Introduce

$$\vartheta \triangleq \arctan\left(\frac{z}{a}\right), \tag{4.270}$$

so that

$$\frac{z}{a} = \tan\vartheta, \qquad 1+\left(\frac{z}{a}\right)^2 = \sec^2\vartheta. \tag{4.271}$$

We also write

$$w(z) = \sec\vartheta\, q(\vartheta). \tag{4.272}$$

The resulting equation is

$$\boxed{\frac{d^2q}{d\vartheta^2} + \nu^2 q = 0,} \tag{4.273}$$

where

$$\boxed{\nu \triangleq \sqrt{1 + \left(\frac{a\kappa_m}{2}\right)^2}.} \tag{4.274}$$

Exact solution of the focusing equation.

From Eq. (4.270),

$$z = a \tan\vartheta, \tag{4.275}$$

and hence

$$\frac{d\vartheta}{dz} = \frac{\cos^2\vartheta}{a}. \tag{4.276}$$

Therefore,

$$\frac{d}{dz} = \frac{\cos^2\vartheta}{a}\frac{d}{d\vartheta}, \tag{4.277}$$

and

$$\frac{d^2}{dz^2} = \frac{\cos^4\vartheta}{a^2}\left(\frac{d^2}{d\vartheta^2} - 2\tan\vartheta\frac{d}{d\vartheta}\right). \tag{4.278}$$

Furthermore,

$$\frac{1}{\left[1 + (z/a)^2\right]^2} = \cos^4\vartheta. \tag{4.279}$$

Substitution into Eq. (4.269) and multiplication by $a^2/\cos^4\vartheta$ gives

$$\frac{d^2w}{d\vartheta^2} - 2\tan\vartheta\frac{dw}{d\vartheta} + \frac{a^2\kappa_m^2}{4}w = 0. \tag{4.280}$$

Now substitute Eq. (4.272),

$$w = \sec\vartheta\, q. \tag{4.281}$$

A direct differentiation gives

$$\frac{d^2w}{d\vartheta^2} - 2\tan\vartheta\frac{dw}{d\vartheta} = \sec\vartheta\left(\frac{d^2q}{d\vartheta^2} + q\right). \tag{4.282}$$

Equation (4.280) therefore becomes

$$\frac{d^2q}{d\vartheta^2} + \left[1 + \frac{a^2\kappa_m^2}{4}\right] q = 0. \tag{4.283}$$

Introducing ν according to Eq. (4.274) gives Eq. (4.273).

The solution is therefore

$$q(\vartheta) = C_1 \cos(\nu\vartheta) + C_2 \sin(\nu\vartheta), \tag{4.284}$$

and hence

$$\boxed{\begin{aligned} w(z) = &\sqrt{1 + \left(\frac{z}{a}\right)^2} \\ &\times \left[C_1 \cos\left(\nu \arctan\frac{z}{a}\right) + C_2 \sin\left(\nu \arctan\frac{z}{a}\right)\right]. \end{aligned}} \tag{4.285}$$

The constants C_1 and C_2 are determined by the transverse position and slope of the incident electron. Since w represents the complex coordinate in the Larmor frame, these constants may themselves be complex; equivalently, Eq. (4.285) can be applied independently to the two orthogonal transverse coordinates.

Trajectory in the laboratory frame

The physical transverse trajectory is recovered using the Larmor transformation, Eq. (4.164),

$$u(z) = x(z) + jy(z) = w(z)e^{j\theta_L(z)}. \tag{4.286}$$

Combining Eqs. (4.260) and (4.285) gives

$$\boxed{\begin{aligned} x(z) + jy(z) = &\sqrt{1 + \left(\frac{z}{a}\right)^2} \\ &\times \left[C_1 \cos\left(\nu \arctan\frac{z}{a}\right) + C_2 \sin\left(\nu \arctan\frac{z}{a}\right)\right] \\ &\times \exp\left\{j\frac{a\kappa_m}{2}\left[\arctan\left(\frac{z}{a}\right) + \frac{\pi}{2}\right]\right\}. \end{aligned}} \tag{4.287}$$

Equation (4.287) is an exact solution *within the paraxial round-lens approximation*. It displays explicitly the two simultaneous effects of a magnetic lens:

$$\boxed{\begin{array}{ccc} \exp[j\theta_L(z)] & \longrightarrow & \text{rotation about the optical axis,} \\ w(z) & \longrightarrow & \text{radial focusing or defocusing evolution.} \end{array}} \tag{4.288}$$

The apparently complicated laboratory trajectory is therefore much simpler when viewed in a coordinate system that rotates with the Larmor angle.

Focal length of the model lens

The general thin-lens result Eq. (4.247) gives

$$\frac{1}{f} = \frac{e^2}{4p^2} \int_{-\infty}^{\infty} B_0^2(z)\, dz. \tag{4.289}$$

For the field in Eq. (4.249),

$$B_0^2(z) = \frac{B_m^2}{[1 + (z/a)^2]^2}. \tag{4.290}$$

The required integral is

$$\boxed{\int_{-\infty}^{\infty} B_0^2(z)\, dz = \frac{\pi a B_m^2}{2}.} \tag{4.291}$$

Consequently,

$$\boxed{\frac{1}{f} = \frac{\pi a e^2 B_m^2}{8p^2} = \frac{\pi a \kappa_m^2}{8}.} \tag{4.292}$$

Focal length of the model lens.

Using $s = z/a$,

$$\int_{-\infty}^{\infty} B_0^2(z)\, dz = a B_m^2 \int_{-\infty}^{\infty} \frac{ds}{(1 + s^2)^2}. \tag{4.293}$$

Let

$$s = \tan\vartheta, \qquad ds = \sec^2\vartheta\, d\vartheta. \tag{4.294}$$

Then

$$\int_{-\infty}^{\infty} \frac{ds}{(1+s^2)^2} = \int_{-\pi/2}^{\pi/2} \cos^2\vartheta \, d\vartheta \tag{4.295}$$

$$= \frac{\pi}{2}. \tag{4.296}$$

Therefore,

$$\int_{-\infty}^{\infty} B_0^2(z)\, dz = \frac{\pi a B_m^2}{2}, \tag{4.297}$$

which proves Eq. (4.291). Substitution into Eq. (4.247) gives Eq. (4.292).

There is an important difference between the dependence of rotation and focusing on the magnetic field. From Eq. (4.267),

$$\Delta\theta_L \propto B_m, \tag{4.298}$$

whereas from Eq. (4.292),

$$\frac{1}{f} \propto B_m^2. \tag{4.299}$$

Thus the rotation angle depends linearly on the magnetic field, while the focusing power depends quadratically on it. If the magnetic field is approximately proportional to the lens current, this also explains why image rotation and focal power have different dependences on the excitation of the magnetic lens.

Orders of magnitude for a 300 kV objective lens

To obtain a feeling for the relevant scales, consider again a 300 kV electron. From Eq. (2.77), with negligible emission energy,

$$\widehat{\Phi} = V(1+\epsilon V) \simeq 3.88 \times 10^5 \text{ V}. \tag{4.300}$$

Using Eq. (2.78), or equivalently Eq. (4.37), a peak field

$$B_m = 2 \text{ T} \tag{4.301}$$

corresponds to

$$\kappa_m \simeq 9.52 \times 10^2 \text{ rad}\,\text{m}^{-1}. \tag{4.302}$$

Let the characteristic half-width of the field be

$$a = 1 \text{ mm}, \tag{4.303}$$

corresponding in this model to a magnetic-field FWHM of 2 mm. The dimensionless strength parameter appearing in Eq. (4.274) is then

$$\frac{a\kappa_m}{2} \simeq 0.476, \tag{4.304}$$

and

$$\nu \simeq 1.108. \tag{4.305}$$

The total Larmor rotation predicted by Eq. (4.267) is

$$\Delta\theta_L = \frac{\pi}{2}(10^{-3})(9.52 \times 10^2) \tag{4.306}$$

$$\simeq 1.50 \text{ rad} \tag{4.307}$$

$$\simeq 86^\circ. \tag{4.308}$$

Thus a strong objective lens can rotate the transverse electron-optical coordinates by an angle of order one radian while the electrons traverse only a few millimetres of significant magnetic field.

Within the thin-lens approximation, Eq. (4.292) gives

$$f = \frac{8}{\pi a \kappa_m^2} \tag{4.309}$$

$$\simeq 2.8 \times 10^{-3} \text{ m} \tag{4.310}$$

$$\simeq 2.8 \text{ mm}. \tag{4.311}$$

The numerical values should not be interpreted as specifications of a particular objective lens: the assumed field profile is deliberately idealized. They nevertheless demonstrate the correct orders of magnitude. A magnetic field of a few tesla acting over a few millimetres is capable of producing both substantial image rotation and millimetre-scale focal lengths for 300 kV electrons.

Connection with the thin-lens description of wave optics

This calculation also provides the physical connection between the magnetic field of an electron lens and the abstract thin-lens operator introduced in Sec. 3.8.

In the present section, the focal length is not introduced as an independent parameter: it emerges from the magnetic-field distribution through Eq. (4.247), or through Eq. (4.292) for the particular model considered here. Once this focal length has been determined, propagation through

the remainder of the optical system can be represented using the thin-lens and ABCD machinery developed previously, in particular the lens matrix of Eq. (3.637).

The chain of approximations and descriptions is therefore

$$\boxed{\begin{array}{c} B_0(z) \\ \Downarrow \\ \text{Lorentz force} \\ \Downarrow \\ \text{Larmor rotation and paraxial focusing} \\ \Downarrow \\ f \\ \Downarrow \\ \text{thin-lens / ABCD representation} \\ \Downarrow \\ \text{wave-optical propagation.} \end{array}} \tag{4.312}$$

This connection will allow us to move naturally from the physical electromagnetic origin of an electron lens to the Fourier-optical description of the complete microscope.

4.3 Geometrical optics: trajectories and wavefronts

The historical debate over whether light should be understood as a stream of particles or as a wave began in the late seventeenth century and continued well into the eighteenth century. In geometrical optics, these two descriptions survive as two complementary mathematical representations of the same propagation process (Houchmandzadeh, 2020).

In the particle description, light propagates along spatial trajectories, or *rays*. In the wave description, one follows surfaces of equal phase, or *wavefronts*. In an isotropic medium these descriptions are closely related: rays are locally orthogonal to the wavefronts, so that, at least locally and away from caustics, either description can be reconstructed from the other.

This ray–wavefront duality will later provide the natural bridge to the Hamilton–Jacobi description of electron trajectories in Sec. 4.6. We first develop it for light, where the geometrical interpretation is particularly transparent.

4.3.1 Wavefronts emitted by a point source

Consider a point source located at $\mathbf{r}_0$ that emits light at time t_0. In a homogeneous and isotropic medium of refractive index n, the speed of propagation is

$$v = \frac{c}{n}. \tag{4.313}$$

After an elapsed time $t - t_0$, every ray has travelled the same distance

$$\frac{c}{n}(t - t_0). \tag{4.314}$$

Consequently, the set of points reached at time t is the sphere

$$\boxed{\|\mathbf{r} - \mathbf{r}_0\| = \frac{c}{n}(t - t_0).} \tag{4.315}$$

It is convenient to express the wavefront as the zero level set of a scalar function,

$$\boxed{D(\mathbf{r}, t; \mathbf{r}_0, t_0) = \|\mathbf{r} - \mathbf{r}_0\| - \frac{c}{n}(t - t_0) = 0.} \tag{4.316}$$

For fixed t, Eq. (4.316) defines a surface in space. As t increases, this surface expands radially.

The corresponding rays are simply

$$\mathbf{r}(t) = \mathbf{r}_0 + \frac{c}{n}(t - t_0)\,\mathbf{u}, \tag{4.317}$$

where $\mathbf{u}$ is any constant unit vector. Thus the ray description consists of radial straight lines, whereas the wave description consists of concentric expanding spheres. These are two geometrical views of the same propagation process.

4.3.2 Huygens' construction

The previous construction started from a single point source. Huygens' principle generalizes it to an arbitrary wavefront.

Suppose the wavefront is known at time t. Every point of that wavefront may be regarded as the centre of a secondary spherical wavelet. During a small interval dt, each wavelet expands by the local distance

$$d\ell = \frac{c}{n}\,dt. \tag{4.318}$$

The wavefront at time $t + dt$ is the forward envelope of these secondary wavelets, as illustrated in Fig. 4.2.

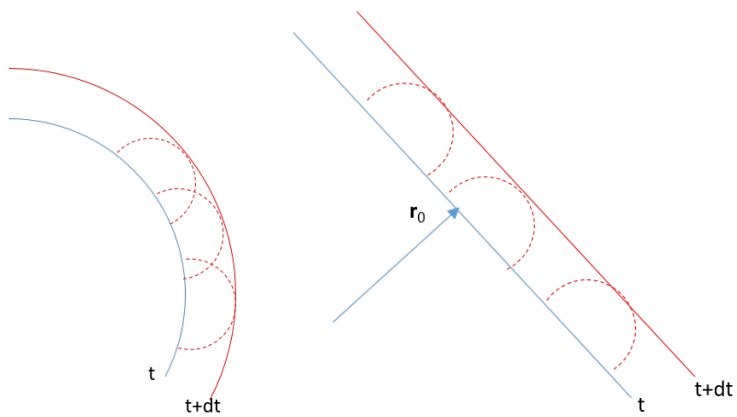


Figure 4.2: Huygens' construction. The wavefront at time t (blue) can be propagated to a later time by allowing every point of the initial wavefront to act as the centre of a secondary spherical wavelet. The new wavefront (red) is the forward envelope of these wavelets. In the far-field limit, the spherical wavefront becomes locally planar.

At the level of geometrical optics, we shall refer to this construction as Huygens' principle. The more complete Huygens–Fresnel principle also assigns phases and amplitudes to the secondary waves and therefore contains interference effects, which are beyond the purely geometrical description considered here.

4.3.3 The plane-wave limit of a spherical wavefront

A spherical wavefront becomes locally indistinguishable from a plane when its radius of curvature is much larger than the spatial region over which the wavefront is observed.

The geometrical arrangement is illustrated in Fig. 4.3. Let $\mathbf{r}_0$ be the position of the point source and let $\mathbf{r}_c$ be a reference point near the centre of the observation region. We define

$$\mathbf{R}_0 \triangleq \mathbf{r}_c - \mathbf{r}_0 \tag{4.319}$$

as the vector from the source to the centre of the observation region, and

$$R_0 \triangleq \|\mathbf{R}_0\|. \tag{4.320}$$

A nearby observation point $\mathbf{r}$ is conveniently written as

$$\boxed{\mathbf{r} = \mathbf{r}_c + \boldsymbol{\rho},} \tag{4.321}$$

where $\boldsymbol{\rho}$ describes the position relative to the centre of the observation region. Hence,

$$\boxed{\mathbf{r} - \mathbf{r}_0 = \mathbf{R}_0 + \boldsymbol{\rho}.} \tag{4.322}$$

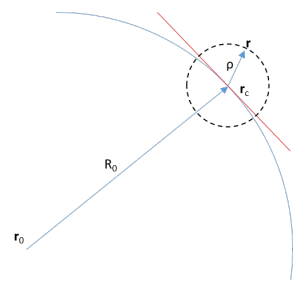


Figure 4.3: Geometrical interpretation of the local plane-wave approximation to a spherical wavefront. A point source at $\mathbf{r}_0$ generates a spherical wavefront. The reference point $\mathbf{r}_c$ lies near the centre of the observation region, and $\mathbf{R}_0 = \mathbf{r}_c - \mathbf{r}_0$ is the vector from the source to this region. A nearby point is written as $\mathbf{r} = \mathbf{r}_c + \boldsymbol{\rho}$, so that $\mathbf{r} - \mathbf{r}_0 = \mathbf{R}_0 + \boldsymbol{\rho}$. When $R_0 \gg \|\boldsymbol{\rho}\|$, the radius of curvature of the spherical wavefront is much larger than the observed region and the wavefront is locally approximated by its tangent plane. The unit vector $\mathbf{u} = \mathbf{R}_0/R_0$ gives the local propagation direction and is normal to that plane.

The condition for local planarity is therefore

$$\boxed{R_0 \gg \|\boldsymbol{\rho}\|.} \tag{4.323}$$

Physically, Eq. (4.323) states that the radius of the spherical wavefront is much larger than the size of the region in which we inspect it. In the formal limit

$$R_0 \to \infty \tag{4.324}$$

with $\boldsymbol{\rho}$ remaining finite, the curvature tends to zero over that local region.

Under this condition, the exact source-to-observation distance $\|\mathbf{r} - \mathbf{r}_0\|$ can be replaced, to first order in $\|\boldsymbol{\rho}\|/R_0$, by

$$\boxed{\|\mathbf{r} - \mathbf{r}_0\| \simeq R_0 + \boldsymbol{\rho} \cdot \mathbf{u},} \tag{4.325}$$

where

$$\boxed{\mathbf{u} \triangleq \frac{\mathbf{R}_0}{R_0} = \frac{\mathbf{r}_c - \mathbf{r}_0}{\|\mathbf{r}_c - \mathbf{r}_0\|}} \tag{4.326}$$

is the unit vector pointing from the source toward the observation region.

Local expansion of the spherical wavefront.

From Eq. (4.322),

$$\mathbf{r} - \mathbf{r}_0 = \mathbf{R}_0 + \boldsymbol{\rho}. \tag{4.327}$$

Its magnitude is

$$\|\mathbf{r} - \mathbf{r}_0\| = \|\mathbf{R}_0 + \boldsymbol{\rho}\| \tag{4.328}$$

$$= \sqrt{(\mathbf{R}_0 + \boldsymbol{\rho}) \cdot (\mathbf{R}_0 + \boldsymbol{\rho})} \tag{4.329}$$

$$= \sqrt{R_0^2 + 2\mathbf{R}_0 \cdot \boldsymbol{\rho} + \|\boldsymbol{\rho}\|^2}. \tag{4.330}$$

Factoring out R_0^2,

$$\|\mathbf{r} - \mathbf{r}_0\| = R_0 \sqrt{1 + 2\frac{\mathbf{R}_0 \cdot \boldsymbol{\rho}}{R_0^2} + \frac{\|\boldsymbol{\rho}\|^2}{R_0^2}}. \tag{4.331}$$

Under the condition Eq. (4.323), the ratio

$$\frac{\|\boldsymbol{\rho}\|}{R_0} \tag{4.332}$$

is small. We therefore use the first-order expansion

$$\sqrt{1 + \varepsilon} = 1 + \frac{\varepsilon}{2} + O(\varepsilon^2). \tag{4.333}$$

In Eq. (4.331), the term

$$2\frac{\mathbf{R}_0 \cdot \boldsymbol{\rho}}{R_0^2} \tag{4.334}$$

is first order in $\|\boldsymbol{\rho}\|/R_0$, whereas

$$\frac{\|\boldsymbol{\rho}\|^2}{R_0^2} \tag{4.335}$$

is second order. Retaining only the first-order contribution,

$$\|\mathbf{r} - \mathbf{r}_0\| \simeq R_0 \left[1 + \frac{\mathbf{R}_0 \cdot \boldsymbol{\rho}}{R_0^2}\right] \tag{4.336}$$

$$= R_0 + \boldsymbol{\rho} \cdot \frac{\mathbf{R}_0}{R_0}. \tag{4.337}$$

Using the definition of $\mathbf{u}$ in Eq. (4.326),

$$\|\mathbf{r} - \mathbf{r}_0\| \simeq R_0 + \boldsymbol{\rho} \cdot \mathbf{u}, \tag{4.338}$$

which is Eq. (4.325).

We can now apply this result directly to the spherical wavefront introduced in Eq. (4.316). That wavefront satisfies

$$\|\mathbf{r} - \mathbf{r}_0\| - \frac{c}{n}(t - t_0) = 0. \tag{4.339}$$

Using Eq. (4.325),

$$\boxed{R_0 + \boldsymbol{\rho} \cdot \mathbf{u} - \frac{c}{n}(t - t_0) \simeq 0.} \tag{4.340}$$

At a fixed time, all terms except $\boldsymbol{\rho} \cdot \mathbf{u}$ are constant. Consequently,

$$\boxed{\boldsymbol{\rho} \cdot \mathbf{u} = \text{constant}.} \tag{4.341}$$

Equation (4.341) is the equation of a plane whose normal vector is $\mathbf{u}$.

Thus, a spherical wavefront does not literally become a plane at any finite distance. Rather, when its radius of curvature is much larger than the region being observed, its variation over that region is, to first order, identical to that of a plane wavefront. The local propagation direction is

$$\mathbf{u} = \frac{\mathbf{r}_c - \mathbf{r}_0}{\|\mathbf{r}_c - \mathbf{r}_0\|}, \tag{4.342}$$

which is normal to the corresponding tangent plane.

4.3.4 Fermat's principle in an inhomogeneous medium

If the refractive index varies with position,

$$n = n(\mathbf{r}), \tag{4.343}$$

light no longer follows straight trajectories and the corresponding wavefronts are no longer spherical or planar.

The travel time along a path $\mathbf{r}$ joining two fixed points $\mathbf{r}_0$ and $\mathbf{r}_1$ is

$$\boxed{T[\mathbf{r}] = \frac{1}{c}\int_{\mathbf{r}_0}^{\mathbf{r}_1} n(\mathbf{r})\, ds,} \tag{4.344}$$

where ds is the differential arc length.

Equivalently, we define the *optical path length*

$$\boxed{\mathcal{O}[\mathbf{r}] \triangleq \int_{\mathbf{r}_0}^{\mathbf{r}_1} n(\mathbf{r})\, ds,} \tag{4.345}$$

so that

$$T = \frac{\mathcal{O}}{c}. \tag{4.346}$$

Fermat's principle states that the physical ray makes Eq. (4.344), or equivalently Eq. (4.345), stationary with respect to infinitesimal variations of the path that keep its endpoints fixed:

$$\boxed{\delta \int_{\mathbf{r}_0}^{\mathbf{r}_1} n(\mathbf{r})\, ds = 0.} \tag{4.347}$$

The word *stationary* is important: the physical path need not always be a strict minimum.

4.3.5 The ray equation

Fermat's principle gives a differential equation for the trajectories.

Let the trial path be parameterized by an arbitrary parameter λ,

$$\mathbf{r} = \mathbf{r}(\lambda). \tag{4.348}$$

Since

$$ds = \left\| \frac{d\mathbf{r}}{d\lambda} \right\| d\lambda, \tag{4.349}$$

the optical path length defined in Eq. (4.345) becomes

$$\mathcal{O}[\mathbf{r}] = \int_{\lambda_0}^{\lambda_1} n(\mathbf{r}) \left\| \frac{d\mathbf{r}}{d\lambda} \right\| d\lambda. \tag{4.350}$$

It is customary in variational problems to denote the integrand by $\mathcal{L}$ and call it the *Lagrangian*. Thus, in the present optical problem,

$$\boxed{\mathcal{L}\left(\mathbf{r}, \dot{\mathbf{r}}\right) \triangleq n(\mathbf{r}) \left\|\dot{\mathbf{r}}\right\|,} \tag{4.351}$$

where, only in this derivation,

$$\dot{\mathbf{r}} \triangleq \frac{d\mathbf{r}}{d\lambda}. \tag{4.352}$$

The word *Lagrangian* should not introduce any additional physics at this point: it is simply a compact notation for the function appearing inside the integral in Eq. (4.350). Fermat's principle, Eq. (4.347), states that the physical ray makes this integral stationary.

We shall now derive the differential equation obeyed by that stationary path. The result is

$$\boxed{\frac{d}{ds}\left(n\mathbf{t}\right) = \nabla n,} \tag{4.353}$$

where

$$\mathbf{t} \triangleq \frac{d\mathbf{r}}{ds} \tag{4.354}$$

is the unit tangent to the ray.

Ray equation from Fermat's principle.

We start from the optical path length Eq. (4.350),

$$\mathcal{O}[\mathbf{r}] = \int_{\lambda_0}^{\lambda_1} \mathcal{L}\left(\mathbf{r}, \dot{\mathbf{r}}\right) d\lambda, \tag{4.355}$$

with the optical Lagrangian defined in Eq. (4.351) as

$$\mathcal{L} = n(\mathbf{r})\|\dot{\mathbf{r}}\|. \tag{4.356}$$

To determine which path makes this integral stationary, consider a small variation of the trajectory,

$$\mathbf{r}_\varepsilon(\lambda) = \mathbf{r}(\lambda) + \varepsilon\, \boldsymbol{\eta}(\lambda), \tag{4.357}$$

where ε is a small scalar and $\boldsymbol{\eta}(\lambda)$ is an arbitrary vector function.

Because the endpoints are fixed in Fermat's principle,

$$\boldsymbol{\eta}(\lambda_0) = \boldsymbol{\eta}(\lambda_1) = \mathbf{0}. \tag{4.358}$$

The corresponding variation of the optical path is

$$\delta\mathcal{O} \triangleq \frac{d}{d\varepsilon}\mathcal{O}[\mathbf{r}_\varepsilon]\bigg|_{\varepsilon=0}. \tag{4.359}$$

Using the chain rule,

$$\delta\mathcal{O} = \int_{\lambda_0}^{\lambda_1} \left[\frac{\partial\mathcal{L}}{\partial\mathbf{r}} \cdot \boldsymbol{\eta} + \frac{\partial\mathcal{L}}{\partial\dot{\mathbf{r}}} \cdot \frac{d\boldsymbol{\eta}}{d\lambda}\right] d\lambda. \tag{4.360}$$

Integrating the second term by parts gives

$$\begin{aligned}\delta\mathcal{O} = &\left[\frac{\partial\mathcal{L}}{\partial\dot{\mathbf{r}}} \cdot \boldsymbol{\eta}\right]_{\lambda_0}^{\lambda_1} \\ &+ \int_{\lambda_0}^{\lambda_1} \left[\frac{\partial\mathcal{L}}{\partial\mathbf{r}} - \frac{d}{d\lambda}\left(\frac{\partial\mathcal{L}}{\partial\dot{\mathbf{r}}}\right)\right] \cdot \boldsymbol{\eta}\, d\lambda.\end{aligned} \tag{4.361}$$

The boundary term vanishes because of Eq. (4.358). Therefore, Eq. (4.360) becomes

$$\delta\mathcal{O} = \int_{\lambda_0}^{\lambda_1} \left[\frac{\partial\mathcal{L}}{\partial\mathbf{r}} - \frac{d}{d\lambda}\left(\frac{\partial\mathcal{L}}{\partial\dot{\mathbf{r}}}\right)\right] \cdot \boldsymbol{\eta}\, d\lambda. \tag{4.362}$$

Fermat's principle requires

$$\delta\mathcal{O} = 0 \tag{4.363}$$

for every admissible variation $\boldsymbol{\eta}(\lambda)$. Since $\boldsymbol{\eta}$ is arbitrary in the interior of the interval, the factor multiplying it must vanish:

$$\boxed{\frac{d}{d\lambda}\left(\frac{\partial\mathcal{L}}{\partial\dot{\mathbf{r}}}\right) - \frac{\partial\mathcal{L}}{\partial\mathbf{r}} = \mathbf{0}.} \tag{4.364}$$

This is the Euler–Lagrange equation. Here it has not been introduced as an independent principle: it follows directly from the stationary-path condition of Fermat's principle.

We now apply Eq. (4.364) to the optical Lagrangian of Eq. (4.351),

$$\mathcal{L} = n(\mathbf{r})\|\dot{\mathbf{r}}\|. \tag{4.365}$$

First,

$$\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{r}}} = n(\mathbf{r})\frac{\dot{\mathbf{r}}}{\|\dot{\mathbf{r}}\|} \tag{4.366}$$

$$= n(\mathbf{r})\mathbf{t}, \tag{4.367}$$

because

$$\mathbf{t} = \frac{\dot{\mathbf{r}}}{\|\dot{\mathbf{r}}\|} = \frac{d\mathbf{r}}{ds}. \tag{4.368}$$

Second,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{r}} = \|\dot{\mathbf{r}}\|\nabla n. \tag{4.369}$$

Substitution into Eq. (4.364) gives

$$\frac{d}{d\lambda}(n\mathbf{t}) = \|\dot{\mathbf{r}}\|\nabla n. \tag{4.370}$$

Finally, because

$$ds = \|\dot{\mathbf{r}}\|\, d\lambda, \tag{4.371}$$

we have

$$\frac{ds}{d\lambda} = \|\dot{\mathbf{r}}\|, \tag{4.372}$$

and therefore, by the chain rule,

$$\frac{d}{d\lambda} = \frac{ds}{d\lambda}\frac{d}{ds} = \|\dot{\mathbf{r}}\|\frac{d}{ds}. \tag{4.373}$$

Using this relation in Eq. (4.370) and cancelling the common factor $\|\dot{\mathbf{r}}\|$ gives

$$\boxed{\frac{d}{ds}(n\mathbf{t}) = \nabla n,} \tag{4.374}$$

which proves Eq. (4.353).

Equation (4.353) immediately reproduces the homogeneous-medium result. If

$$\nabla n = \mathbf{0}, \tag{4.375}$$

then

$$\frac{d\mathbf{t}}{ds} = \mathbf{0}, \tag{4.376}$$

and the ray is a straight line.

4.3.6 The eikonal and the wavefront

We now establish explicitly the connection between the ray trajectory and the wavefront.

Choose an initial point or initial wavefront and define

$$\boxed{W(\mathbf{r}) = \int_{\text{physical ray}}^{\mathbf{r}} n(\mathbf{r}')\, ds'} \tag{4.377}$$

as the stationary optical path length required to reach the point $\mathbf{r}$. The function W is called the *eikonal*.

For a point source emitted at time t_0, the arrival time at $\mathbf{r}$ is

$$\boxed{t(\mathbf{r}) = t_0 + \frac{W(\mathbf{r})}{c}.} \tag{4.378}$$

Therefore all points reached at the same time satisfy

$$\boxed{W(\mathbf{r}) = c(t - t_0) = \text{constant}.} \tag{4.379}$$

Wavefronts are consequently the level surfaces of the eikonal.

For a homogeneous medium,

$$W(\mathbf{r}) = n\|\mathbf{r} - \mathbf{r}_0\|, \tag{4.380}$$

and Eq. (4.379) immediately reproduces Eq. (4.315).

4.3.7 From trajectories to wavefronts

We have so far used Fermat's principle to determine the trajectories of light rays. We now ask the converse question: if the family of stationary rays is known, can we construct the corresponding wavefronts?

The quantity that connects the two descriptions is the eikonal $W(\mathbf{r})$ introduced in Eq. (4.377). For a fixed initial point $\mathbf{r}_0$, it is the optical path length accumulated along the physical ray that reaches the final point $\mathbf{r}$:

$$W(\mathbf{r}) = \int_{\mathbf{r}_0}^{\mathbf{r}} n(\mathbf{r}')\, ds'. \tag{4.381}$$

Locally, and provided that a unique smooth stationary ray reaches $\mathbf{r}$, this defines W as an ordinary scalar function of the final position. If several rays reach the same point, as can occur near caustics, different branches of W must be considered separately.

We shall prove that the gradient of this scalar function is

$$\boxed{\nabla W = n(\mathbf{r})\mathbf{t},} \tag{4.382}$$

where $\mathbf{t}$ is the unit tangent to the ray defined in Eq. (4.354).

Equation (4.382) is the fundamental local relation between trajectories and wavefronts. The essential point in its derivation is to determine how the stationary optical path changes when its *final endpoint* is displaced.

The following calculation is not a new variational argument. It is the same first-variation calculation used to derive the ray equation in Sec. 4.3.5, but with one important difference: the final endpoint is now allowed to move.

Variation of the eikonal when the final endpoint moves.

Let the physical ray joining the fixed initial point $\mathbf{r}_0$ to the final point $\mathbf{r}_1$ be parameterized by λ. Its optical path is

$$W(\mathbf{r}_1) = \int_{\lambda_0}^{\lambda_1} \mathcal{L}\left(\mathbf{r}, \dot{\mathbf{r}}\right) d\lambda, \tag{4.383}$$

where the optical Lagrangian is given by Eq. (4.351).

In the derivation of the ray equation we obtained the general first variation Eq. (4.361),

$$\begin{aligned} \delta W = & \left[\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{r}}} \cdot \boldsymbol{\eta}\right]_{\lambda_0}^{\lambda_1} \\ & + \int_{\lambda_0}^{\lambda_1} \left[\frac{\partial \mathcal{L}}{\partial \mathbf{r}} - \frac{d}{d\lambda}\left(\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{r}}}\right)\right] \cdot \boldsymbol{\eta}\, d\lambda. \end{aligned} \tag{4.384}$$

The important difference is that, in the present problem, the final endpoint is allowed to move. The initial point remains fixed,

$$\boldsymbol{\eta}(\lambda_0) = \mathbf{0}, \tag{4.385}$$

whereas

$$\boldsymbol{\eta}(\lambda_1) = d\mathbf{r}_1. \tag{4.386}$$

Because the unperturbed path is the physical ray, it satisfies the Euler–Lagrange equation Eq. (4.364). Therefore the integral term in

Eq. (4.361) vanishes, leaving only the endpoint contribution,

$$\boxed{dW = \left.\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{r}}}\right|_{\mathbf{r}_1} \cdot d\mathbf{r}_1.} \tag{4.387}$$

For the optical Lagrangian, we already showed in Eq. (4.367) that

$$\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{r}}} = n\mathbf{t}. \tag{4.388}$$

Hence,

$$\boxed{dW = n\mathbf{t} \cdot d\mathbf{r}.} \tag{4.389}$$

On the other hand, because $W(\mathbf{r})$ is a scalar function of its endpoint,

$$dW = \nabla W \cdot d\mathbf{r}. \tag{4.390}$$

Since $d\mathbf{r}$ is arbitrary, comparison with Eq. (4.389) gives

$$\boxed{\nabla W = n\mathbf{t},} \tag{4.391}$$

which proves Eq. (4.382).

Equation (4.382) can now be interpreted geometrically. The vector ∇W points in the direction in which the optical path increases most rapidly. In an isotropic medium it is parallel to the unit tangent $\mathbf{t}$ of the ray:

$$\mathbf{t} = \frac{\nabla W}{n}. \tag{4.392}$$

At the same time, the wavefronts are the level surfaces

$$W(\mathbf{r}) = \text{constant}, \tag{4.393}$$

as established in Eq. (4.379). The gradient of a scalar function is perpendicular to its level surfaces. Therefore ∇W is normal to the wavefront, while Eq. (4.382) shows that the ray tangent is parallel to this same vector.

Consequently,

$$\boxed{\text{rays are locally perpendicular to wavefronts in an isotropic medium.}} \tag{4.394}$$

The correspondence works in both directions. If the eikonal W is known, Eq. (4.392) determines the local direction of the rays. Conversely, if the

complete family of rays is known, the relation

$$dW = n\mathbf{t} \cdot d\mathbf{r} \tag{4.395}$$

allows the optical path to be integrated and the level surfaces of W to be constructed.

Thus trajectories and wavefronts provide two local descriptions of the same geometrical-optics propagation:

$$\boxed{\begin{array}{ccc} \text{ray trajectory} & \longleftrightarrow & \text{wavefront} \\ \mathbf{t} & \longleftrightarrow & \nabla W \\ & & \nabla W = n\mathbf{t}. \end{array}} \tag{4.396}$$

4.3.8 The eikonal equation

Taking the norm of Eq. (4.382), and using $\|\mathbf{t}\| = 1$, gives

$$\boxed{\|\nabla W\| = n(\mathbf{r}).} \tag{4.397}$$

Equivalently,

$$\boxed{\nabla W \cdot \nabla W = n^2(\mathbf{r}).} \tag{4.398}$$

This is the *eikonal equation*, the fundamental partial differential equation of geometrical optics.

Equations (4.353) and (4.398) are two equivalent descriptions of the same physics:

$$\boxed{\begin{array}{ccc} \dfrac{d}{ds}(n\mathbf{t}) = \nabla n & \Longleftrightarrow & \text{ray trajectories,} \\ \|\nabla W\|^2 = n^2 & \Longleftrightarrow & \text{wavefronts.} \end{array}} \tag{4.399}$$

The relation Eq. (4.382) connects the two descriptions.

4.3.9 Why this becomes more subtle for electrons

The orthogonality expressed in Eq. (4.394) depends on an important property of isotropic geometrical optics: the quantity normal to the wavefront, ∇W, is parallel to the physical ray momentum and therefore to the ray velocity.

This correspondence need not hold in more general systems. In anisotropic optical media, for example, the direction associated with the phase gradient need not coincide with the direction in which energy propagates.

An even more important example for the present book occurs for electrons moving in a magnetic field. As will be derived in Sec. 4.6, the gradient of the corresponding Hamilton–Jacobi function determines the *canonical* momentum, whereas the electron velocity is parallel to the *mechanical* momentum. In the presence of a magnetic vector potential these two momenta are not identical.

Consequently, the trajectories of electrons in a magnetic lens are not, in general, simply the normals to the corresponding constant-action surfaces. The geometrical ray–wavefront duality survives, but the relation between the tangent to a trajectory and the normal to a wavefront must be generalized. Developing that generalization is the purpose of Sec. 4.6.

4.4 Analytical mechanics

In Sec. 4.3 we saw that geometrical optics admits two complementary descriptions. Light may be described by individual trajectories, or rays, and it may also be described by scalar functions whose level surfaces are wavefronts. In an isotropic medium the two descriptions are connected by Eq. (4.382),

$$\nabla W_{\text{opt}} = n\mathbf{t}. \tag{4.400}$$

The purpose of the present section is to develop the more general mathematical framework behind this duality. Analytical mechanics provides three increasingly global descriptions of a dynamical system:

1. the *Lagrangian formulation*, in which the physical trajectory is obtained from a variational principle;

2. the *Hamiltonian formulation*, in which the motion is described in terms of positions and canonical momenta;

3. the *Hamilton–Jacobi formulation*, in which an entire family of trajectories is encoded by a scalar function whose level surfaces play a role analogous to wavefronts.

This last formulation will be particularly useful for electron optics. For electrons moving in magnetic fields, however, an important complication will arise: the momentum normal to the Hamilton–Jacobi surfaces is the *canonical* momentum, whereas the tangent to the physical trajectory is determined by the *mechanical* momentum. We shall derive this distinction explicitly in Sec. 4.6.

4.4.1 Action and the principle of stationary action

The variational machinery required here has already appeared in Sec. 4.3.5. There, Fermat's principle selected the physical light ray by requiring the optical path length to be stationary.

In mechanics, the analogous quantity is the *action*. For a trajectory $\mathbf{r}(t)$ joining the spacetime points $(\mathbf{r}_0, t_0)$ and $(\mathbf{r}_1, t_1)$, define

$$\boxed{\mathcal{A}[\mathbf{r}] \triangleq \int_{t_0}^{t_1} L(\mathbf{r}, \dot{\mathbf{r}}, t)\, dt,} \tag{4.401}$$

where

$$\dot{\mathbf{r}} \triangleq \frac{d\mathbf{r}}{dt}, \tag{4.402}$$

and $L(\mathbf{r}, \dot{\mathbf{r}}, t)$ is the *Lagrangian*.

The notation $\mathcal{A}[\mathbf{r}]$ emphasizes that the action is a *functional*: its argument is an entire trajectory rather than a single number or vector.

The principle of stationary action states that the physical trajectory makes Eq. (4.401) stationary with respect to infinitesimal variations of the path that leave the initial and final spacetime points fixed:

$$\boxed{\delta\mathcal{A} = 0.} \tag{4.403}$$

The corresponding Euler–Lagrange equation was derived from the same variational argument in Eq. (4.364). With time as the trajectory parameter, it becomes

$$\boxed{\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\mathbf{r}}}\right) - \frac{\partial L}{\partial \mathbf{r}} = \mathbf{0}.} \tag{4.404}$$

4.4.2 A familiar example: recovery of Newton's equation

For a non-relativistic particle moving in a scalar potential $V(\mathbf{r})$, the Lagrangian is

$$\boxed{L = \frac{1}{2}m\dot{\mathbf{r}}^{\,2} - V(\mathbf{r}).} \tag{4.405}$$

The Euler–Lagrange equation then reproduces Newton's second law.

Newton's equation from stationary action.

For Eq. (4.405),

$$\frac{\partial L}{\partial \dot{\mathbf{r}}} = m\dot{\mathbf{r}}, \tag{4.406}$$

while

$$\frac{\partial L}{\partial \mathbf{r}} = -\nabla V. \tag{4.407}$$

Substitution into Eq. (4.404) gives

$$\frac{d}{dt}(m\dot{\mathbf{r}}) + \nabla V = \mathbf{0}. \tag{4.408}$$

For constant mass,

$$\boxed{m\ddot{\mathbf{r}} = -\nabla V.} \tag{4.409}$$

Since the force associated with the potential is

$$\mathbf{F} = -\nabla V, \tag{4.410}$$

Eq. (4.409) is Newton's second law.

The Lagrangian formulation therefore does not describe different physics. It reformulates the equations of motion in a way that is particularly well suited to changes of coordinates, symmetry arguments, and the connection between trajectories and wavefronts.

4.4.3 Canonical momentum

For every generalized velocity, the Lagrangian defines a corresponding *canonical momentum*. In Cartesian vector notation,

$$\boxed{\mathbf{P} \triangleq \frac{\partial L}{\partial \dot{\mathbf{r}}}.} \tag{4.411}$$

We deliberately use the capital symbol $\mathbf{P}$ for canonical momentum. Throughout this book,

$$\boxed{\mathbf{p} = \gamma m_0 \mathbf{v}} \tag{4.412}$$

denotes the mechanical, or kinetic, momentum introduced previously.

For the elementary non-relativistic Lagrangian Eq. (4.405),

$$\mathbf{P} = m\dot{\mathbf{r}}, \tag{4.413}$$

so canonical and mechanical momentum coincide.

This equality is not general. In particular, for a charged particle moving in an electromagnetic field, we shall show in Sec. 4.6 that

$$\mathbf{P} = \mathbf{p} + Q\mathbf{A}, \tag{4.414}$$

where Q is the particle charge and $\mathbf{A}$ the magnetic vector potential. Thus, in magnetic electron optics, the distinction between $\mathbf{P}$ and $\mathbf{p}$ is essential.

4.4.4 The Legendre transformation and the Hamiltonian

The Lagrangian uses position and velocity,

$$L = L(\mathbf{r}, \dot{\mathbf{r}}, t), \tag{4.415}$$

as independent dynamical variables. Hamiltonian mechanics replaces the velocity by its conjugate canonical momentum.

The mathematical operation that performs this exchange is a *Legendre transformation*. For an ordinary scalar function $f(v)$, one defines

$$p = \frac{df}{dv}, \tag{4.416}$$

and constructs the transformed function

$$g(p) = pv - f(v), \tag{4.417}$$

where v is subsequently expressed as a function of p.

Applying the same construction to the Lagrangian gives the *Hamiltonian*,

$$\boxed{H(\mathbf{r}, \mathbf{P}, t) \triangleq \mathbf{P} \cdot \dot{\mathbf{r}} - L(\mathbf{r}, \dot{\mathbf{r}}, t).} \tag{4.418}$$

The velocity $\dot{\mathbf{r}}$ appearing on the right-hand side is to be eliminated in favour of $\mathbf{P}$ using Eq. (4.411). This requires that the momentum–velocity relation be locally invertible.

The Hamiltonian introduced here is a classical function of the phase-space variables,

$$H = H(\mathbf{r}, \mathbf{P}, t). \tag{4.419}$$

It should not be confused with the quantum-mechanical Hamiltonian operator H_D introduced in Sec. 3.4. Although the two concepts are closely related, the former is an ordinary function on phase space, whereas the latter is an operator acting on a quantum state.

4.4.5 Hamilton's equations

The Hamiltonian formulation replaces the Euler–Lagrange equation by a pair of coupled first-order equations for the position $\mathbf{r}(t)$ and the canonical momentum $\mathbf{P}(t)$.

Starting from the Hamiltonian defined in Eq. (4.418),

$$H = \mathbf{P} \cdot \dot{\mathbf{r}} - L, \tag{4.420}$$

one obtains Hamilton's canonical equations,

$$\boxed{\begin{aligned} \dot{\mathbf{r}} &= \frac{\partial H}{\partial \mathbf{P}}, \\ \dot{\mathbf{P}} &= -\frac{\partial H}{\partial \mathbf{r}}. \end{aligned}} \tag{4.421}$$

The same derivation also gives the relation

$$\boxed{\frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t}.} \tag{4.422}$$

Derivation of Hamilton's equations.

We begin with the definition of the Hamiltonian, Eq. (4.418),

$$H = \mathbf{P} \cdot \dot{\mathbf{r}} - L. \tag{4.423}$$

Taking its total differential,

$$dH = d\left(\mathbf{P} \cdot \dot{\mathbf{r}}\right) - dL. \tag{4.424}$$

Using the product rule for the scalar product,

$$d\left(\mathbf{P} \cdot \dot{\mathbf{r}}\right) = \dot{\mathbf{r}} \cdot d\mathbf{P} + \mathbf{P} \cdot d\dot{\mathbf{r}}, \tag{4.425}$$

so that

$$dH = \dot{\mathbf{r}} \cdot d\mathbf{P} + \mathbf{P} \cdot d\dot{\mathbf{r}} - dL. \tag{4.426}$$

The Lagrangian is a function

$$L = L\left(\mathbf{r}, \dot{\mathbf{r}}, t\right), \tag{4.427}$$

and therefore its total differential is

$$dL = \frac{\partial L}{\partial \mathbf{r}} \cdot d\mathbf{r} + \frac{\partial L}{\partial \dot{\mathbf{r}}} \cdot d\dot{\mathbf{r}} + \frac{\partial L}{\partial t}\, dt. \tag{4.428}$$

Substituting Eq. (4.428) into Eq. (4.426),

$$\begin{aligned} dH = {} & \dot{\mathbf{r}} \cdot d\mathbf{P} + \mathbf{P} \cdot d\dot{\mathbf{r}} \\ & - \frac{\partial L}{\partial \mathbf{r}} \cdot d\mathbf{r} - \frac{\partial L}{\partial \dot{\mathbf{r}}} \cdot d\dot{\mathbf{r}} - \frac{\partial L}{\partial t}\, dt. \end{aligned} \tag{4.429}$$

By the definition of canonical momentum, Eq. (4.411),

$$\mathbf{P} = \frac{\partial L}{\partial \dot{\mathbf{r}}}. \tag{4.430}$$

Hence, the two terms multiplying $d\dot{\mathbf{r}}$ cancel:

$$\mathbf{P} \cdot d\dot{\mathbf{r}} - \frac{\partial L}{\partial \dot{\mathbf{r}}} \cdot d\dot{\mathbf{r}} = 0. \tag{4.431}$$

We are left with

$$dH = \dot{\mathbf{r}} \cdot d\mathbf{P} - \frac{\partial L}{\partial \mathbf{r}} \cdot d\mathbf{r} - \frac{\partial L}{\partial t}\, dt. \tag{4.432}$$

The Euler–Lagrange equation, Eq. (4.404), gives

$$\frac{\partial L}{\partial \mathbf{r}} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\mathbf{r}}}\right). \tag{4.433}$$

Using again the definition of canonical momentum,

$$\frac{\partial L}{\partial \mathbf{r}} = \frac{d\mathbf{P}}{dt} = \dot{\mathbf{P}}. \tag{4.434}$$

Therefore, Eq. (4.432) becomes

$$\boxed{dH = \dot{\mathbf{r}} \cdot d\mathbf{P} - \dot{\mathbf{P}} \cdot d\mathbf{r} - \frac{\partial L}{\partial t}\, dt.} \tag{4.435}$$

On the other hand, after the Legendre transformation the Hamiltonian is regarded as a function

$$H = H\left(\mathbf{r}, \mathbf{P}, t\right). \tag{4.436}$$

Its total differential is therefore

$$dH = \frac{\partial H}{\partial \mathbf{r}} \cdot d\mathbf{r} + \frac{\partial H}{\partial \mathbf{P}} \cdot d\mathbf{P} + \frac{\partial H}{\partial t}\, dt. \tag{4.437}$$

Equations (4.435) and (4.437) are two expressions for the same differential dH. Since $d\mathbf{r}$, $d\mathbf{P}$, and dt are independent increments, their coefficients must be equal.

Comparing the coefficients multiplying $d\mathbf{P}$,

$$\frac{\partial H}{\partial \mathbf{P}} = \dot{\mathbf{r}}. \tag{4.438}$$

Comparing those multiplying $d\mathbf{r}$,

$$\frac{\partial H}{\partial \mathbf{r}} = -\dot{\mathbf{P}}. \tag{4.439}$$

Finally, comparison of the coefficients multiplying dt gives

$$\frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t}. \tag{4.440}$$

Thus,

$$\boxed{\begin{aligned} \dot{\mathbf{r}} &= \frac{\partial H}{\partial \mathbf{P}}, \\ \dot{\mathbf{P}} &= -\frac{\partial H}{\partial \mathbf{r}}, \end{aligned}} \tag{4.441}$$

which are Hamilton's canonical equations, Eq. (4.421).

The first Hamilton equation determines how the position changes from the dependence of the Hamiltonian on canonical momentum,

$$\dot{\mathbf{r}} = \frac{\partial H}{\partial \mathbf{P}}, \tag{4.442}$$

whereas the second determines how the canonical momentum changes from the spatial dependence of the Hamiltonian,

$$\dot{\mathbf{P}} = -\frac{\partial H}{\partial \mathbf{r}}. \tag{4.443}$$

The Lagrangian formulation describes the trajectory through a second-order differential equation for $\mathbf{r}(t)$. The Hamiltonian formulation instead describes the same dynamics through two coupled first-order equations for

$$\left(\mathbf{r}(t), \mathbf{P}(t)\right). \tag{4.444}$$

The combined space of positions and canonical momenta is called *phase space*. A physical trajectory is therefore represented in Hamiltonian mechanics by a curve through phase space rather than only by a curve through ordinary position space.

4.4.6 Conservation of the Hamiltonian

Hamilton's equations also reveal an important conservation law. Along a physical trajectory, the total time derivative of the Hamiltonian satisfies

$$\boxed{\frac{dH}{dt} = \frac{\partial H}{\partial t}.} \tag{4.445}$$

Using the relation between the explicit time dependences of the Hamiltonian and Lagrangian obtained in Eq. (4.422),

$$\frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t}, \tag{4.446}$$

we therefore have

$$\boxed{\frac{dH}{dt} = \frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t}.} \tag{4.447}$$

Time evolution of the Hamiltonian along a trajectory.

Since the Hamiltonian is a function

$$H = H\left(\mathbf{r}, \mathbf{P}, t\right), \tag{4.448}$$

its total derivative along the trajectory $\mathbf{r}(t), \mathbf{P}(t)$ is

$$\frac{dH}{dt} = \frac{\partial H}{\partial \mathbf{r}} \cdot \dot{\mathbf{r}} + \frac{\partial H}{\partial \mathbf{P}} \cdot \dot{\mathbf{P}} + \frac{\partial H}{\partial t}. \tag{4.449}$$

Using Hamilton's equations, Eq. (4.421),

$$\frac{\partial H}{\partial \mathbf{r}} = -\dot{\mathbf{P}}, \qquad \frac{\partial H}{\partial \mathbf{P}} = \dot{\mathbf{r}}, \tag{4.450}$$

Eq. (4.449) becomes

$$\frac{dH}{dt} = -\dot{\mathbf{P}} \cdot \dot{\mathbf{r}} + \dot{\mathbf{r}} \cdot \dot{\mathbf{P}} + \frac{\partial H}{\partial t}. \tag{4.451}$$

The first two terms are equal and opposite because the scalar product is commutative,

$$\dot{\mathbf{P}} \cdot \dot{\mathbf{r}} = \dot{\mathbf{r}} \cdot \dot{\mathbf{P}}. \tag{4.452}$$

They therefore cancel, leaving

$$\boxed{\frac{dH}{dt} = \frac{\partial H}{\partial t}.} \tag{4.453}$$

Finally, Eq. (4.422) gives

$$\frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t}, \tag{4.454}$$

and hence

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t}, \tag{4.455}$$

which is Eq. (4.447).

If the Lagrangian has no explicit dependence on time,

$$\frac{\partial L}{\partial t} = 0, \tag{4.456}$$

then Eq. (4.447) gives

$$\frac{dH}{dt} = 0. \tag{4.457}$$

Therefore,

$$\boxed{H = \text{constant.}} \tag{4.458}$$

This is a direct consequence of time-translation symmetry: if the Lagrangian does not change explicitly with time, the corresponding Hamiltonian is conserved along the motion.

The conserved Hamiltonian is often, but not by definition, the physical energy of the system. This can be seen in the simple non-relativistic example introduced in Eq. (4.405),

$$L = \frac{1}{2} m \dot{\mathbf{r}}^{\,2} - V(\mathbf{r}). \tag{4.459}$$

The canonical momentum is

$$\mathbf{P} = m\dot{\mathbf{r}}, \tag{4.460}$$

so the Hamiltonian defined in Eq. (4.418) is

$$H = \mathbf{P} \cdot \dot{\mathbf{r}} - L \tag{4.461}$$

$$= m\dot{\mathbf{r}}^{\,2} - \left[\frac{1}{2}m\dot{\mathbf{r}}^{\,2} - V\right] \tag{4.462}$$

$$= \frac{1}{2}m\dot{\mathbf{r}}^{\,2} + V. \tag{4.463}$$

Thus,

$$\boxed{H = T + V = E.} \tag{4.464}$$

For this system, the Hamiltonian is therefore the total mechanical energy. The identification of H with energy follows from the physical form of the Lagrangian; it is not part of the mathematical definition of the Hamiltonian in Eq. (4.418).

4.4.7 The reduced action in a time-independent system

When the system has no explicit time dependence, the Hamiltonian is conserved along every physical trajectory, as established in Eq. (4.458). We may therefore write

$$H = E, \tag{4.465}$$

where E is the constant energy associated with the trajectory.

In this case, the usual action can be separated into a purely spatial contribution and a term containing the elapsed time:

$$\boxed{\mathcal{A} = W - E(t_1 - t_0),} \tag{4.466}$$

where

$$\boxed{W \triangleq \int_{\mathbf{r}_0}^{\mathbf{r}_1} \mathbf{P} \cdot d\mathbf{r}.} \tag{4.467}$$

The quantity W is called the *reduced action* or *abbreviated action*. We shall see in Sec. 4.4.11 that, for a time-independent system, this same quantity is Hamilton's characteristic function.

Reduction of the action in a time-independent system.

From the definition of the Hamiltonian, Eq. (4.418),

$$H = \mathbf{P} \cdot \dot{\mathbf{r}} - L. \tag{4.468}$$

Solving for the Lagrangian,

$$L = \mathbf{P} \cdot \dot{\mathbf{r}} - H. \tag{4.469}$$

For a time-independent system, the Hamiltonian is conserved according to Eq. (4.458), so along a physical trajectory

$$H = E. \tag{4.470}$$

The action functional of Eq. (4.401) therefore becomes

$$\mathcal{A} = \int_{t_0}^{t_1} L \, dt \tag{4.471}$$

$$= \int_{t_0}^{t_1} (\mathbf{P} \cdot \dot{\mathbf{r}} - E) \, dt \tag{4.472}$$

$$= \int_{t_0}^{t_1} \mathbf{P} \cdot \dot{\mathbf{r}} \, dt - E(t_1 - t_0). \tag{4.473}$$

Since

$$\dot{\mathbf{r}} = \frac{d\mathbf{r}}{dt}, \tag{4.474}$$

we have

$$\dot{\mathbf{r}} \, dt = d\mathbf{r}. \tag{4.475}$$

Hence,

$$\int_{t_0}^{t_1} \mathbf{P} \cdot \dot{\mathbf{r}} \, dt = \int_{\mathbf{r}_0}^{\mathbf{r}_1} \mathbf{P} \cdot d\mathbf{r}. \tag{4.476}$$

Defining this spatial line integral as

$$W \triangleq \int_{\mathbf{r}_0}^{\mathbf{r}_1} \mathbf{P} \cdot d\mathbf{r}, \tag{4.477}$$

we obtain

$$\mathcal{A} = W - E(t_1 - t_0), \tag{4.478}$$

which is Eq. (4.466).

The important change is that W depends only on the path through configuration space; time no longer appears explicitly in the integral. At fixed energy and fixed spatial endpoints, the physical trajectory makes the

reduced action stationary:

$$\boxed{\delta \int_{\mathbf{r}_0}^{\mathbf{r}_1} \mathbf{P} \cdot d\mathbf{r} = 0, \qquad E = \text{constant}.} \tag{4.479}$$

This is the *reduced principle of stationary action*, usually associated with Maupertuis' principle.

An important consequence is that the trajectory no longer needs to be parameterized by time. Let

$$\mathbf{r} = \mathbf{r}(u), \tag{4.480}$$

where u is any monotonic parameter along the trajectory. Then

$$d\mathbf{r} = \frac{d\mathbf{r}}{du}\, du, \tag{4.481}$$

and Eq. (4.467) becomes

$$\boxed{W = \int \mathbf{P} \cdot \frac{d\mathbf{r}}{du}\, du.} \tag{4.482}$$

The value of the line integral is independent of the particular parameter used to traverse the same geometrical path. The parameter only determines how points along the path are labelled.

For an optical system with a straight optical axis, the longitudinal coordinate z is therefore a particularly natural choice. Setting

$$u = z, \tag{4.483}$$

and defining

$$\mathbf{r}'(z) \triangleq \frac{d\mathbf{r}}{dz}, \tag{4.484}$$

we obtain

$$\boxed{W = \int_{z_0}^{z_1} \mathbf{P} \cdot \mathbf{r}'(z)\, dz.} \tag{4.485}$$

Equation (4.485) provides the variational justification for using the optical-axis coordinate z, rather than time, as the independent variable in geometrical and electron optics.

4.4.8 From the action functional to a characteristic function

There is an important distinction between the action functional $\mathcal{A}[\mathbf{r}]$ and its value along a physical trajectory.

Fix the initial spacetime point $(\mathbf{r}_0, t_0)$ and consider the physical trajectory that reaches a final point $(\mathbf{r}, t)$. We define *Hamilton's principal function* as

$$\boxed{S(\mathbf{r}_0, t_0; \mathbf{r}, t) \triangleq \int_{\text{physical trajectory}} L\, dt.} \tag{4.486}$$

The path has now been selected by the Euler–Lagrange equation, so S is no longer a functional of arbitrary curves. Locally, provided that the physical trajectory is unique, it is an ordinary scalar function of its endpoints.

We next determine how S changes when those endpoints are varied.

4.4.9 Hamilton's central equation

The endpoint variation is the same variational calculation already used in Sec. 4.3.7, but now both space and time may vary.

Variation of the stationary action with its endpoints.

Consider first a variation of the final spacetime point,

$$(\mathbf{r}, t) \longrightarrow (\mathbf{r} + d\mathbf{r}, t + dt). \tag{4.487}$$

At the original final time t, let the variation of the trajectory itself be $\boldsymbol{\eta}(t)$. The additional evolution during the interval dt contributes the displacement

$$\dot{\mathbf{r}}\, dt. \tag{4.488}$$

Therefore the total displacement of the final spatial point is

$$d\mathbf{r} = \boldsymbol{\eta}(t) + \dot{\mathbf{r}}\, dt, \tag{4.489}$$

so

$$\boldsymbol{\eta}(t) = d\mathbf{r} - \dot{\mathbf{r}}\, dt. \tag{4.490}$$

The general first-variation formula derived in Eq. (4.361) contains an interior term and a boundary term. Since the unperturbed path is a physical trajectory, it satisfies the Euler–Lagrange equation Eq. (4.404), and the interior term vanishes.

The boundary contribution at the final point is therefore

$$\mathbf{P} \cdot \boldsymbol{\eta}(t), \tag{4.491}$$

whereas extending the upper integration limit by dt contributes

$$L\,dt. \tag{4.492}$$

Consequently,

$$\begin{aligned} dS &= \mathbf{P} \cdot \boldsymbol{\eta}(t) + L\,dt && (4.493)\\ &= \mathbf{P} \cdot (d\mathbf{r} - \dot{\mathbf{r}}\,dt) + L\,dt && (4.494)\\ &= \mathbf{P} \cdot d\mathbf{r} - (\mathbf{P} \cdot \dot{\mathbf{r}} - L)\,dt. && (4.495) \end{aligned}$$

Using Eq. (4.418),

$$\boxed{dS = \mathbf{P} \cdot d\mathbf{r} - H\,dt.} \tag{4.496}$$

Applying the same argument to the initial endpoint introduces the opposite sign. Therefore the most general endpoint variation is

$$\boxed{dS = [\mathbf{P} \cdot d\mathbf{r} - H\,dt]_0^1\,.} \tag{4.497}$$

Explicitly,

$$\begin{aligned} dS = \mathbf{P}_1 \cdot d\mathbf{r}_1 - \mathbf{P}_0 \cdot d\mathbf{r}_0 \\ - H_1\,dt_1 + H_0\,dt_0. \end{aligned} \tag{4.498}$$

Equation (4.497) is sometimes called *Hamilton's central equation.*

If the initial endpoint is fixed, then

$$d\mathbf{r}_0 = \mathbf{0}, \qquad dt_0 = 0, \tag{4.499}$$

and Eq. (4.497) reduces to

$$dS = \mathbf{P} \cdot d\mathbf{r} - H\,dt. \tag{4.500}$$

But $S(\mathbf{r}, t)$ is an ordinary scalar function of its final coordinates, so

$$dS = \nabla S \cdot d\mathbf{r} + \frac{\partial S}{\partial t}\,dt. \tag{4.501}$$

Comparison with Eq. (4.500) gives

$$\boxed{\begin{aligned} &\nabla S = \mathbf{P}, \\ &\frac{\partial S}{\partial t} = -H. \end{aligned}} \tag{4.502}$$

These relations are the central bridge between trajectory mechanics and the wavefront-like description that follows.

4.4.10 The Hamilton–Jacobi equation

The Hamiltonian is a function

$$H = H(\mathbf{r}, \mathbf{P}, t). \tag{4.503}$$

The first relation in Eq. (4.502) allows us to replace the canonical momentum by the spatial gradient of the principal function,

$$\mathbf{P} = \nabla S. \tag{4.504}$$

Substituting this into the second relation gives

$$\boxed{\frac{\partial S}{\partial t} + H\left(\mathbf{r}, \nabla S, t\right) = 0.} \tag{4.505}$$

This is the *Hamilton–Jacobi equation.*

Hamilton's equations, Eq. (4.421), describe individual trajectories in phase space. The Hamilton–Jacobi equation instead describes a scalar function $S(\mathbf{r}, t)$. Surfaces

$$S(\mathbf{r}, t) = \text{constant} \tag{4.506}$$

therefore provide a wavefront-like description of the dynamics.

The analogy with geometrical optics is now apparent: instead of following one trajectory at a time, we may encode an entire family of trajectories in a scalar function.

4.4.11 Static systems and Hamilton's characteristic function

For a time-independent system, the Hamiltonian is conserved along the physical trajectory, as established in Eq. (4.458),

$$H = E. \tag{4.507}$$

In this case, Hamilton's principal function can be separated into a spatial part and a temporal part:

$$\boxed{S(\mathbf{r}_0, t_0; \mathbf{r}, t) = W(\mathbf{r}_0; \mathbf{r}) - E(t - t_0).} \tag{4.508}$$

The spatial function W is called *Hamilton's characteristic function.*

Separation of Hamilton's principal function in a static system.

Hamilton's principal function satisfies the Hamilton–Jacobi equation derived in Eq. (4.505),

$$\frac{\partial S}{\partial t} + H\left(\mathbf{r}, \nabla S, t\right) = 0. \tag{4.509}$$

If the Hamiltonian has no explicit time dependence, we seek a separated solution of the form

$$S(\mathbf{r}_0, t_0; \mathbf{r}, t) = W(\mathbf{r}_0; \mathbf{r}) + T(t). \tag{4.510}$$

Since W is independent of time,

$$\frac{\partial S}{\partial t} = \frac{dT}{dt}, \tag{4.511}$$

and since T is independent of position,

$$\nabla S = \nabla W. \tag{4.512}$$

Substitution into the Hamilton–Jacobi equation gives

$$\frac{dT}{dt} + H\left(\mathbf{r}, \nabla W\right) = 0. \tag{4.513}$$

The first term in Eq. (4.513) depends only on time, whereas the second depends only on position. Since their sum must vanish for every $\mathbf{r}$ and t, each must be equal to a constant.

For a static system, the value of the Hamiltonian is the conserved energy E. Therefore,

$$H\left(\mathbf{r}, \nabla W\right) = E, \tag{4.514}$$

and consequently

$$\frac{dT}{dt} = -E. \tag{4.515}$$

Integrating,

$$T(t) = -Et + C. \tag{4.516}$$

Choosing the additive constant as

$$C = Et_0 \tag{4.517}$$

gives

$$T(t) = -E(t - t_0). \tag{4.518}$$

Hence,

$$S(\mathbf{r}_0, t_0; \mathbf{r}, t) = W(\mathbf{r}_0; \mathbf{r}) - E(t - t_0), \tag{4.519}$$

which is Eq. (4.508).

Substituting Eq. (4.508) into the Hamilton–Jacobi equation leaves a purely spatial equation,

$$\boxed{H(\mathbf{r}, \nabla W) = E.} \tag{4.520}$$

This is the *time-independent Hamilton–Jacobi equation.*

Moreover, the endpoint relation for Hamilton's principal function, Eq. (4.502), gives

$$\mathbf{P} = \nabla S. \tag{4.521}$$

Since the time-dependent term in Eq. (4.508) has no spatial dependence,

$$\nabla S = \nabla W. \tag{4.522}$$

Therefore,

$$\boxed{\mathbf{P}(\mathbf{r}) = \nabla W(\mathbf{r}).} \tag{4.523}$$

Thus, the characteristic function determines not only the energy through Eq. (4.520), but also the local canonical momentum through Eq. (4.523). A single scalar function $W(\mathbf{r})$ can therefore encode an entire family of trajectories.

The characteristic function is the reduced action.

The reduced action was introduced in Eq. (4.467) as

$$W_{\text{red}} = \int_{\mathbf{r}_0}^{\mathbf{r}_1} \mathbf{P} \cdot d\mathbf{r}. \tag{4.524}$$

In Sec. 4.4.7, we showed that for a time-independent system the ordinary action can be written as

$$\mathcal{A} = W_{\text{red}} - E(t_1 - t_0) \tag{4.525}$$

according to Eq. (4.466).

Hamilton's principal function is the stationary value of this action evaluated along the physical trajectory. Therefore,

$$S(\mathbf{r}_0, t_0; \mathbf{r}_1, t_1) = W_{\text{red}}(\mathbf{r}_0; \mathbf{r}_1) - E(t_1 - t_0). \tag{4.526}$$

On the other hand, the separation obtained in Eq. (4.508) gives

$$S(\mathbf{r}_0, t_0; \mathbf{r}_1, t_1) = W(\mathbf{r}_0; \mathbf{r}_1) - E(t_1 - t_0). \tag{4.527}$$

Comparing the two expressions, we obtain

$$W(\mathbf{r}_0; \mathbf{r}_1) = W_{\text{red}}(\mathbf{r}_0; \mathbf{r}_1). \tag{4.528}$$

Hence,

$$\boxed{W(\mathbf{r}_0; \mathbf{r}_1) = \int_{\mathbf{r}_0}^{\mathbf{r}_1} \mathbf{P} \cdot d\mathbf{r},} \tag{4.529}$$

where the integral is evaluated along the physical trajectory.

Hamilton's characteristic function is therefore not a new quantity introduced independently of the reduced variational principle. It is precisely the reduced action of Eq. (4.467), now interpreted as a function of the trajectory endpoints.

The three central relations for a static system are consequently

$$\boxed{\begin{aligned} S &= W - E(t - t_0), \\ H(\mathbf{r}, \nabla W) &= E, \\ \mathbf{P} &= \nabla W. \end{aligned}} \tag{4.530}$$

4.4.12 Connection with geometrical optics

We can now reinterpret the geometrical-optics results of Sec. 4.3 in the language of Hamilton–Jacobi theory.

For light propagating in an isotropic medium, the optical eikonal was defined in Eq. (4.377) as the stationary optical path length,

$$\boxed{W_{\text{opt}}(\mathbf{r}) = \int n(\mathbf{r})\, ds,} \tag{4.531}$$

where the integral is evaluated along the physical ray reaching $\mathbf{r}$.

The same optical path was written in variational form in Sec. 4.3, using the optical Lagrangian of Eq. (4.351),

$$\mathcal{L}_{\text{opt}} = n(\mathbf{r})\|\dot{\mathbf{r}}\|. \tag{4.532}$$

The corresponding canonical momentum is

$$\boxed{\mathbf{P}_{\text{opt}} = n(\mathbf{r})\mathbf{t},} \tag{4.533}$$

where

$$\mathbf{t} = \frac{d\mathbf{r}}{ds} \tag{4.534}$$

is the unit tangent to the ray.

Canonical momentum of the optical variational problem.

By the general definition of canonical momentum, Eq. (4.411),

$$\mathbf{P}_{\text{opt}} = \frac{\partial \mathcal{L}_{\text{opt}}}{\partial \dot{\mathbf{r}}}. \tag{4.535}$$

Using Eq. (4.351),

$$\mathcal{L}_{\text{opt}} = n(\mathbf{r})\|\dot{\mathbf{r}}\|. \tag{4.536}$$

When differentiating with respect to $\dot{\mathbf{r}}$, the position $\mathbf{r}$ is held fixed, so $n(\mathbf{r})$ is constant with respect to this derivative. Hence,

$$\mathbf{P}_{\text{opt}} = n(\mathbf{r})\frac{\partial \|\dot{\mathbf{r}}\|}{\partial \dot{\mathbf{r}}}. \tag{4.537}$$

Since

$$\|\dot{\mathbf{r}}\| = \sqrt{\dot{\mathbf{r}} \cdot \dot{\mathbf{r}}}, \tag{4.538}$$

we have

$$\frac{\partial\|\dot{\mathbf{r}}\|}{\partial\dot{\mathbf{r}}} = \frac{\dot{\mathbf{r}}}{\|\dot{\mathbf{r}}\|}. \tag{4.539}$$

Therefore,

$$\mathbf{P}_{\text{opt}} = n(\mathbf{r})\frac{\dot{\mathbf{r}}}{\|\dot{\mathbf{r}}\|}. \tag{4.540}$$

But

$$\frac{\dot{\mathbf{r}}}{\|\dot{\mathbf{r}}\|} = \mathbf{t}, \tag{4.541}$$

the unit tangent to the ray. Thus,

$$\mathbf{P}_{\text{opt}} = n(\mathbf{r})\mathbf{t}, \tag{4.542}$$

which is Eq. (4.533).

Independently, the endpoint-variation argument of Sec. 4.3.7 gave

$$\nabla W_{\text{opt}} = n(\mathbf{r})\mathbf{t} \tag{4.543}$$

in Eq. (4.382).

Comparing this result with Eq. (4.533), we obtain

$$\boxed{\nabla W_{\text{opt}} = \mathbf{P}_{\text{opt}}.} \tag{4.544}$$

This is precisely the optical counterpart of the Hamilton–Jacobi relation derived for a time-independent mechanical system in Eq. (4.523),

$$\mathbf{P} = \nabla W. \tag{4.545}$$

The correspondence is therefore

$$\boxed{\begin{array}{cc} \text{mechanics} & \text{geometrical optics} \\ W(\mathbf{r}) & W_{\text{opt}}(\mathbf{r}) \\ \mathbf{P} = \nabla W & \mathbf{P}_{\text{opt}} = \nabla W_{\text{opt}} = n\mathbf{t}. \end{array}} \tag{4.546}$$

The analogy is mathematical rather than dimensional. The mechanical quantity W has units of action, whereas W_{opt} has units of length. What is common to the two theories is their Hamilton–Jacobi structure: the gradient of the corresponding characteristic function gives the canonical momentum of the variational problem.

The usual eikonal equation follows immediately from this identification:

$$\boxed{\|\nabla W_{\text{opt}}\|^2 = n^2(\mathbf{r}).} \tag{4.547}$$

Recovery of the eikonal equation from Hamilton–Jacobi theory.

From Eq. (4.544),

$$\nabla W_{\text{opt}} = \mathbf{P}_{\text{opt}}. \tag{4.548}$$

Using Eq. (4.533),

$$\nabla W_{\text{opt}} = n(\mathbf{r})\mathbf{t}. \tag{4.549}$$

Taking the squared norm,

$$\|\nabla W_{\text{opt}}\|^2 = n^2(\mathbf{r})\|\mathbf{t}\|^2. \tag{4.550}$$

Because $\mathbf{t}$ is a unit vector,

$$\|\mathbf{t}\| = 1. \tag{4.551}$$

Therefore,

$$\|\nabla W_{\text{opt}}\|^2 = n^2(\mathbf{r}), \tag{4.552}$$

which is Eq. (4.547) and reproduces the eikonal equation already obtained in Eq. (4.398).

The geometrical interpretation also follows directly. The wavefronts are the level surfaces

$$W_{\text{opt}}(\mathbf{r}) = \text{constant}. \tag{4.553}$$

The gradient of a scalar function is normal to its level surfaces, so ∇W_{opt} is normal to the wavefronts.

At the same time, Eq. (4.382) gives

$$\nabla W_{\text{opt}} = n(\mathbf{r})\mathbf{t}, \tag{4.554}$$

so the same vector is parallel to the ray tangent. Consequently,

$$\boxed{\text{rays are orthogonal to wavefronts in an isotropic medium.}} \tag{4.555}$$

This is the Hamilton–Jacobi interpretation of the geometrical result already established in Eq. (4.394).

Thus, the eikonal W_{opt} plays in geometrical optics the same mathematical role that Hamilton's characteristic function W plays in mechanics. In both cases, a scalar function encodes an entire family of trajectories through the direction field given by its gradient.

4.4.13 A z-parameterized formulation of geometrical optics

For systems possessing a straight optical axis, it is useful to express the same theory using z as the independent coordinate, anticipating the form used later in electron optics.

Write the ray as

$$\mathbf{r}(z) = (x(z), y(z), z)\,. \tag{4.556}$$

The differential arc length is

$$ds = \sqrt{1 + x'^2 + y'^2}\, dz, \tag{4.557}$$

where

$$x' = \frac{dx}{dz}, \qquad y' = \frac{dy}{dz}. \tag{4.558}$$

The optical path is therefore

$$W_{\text{opt}} = \int \mathcal{L}_{\text{opt}}\, dz, \tag{4.559}$$

with the z-parameterized optical Lagrangian

$$\boxed{\mathcal{L}_{\text{opt}} = n(x, y, z)\sqrt{1 + x'^2 + y'^2}.} \tag{4.560}$$

The transverse canonical momenta are

$$\boxed{\begin{aligned} P_x &= \frac{\partial \mathcal{L}_{\text{opt}}}{\partial x'} = \frac{nx'}{\sqrt{1 + x'^2 + y'^2}}, \\ P_y &= \frac{\partial \mathcal{L}_{\text{opt}}}{\partial y'} = \frac{ny'}{\sqrt{1 + x'^2 + y'^2}}. \end{aligned}} \tag{4.561}$$

The corresponding Hamiltonian for evolution along z is

$$\mathcal{H}_{\text{opt}} = P_x x' + P_y y' - \mathcal{L}_{\text{opt}}. \tag{4.562}$$

It may be written entirely in terms of the canonical momenta as

$$\boxed{\mathcal{H}_{\text{opt}} = -\sqrt{n^2 - P_x^2 - P_y^2}.} \tag{4.563}$$

Hamiltonian of z-parameterized geometrical optics.

Define

$$q^2 \triangleq x'^2 + y'^2. \tag{4.564}$$

From Eq. (4.561),

$$P_x^2 + P_y^2 = \frac{n^2 q^2}{1+q^2}. \tag{4.565}$$

Hence,

$$n^2 - P_x^2 - P_y^2 = n^2 \left(1 - \frac{q^2}{1+q^2}\right) \tag{4.566}$$

$$= \frac{n^2}{1+q^2}, \tag{4.567}$$

and therefore

$$\sqrt{n^2 - P_x^2 - P_y^2} = \frac{n}{\sqrt{1+q^2}}. \tag{4.568}$$

Furthermore,

$$P_x x' + P_y y' = \frac{nq^2}{\sqrt{1+q^2}}. \tag{4.569}$$

Thus,

$$\mathcal{H}_{\rm opt} = \frac{nq^2}{\sqrt{1+q^2}} - n\sqrt{1+q^2} \tag{4.570}$$

$$= -\frac{n}{\sqrt{1+q^2}} \tag{4.571}$$

$$= -\sqrt{n^2 - P_x^2 - P_y^2}. \tag{4.572}$$

This proves Eq. (4.563).

The corresponding Hamilton–Jacobi equation is

$$\frac{\partial W_{\rm opt}}{\partial z} + \mathcal{H}_{\rm opt}\left(x, y, \frac{\partial W_{\rm opt}}{\partial x}, \frac{\partial W_{\rm opt}}{\partial y}, z\right) = 0. \tag{4.573}$$

Substituting Eq. (4.563),

$$\frac{\partial W_{\text{opt}}}{\partial z} = \sqrt{n^2 - \left(\frac{\partial W_{\text{opt}}}{\partial x}\right)^2 - \left(\frac{\partial W_{\text{opt}}}{\partial y}\right)^2}. \tag{4.574}$$

Squaring and rearranging,

$$\boxed{\left(\frac{\partial W_{\text{opt}}}{\partial x}\right)^2 + \left(\frac{\partial W_{\text{opt}}}{\partial y}\right)^2 + \left(\frac{\partial W_{\text{opt}}}{\partial z}\right)^2 = n^2,} \tag{4.575}$$

or

$$\|\nabla W_{\text{opt}}\|^2 = n^2, \tag{4.576}$$

which reproduces Eq. (4.547).

4.4.14 Geometrical optics as the short-wavelength limit of wave optics

We have now obtained the eikonal equation from two apparently different starting points: Fermat's principle and Hamilton–Jacobi theory. We shall now show that the same equation also emerges directly from wave optics when the wavelength is small compared with the characteristic spatial scales of the problem.

Consider the scalar wave equation introduced in Eq. (3.20),

$$\nabla^2 U = \frac{n^2(\mathbf{r})}{c^2}\frac{\partial^2 U}{\partial t^2}. \tag{4.577}$$

For a monochromatic wave, we write the field in the form

$$\boxed{U(\mathbf{r}, t) = A(\mathbf{r}) \exp\left\{jk_0\left[W_{\text{opt}}(\mathbf{r}) - ct\right]\right\},} \tag{4.578}$$

where

$$k_0 = \frac{2\pi}{\lambda_0} \tag{4.579}$$

is the vacuum wavenumber, $W_{\text{opt}}(\mathbf{r})$ has units of length, and $A(\mathbf{r})$ is a real spatially varying amplitude.

Substitution of Eq. (4.578) into the wave equation gives two coupled equations:

$$\boxed{\|\nabla W_{\text{opt}}\|^2 - \frac{1}{k_0^2}\frac{\nabla^2 A}{A} = n^2(\mathbf{r}),} \tag{4.580}$$

and

$$\boxed{2\nabla A \cdot \nabla W_{\text{opt}} + A\nabla^2 W_{\text{opt}} = 0.} \tag{4.581}$$

The first determines the phase geometry of the wave, while the second determines how its amplitude is transported through space.

Short-wavelength expansion of the scalar wave equation.

Starting from Eq. (4.578), let

$$\varphi(\mathbf{r}, t) = k_0 \left[W_{\text{opt}}(\mathbf{r}) - ct \right], \tag{4.582}$$

so that

$$U = Ae^{j\varphi}. \tag{4.583}$$

Since A and W_{opt} are independent of time,

$$\frac{\partial U}{\partial t} = -jk_0 c\, U, \tag{4.584}$$

and hence

$$\frac{\partial^2 U}{\partial t^2} = -k_0^2 c^2 U. \tag{4.585}$$

For the spatial derivative,

$$\nabla U = \nabla \left[Ae^{jk_0(W_{\text{opt}} - ct)} \right] \tag{4.586}$$

$$= \left[\nabla A + jk_0 A \nabla W_{\text{opt}} \right] e^{jk_0(W_{\text{opt}} - ct)}. \tag{4.587}$$

Taking the divergence once more,

$$\nabla^2 U = \Big[\nabla^2 A - k_0^2 A \|\nabla W_{\text{opt}}\|^2 + jk_0 \left(2\nabla A \cdot \nabla W_{\text{opt}} + A \nabla^2 W_{\text{opt}} \right) \Big] e^{jk_0(W_{\text{opt}} - ct)}. \tag{4.588}$$

Using Eq. (4.585), the right-hand side of Eq. (4.577) becomes

$$\frac{n^2}{c^2} \frac{\partial^2 U}{\partial t^2} = -k_0^2 n^2 A e^{jk_0(W_{\text{opt}} - ct)}. \tag{4.589}$$

Equating both sides and cancelling the common exponential factor gives

$$0 = \nabla^2 A - k_0^2 A \left(\|\nabla W_{\text{opt}}\|^2 - n^2 \right) + jk_0 \left(2\nabla A \cdot \nabla W_{\text{opt}} + A \nabla^2 W_{\text{opt}} \right). \tag{4.590}$$

Because A and W_{opt} are real, the real and imaginary parts must vanish independently. The real part gives

$$\nabla^2 A - k_0^2 A \left(\|\nabla W_{\text{opt}}\|^2 - n^2 \right) = 0, \tag{4.591}$$

or, after division by $k_0^2 A$,

$$\|\nabla W_{\text{opt}}\|^2 - \frac{1}{k_0^2} \frac{\nabla^2 A}{A} = n^2, \tag{4.592}$$

which is Eq. (4.580).

The imaginary part gives

$$2\nabla A \cdot \nabla W_{\text{opt}} + A \nabla^2 W_{\text{opt}} = 0, \tag{4.593}$$

which is Eq. (4.581).

Multiplying this last equation by A,

$$2A\nabla A \cdot \nabla W_{\text{opt}} + A^2 \nabla^2 W_{\text{opt}} = 0. \tag{4.594}$$

Since

$$\nabla(A^2) = 2A\nabla A, \tag{4.595}$$

the product rule for the divergence gives

$$\nabla \cdot \left(A^2 \nabla W_{\text{opt}} \right) = \nabla(A^2) \cdot \nabla W_{\text{opt}} + A^2 \nabla^2 W_{\text{opt}} \tag{4.596}$$

$$= 0. \tag{4.597}$$

Therefore,

$$\boxed{\nabla \cdot \left(A^2 \nabla W_{\text{opt}} \right) = 0.} \tag{4.598}$$

Equation (4.581), or equivalently Eq. (4.598), is called the *transport equation*. It determines how the amplitude changes as a bundle of rays expands or contracts. In this sense, geometrical optics does not discard the amplitude: the ray geometry is determined first by the eikonal equation, and the amplitude is then transported along that geometry.

The geometrical-optics approximation appears when the wavelength is much smaller than the spatial scales over which the amplitude and refractive index vary. In particular, the amplitude correction in Eq. (4.580) must

satisfy

$$\boxed{\left|\frac{1}{k_0^2}\frac{\nabla^2 A}{A}\right| \ll n^2(\mathbf{r}).} \tag{4.599}$$

Since

$$k_0 = \frac{2\pi}{\lambda_0}, \tag{4.600}$$

the factor $1/k_0^2$ scales as λ_0^2. Thus, as the wavelength becomes small compared with the characteristic spatial scales of the field, the correction term becomes negligible. Equation (4.580) then reduces to

$$\boxed{\|\nabla W_{\text{opt}}\|^2 = n^2(\mathbf{r}),} \tag{4.601}$$

which is precisely the eikonal equation previously obtained in Eq. (4.547).

This result gives a particularly important interpretation of geometrical optics. Rays are not an independent physical description unrelated to waves. They are the leading-order description of wave propagation when the wavelength is sufficiently small compared with the spatial scales over which the wave and the medium change.

We have therefore reached the same eikonal equation by three complementary routes:

$$\boxed{\begin{array}{ccc} \text{Fermat's principle} & \Longrightarrow & \text{stationary optical paths,} \\ \text{Hamilton–Jacobi theory} & \Longrightarrow & \text{characteristic surfaces,} \\ \text{short-wavelength wave theory} & \Longrightarrow & \text{geometrical optics.} \end{array}} \tag{4.602}$$

These three routes describe different aspects of the same underlying structure. Fermat's principle determines the physical ray by a variational principle. Hamilton–Jacobi theory encodes a family of such rays in the scalar characteristic function, as discussed in Sec. 4.4.12. Wave theory explains why this trajectory-based description emerges when the wavelength becomes small.

A complementary connection with electron wave mechanics was established earlier in Sec. 3.4. There, starting from relativistic quantum mechanics, we showed that the spatial part of a stationary free-electron state satisfies the Helmholtz equation Eq. (3.159).

We have therefore approached electron optics from two complementary directions:

variational mechanics	$\longrightarrow$	Hamilton–Jacobi theory	$\longrightarrow$	geometrical optics,
relativistic quantum mechanics	$\longrightarrow$	electron Helmholtz equation	$\longrightarrow$	wave optics.

(4.603)

In Sec. 4.6, we shall apply the Hamiltonian route specifically to electrons moving through the electric and magnetic fields of the electron microscope.

4.5 Reconnecting geometrical optics to quantum mechanics

We have reached the description of electron propagation by two different routes.

In Sec. 3.4, starting from the relativistic Dirac equation, we showed that the spatial part of a stationary free-electron state satisfies the Helmholtz equation

$$\left(\nabla^2 + k^2\right)\psi_E = 0 \tag{4.604}$$

with

$$k = \frac{p}{\hbar} \tag{4.605}$$

according to Eqs. (3.159) and (3.163). This justified the use of scalar wave-optical propagation for electrons in homogeneous, field-free regions.

In Sec. 4.4, by contrast, we started from a variational description of trajectories. Hamilton–Jacobi theory showed that a family of classical trajectories can be encoded by Hamilton's principal function $S(\mathbf{r}, t)$ or, in a time-independent system, by Hamilton's characteristic function $W(\mathbf{r})$.

These two descriptions must be related. The connection is provided by the *semiclassical limit*: when the phase of the quantum wave varies on a much shorter spatial scale than its amplitude and than the characteristic scales of the system, the phase obeys the Hamilton–Jacobi equation to leading order. Classical rays then emerge as the trajectories associated with that rapidly varying phase.

4.5.1 The classical action as the phase of the quantum wave

Away from points at which the wave function vanishes, any complex scalar wave can locally be written in polar form as

$$\psi(\mathbf{r}, t) = A(\mathbf{r}, t) \exp\left[j\Theta(\mathbf{r}, t)\right], \tag{4.606}$$

where $A(\mathbf{r}, t) \geq 0$ is the amplitude and $\Theta(\mathbf{r}, t)$ is the phase.

Because a phase is dimensionless, it is useful to express it in units of action by defining

$$\boxed{S_{\mathrm{q}}(\mathbf{r}, t) \triangleq \hbar\Theta(\mathbf{r}, t).} \tag{4.607}$$

The wave function may then be written as

$$\boxed{\psi(\mathbf{r}, t) = A(\mathbf{r}, t) \exp\left[\frac{j}{\hbar} S_{\mathrm{q}}(\mathbf{r}, t)\right].} \tag{4.608}$$

At this stage, Eq. (4.608) is only a reparameterization of the quantum phase. The quantity S_{q} has the same physical dimensions as an action, but this does not yet imply that it is Hamilton's principal function S introduced in Sec. 4.4. That connection emerges in the semiclassical limit.

The relationship can already be anticipated by comparing the phase of a stationary quantum state with the form of Hamilton's principal function for a time-independent classical system. In Sec. 3.4.1, a stationary state of definite energy E was written in the form

$$\psi(\mathbf{r}, t) = \psi_E(\mathbf{r}) \exp\left(-j\frac{E}{\hbar}t\right) \tag{4.609}$$

according to Eq. (3.155).

On the classical side, for a time-independent system Hamilton's principal function separates according to Eq. (4.508),

$$\boxed{S(\mathbf{r}_0, t_0; \mathbf{r}, t) = W(\mathbf{r}_0; \mathbf{r}) - E(t - t_0),} \tag{4.610}$$

where W is Hamilton's characteristic function.

The corresponding phase factor separates as

$$\boxed{\exp\left(\frac{jS}{\hbar}\right) = \exp\left(\frac{jW}{\hbar}\right) \exp\left[-j\frac{E}{\hbar}(t - t_0)\right].} \tag{4.611}$$

Separation of the classical action in the quantum phase factor.

Starting from the time-independent Hamilton–Jacobi result Eq. (4.508),

$$S = W - E(t - t_0), \tag{4.612}$$

we divide by $\hbar$ and multiply by j,

$$\frac{jS}{\hbar} = \frac{jW}{\hbar} - j\frac{E}{\hbar}(t - t_0). \tag{4.613}$$

Exponentiating both sides,

$$\exp\left(\frac{jS}{\hbar}\right) = \exp\left[\frac{jW}{\hbar} - j\frac{E}{\hbar}(t - t_0)\right] \tag{4.614}$$

$$= \exp\left(\frac{jW}{\hbar}\right)\exp\left[-j\frac{E}{\hbar}(t - t_0)\right], \tag{4.615}$$

which gives Eq. (4.611).

The second factor has the same time dependence as the stationary quantum state of Eq. (3.155). The remaining factor therefore has precisely the form required for a spatial phase determined by Hamilton's characteristic function.

The factor

$$\exp\left[-j\frac{E}{\hbar}(t - t_0)\right] \tag{4.616}$$

differs from the usual stationary-state time factor only by the time-independent phase

$$\exp\left(j\frac{Et_0}{\hbar}\right), \tag{4.617}$$

which can be absorbed into the spatial wave function. Thus the temporal part of the classical Hamilton–Jacobi phase has exactly the structure of the temporal phase of a stationary quantum state.

This correspondence suggests that, in the semiclassical regime, the spatial phase of the stationary wave should be governed by Hamilton's characteristic function:

$$\boxed{\psi_E(\mathbf{r}) \simeq A(\mathbf{r})\exp\left[\frac{j}{\hbar}W(\mathbf{r})\right].} \tag{4.618}$$

Equation (4.618) is the stationary WKB form. In this approximation, W controls the rapidly varying phase of the wave, whereas A describes its more slowly varying amplitude.

It is important that the argument above has established only the *structural compatibility* between the quantum phase and Hamilton's principal function. The actual semiclassical identification

$$\boxed{S_{\mathrm{q}} \longrightarrow S} \tag{4.619}$$

does not follow merely from giving the quantum phase the dimensions of an action. It follows because, in the short-wavelength or semiclassical limit, the equation governing the quantum phase reduces to the Hamilton–Jacobi equation. We derive this result next.

4.5.2 From the electron wave to the Hamilton–Jacobi equation

The appearance of Hamilton's characteristic function W in the semiclassical wave function, Eq. (4.618), is not merely a formal analogy. The electron wave equation itself shows that, in the appropriate limit, the phase W satisfies the classical Hamilton–Jacobi equation.

For a free stationary electron, the Helmholtz equation derived in Eq. (3.159) may be written as

$$\boxed{\nabla^2\psi_E + \frac{p^2}{\hbar^2}\psi_E = 0,} \tag{4.620}$$

where p is the magnitude of the mechanical momentum.

We now use the semiclassical form introduced in Eq. (4.618),

$$\psi_E(\mathbf{r}) = A(\mathbf{r})\exp\left[\frac{j}{\hbar}W(\mathbf{r})\right], \tag{4.621}$$

with A and W taken to be real.

Substitution into Eq. (4.620) separates the wave equation into two coupled equations:

$$\boxed{\|\nabla W\|^2 - \hbar^2\frac{\nabla^2 A}{A} = p^2,} \tag{4.622}$$

and

$$\boxed{2\nabla A\cdot\nabla W + A\nabla^2 W = 0.} \tag{4.623}$$

The second equation may equivalently be written as

$$\boxed{\nabla\cdot\left(A^2\nabla W\right) = 0.} \tag{4.624}$$

Amplitude–phase decomposition of the electron wave equation.

Starting from Eq. (4.618),

$$\psi_E = Ae^{jW/\hbar}, \tag{4.625}$$

the gradient is

$$\nabla\psi_E = \nabla\left(Ae^{jW/\hbar}\right) \tag{4.626}$$

$$= \left[\nabla A + \frac{j}{\hbar}A\nabla W\right]e^{jW/\hbar}. \tag{4.627}$$

Taking the divergence once more gives

$$\nabla^2\psi_E = \left[\nabla^2 A - \frac{A}{\hbar^2}\|\nabla W\|^2 + \frac{j}{\hbar}\left(2\nabla A\cdot\nabla W + A\nabla^2 W\right)\right]e^{jW/\hbar}. \tag{4.628}$$

Substituting Eq. (4.628) into Eq. (4.620) and cancelling the non-zero factor $e^{jW/\hbar}$ gives

$$0 = \nabla^2 A - \frac{A}{\hbar^2}\|\nabla W\|^2 + \frac{p^2}{\hbar^2}A + \frac{j}{\hbar}\left(2\nabla A\cdot\nabla W + A\nabla^2 W\right). \tag{4.629}$$

Because A and W are real, the real and imaginary parts must vanish independently. The real part gives

$$\nabla^2 A - \frac{A}{\hbar^2}\|\nabla W\|^2 + \frac{p^2}{\hbar^2}A = 0. \tag{4.630}$$

Multiplying by $\hbar^2/A$,

$$\hbar^2\frac{\nabla^2 A}{A} - \|\nabla W\|^2 + p^2 = 0, \tag{4.631}$$

or equivalently

$$\|\nabla W\|^2 - \hbar^2\frac{\nabla^2 A}{A} = p^2, \tag{4.632}$$

which is Eq. (4.622).

The imaginary part gives

$$2\nabla A\cdot\nabla W + A\nabla^2 W = 0, \tag{4.633}$$

which is Eq. (4.623).

Multiplying this equation by A,

$$2A\nabla A \cdot \nabla W + A^2 \nabla^2 W = 0. \tag{4.634}$$

Since

$$\nabla(A^2) = 2A\nabla A, \tag{4.635}$$

the product rule gives

$$\nabla \cdot \left(A^2 \nabla W\right) = 0, \tag{4.636}$$

which is Eq. (4.624).

The structure is exactly analogous to that obtained for an ordinary scalar wave in Sec. 4.4.14. Equation (4.622) controls the phase, whereas Eq. (4.623) controls the evolution of the amplitude.

The classical limit appears when the amplitude varies slowly compared with the spatial scale over which the phase oscillates. More precisely, the amplitude-dependent term in Eq. (4.622) must satisfy

$$\boxed{\left|\hbar^2 \frac{\nabla^2 A}{A}\right| \ll p^2.} \tag{4.637}$$

Under this condition, Eq. (4.622) reduces to

$$\boxed{\|\nabla W\|^2 = p^2.} \tag{4.638}$$

This is precisely the time-independent Hamilton–Jacobi equation for a free particle. Indeed, Hamilton–Jacobi theory gave Eq. (4.523),

$$\mathbf{P} = \nabla W. \tag{4.639}$$

For a free particle there is no vector potential, so the canonical and mechanical momenta coincide,

$$\mathbf{P} = \mathbf{p}. \tag{4.640}$$

Consequently,

$$\boxed{\nabla W = \mathbf{p}.} \tag{4.641}$$

Taking the squared norm of Eq. (4.641) immediately gives Eq. (4.638).

The quantum phase therefore has a direct geometrical meaning. Since the phase of the stationary wave is

$$\Theta(\mathbf{r}) = \frac{W(\mathbf{r})}{\hbar}, \tag{4.642}$$

its local wavevector is the gradient of the phase,

$$\mathbf{k} = \nabla\Theta = \frac{1}{\hbar}\nabla W. \tag{4.643}$$

Using Eq. (4.641),

$$\boxed{\mathbf{k} = \frac{1}{\hbar}\nabla W = \frac{\mathbf{p}}{\hbar}.} \tag{4.644}$$

Its magnitude reproduces the de Broglie relation already established in Eq. (3.163),

$$k = \frac{p}{\hbar}. \tag{4.645}$$

Thus, in the semiclassical limit, quantities introduced independently in wave mechanics and analytical mechanics become different representations of the same underlying structure:

$$\boxed{\begin{array}{ccc} \text{quantum phase} & \longleftrightarrow & W/\hbar, \\ \text{local wavevector} & \longleftrightarrow & \nabla W/\hbar, \\ \text{classical momentum} & \longleftrightarrow & \nabla W. \end{array}} \tag{4.646}$$

This provides the wave-mechanical origin of the classical ray description in a field-free region. Surfaces

$$W(\mathbf{r}) = \text{constant} \tag{4.647}$$

are surfaces of constant quantum phase. Their normal is ∇W, which according to Eq. (4.641) is parallel to the classical momentum and therefore to the electron trajectory.

In the semiclassical regime, the classical ray is therefore locally normal to the quantum phase fronts.

4.5.3 Feynman's path integral and the emergence of classical rays

There is another, particularly suggestive way of understanding why a quantum electron can nevertheless be described approximately by a classical ray. It is provided by Feynman's path-integral formulation of quantum mechanics.

In classical mechanics, once the initial conditions are specified, the particle follows a particular trajectory. Feynman's formulation starts from a

radically different picture. The quantum amplitude for propagation from an initial spacetime point $(\mathbf{r}_a, t_a)$ to a final point $(\mathbf{r}_b, t_b)$ is obtained by coherently adding contributions associated with *all possible paths* joining the two endpoints.

Schematically,

$$\boxed{K(\mathbf{r}_b, t_b; \mathbf{r}_a, t_a) = \int \mathcal{D}\mathbf{r} \, \exp\left[\frac{j}{\hbar}\mathcal{A}[\mathbf{r}]\right],} \tag{4.648}$$

where $\mathcal{A}[\mathbf{r}]$ is the action functional defined in Eq. (4.401). The notation

$$\int \mathcal{D}\mathbf{r} \tag{4.649}$$

indicates that the integration is not over points in ordinary space, but over an entire space of possible paths.

Equation (4.648) should not be interpreted literally as saying that an electron possesses all these trajectories simultaneously in the classical sense. Rather, the quantum propagator is constructed by assigning an amplitude to every possible path and adding all those amplitudes coherently.

Each path contributes the complex factor

$$\exp\left[\frac{j}{\hbar}\mathcal{A}[\mathbf{r}]\right]. \tag{4.650}$$

Its magnitude is the same, but its phase depends on the action of that particular path:

$$\phi[\mathbf{r}] = \frac{\mathcal{A}[\mathbf{r}]}{\hbar}. \tag{4.651}$$

Thus, the essential question is not simply which paths are possible, but whether the phases associated with neighbouring paths reinforce or cancel one another.

Strictly speaking, Eq. (4.648) is a schematic representation of the path-integral idea. A fully relativistic treatment of a spin-1/2 particle requires additional structure. Here we use the path integral only to expose the stationary-phase mechanism by which classical trajectories emerge from quantum propagation.

Why most paths cancel

Consider two nearby paths. If their actions differ by $\Delta\mathcal{A}$, their relative quantum phase is

$$\boxed{\Delta\phi = \frac{\Delta\mathcal{A}}{\hbar}.} \tag{4.652}$$

When

$$|\Delta\mathcal{A}| \gg \hbar, \tag{4.653}$$

even a very small change of trajectory produces a large change of phase.

The corresponding complex contributions then point in rapidly changing directions in the complex plane. When the contributions from a continuous family of such paths are added, positive and negative phases largely cancel one another. This is the same mathematical phenomenon that occurs in an ordinary rapidly oscillating integral such as

$$I = \int e^{jf(x)/\epsilon}\, dx \tag{4.654}$$

when ϵ is very small: regions in which $f(x)$ changes rapidly contribute little to the final integral because neighbouring oscillations cancel.

For electrons in an electron microscope this phase oscillation is extremely rapid. To see why, consider for simplicity a field-free region, where canonical and mechanical momentum coincide. For a stationary electron, the reduced action of Eq. (4.467) is

$$W = \int \mathbf{p} \cdot d\mathbf{r}. \tag{4.655}$$

For two nearby paths whose lengths differ locally by an amount $\Delta\ell$, and for approximately constant momentum,

$$\Delta W \simeq p\,\Delta\ell. \tag{4.656}$$

The corresponding phase difference is therefore

$$\Delta\phi \simeq \frac{p}{\hbar}\Delta\ell. \tag{4.657}$$

Using Eq. (3.163),

$$\frac{p}{\hbar} = k = \frac{2\pi}{\lambda}, \tag{4.658}$$

we obtain

$$\boxed{\Delta\phi \simeq 2\pi\frac{\Delta\ell}{\lambda}.} \tag{4.659}$$

Equation (4.659) makes the semiclassical limit particularly intuitive. At an electron energy of approximately 300 keV, the relativistic de Broglie wavelength derived in Sec. 2.4 is of order

$$\lambda \simeq 2 \text{ pm}. \tag{4.660}$$

A difference in path length of only

$$\Delta\ell = 1 \text{ nm} \tag{4.661}$$

already corresponds to roughly five hundred electron wavelengths and therefore to a phase change of thousands of radians.

For path differences on micrometre scales, the phase winds around the complex plane hundreds of thousands of times. Contributions from generic neighbouring paths therefore oscillate so rapidly that their coherent sum is strongly suppressed.

This is the physical meaning, in the present context, of the short-wavelength or semiclassical regime. It is not merely that the absolute value of the action is "large". What matters is that *small changes of path generally produce changes of action much larger than* $\hbar$, and hence very large changes of quantum phase.

Why stationary paths survive

There is, however, an important exception.

Consider a path close to some trajectory $\mathbf{r}_{\mathrm{c}}(t)$,

$$\mathbf{r}(t) = \mathbf{r}_{\mathrm{c}}(t) + \delta\mathbf{r}(t). \tag{4.662}$$

The action may be expanded as

$$\mathcal{A}[\mathbf{r}] = \mathcal{A}[\mathbf{r}_{\mathrm{c}}] + \delta\mathcal{A} + \frac{1}{2}\delta^2\mathcal{A} + \cdots . \tag{4.663}$$

For an arbitrary reference path,

$$\delta\mathcal{A} \neq 0. \tag{4.664}$$

The action therefore changes to first order when the path is displaced. Because this change is divided by the very small quantity $\hbar$ in Eq. (4.651), the phases of neighbouring paths rapidly lose coherence and their contributions largely cancel.

A classical trajectory is different. It satisfies the stationary-action condition

$$\boxed{\delta\mathcal{A} = 0,} \tag{4.665}$$

which is precisely the principle introduced in Eq. (4.403).

Near such a path, the linear term in Eq. (4.663) disappears:

$$\mathcal{A}[\mathbf{r}] = \mathcal{A}[\mathbf{r}_{\mathrm{c}}] + \frac{1}{2}\delta^2\mathcal{A} + \cdots . \tag{4.666}$$

The action therefore changes only to second order for sufficiently small deformations of the path.

Consequently, the phases of neighbouring paths vary much more slowly near a stationary trajectory than they do around a generic one. Their contributions to the Feynman integral can therefore add coherently instead of immediately cancelling.

This produces the remarkably suggestive picture illustrated by the path integral:

$$\boxed{\begin{array}{c}\text{all possible paths contribute quantum amplitudes}\\ \downarrow\\ \text{most paths have rapidly varying phases}\\ \downarrow\\ \text{their contributions largely cancel}\\ \downarrow\\ \delta\mathcal{A}=0 \text{ near stationary paths}\\ \downarrow\\ \text{neighbouring contributions remain coherent}\\ \downarrow\\ \text{classical trajectories dominate the semiclassical limit.}\end{array}} \tag{4.667}$$

The stationary paths are exactly those that satisfy the Euler–Lagrange equation Eq. (4.404). Thus the same trajectories that emerged from the classical variational principle in Sec. 4.4 also emerge from Feynman's quantum description through stationary-phase interference.

This provides a particularly intuitive interpretation of the classical ray. Quantum mechanics does not begin by postulating that the electron travels along one infinitesimally thin trajectory. Instead, amplitudes associated with all possible paths enter the propagator. In the short-wavelength limit relevant to high-energy electrons, however, the phase varies extraordinarily rapidly from path to path. Almost all these contributions cancel by destructive interference, while neighbourhoods of stationary-action trajectories contribute coherently. These surviving stationary paths are the trajectories that appear as rays in geometrical electron optics.

The word "surviving" should not be interpreted as meaning that all non-stationary paths individually contribute exactly zero. Their *combined* contribution becomes small because of rapid phase cancellation.

Neither does the argument require a unique stationary trajectory. Several stationary paths may connect the same endpoints. Each then con-

tributes to the semiclassical propagator, with its own amplitude and phase, and the different stationary contributions may themselves interfere. This is one reason why caustics, multiple-path propagation, and diffraction require a wave-optical treatment beyond a single geometrical ray.

4.5.4 Meaning and validity of the geometrical-ray description

We can now state more precisely what a geometrical ray represents in electron optics.

The path-integral argument of Sec. 4.5.3 showed how classical trajectories emerge in the semiclassical regime: although the quantum propagator contains contributions from all possible paths, rapid phase cancellation suppresses the combined contribution of generic non-stationary paths, while neighbourhoods of stationary-action paths contribute coherently.

A geometrical ray represents one of these stationary trajectories. It should therefore not be interpreted as the assertion that the quantum electron is, at every instant, an exactly localized particle following an infinitesimally thin line. Rather, the ray is the trajectory that emerges from the leading-order semiclassical description of the quantum wave.

This interpretation can also be expressed locally in terms of the phase. For a stationary electron in a field-free region, we obtained in Eq. (4.618)

$$\psi_E(\mathbf{r}) \simeq A(\mathbf{r}) \exp\left[\frac{j}{\hbar} W(\mathbf{r})\right]. \tag{4.668}$$

The local phase is therefore

$$\Theta(\mathbf{r}) = \frac{W(\mathbf{r})}{\hbar}, \tag{4.669}$$

and its gradient defines the local wavevector,

$$\mathbf{k} = \nabla\Theta = \frac{1}{\hbar}\nabla W. \tag{4.670}$$

Using the Hamilton–Jacobi relation Eq. (4.523), and noting that canonical and mechanical momentum coincide in a field-free region, we obtain

$$\boxed{\mathbf{k} = \frac{\mathbf{p}}{\hbar}.} \tag{4.671}$$

This is the vector form of the de Broglie relation already obtained in Eq. (3.163).

Thus, in a field-free region,

$$\boxed{\mathbf{k} \parallel \mathbf{p} \parallel \mathbf{t},} \tag{4.672}$$

where $\mathbf{t}$ is the tangent to the classical trajectory. The local wavevector of the quantum wave therefore points along the geometrical ray.

This result connects the two descriptions developed earlier in the book. The wave-mechanical route of Sec. 3.4 led from the Dirac equation to the stationary electron Helmholtz equation, whereas the variational route of Sec. 4.4 led from stationary action to Hamilton–Jacobi theory. In the semiclassical limit these two descriptions meet:

$$\boxed{\begin{array}{ccccc} \text{Dirac equation} & \longrightarrow & \text{electron Helmholtz equation} & \longrightarrow & \text{quantum phase} \\ & & & & \Theta = \dfrac{W}{\hbar} \\ & & & & \downarrow \\ & & & & \mathbf{k} = \dfrac{\nabla W}{\hbar} = \dfrac{\mathbf{p}}{\hbar} \\ & & & & \downarrow \\ & & & & \text{geometrical ray.} \end{array}} \tag{4.673}$$

The ray description is therefore a *short-wavelength approximation*. It is appropriate when the electron wavelength is much smaller than the characteristic length scales over which the wave amplitude, the fields, and the geometry of the optical system vary. This is precisely the regime in which the rapidly varying phase can be separated from the more slowly varying structure of the wave, as discussed in Sec. 4.5.2.

When this separation of scales breaks down, the description by isolated rays is no longer sufficient. Diffraction, interference, caustics, and situations in which several stationary trajectories contribute significantly require the wave-optical description developed in Chap. 3.

Thus geometrical electron optics and electron wave optics should not be regarded as competing descriptions. Geometrical optics is the semiclassical, short-wavelength limit of the underlying quantum-wave description, while wave optics retains the phase relationships that are lost when propagation is reduced to individual rays.

4.6 Hamiltonian dynamics of electrons in electromagnetic fields

The electron trajectories studied in Sec. 2.3 were obtained directly from the Lorentz force, Eq. (2.48). We now describe the same dynamics using the analytical-mechanics framework developed in Sec. 4.4.

The advantage is not that the physical trajectories change. They do not. Rather, the Lagrangian and Hamiltonian formulations introduce canonical variables and characteristic functions that will allow us to describe whole families of electron trajectories and, later, their aberrations.

Throughout this section,

$$\mathbf{p} = \gamma m_0 \mathbf{v} \tag{4.674}$$

denotes the *mechanical* momentum, as in Eq. (4.412), whereas

$$\mathbf{P} \tag{4.675}$$

will denote the *canonical* momentum introduced in Eq. (4.411).

4.6.1 Electromagnetic potentials

It is convenient to describe the electromagnetic fields in terms of a scalar potential $\Phi(\mathbf{r}, t)$ and a vector potential $\mathbf{A}(\mathbf{r}, t)$, defined through

$$\boxed{\begin{aligned} &\mathbf{B} = \nabla \times \mathbf{A}, \\ &\mathbf{E} = -\nabla\Phi - \frac{\partial \mathbf{A}}{\partial t}. \end{aligned}} \tag{4.676}$$

The potentials are not uniquely determined by the physical fields: different choices of Φ and $\mathbf{A}$ may produce exactly the same electric and magnetic fields. We shall return to this freedom in Sec. 4.8.5, where we shall see that a change of electromagnetic potentials is accompanied by a corresponding change in the phase of the electron wave function. For the moment, their importance is that the dynamics of a charged particle can be written in a particularly compact variational form.

4.6.2 The relativistic electromagnetic Lagrangian

For a relativistic particle of charge Q moving in electromagnetic potentials $\Phi(\mathbf{r}, t)$ and $\mathbf{A}(\mathbf{r}, t)$, a convenient Lagrangian is

$$\boxed{L = m_0 c^2 \left(1 - \frac{1}{\gamma}\right) + Q\left(\mathbf{v} \cdot \mathbf{A} - \Phi\right).} \tag{4.677}$$

The first term differs from the more usual relativistic free-particle Lagrangian only by the additive constant m_0c^2. Since adding a constant to a Lagrangian does not change its derivatives with respect to position or velocity, it does not change the Euler–Lagrange equations.

For an electron,

$$Q = -e, \tag{4.678}$$

and therefore

$$\boxed{L_e = m_0c^2\left(1 - \frac{1}{\gamma}\right) - e\,\mathbf{v}\cdot\mathbf{A} + e\Phi.} \tag{4.679}$$

Although Eq. (4.679) may look less familiar than the Lorentz force, it contains exactly the same dynamics. We shall first determine the canonical momentum associated with this Lagrangian and then verify explicitly that its Euler–Lagrange equation reproduces the Lorentz equation derived previously in Eq. (2.48).

Canonical momentum

The canonical momentum was defined generally in Eq. (4.411) as

$$\mathbf{P} = \frac{\partial L}{\partial \dot{\mathbf{r}}}. \tag{4.680}$$

Here

$$\dot{\mathbf{r}} = \mathbf{v}, \tag{4.681}$$

so for the electron

$$\mathbf{P} = \frac{\partial L_e}{\partial \mathbf{v}}. \tag{4.682}$$

The derivative is taken with respect to velocity while keeping $\mathbf{r}$ and t fixed. Thus, during this differentiation, $\Phi(\mathbf{r}, t)$ and $\mathbf{A}(\mathbf{r}, t)$ are regarded as independent of $\mathbf{v}$.

Canonical momentum of an electron.

We first consider the relativistic kinetic part of Eq. (4.679),

$$L_{\text{kin}} = m_0c^2\left(1 - \frac{1}{\gamma}\right). \tag{4.683}$$

Using the Lorentz factor of Eq. (2.47),

$$\frac{1}{\gamma} = \sqrt{1 - \frac{v^2}{c^2}}, \tag{4.684}$$

we may write

$$L_{\text{kin}} = m_0 c^2 \left[1 - \sqrt{1 - \frac{v_x^2 + v_y^2 + v_z^2}{c^2}} \right]. \tag{4.685}$$

Consider one Cartesian component v_i. Differentiating,

$$\frac{\partial L_{\text{kin}}}{\partial v_i} = -m_0 c^2 \frac{1}{2} \left(1 - \frac{v^2}{c^2} \right)^{-1/2} \left(-\frac{2 v_i}{c^2} \right) \tag{4.686}$$

$$= m_0 v_i \left(1 - \frac{v^2}{c^2} \right)^{-1/2} \tag{4.687}$$

$$= \gamma m_0 v_i. \tag{4.688}$$

Since this holds for every Cartesian component,

$$\boxed{\frac{\partial L_{\text{kin}}}{\partial \mathbf{v}} = \gamma m_0 \mathbf{v} = \mathbf{p},} \tag{4.689}$$

where $\mathbf{p}$ is the mechanical momentum introduced previously in Eq. (4.412).

We next consider the electromagnetic terms,

$$L_{\text{em}} = -e\, \mathbf{v} \cdot \mathbf{A} + e\Phi. \tag{4.690}$$

Because Φ does not depend on velocity,

$$\frac{\partial \Phi}{\partial \mathbf{v}} = \mathbf{0}. \tag{4.691}$$

Furthermore,

$$\mathbf{v} \cdot \mathbf{A} = v_x A_x + v_y A_y + v_z A_z, \tag{4.692}$$

and therefore

$$\frac{\partial}{\partial \mathbf{v}} (\mathbf{v} \cdot \mathbf{A}) = \mathbf{A}. \tag{4.693}$$

Hence,

$$\frac{\partial L_{\text{em}}}{\partial \mathbf{v}} = -e\mathbf{A}. \tag{4.694}$$

Adding the kinetic and electromagnetic contributions gives

$$\boxed{\mathbf{P} = \frac{\partial L_e}{\partial \mathbf{v}} = \gamma m_0 \mathbf{v} - e\mathbf{A}.} \tag{4.695}$$

Using

$$\mathbf{p} = \gamma m_0 \mathbf{v}, \tag{4.696}$$

we obtain the particularly important relation

$$\boxed{\mathbf{P} = \mathbf{p} - e\mathbf{A}.} \tag{4.697}$$

Equation (4.697) distinguishes two different notions of momentum. The mechanical momentum

$$\mathbf{p} = \gamma m_0 \mathbf{v} \tag{4.698}$$

is parallel to the electron velocity and therefore determines the direction of the physical trajectory. The canonical momentum $\mathbf{P}$, by contrast, contains the additional contribution $-e\mathbf{A}$ and is the momentum that enters naturally in Hamiltonian mechanics.

When

$$\mathbf{A} = \mathbf{0}, \tag{4.699}$$

the two momenta coincide. In general they do not.

Recovery of the Lorentz equation

The electromagnetic Lagrangian of Eq. (4.679) was constructed so that its Euler–Lagrange equation reproduces the motion of a relativistic electron in electromagnetic fields. We now verify this explicitly.

Using the canonical momentum obtained in Eq. (4.695), the Euler–Lagrange equation leads to

$$\boxed{\frac{d\mathbf{p}}{dt} = -e\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right),} \tag{4.700}$$

which is precisely the relativistic Lorentz equation already introduced in Eq. (2.48).

Recovery of the Lorentz force from the electron Lagrangian.

The Euler–Lagrange equation, Eq. (4.404), applied to the electron Lagrangian gives

$$\frac{d}{dt}\left(\frac{\partial L_e}{\partial \mathbf{v}}\right) - \frac{\partial L_e}{\partial \mathbf{r}} = \mathbf{0}. \tag{4.701}$$

From Eq. (4.695),

$$\frac{\partial L_e}{\partial \mathbf{v}} = \mathbf{p} - e\mathbf{A}. \tag{4.702}$$

We next evaluate the spatial derivative. The kinetic term in Eq. (4.679) depends only on $\mathbf{v}$, so its partial derivative with respect to position vanishes. Holding $\mathbf{v}$ fixed during this differentiation,

$$\frac{\partial L_e}{\partial \mathbf{r}} = -e\nabla\left(\mathbf{v}\cdot\mathbf{A}\right) + e\nabla\Phi. \tag{4.703}$$

Substituting these two results into Eq. (4.701),

$$\frac{d}{dt}\left(\mathbf{p} - e\mathbf{A}\right) + e\nabla\left(\mathbf{v}\cdot\mathbf{A}\right) - e\nabla\Phi = \mathbf{0}. \tag{4.704}$$

Hence,

$$\frac{d\mathbf{p}}{dt} = e\frac{d\mathbf{A}}{dt} - e\nabla\left(\mathbf{v}\cdot\mathbf{A}\right) + e\nabla\Phi. \tag{4.705}$$

The derivative $d\mathbf{A}/dt$ is the total derivative of the vector potential evaluated along the electron trajectory,

$$\mathbf{A} = \mathbf{A}\big(\mathbf{r}(t), t\big). \tag{4.706}$$

For each Cartesian component A_i, the chain rule gives

$$\frac{dA_i}{dt} = \frac{\partial A_i}{\partial t} + \frac{\partial A_i}{\partial x}v_x + \frac{\partial A_i}{\partial y}v_y + \frac{\partial A_i}{\partial z}v_z. \tag{4.707}$$

Therefore,

$$\boxed{\frac{d\mathbf{A}}{dt} = \frac{\partial \mathbf{A}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{A}.} \tag{4.708}$$

We also use the vector identity

$$\boxed{\nabla\left(\mathbf{v}\cdot\mathbf{A}\right) = (\mathbf{v}\cdot\nabla)\mathbf{A} + \mathbf{v}\times\left(\nabla\times\mathbf{A}\right),} \tag{4.709}$$

where $\mathbf{v}$ is treated as constant with respect to the spatial derivative.

Substituting Eqs. (4.708) and (4.709) into Eq. (4.705),

$$\begin{aligned}\frac{d\mathbf{p}}{dt} &= e\left[\frac{\partial \mathbf{A}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{A}\right] \\ &\quad - e\left[(\mathbf{v}\cdot\nabla)\mathbf{A} + \mathbf{v}\times(\nabla\times\mathbf{A})\right] + e\nabla\Phi. \end{aligned} \tag{4.710}$$

The two convective-derivative terms cancel:

$$\frac{d\mathbf{p}}{dt} = e\frac{\partial \mathbf{A}}{\partial t} + e\nabla\Phi - e\mathbf{v}\times(\nabla\times\mathbf{A})\,. \tag{4.711}$$

Finally, using the definitions of the electromagnetic fields from Eq. (4.676),

$$\mathbf{B} = \nabla\times\mathbf{A}, \qquad \mathbf{E} = -\nabla\Phi - \frac{\partial \mathbf{A}}{\partial t}, \tag{4.712}$$

we have

$$e\left(\frac{\partial \mathbf{A}}{\partial t} + \nabla\Phi\right) = -e\mathbf{E}. \tag{4.713}$$

Therefore,

$$\frac{d\mathbf{p}}{dt} = -e\mathbf{E} - e\mathbf{v}\times\mathbf{B}, \tag{4.714}$$

or

$$\boxed{\frac{d\mathbf{p}}{dt} = -e\left(\mathbf{E} + \mathbf{v}\times\mathbf{B}\right),} \tag{4.715}$$

which is Eq. (4.700).

Thus, the relativistic electromagnetic Lagrangian of Eq. (4.679) is dynamically equivalent to the Lorentz-force description of Eq. (2.48). The Lagrangian formulation does not introduce a different electron dynamics; rather, it expresses the same dynamics in a variational form that will allow us to construct the corresponding Hamiltonian and Hamilton–Jacobi descriptions.

4.6.3 The relativistic electron Hamiltonian

The general Hamiltonian was defined through the Legendre transformation in Eq. (4.418),

$$H = \mathbf{P}\cdot\mathbf{v} - L. \tag{4.716}$$

For the relativistic electron Lagrangian of Eq. (4.679), this transforma-

tion gives

$$\boxed{H = E_k - e\Phi,} \tag{4.717}$$

where E_k is the relativistic kinetic energy defined in Eq. (2.52).

Thus the Hamiltonian is the kinetic energy plus the electrostatic potential energy of an electron, whose charge is $-e$. For static electromagnetic fields, this is precisely the conserved energy combination introduced in Sec. 2.3.1.

To use Hamilton's equations, however, the Hamiltonian must be expressed in terms of the canonical variables $(\mathbf{r}, \mathbf{P})$ rather than the velocity. The resulting exact relativistic Hamiltonian is

$$\boxed{H(\mathbf{r}, \mathbf{P}, t) = c\sqrt{(m_0 c)^2 + \|\mathbf{P} + e\mathbf{A}(\mathbf{r}, t)\|^2} - m_0 c^2 - e\Phi(\mathbf{r}, t).} \tag{4.718}$$

Relativistic electron Hamiltonian from the Legendre transformation.

Starting from the definition of the Hamiltonian, Eq. (4.418),

$$H = \mathbf{P} \cdot \mathbf{v} - L_e, \tag{4.719}$$

and substituting the electron Lagrangian of Eq. (4.679),

$$L_e = m_0 c^2 \left(1 - \frac{1}{\gamma}\right) - e\mathbf{v} \cdot \mathbf{A} + e\Phi, \tag{4.720}$$

we obtain

$$H = \mathbf{P} \cdot \mathbf{v} - m_0 c^2 \left(1 - \frac{1}{\gamma}\right) + e\mathbf{v} \cdot \mathbf{A} - e\Phi \tag{4.721}$$

$$= (\mathbf{P} + e\mathbf{A}) \cdot \mathbf{v} - m_0 c^2 \left(1 - \frac{1}{\gamma}\right) - e\Phi. \tag{4.722}$$

The relation between canonical and mechanical momentum, Eq. (4.697), is

$$\mathbf{P} + e\mathbf{A} = \mathbf{p}. \tag{4.723}$$

Using the relativistic mechanical momentum

$$\mathbf{p} = \gamma m_0 \mathbf{v}, \tag{4.724}$$

Eq. (4.722) becomes

$$H = \gamma m_0 v^2 - m_0 c^2 + \frac{m_0 c^2}{\gamma} - e\Phi. \tag{4.725}$$

From the definition of the Lorentz factor, Eq. (2.47),

$$\frac{1}{\gamma^2} = 1 - \frac{v^2}{c^2}, \tag{4.726}$$

and therefore

$$v^2 = c^2 \left(1 - \frac{1}{\gamma^2}\right). \tag{4.727}$$

Consequently,

$$\gamma m_0 v^2 = \gamma m_0 c^2 \left(1 - \frac{1}{\gamma^2}\right) \tag{4.728}$$

$$= \gamma m_0 c^2 - \frac{m_0 c^2}{\gamma}. \tag{4.729}$$

Substitution gives

$$H = \gamma m_0 c^2 - \frac{m_0 c^2}{\gamma} - m_0 c^2 + \frac{m_0 c^2}{\gamma} - e\Phi \tag{4.730}$$

$$= \gamma m_0 c^2 - m_0 c^2 - e\Phi. \tag{4.731}$$

The terms containing $1/\gamma$ cancel. Since Eq. (2.52) gives

$$E_k = \gamma m_0 c^2 - m_0 c^2, \tag{4.732}$$

we obtain

$$H = E_k - e\Phi, \tag{4.733}$$

which is Eq. (4.717).

We now eliminate the velocity and express the Hamiltonian entirely in canonical variables. The relativistic energy–momentum relation, Eq. (2.68), gives

$$\gamma m_0 c^2 = c\sqrt{(m_0 c)^2 + p^2}. \tag{4.734}$$

Using again Eq. (4.697),

$$\mathbf{p} = \mathbf{P} + e\mathbf{A}, \tag{4.735}$$

so that

$$p^2 = \|\mathbf{P} + e\mathbf{A}\|^2 . \tag{4.736}$$

Substitution into

$$H = \gamma m_0 c^2 - m_0 c^2 - e\Phi \tag{4.737}$$

therefore gives

$$H(\mathbf{r}, \mathbf{P}, t) = c\sqrt{(m_0 c)^2 + \|\mathbf{P} + e\mathbf{A}(\mathbf{r}, t)\|^2} - m_0 c^2 - e\Phi(\mathbf{r}, t), \tag{4.738}$$

which is Eq. (4.718).

Equation (4.718) is the exact relativistic Hamiltonian for an electron in an electromagnetic field, expressed in canonical phase-space variables. The vector potential enters through the combination

$$\mathbf{P} + e\mathbf{A} = \mathbf{p}, \tag{4.739}$$

which is the mechanical momentum, while the scalar potential contributes the electrostatic potential energy $-e\Phi$.

The subtraction of $m_0 c^2$ in Eq. (4.718) follows from the convention adopted in the Lagrangian of Eq. (4.679). It removes the constant rest energy without changing the classical equations of motion. Had we used the more conventional relativistic Lagrangian without this additive constant, the Hamiltonian would differ by $m_0 c^2$, but the electron trajectory would be identical.

4.6.4 Static fields and the reduced electron action

For static electric and magnetic fields,

$$\frac{\partial \Phi}{\partial t} = 0, \qquad \frac{\partial \mathbf{A}}{\partial t} = \mathbf{0}, \tag{4.740}$$

the electron Hamiltonian of Eq. (4.718) has no explicit time dependence. Therefore, according to Eq. (4.458),

$$\boxed{H = E_0 = \text{constant},} \tag{4.741}$$

where E_0 denotes the conserved energy with the rest energy removed, consistently with the convention adopted in Eq. (4.717).

The reduction from Hamilton's principal function S to Hamilton's characteristic function W for a time-independent system was derived generally

in Secs. 4.4.7 and 4.4.11. In particular, Eq. (4.529) gives

$$W = \int_{\mathbf{r}_0}^{\mathbf{r}_1} \mathbf{P} \cdot d\mathbf{r}, \tag{4.742}$$

where the integral is evaluated along the physical trajectory.

For an electron, canonical and mechanical momenta are related by Eq. (4.697),

$$\mathbf{P} = \mathbf{p} - e\mathbf{A}. \tag{4.743}$$

Hence,

$$\boxed{W(\mathbf{r}_0; \mathbf{r}_1) = \int_{\mathbf{r}_0}^{\mathbf{r}_1} (\mathbf{p} - e\mathbf{A}) \cdot d\mathbf{r}.} \tag{4.744}$$

This expression is the reduced action appropriate to an electron moving through static electromagnetic fields.

Mechanical momentum along the trajectory

The mechanical momentum is tangent to the physical electron trajectory. If s denotes arc length and

$$\boxed{\mathbf{t} \triangleq \frac{d\mathbf{r}}{ds}} \tag{4.745}$$

is the unit tangent, then

$$\boxed{\mathbf{p} = p\mathbf{t}} \tag{4.746}$$

and consequently

$$\boxed{\mathbf{p} \cdot d\mathbf{r} = p\, ds.} \tag{4.747}$$

Mechanical momentum and the trajectory tangent.

By definition of the arc-length parameter,

$$\mathbf{t} = \frac{d\mathbf{r}}{ds}, \qquad \|\mathbf{t}\| = 1, \tag{4.748}$$

so that

$$\boxed{d\mathbf{r} = \mathbf{t}\, ds.} \tag{4.749}$$

Since the trajectory may also be regarded as $\mathbf{r} = \mathbf{r}(s(t))$, the chain rule gives

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{d\mathbf{r}}{ds}\frac{ds}{dt}. \tag{4.750}$$

Because $ds/dt = v$, where $v = \|\mathbf{v}\|$, we obtain

$$\boxed{\mathbf{v} = v\mathbf{t}.} \tag{4.751}$$

The relativistic mechanical momentum is

$$\mathbf{p} = \gamma m_0 \mathbf{v}, \tag{4.752}$$

and therefore

$$\mathbf{p} = \gamma m_0 v\, \mathbf{t}. \tag{4.753}$$

Since

$$p = \|\mathbf{p}\| = \gamma m_0 v, \tag{4.754}$$

we obtain

$$\mathbf{p} = p\mathbf{t}, \tag{4.755}$$

which is Eq. (4.746).

Finally, using Eq. (4.749),

$$\mathbf{p} \cdot d\mathbf{r} = (p\mathbf{t}) \cdot (\mathbf{t}\, ds) \tag{4.756}$$

$$= p\,(\mathbf{t} \cdot \mathbf{t})\, ds \tag{4.757}$$

$$= p\, ds, \tag{4.758}$$

because $\|\mathbf{t}\| = 1$. This gives Eq. (4.747).

Using the optical-axis coordinate as parameter

In an electron-optical system with a straight optical axis, it is natural to use the longitudinal coordinate z as the parameter along any trajectory for which z varies monotonically. This choice was justified generally in Eq. (4.485).

We therefore write

$$\mathbf{r}(z) = \begin{pmatrix} x(z) \\ y(z) \\ z \end{pmatrix}, \qquad x' \triangleq \frac{dx}{dz}, \qquad y' \triangleq \frac{dy}{dz}. \tag{4.759}$$

The electron reduced action can then be written as

$$\boxed{W = \int_{z_0}^{z_1} \overline{M}\,(x, y, z, x', y')\, dz,} \tag{4.760}$$

where the corresponding z-parameterized electron-optical Lagrangian is

$$\boxed{\begin{aligned}\overline{M}(x,y,z,x',y') = p(\mathbf{r})\sqrt{1+x'^2+y'^2} \\ - e\left(x'A_x + y'A_y + A_z\right).\end{aligned}} \tag{4.761}$$

The z-parameterized electron-optical Lagrangian.

The trajectory differential is

$$d\mathbf{r} = \frac{d\mathbf{r}}{dz}\,dz \tag{4.762}$$

$$= \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} dz. \tag{4.763}$$

Thus,

$$\boxed{d\mathbf{r} = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} dz.} \tag{4.764}$$

Its magnitude gives the arc-length element,

$$\boxed{ds = \sqrt{1+x'^2+y'^2}\,dz.} \tag{4.765}$$

Using Eq. (4.747),

$$\mathbf{p}\cdot d\mathbf{r} = p\,ds \tag{4.766}$$

$$= p(\mathbf{r})\sqrt{1+x'^2+y'^2}\,dz. \tag{4.767}$$

Therefore,

$$\boxed{\mathbf{p}\cdot d\mathbf{r} = p(\mathbf{r})\sqrt{1+x'^2+y'^2}\,dz.} \tag{4.768}$$

For the vector-potential contribution,

$$\mathbf{A} \cdot d\mathbf{r} = \begin{pmatrix} A_x & A_y & A_z \end{pmatrix} \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} dz \tag{4.769}$$

$$= (x' A_x + y' A_y + A_z)\, dz. \tag{4.770}$$

Hence,

$$\boxed{\mathbf{A} \cdot d\mathbf{r} = (x' A_x + y' A_y + A_z)\, dz.} \tag{4.771}$$

The coefficient of A_z is unity because z itself is being used as the trajectory parameter, so that $dz/dz = 1$.

Substituting Eqs. (4.768) and (4.771) into Eq. (4.744),

$$W = \int_{z_0}^{z_1} \Bigg[p(\mathbf{r}) \sqrt{1 + x'^2 + y'^2} - e\,(x' A_x + y' A_y + A_z) \Bigg] dz. \tag{4.772}$$

The coefficient of dz is precisely $\overline{M}$, giving Eq. (4.761) and therefore Eq. (4.760).

The function $\overline{M}$ plays, with z as the independent variable, the same variational role that an ordinary Lagrangian plays when time is used as the independent variable. The electron trajectory can therefore be determined directly in terms of the transverse functions $x(z)$ and $y(z)$.

No paraxial approximation has been introduced in Eq. (4.761). In particular,

$$\sqrt{1 + x'^2 + y'^2} \tag{4.773}$$

is the exact geometrical factor relating ds to dz. The paraxial approximation will enter only when this expression, together with the electromagnetic potentials, is expanded for small transverse coordinates and small slopes.

The electron Hamilton–Jacobi equation

The characteristic function also provides a local description of the electron trajectory. For any time-independent Hamiltonian system, Eq. (4.523) gives

$$\mathbf{P} = \nabla W. \tag{4.774}$$

Combining this with the electron canonical momentum, Eq. (4.697), gives

$$\boxed{\mathbf{p} = \nabla W + e\mathbf{A}.} \tag{4.775}$$

Using the relativistic relation between momentum and acceleration potential derived in Eq. (2.78), the time-independent Hamilton–Jacobi equation for the electron becomes

$$\boxed{\|\nabla W + e\mathbf{A}\|^2 = 2m_0 e\widehat{\Phi}.} \tag{4.776}$$

Hamilton–Jacobi equation for an electron.

For a static system, Eq. (4.523) gives

$$\mathbf{P} = \nabla W. \tag{4.777}$$

The canonical and mechanical momenta of an electron are related by Eq. (4.697),

$$\mathbf{P} = \mathbf{p} - e\mathbf{A}. \tag{4.778}$$

Therefore,

$$\mathbf{p} = \nabla W + e\mathbf{A}, \tag{4.779}$$

which is Eq. (4.775).

Taking the squared norm,

$$p^2 = \|\nabla W + e\mathbf{A}\|^2 . \tag{4.780}$$

But Eq. (2.78) gives

$$p^2 = 2m_0 e\widehat{\Phi}. \tag{4.781}$$

Equating the two expressions for p^2 yields

$$\|\nabla W + e\mathbf{A}\|^2 = 2m_0 e\widehat{\Phi}, \tag{4.782}$$

which is Eq. (4.776).

Equation (4.776) is the electron-optical counterpart of the eikonal equation of geometrical light optics, Eq. (4.398). There is, however, an important difference when a magnetic field is present. In geometrical light optics, the gradient of the eikonal is parallel to the ray direction. For the electron,

$$\nabla W = \mathbf{P} = \mathbf{p} - e\mathbf{A}, \tag{4.783}$$

so ∇W gives the canonical momentum rather than directly the mechanical momentum tangent to the physical trajectory. Only when the vector-potential contribution vanishes do the two directions coincide.

4.7 Electron lenses and multipole elements

In Sec. 4.6 we established how an electron moves through prescribed electromagnetic potentials. In particular, the time-independent Hamilton–Jacobi equation, Eq. (4.776), relates the electron trajectories to the scalar and vector potentials $\Phi(\mathbf{r})$ and $\mathbf{A}(\mathbf{r})$. We have therefore obtained the general dynamics, but we have not yet asked a much more practical question: *what kinds of electromagnetic fields are deliberately created inside an electron microscope, and what optical functions do they perform?*

An electron lens is not a transparent piece of matter analogous to a glass lens. It is a region of electrostatic or magnetic field whose spatial variation changes the direction of the electron trajectories. Different symmetries of the field produce qualitatively different optical actions. Rotationally symmetric fields produce the familiar round lenses; fields invariant along one transverse direction produce cylindrical lenses; and fields with progressively higher angular symmetries give quadrupoles, hexapoles, octupoles, and higher multipoles (Hawkes and Kasper, 2017, Chaps. 19–20).

The purpose of this section is not to develop the theory required to design these elements. That would require detailed calculations of pole pieces, electrodes, fringe fields, transfer maps, and aberrations of the multipoles themselves. Instead, our objective is to understand what the principal elements do and, especially, how their collective action is eventually absorbed into the image-formation model developed later in this chapter.

Thus, although a modern microscope may contain many physically distinct lenses, stigmators, deflectors, and aberration-correcting multipoles, the image-formation theory will ultimately compress their deterministic effect into a comparatively small number of parameters describing the net electron-optical system.

4.7.1 Round lenses

The principal imaging lenses of a conventional transmission electron microscope are *round lenses*: their electromagnetic configuration is rotationally symmetric about the optical axis. If

$$r_\perp = \sqrt{x^2 + y^2}, \tag{4.784}$$

the field strength and potential of an ideal round lens depend on the transverse coordinates only through $r_\perp$.

This rotational symmetry has an important consequence. To first order in the transverse coordinates, the optical action is the same in every azimuthal direction. A round lens therefore has a single paraxial focal strength rather than different focal strengths along x and y.

For magnetic round lenses, the physical mechanism is subtler than a direct radial force. The axial magnetic field generates an azimuthal motion of the electron, and the Lorentz force acting on that motion produces the net focusing effect. This is the origin of the Larmor rotation and focusing derived earlier in this chapter.

Round magnetic lenses form the backbone of the TEM column. The condenser system controls illumination at the specimen, the objective lens performs the first and most resolution-critical imaging operation, and intermediate and projector lenses transport and magnify either image planes or diffraction planes toward the detector.

At the Gaussian level, all of these complicated field distributions can eventually be represented by quantities such as focal lengths, magnifications, and conjugate planes. Their departures from ideal Gaussian behaviour will later appear as aberrations.

This distinction between the *nominal imaging action* and the *departure from that action* will be useful throughout the rest of the chapter:

$$\boxed{\text{round-lens field} = \text{Gaussian focusing} + \text{aberrations.}} \tag{4.785}$$

4.7.2 Cylindrical lenses

A cylindrical electron lens has a lower symmetry. Its electrostatic potential or magnetic field is invariant along one direction transverse to the optical axis (Hawkes and Kasper, 2017, Chap. 20). Suppose, for example, that an electrostatic cylindrical lens is invariant along y. Then

$$\frac{\partial \Phi}{\partial y} = 0, \tag{4.786}$$

and the potential depends only on x and z,

$$\Phi = \Phi(x, z). \tag{4.787}$$

The name "cylindrical" refers to this optical symmetry, by analogy with a cylindrical glass lens. It should not be confused with the presence of cylindrical electrodes, which may also occur in rotationally symmetric electron lenses.

In a charge-free region, the electrostatic potential satisfies Laplace's equation,

$$\nabla^2\Phi = 0. \tag{4.788}$$

Because the potential is independent of y, this becomes

$$\frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial z^2} = 0. \tag{4.789}$$

If the lens is also symmetric under $x \to -x$, the potential near the optical axis has the expansion

$$\boxed{\Phi(x,z) = \phi(z) - \frac{1}{2}\phi''(z)x^2 + \frac{1}{24}\phi^{(4)}(z)x^4 + \cdots .} \tag{4.790}$$

Near-axis potential of an electrostatic cylindrical lens.

Because the potential is symmetric under $x \to -x$, its Taylor series contains only even powers of x:

$$\Phi(x,z) = \phi(z) + a_2(z)x^2 + a_4(z)x^4 + \cdots . \tag{4.791}$$

Substituting this expansion into Eq. (4.789) gives

$$2a_2 + 12a_4x^2 + \cdots + \phi'' + a_2''x^2 + \cdots = 0. \tag{4.792}$$

Terms containing equal powers of x must vanish independently. At zeroth order,

$$2a_2 + \phi'' = 0, \tag{4.793}$$

so

$$a_2 = -\frac{1}{2}\phi''. \tag{4.794}$$

At second order,

$$12a_4 + a_2'' = 0, \tag{4.795}$$

and therefore

$$a_4 = -\frac{a_2''}{12} = \frac{\phi^{(4)}}{24}. \tag{4.796}$$

Thus,

$$\Phi(x,z) = \phi(z) - \frac{1}{2}\phi''(z)x^2 + \frac{1}{24}\phi^{(4)}(z)x^4 + \cdots , \tag{4.797}$$

which is Eq. (4.790).

The transverse electric field follows from

$$\mathbf{E} = -\nabla\Phi. \tag{4.798}$$

To lowest order,

$$E_x = -\frac{\partial\Phi}{\partial x} \simeq \phi''(z)x, \tag{4.799}$$

$$E_y = -\frac{\partial\Phi}{\partial y} = 0. \tag{4.800}$$

Thus the electrostatic cylindrical lens exerts, to first order, a restoring or diverging force in one transverse direction but no analogous force in the orthogonal direction.

The resulting optical system therefore has two very different transverse behaviours. In the x–z plane it behaves as a genuine focusing lens; in the y–z plane there is no corresponding transverse focusing field. This is the electron-optical analogue of an ordinary cylindrical glass lens, which transforms a point into a line rather than into a stigmatic point image.

Magnetic cylindrical lenses are more complicated because the Lorentz force can couple the two transverse coordinates. In fact, unlike the electrostatic case, a general magnetic cylindrical lens is not an orthogonal system (Hawkes and Kasper, 2017, Chap. 20). That additional complexity is not required for the image-formation model developed here.

Cylindrical lenses are therefore useful mainly as a conceptual bridge: they demonstrate that there is no requirement that an electron lens focus equally in the two transverse directions. Once rotational symmetry is abandoned, anisotropic focusing becomes possible, and this idea leads naturally to quadrupoles.

4.7.3 Dipoles

The simplest transverse multipole is the *dipole*. Its principal optical function is not to focus the electron beam, but to change its mean propagation direction. Dipole fields are therefore used for beam steering, alignment, beam tilt and shift, scanning, and other operations in which the optical axis of the electron beam must be displaced or redirected.

Consider first an electrostatic dipole whose deflection direction is the x axis. Close to the optical axis, its leading potential has the form

$$\boxed{\Phi_1(x, y, z) = G_1(z)x + \cdots,} \tag{4.801}$$

where $G_1(z)$ describes the longitudinal variation of the dipole strength. This is the $\ell = 1$ member of the multipole expansion introduced later in Eq. (4.830), since

$$r_\perp \cos\varphi = x. \tag{4.802}$$

The corresponding transverse electric field is

$$\boxed{E_x = -\frac{\partial \Phi_1}{\partial x} \simeq -G_1(z), \qquad E_y \simeq 0.} \tag{4.803}$$

Thus, to lowest order, the transverse electric field does not depend on the distance of the electron from the optical axis.

This is the essential difference between a dipole and a cylindrical lens. For the cylindrical lens, Eq. (4.790) gives a transverse field proportional to position,

$$E_x \propto x, \tag{4.804}$$

so electrons on opposite sides of the axis experience forces pointing in opposite directions and the beam is focused. For a dipole, Eq. (4.803) instead gives

$$E_x \simeq \text{constant}, \tag{4.805}$$

so nearby electrons receive approximately the same transverse momentum change. The beam is therefore translated in angle rather than focused.

Why a dipole deflects the beam but does not focus it.

Let the electron have charge $q = -e$, longitudinal speed v, and relativistic momentum magnitude p. In the paraxial regime,

$$p_x \simeq p\, x', \tag{4.806}$$

where

$$x' = \frac{dx}{dz}. \tag{4.807}$$

The transverse equation of motion is

$$\frac{dp_x}{dt} = qE_x. \tag{4.808}$$

Since

$$\frac{dz}{dt} \simeq v, \tag{4.809}$$

we may write

$$\frac{dp_x}{dz} = \frac{qE_x}{v}. \tag{4.810}$$

If the electron energy is approximately constant through the deflector, p may be regarded as constant. Therefore,

$$p\,x'' = \frac{qE_x}{v}, \tag{4.811}$$

or

$$\boxed{x'' = \frac{qE_x}{pv}.} \tag{4.812}$$

Using Eq. (4.803),

$$x'' \simeq -\frac{qG_1(z)}{pv}. \tag{4.813}$$

The right-hand side is independent of x. Consequently, two neighboring rays initially separated by a small distance δx satisfy

$$\delta x'' = 0. \tag{4.814}$$

Their relative separation is therefore not driven toward or away from zero by the dipole field. The field changes the common trajectory of the rays, but not their first-order separation.

This is precisely the distinction between deflection and focusing:

$$\boxed{\begin{aligned} x'' = \text{constant} \quad &\Rightarrow \quad \text{beam deflection,} \\ x'' \propto -x \quad &\Rightarrow \quad \text{focusing.} \end{aligned}} \tag{4.815}$$

For an electrostatic dipole extending from z_1 to z_2, integration of Eq. (4.813) gives the change in propagation angle,

$$\boxed{\Delta x' = -\frac{q}{pv}\int_{z_1}^{z_2} G_1(z)\,dz.} \tag{4.816}$$

Thus the integrated transverse electric field, rather than its detailed longitudinal distribution, determines the first-order deflection angle.

Magnetic dipoles. Beam steering can also be produced magnetically. Consider a transverse magnetic field

$$\mathbf{B} = B_y(z)\,\hat{\mathbf{y}}, \tag{4.817}$$

acting on an electron whose velocity is predominantly along $\hat{\mathbf{z}}$. The Lorentz force introduced in Sec. 4.6 is

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}. \tag{4.818}$$

Since

$$\hat{\mathbf{z}} \times \hat{\mathbf{y}} = -\hat{\mathbf{x}}, \tag{4.819}$$

the transverse force is

$$F_x \simeq -qvB_y. \tag{4.820}$$

Proceeding as in Eq. (4.812),

$$\boxed{x'' = -\frac{q}{p} B_y(z).} \tag{4.821}$$

For an electron, $q = -e$, and therefore

$$x'' = \frac{e}{p} B_y(z) \tag{4.822}$$

for the orientation chosen here.

The corresponding angular deflection is

$$\boxed{\Delta x' = -\frac{q}{p} \int_{z_1}^{z_2} B_y(z)\, dz.} \tag{4.823}$$

For an approximately uniform field B_y over a length L,

$$|\Delta x'| \simeq \frac{eB_y L}{p}. \tag{4.824}$$

This expression can also be understood from the circular trajectory of a charged particle in a transverse magnetic field. Its radius of curvature is

$$\rho = \frac{p}{eB_y}, \tag{4.825}$$

so traversing a short magnet of length L changes the propagation angle by

$$\Delta\theta \simeq \frac{L}{\rho} = \frac{eB_y L}{p}, \tag{4.826}$$

in agreement with Eq. (4.824).

It is important to distinguish the magnetic scalar potential used here to classify the multipole symmetry from the vector potential $\mathbf{A}$ that enters the

electron Hamiltonian. For electrostatic elements, the potential Φ introduced above is directly the scalar potential appearing in Eq. (4.718). For magnetic elements, by contrast, the local scalar potential Ψ_m is only an auxiliary representation of the magnetic field,

$$\mathbf{B} = -\nabla\Psi_m. \tag{4.827}$$

To use this field in the Hamiltonian formalism of Sec. 4.6, one must choose a vector potential $\mathbf{A}$ satisfying

$$\boxed{\nabla \times \mathbf{A} = \mathbf{B}.} \tag{4.828}$$

The choice of $\mathbf{A}$ is not unique because of gauge freedom, but all gauge-equivalent choices produce the same magnetic field and the same physical electron trajectories and image phase.

Dipoles in an electron microscope. Dipolar fields occur repeatedly in an electron microscope. Small electrostatic or magnetic deflectors can be used to change the beam position or direction, while pairs of deflectors separated along the column allow beam shift and beam tilt to be controlled approximately independently. Scanning transmission electron microscopes use rapidly varying dipolar fields to scan the focused probe over the specimen.

Dipole fields also arise intentionally in more complex electron-optical systems, for example in magnetic prisms and energy-dispersive devices. Their detailed theory is beyond the purpose of this section, but the basic optical action remains the same: a dipole changes the central trajectory rather than producing first-order focusing about it.

This distinction can be summarized as

$$\boxed{\begin{array}{ccc} \text{dipole} & : & F_\perp = O(r_\perp^0), \\ \text{cylindrical lens or quadrupole} & : & F_\perp = O(r_\perp). \end{array}} \tag{4.829}$$

The dipole is therefore the natural $\ell = 1$ starting point of the multipole hierarchy developed in Sec. 4.7.4.

Connection with image formation. An ideal static dipole does not introduce a new focusing term into the Gaussian image-formation model. Its first-order action is instead to change the reference trajectory of the beam. If the downstream optics are perfectly aligned with this new trajectory,

the effect can often be regarded simply as a change of beam position or illumination angle.

In practice, however, steering and aberrations are coupled through the rest of the column. A beam that enters a subsequent round lens away from its intended axis or with an incorrect tilt samples the lens off-axis and may acquire aberrations such as coma. Dipole alignment elements are therefore important not only for positioning the beam but also for ensuring that the assumptions underlying the aberration function $\chi(\mathbf{R})$ remain valid.

Unlike quadrupoles, hexapoles, and octupoles, its principal purpose is not to generate or compensate focusing and aberration terms directly, but to establish the correct central trajectory through the electron-optical system.

4.7.4 The transverse multipole hierarchy

Quadrupoles, hexapoles, and octupoles can be understood most easily as members of a common multipole hierarchy. We shall denote the multipole order by ℓ in order to avoid confusion with the radial index n used later for Zernike polynomials. Quadrupoles, hexapoles, and octupoles can be understood most easily as members of a common multipole hierarchy. We shall denote the multipole order by ℓ in order to avoid confusion with the radial index n used later for Zernike polynomials. Figure 4.4 summarizes the corresponding transverse symmetries. A round lens is rotationally symmetric, whereas a cylindrical lens singles out one transverse direction. Multipoles are characterized by 2ℓ alternating poles around the optical axis.

Figure 4.4 suggests that dipoles, quadrupoles, hexapoles, and octupoles belong to a common hierarchy. The dipole discussed in Sec. 4.7.3 is the $\ell = 1$ member of this family. We now derive the general form and use it to understand the higher multipoles.

Consider first an electrostatic field in a charge-free region around the electron beam. Close to the optical axis, an angular multipole of order ℓ has a leading transverse dependence

$$\boxed{\Phi_\ell(r_\perp, \varphi, z) = C_\ell(z) r_\perp^\ell \cos\left[\ell(\varphi - \varphi_\ell)\right] + O(r_\perp^{\ell+2}),} \tag{4.830}$$

where φ_ℓ specifies the orientation of the multipole and $C_\ell(z)$ describes how its strength varies along the optical axis.

The corresponding sine dependence does not define a different kind of multipole: it is obtained by rotating the element. Thus a given multipole order ℓ has one characteristic angular periodicity, while its orientation in the transverse plane is arbitrary.

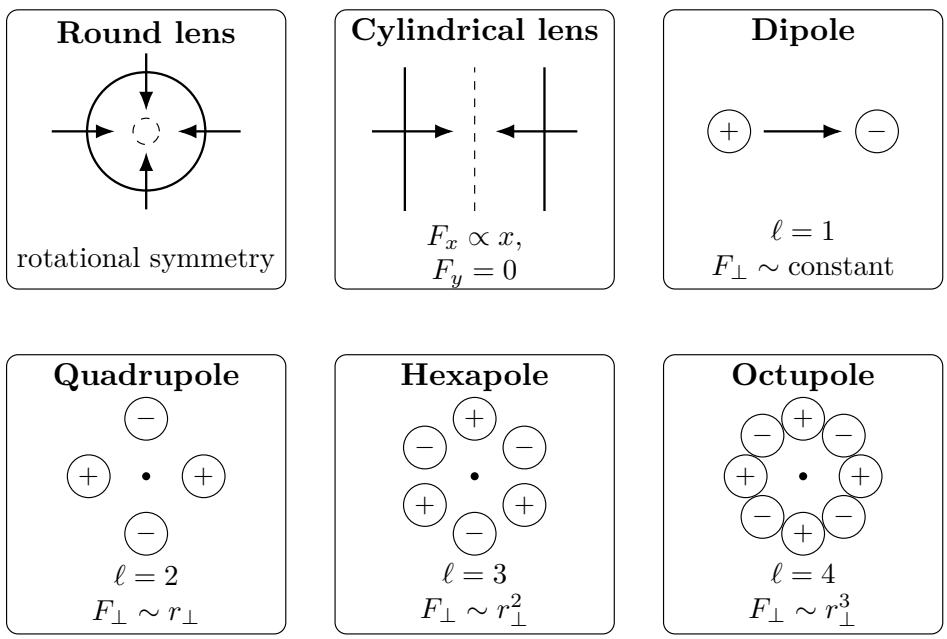


Figure 4.4: Schematic comparison of the transverse symmetries of common electron-optical elements. A round lens is rotationally symmetric and produces the same first-order focusing in every transverse direction. A cylindrical lens focuses in one transverse direction while leaving the orthogonal direction unaffected in the ideal electrostatic case. Dipoles, quadrupoles, hexapoles, and octupoles form a multipole hierarchy with 2ℓ alternating poles. Their leading transverse forces scale respectively as $r_\perp^0$, $r_\perp^1$, $r_\perp^2$, and $r_\perp^3$. The arrows indicate the qualitative first-order optical action rather than the detailed local electric or magnetic force field.

Why multipole fields contain $r_\perp^\ell \cos(\ell\varphi)$.

To determine the leading transverse dependence, consider first a region over which the longitudinal variation of the field may be neglected locally. The potential then satisfies the transverse Laplace equation

$$\nabla_\perp^2 \Phi = \frac{1}{r_\perp}\frac{\partial}{\partial r_\perp}\left(r_\perp \frac{\partial \Phi}{\partial r_\perp}\right) + \frac{1}{r_\perp^2}\frac{\partial^2 \Phi}{\partial \varphi^2} = 0. \tag{4.831}$$

Seek a separated solution

$$\Phi(r_\perp, \varphi) = F(r_\perp)\Theta(\varphi). \tag{4.832}$$

Substitution into Eq. (4.831) gives

$$\frac{r_\perp^2 F'' + r_\perp F'}{F} = -\frac{\Theta''}{\Theta}. \tag{4.833}$$

The left-hand side depends only on $r_\perp$, whereas the right-hand side depends only on φ. Therefore both must be equal to a constant. We write this constant as ℓ^2,

$$\frac{r_\perp^2 F'' + r_\perp F'}{F} = -\frac{\Theta''}{\Theta} = \ell^2. \tag{4.834}$$

The angular equation is then

$$\Theta'' + \ell^2 \Theta = 0, \tag{4.835}$$

whose solutions are

$$\Theta(\varphi) = A\cos(\ell\varphi) + B\sin(\ell\varphi). \tag{4.836}$$

Because the potential must be single-valued when $\varphi \to \varphi + 2\pi$, the number ℓ must be an integer.

The radial equation is

$$r_\perp^2 F'' + r_\perp F' - \ell^2 F = 0, \tag{4.837}$$

with solutions

$$F(r_\perp) = C r_\perp^\ell + D r_\perp^{-\ell}. \tag{4.838}$$

The term $r_\perp^{-\ell}$ diverges on the optical axis and is therefore excluded from a regular field in the beam region. Hence,

$$F(r_\perp) \propto r_\perp^{\ell}. \tag{4.839}$$

The most general regular angular dependence can therefore be written as

$$\Phi_\ell \propto r_\perp^{\ell} \left[A\cos(\ell\varphi) + B\sin(\ell\varphi)\right]. \tag{4.840}$$

The two coefficients may be combined into an amplitude and an orientation,

$$A\cos(\ell\varphi) + B\sin(\ell\varphi) = C\cos[\ell(\varphi - \varphi_\ell)], \tag{4.841}$$

which gives the angular dependence used in Eq. (4.830).

In a real three-dimensional element, the strength $C_\ell(z)$ varies along the optical axis. The complete three-dimensional Laplace equation then generates additional terms of orders

$$r_\perp^{\ell+2}, \quad r_\perp^{\ell+4}, \quad \ldots, \tag{4.842}$$

associated particularly with the entrance and exit fringe fields. Equation (4.830) represents only the leading near-axis term.

A convenient Cartesian representation follows from

$$r_\perp^{\ell} e^{j\ell\varphi} = (x + jy)^{\ell}. \tag{4.843}$$

Taking the real part for the orientation $\varphi_\ell = 0$ gives

$$\ell = 1: \qquad r_\perp \cos\varphi = x, \tag{4.844}$$

$$\ell = 2: \qquad r_\perp^2 \cos 2\varphi = x^2 - y^2, \tag{4.845}$$

$$\ell = 3: \qquad r_\perp^3 \cos 3\varphi = x^3 - 3xy^2, \tag{4.846}$$

$$\ell = 4: \qquad r_\perp^4 \cos 4\varphi = x^4 - 6x^2y^2 + y^4. \tag{4.847}$$

The first of these is precisely the dipole potential introduced in Eq. (4.801). The remaining three give the leading transverse structures of quadrupoles, hexapoles, and octupoles.

The terminology follows from the number of alternating poles surrounding the beam:

$$\boxed{\text{number of poles} = 2\ell.} \tag{4.848}$$

Thus

ℓ	name	number of poles
1	dipole	2
2	quadrupole	4
3	hexapole	6
4	octupole	8

(4.849)

The term *sextupole* is also commonly used for the six-pole $\ell = 3$ element.

The transverse force follows from the gradient of the potential. For a particle of charge q,

$$\mathbf{F}_\perp = -q\nabla_\perp \Phi_\ell. \tag{4.850}$$

Since Φ_ℓ is, to leading order, a homogeneous polynomial of degree ℓ, differentiation lowers the degree by one:

$$\boxed{\|\mathbf{F}_\perp\| \propto r_\perp^{\ell-1}.} \tag{4.851}$$

This reproduces the dipole result of Sec. 4.7.3 and gives the hierarchy

$$\boxed{\begin{array}{ccc} \text{dipole} & : & F_\perp = O(r_\perp^0), \\ \text{quadrupole} & : & F_\perp = O(r_\perp), \\ \text{hexapole} & : & F_\perp = O(r_\perp^2), \\ \text{octupole} & : & F_\perp = O(r_\perp^3). \end{array}} \tag{4.852}$$

The distinction between these force orders is more important for electron optics than the number of physical poles itself. A dipole produces an approximately common transverse force and therefore mainly steers the beam, as shown in Eq. (4.815). A quadrupole produces a force linear in displacement and therefore modifies the first-order, Gaussian trajectories. Hexapoles and octupoles begin at quadratic and cubic order, respectively, and therefore naturally act on nonlinear properties of the ray system.

This is why quadrupoles can participate directly in focusing and first-order transport, whereas hexapoles and octupoles are especially useful for aberration correction.

A cylindrical lens is not a quadrupole. At first sight, the quadratic dependence of a cylindrical lens in Eq. (4.790) might suggest that it should simply be regarded as a quadrupole. This is not the case.

A pure transverse quadrupole has the leading harmonic

$$\Phi_2 \propto x^2 - y^2. \tag{4.853}$$

The opposite curvatures in x and y are required by the transverse Laplace equation:

$$\frac{\partial^2 \Phi_2}{\partial x^2} + \frac{\partial^2 \Phi_2}{\partial y^2} = 0. \tag{4.854}$$

By contrast, an electrostatic cylindrical lens has

$$\Phi(x, z) = \phi(z) - \frac{1}{2}\phi''(z)x^2 + \cdots , \tag{4.855}$$

with no dependence on y. Its quadratic transverse curvature is balanced not by an opposite curvature in y, but by the longitudinal variation of the axial potential:

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial z^2} = 0, \tag{4.856}$$

as derived in Eq. (4.789).

Thus,

$$\boxed{\begin{array}{ccl} \text{quadrupole} & : & \text{transverse harmonic } x^2 - y^2, \\ \text{cylindrical lens} & : & \text{one-dimensional transverse curvature coupled to } z. \end{array}} \tag{4.857}$$

Both can produce linear transverse forces, but their symmetries and three-dimensional field structures are different.

Magnetic multipoles. The same angular classification applies to magnetic multipoles. In a current-free region around the electron beam,

$$\nabla \times \mathbf{B} = 0, \tag{4.858}$$

so locally one may introduce a magnetic scalar potential Ψ_m such that

$$\mathbf{B} = -\nabla \Psi_m. \tag{4.859}$$

Because

$$\nabla \cdot \mathbf{B} = 0, \tag{4.860}$$

the scalar potential satisfies

$$\nabla^2 \Psi_m = 0. \tag{4.861}$$

It therefore admits the same leading transverse angular harmonics

$$\Psi_{m,\ell} \propto r_\perp^\ell \cos[\ell(\varphi - \varphi_\ell)]. \tag{4.862}$$

This scalar potential is useful for classifying the symmetry of a magnetic multipole, but it should not be confused with the vector potential that enters the electron dynamics. The trajectory and phase of the electron remain governed by the electromagnetic Hamiltonian derived in Sec. 4.6, and in particular by the vector potential $\mathbf{A}$ appearing in Eq. (4.776).

4.7.5 Quadrupoles: focusing one direction while defocusing the other

Choose an electrostatic quadrupole whose principal axes coincide with x and y. Its leading potential may be written

$$\boxed{\Phi_2(x, y, z) = \frac{G_2(z)}{2}\left(x^2 - y^2\right).} \tag{4.863}$$

The transverse electric field is

$$E_x = -\frac{\partial \Phi_2}{\partial x} = -G_2(z)x, \tag{4.864}$$

$$E_y = -\frac{\partial \Phi_2}{\partial y} = +G_2(z)y. \tag{4.865}$$

The two components therefore have opposite signs. For an electron, whose charge is $-e$, the corresponding forces are also opposite in the two principal transverse directions.

After choosing the sign of the quadrupole strength so that the x direction is focusing, the paraxial equations have the generic form

$$x'' + \kappa_q(z)x = 0, \tag{4.866}$$

$$y'' - \kappa_q(z)y = 0, \tag{4.867}$$

where $\kappa_q(z) > 0$ represents the local quadrupole strength in an appropriate electron-optical normalization.

Equations (4.866) and (4.867) display the defining property of a quadrupole: a pure quadrupole focuses trajectories in one transverse plane while defocusing them in the perpendicular plane (Hawkes and Kasper, 2017, Chap. 19). Rotating the quadrupole by 90° interchanges the two roles, while a rotation by 45° generates the complementary orientation proportional to $2xy$.

A single quadrupole therefore does not usually form a stigmatic image of a point. The x and y coordinates obey different equations and generally have different focal planes. One obtains two orthogonal line foci rather than a unique point focus.

This does not make quadrupoles undesirable. On the contrary, their anisotropy is precisely what makes them useful. Several quadrupoles can be combined into a *multiplet* in such a way that the convergent action of one element compensates the divergent action of another. The resulting system can have a desired net first-order transformation even though none of its individual components is rotationally symmetric.

Quadrupolar fields are also the natural means of correcting twofold astigmatism. In practical microscopes, a stigmator is often constructed from several independently driven pole pieces—frequently an eight-pole assembly—from which a quadrupolar field of arbitrary strength and orientation can be synthesized. The number of physical pole pieces in such an adjustable device should therefore not be confused with the multipole order of the field that it is being driven to produce.

4.7.6 Hexapoles: nonlinear control with six-pole symmetry

The leading electrostatic potential of a hexapole oriented along the x axis can be written

$$\boxed{\Phi_3(x,y,z) = \frac{G_3(z)}{3}\left(x^3 - 3xy^2\right).} \tag{4.868}$$

The corresponding field is

$$E_x = -G_3(z)\left(x^2 - y^2\right), \tag{4.869}$$

$$E_y = 2G_3(z)xy. \tag{4.870}$$

Thus the transverse force is quadratic in the displacement from the axis.

A hexapole is therefore fundamentally different from a quadrupole. Very close to the axis,

$$\left.\frac{\partial \mathbf{F}_\perp}{\partial \mathbf{r}_\perp}\right|_{\mathbf{r}_\perp=0} = 0, \tag{4.871}$$

so a pure hexapole has no ordinary first-order focusing power about the central trajectory. Its main action begins at nonlinear order.

This property makes hexapoles especially valuable in aberration correctors. A suitably designed sequence of round transfer lenses and hexapoles can be arranged so that unwanted lower-order effects generated by one hexapole are cancelled by another, while the compound system produces the desired correction to the third-order geometrical aberrations of the objective lens. Successful hexapole correctors were central to the practical correction of spherical aberration in TEM (Haider et al., 1998; Hawkes, 2015).

The important point for the present chapter is not the detailed geometry of such a corrector, but the principle:

$$\boxed{\text{a nonlinear multipole} + \text{appropriate transfer optics} \longrightarrow \text{controlled net aberration.}} \tag{4.872}$$

Consequently, one should not think of a hexapole as simply "adding a hexapolar aberration" to the final image. The effect of the complete corrector depends on where the hexapoles are placed, the trajectories entering them, their relative orientations and strengths, and the lenses between them.

4.7.7 Octupoles: cubic transverse forces

For an octupole, the leading potential can similarly be written

$$\boxed{\Phi_4(x, y, z) = \frac{G_4(z)}{4}\left(x^4 - 6x^2y^2 + y^4\right).} \tag{4.873}$$

Differentiation gives

$$E_x = -G_4(z)\left(x^3 - 3xy^2\right), \tag{4.874}$$

$$E_y = G_4(z)\left(3x^2y - y^3\right). \tag{4.875}$$

The transverse force is now cubic in the ray displacement:

$$F_\perp = O(r_\perp^3). \tag{4.876}$$

Octupoles can therefore act directly on third-order geometrical deviations of the ray trajectories.

A particularly important corrector architecture combines quadrupoles and octupoles. The quadrupoles control the first-order, anisotropic transport of the beam and establish suitable intermediate ray configurations, whereas octupoles provide the nonlinear corrections. Quadrupole–octupole systems played a central role in the development of aberration-corrected STEM instruments (Krivanek et al., 1999; Hawkes, 2015).

Again, the full corrector must be treated as a compound optical system. The action of an octupole cannot in general be inferred simply from the angular appearance of its isolated field.

4.7.8 Multipole order is not aberration order

At this point it is important to distinguish two expansions that will both appear in electron optics.

A *field multipole expansion* describes the physical electromagnetic field generated by electrodes or pole pieces. Its order ℓ refers to the angular harmonic of the field,

$$\Phi_\ell \propto r_\perp^\ell \cos(\ell\varphi), \tag{4.877}$$

and hence to the number 2ℓ of alternating poles.

An *aberration expansion*, by contrast, describes the accumulated departure of the wavefront from the ideal reference wave after the electron has traversed the complete optical system. Its variables are typically pupil coordinates or, equivalently, propagation angles or spatial frequencies.

There is therefore no one-to-one identification

$$\boxed{\text{multipole symmetry} \not\equiv \text{final aberration symmetry.}} \tag{4.878}$$

For example, a hexapole field possesses the angular harmonic $\cos 3\varphi$, but a *pair* of suitably arranged hexapoles can be used to correct the rotationally symmetric spherical aberration of a round objective lens. Likewise, an octupole is not merely a device for creating a fourfold aberration.

The reason is that the final wave aberration depends on the complete trajectory through the column. The transverse coordinates at one multipole depend on all preceding lenses; a subsequent multipole acts on those transformed coordinates; and the resulting contributions can cancel or reinforce one another. Corrector design exploits precisely this composition.

This distinction will become especially important when the final wave-aberration function is expanded in Zernike polynomials. The multipole index ℓ used in the present section describes a *physical field*. The Zernike indices used later describe a *function over the objective pupil*. They should not be identified.

4.7.9 How these elements enter the image-formation model

We can now connect the physical electron-optical elements to the main objective of this chapter.

A real microscope column contains many field regions. Schematically, the electron traverses a sequence

$$\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N, \tag{4.879}$$

separated by propagation regions. Each element changes the trajectories and, equivalently, the phase accumulated by the electron wave.

At first order in position and angle, the combined system defines the Gaussian optics: focal lengths, magnification, image planes, diffraction planes,

and rotations. This first-order description will be developed explicitly in the paraxial treatment later in this chapter.

Terms beyond first order generate departures from ideal Gaussian imaging. After propagation through the whole column, these departures can be expressed as a wave-aberration function or, equivalently, as a phase function

$$\chi(\mathbf{R}). \tag{4.880}$$

The coherent wave-transfer function of the complete objective system is then written in the form

$$\boxed{H_{\mathrm{coh}}(\mathbf{R}) = P(\mathbf{R}) \exp\left[-j\chi(\mathbf{R})\right],} \tag{4.881}$$

where $P(\mathbf{R})$ represents the pupil or aperture.

Equation (4.881) is the crucial reduction. The image-formation model does not normally contain a separate "quadrupole transfer function", "hexapole transfer function", and "octupole transfer function". Instead, all deterministic optical elements contribute to the *net* aberration phase $\chi(\mathbf{R})$.

A corrector works by changing the coefficients entering this net phase. For example, if an uncorrected objective contributes a positive spherical-aberration coefficient C_s, the corrector is adjusted to generate an opposite contribution so that the residual coefficient

$$C_s^{\mathrm{res}} = C_s^{\mathrm{obj}} + C_s^{\mathrm{corr}} \tag{4.882}$$

is much smaller than either contribution individually.

The same principle applies to other aberrations:

$$\boxed{a_k^{\mathrm{res}} = \sum_j a_{k,j},} \tag{4.883}$$

where $a_{k,j}$ denotes the contribution of the jth optical element, including the effects induced by its position within the complete system.

This also clarifies the distinction between deterministic aberrations and coherence envelopes. A static quadrupole, hexapole, or octupole changes the phase $\chi(\mathbf{R})$; it does not by itself generate a statistical envelope. An envelope appears only when the relevant optical condition fluctuates across the ensemble of detected electrons—for example because of energy spread, source extent, field instability, or specimen motion.

Thus the complicated hardware of the electron microscope is compressed twice. First,

$$\boxed{\{\text{round lenses, multipoles, alignments}\} \longrightarrow \{\text{Gaussian optics}, \chi(\mathbf{R})\},} \tag{4.884}$$

and then the weak-object image-formation theory reduces these quantities to the contrast transfer function.

4.7.10 Summary

The main elements discussed in this section can be summarized as follows.

Element	Transverse dependence	Force order	Principal optical role
Round lens	rotationally symmetric	linear	Gaussian focusing, magnification, and formation of image and diffraction planes
Cylindrical lens	x^2 in one transverse direction	linear	One-dimensional focusing and formation of line images
Dipole	x	constant	Beam deflection, steering, and alignment
Quadrupole	$x^2 - y^2$	linear	Anisotropic focusing, astigmatism control, and first-order transport in multipole correctors
Hexapole	$x^3 - 3xy^2$	quadratic	Nonlinear aberration correction; particularly important in spherical-aberration correctors
Octupole	$x^4 - 6x^2y^2 + y^4$	cubic	Third-order ray correction, particularly in quadrupole–octupole correctors

For the remainder of this chapter, we shall not need to follow the detailed trajectory through every one of these physical elements. Their purpose has been to establish the physical origin of the electron-optical transformations that will subsequently be represented by Gaussian propagation and by the aberration phase $\chi(\mathbf{R})$.

The central message is therefore

$$\boxed{\begin{array}{c}\text{many physical electron-optical elements}\\ \Downarrow\\ \text{one effective Gaussian imaging system}\\ +\\ \text{one residual aberration function}\\ \Downarrow\\ \mathrm{CTF}(\mathbf{R}).\end{array}} \tag{4.885}$$

This reduction is what allows the detailed physics of an electron microscope column to be incorporated into a tractable model of cryo-EM image formation.

4.8 Wave mechanics in an electromagnetic field

In Sec. 3.4, we showed that every component of a free Dirac spinor satisfies the scalar Klein–Gordon equation, Eq. (3.130), and that a stationary free electron therefore satisfies the Helmholtz equation Eq. (3.159), with

$$k = \frac{p}{\hbar} \tag{4.886}$$

according to Eq. (3.163).

That analysis applied to a free electron. We have since developed the classical dynamics of an electron moving in electric and magnetic fields. In particular, Sec. 4.6 showed that the exact relativistic Hamiltonian is

$$H = c\sqrt{(m_0c)^2 + \|\mathbf{P} + e\mathbf{A}\|^2} - m_0c^2 - e\Phi, \tag{4.887}$$

as given in Eq. (4.718).

We now reconnect this classical Hamiltonian description with wave mechanics and obtain the scalar wave equation for an electron propagating through the electromagnetic fields of the microscope.

4.8.1 The relativistic scalar wave equation

There is a small notational point that must first be addressed. The Hamiltonian H of Eq. (4.718) was defined with the rest energy m_0c^2 subtracted.

For the connection with the relativistic wave equation, it is more convenient to introduce the corresponding rest-energy-inclusive Hamiltonian

$$\boxed{\mathcal{H} \triangleq H + m_0 c^2.} \tag{4.888}$$

From Eq. (4.718),

$$\boxed{\mathcal{H} = c\sqrt{(m_0 c)^2 + \|\mathbf{P} + e\mathbf{A}\|^2} - e\Phi.} \tag{4.889}$$

The corresponding spin-independent relativistic wave equation is

$$\boxed{\left[\left(j\hbar\frac{\partial}{\partial t} + e\Phi\right)^2 - c^2\left(-j\hbar\nabla + e\mathbf{A}\right)^2 - m_0^2 c^4\right]\Psi(\mathbf{r}, t) = 0.} \tag{4.890}$$

This is the electromagnetic generalization of the free scalar equation derived previously in Eq. (3.130). The derivation below shows how it follows directly from the relativistic Hamiltonian.

From the relativistic Hamiltonian to the scalar wave equation.

Starting from Eq. (4.889),

$$\mathcal{H} = c\sqrt{(m_0 c)^2 + \|\mathbf{P} + e\mathbf{A}\|^2} - e\Phi, \tag{4.891}$$

we first move the electrostatic term to the left:

$$\mathcal{H} + e\Phi = c\sqrt{(m_0 c)^2 + \|\mathbf{P} + e\mathbf{A}\|^2}. \tag{4.892}$$

Squaring both sides gives

$$(\mathcal{H} + e\Phi)^2 = m_0^2 c^4 + c^2\|\mathbf{P} + e\mathbf{A}\|^2. \tag{4.893}$$

Thus,

$$(\mathcal{H} + e\Phi)^2 - c^2\|\mathbf{P} + e\mathbf{A}\|^2 - m_0^2 c^4 = 0. \tag{4.894}$$

We now introduce the quantum operators corresponding to the canonical energy and canonical momentum,

$$\boxed{\mathcal{H} \longrightarrow j\hbar\frac{\partial}{\partial t}, \qquad \mathbf{P} \longrightarrow -j\hbar\nabla.} \tag{4.895}$$

Applying the resulting operator equation to the electron wave function $\Psi(\mathbf{r}, t)$ transforms Eq. (4.894) into

$$\left[\left(j\hbar\frac{\partial}{\partial t} + e\Phi\right)^2 - c^2\left(-j\hbar\nabla + e\mathbf{A}\right)^2 - m_0^2c^4\right]\Psi = 0, \tag{4.896}$$

which is Eq. (4.890).

The two combinations

$$j\hbar\frac{\partial}{\partial t} + e\Phi \tag{4.897}$$

and

$$-j\hbar\nabla + e\mathbf{A} \tag{4.898}$$

are the wave-mechanical counterparts of the mechanical energy and mechanical momentum, respectively. In particular, Eq. (4.697) gave

$$\mathbf{p} = \mathbf{P} + e\mathbf{A}. \tag{4.899}$$

If $\Phi = \mathbf{A} = \mathbf{0}$, Eq. (4.890) reduces immediately to the free Klein–Gordon equation Eq. (3.130), thereby recovering the result of Sec. 3.4.

It is important, however, to state what has been omitted. For an electron, the complete relativistic theory is the Dirac theory introduced in Sec. 3.4. In electromagnetic fields, eliminating the spinor structure of the Dirac equation generates additional spin-dependent interactions. Equation (4.890) retains the relativistic orbital dynamics but neglects these spin degrees of freedom. This is the scalar approximation that we shall use for electron-optical image formation.

4.8.2 Stationary electromagnetic fields

Electron-optical fields will normally be treated as static,

$$\frac{\partial\Phi}{\partial t} = 0, \qquad \frac{\partial\mathbf{A}}{\partial t} = \mathbf{0}. \tag{4.900}$$

The Hamiltonian is then conserved, as established in Eq. (4.458).

Let

$$H = E_0. \tag{4.901}$$

Since $\mathcal{H} = H + m_0c^2$ according to Eq. (4.888), the conserved total energy including the rest energy is

$$\boxed{\mathcal{E}_0 \triangleq E_0 + m_0c^2.} \tag{4.902}$$

The stationary-state separation already derived in Sec. 3.4.1 therefore takes the form

$$\boxed{\Psi(\mathbf{r},t) = \psi(\mathbf{r}) \exp\left(-j\frac{\mathcal{E}_0}{\hbar}t\right).} \tag{4.903}$$

Substitution into the relativistic scalar equation gives the time-independent electron wave equation

$$\boxed{\left(-j\hbar\nabla + e\mathbf{A}\right)^2 \psi = p^2(\mathbf{r})\psi,} \tag{4.904}$$

where $p(\mathbf{r})$ is the local magnitude of the relativistic mechanical momentum.

Stationary relativistic electron wave equation.

From Eq. (4.903),

$$j\hbar\frac{\partial\Psi}{\partial t} = \mathcal{E}_0\Psi. \tag{4.905}$$

Since the electromagnetic potentials are independent of time, Eq. (4.890) becomes

$$\left[\left(\mathcal{E}_0 + e\Phi\right)^2 - c^2\left(-j\hbar\nabla + e\mathbf{A}\right)^2 - m_0^2c^4\right]\psi = 0. \tag{4.906}$$

Rearranging,

$$\left(-j\hbar\nabla + e\mathbf{A}\right)^2 \psi = \frac{\left(\mathcal{E}_0 + e\Phi\right)^2 - m_0^2c^4}{c^2}\psi. \tag{4.907}$$

The quantity multiplying ψ on the right is precisely the squared mechanical momentum. Indeed, the total relativistic energy at position $\mathbf{r}$ is

$$E_{\text{tot}}(\mathbf{r}) = \mathcal{E}_0 + e\Phi(\mathbf{r}), \tag{4.908}$$

and the relativistic energy–momentum relation Eq. (2.68) gives

$$E_{\text{tot}}^2 = p^2c^2 + m_0^2c^4. \tag{4.909}$$

Hence,

$$p^2(\mathbf{r}) = \frac{\left(\mathcal{E}_0 + e\Phi\right)^2 - m_0^2c^4}{c^2}. \tag{4.910}$$

Substitution into Eq. (4.907) gives Eq. (4.904).

The same mechanical momentum was expressed previously in terms of the relativistic acceleration potential. From Eq. (2.78),

$$\boxed{p^2(\mathbf{r}) = 2m_0 e\widehat{\Phi}(\mathbf{r}).} \tag{4.911}$$

Consequently, Eq. (4.904) may also be written as

$$\boxed{\left[(-j\hbar\nabla + e\mathbf{A})^2 - 2m_0 e\widehat{\Phi}\right]\psi = 0.} \tag{4.912}$$

This equation has a particularly important connection with the classical Hamilton–Jacobi equation derived in Eq. (4.776),

$$\|\nabla W + e\mathbf{A}\|^2 = 2m_0 e\widehat{\Phi}. \tag{4.913}$$

The two equations have the same underlying momentum structure:

$$\boxed{\nabla W \quad \longleftrightarrow \quad -j\hbar\nabla.} \tag{4.914}$$

The Hamilton–Jacobi equation describes the semiclassical phase, while Eq. (4.912) retains the wave character of the electron.

4.8.3 Electrostatic and field-free regions

The connection with the Helmholtz equation becomes transparent in a region where the vector potential can be taken to vanish,

$$\mathbf{A} = \mathbf{0}. \tag{4.915}$$

Equation (4.912) then reduces to

$$\boxed{\nabla^2\psi + \frac{2m_0 e\widehat{\Phi}(\mathbf{r})}{\hbar^2}\psi = 0.} \tag{4.916}$$

This has the Helmholtz form

$$\nabla^2\psi + k^2(\mathbf{r})\psi = 0 \tag{4.917}$$

with local wavenumber

$$\boxed{k(\mathbf{r}) = \frac{\sqrt{2m_0 e\widehat{\Phi}(\mathbf{r})}}{\hbar} = \frac{p(\mathbf{r})}{\hbar}.} \tag{4.918}$$

Reduction to the local Helmholtz equation.

Setting $\mathbf{A} = \mathbf{0}$ in Eq. (4.912) gives

$$\left[(-j\hbar\nabla)^2 - 2m_0 e\widehat{\Phi}\right]\psi = 0. \tag{4.919}$$

Since

$$(-j\hbar\nabla)^2 = -\hbar^2\nabla^2, \tag{4.920}$$

we obtain

$$-\hbar^2\nabla^2\psi - 2m_0 e\widehat{\Phi}\psi = 0. \tag{4.921}$$

Dividing by $-\hbar^2$,

$$\nabla^2\psi + \frac{2m_0 e\widehat{\Phi}}{\hbar^2}\psi = 0, \tag{4.922}$$

which is Eq. (4.916).

From Eq. (4.911),

$$p^2 = 2m_0 e\widehat{\Phi}, \tag{4.923}$$

so

$$k^2 = \frac{p^2}{\hbar^2}, \tag{4.924}$$

and therefore

$$k = \frac{p}{\hbar}. \tag{4.925}$$

Thus the de Broglie relation established for a free electron in Eq. (3.163) remains locally valid:

$$\boxed{\lambda(\mathbf{r}) = \frac{h}{p(\mathbf{r})}.} \tag{4.926}$$

The wavelength therefore changes while an electron is being accelerated through a spatially varying electrostatic potential. Once the electron enters a region of constant potential, p, k, and λ become constant, and Eq. (4.916) reduces to the homogeneous Helmholtz equation Eq. (3.159).

The explicit relativistic wavelength as a function of accelerating voltage, including its values at the standard microscope voltages, was already derived in Sec. 2.4 and will not be repeated here.

4.8.4 The effect of a magnetic vector potential

Inside a magnetic lens the vector potential cannot, in general, be discarded. The full stationary equation Eq. (4.904) must then be retained:

$$(-j\hbar\nabla + e\mathbf{A})^2\,\psi = p^2(\mathbf{r})\psi. \tag{4.927}$$

The operator on the left is not simply a Laplacian plus a position-dependent scalar term. Its explicit expansion is

$$\boxed{\begin{aligned}(-j\hbar\nabla + e\mathbf{A})^2\,\psi = &-\hbar^2\nabla^2\psi \\ &- je\hbar\left[2\mathbf{A}\cdot\nabla\psi + (\nabla\cdot\mathbf{A})\psi\right] \\ &+ e^2A^2\psi.\end{aligned}} \tag{4.928}$$

Expansion of the magnetic momentum operator.

Let

$$\widehat{\boldsymbol{\pi}} \triangleq -j\hbar\nabla + e\mathbf{A}. \tag{4.929}$$

Then

$$\begin{aligned}\widehat{\boldsymbol{\pi}}^2\psi &= (-j\hbar\nabla + e\mathbf{A})\cdot(-j\hbar\nabla + e\mathbf{A})\,\psi \qquad (4.930)\\ &= (-j\hbar\nabla)^2\psi + (-j\hbar\nabla)\cdot(e\mathbf{A}\psi) \\ &\quad + e\mathbf{A}\cdot(-j\hbar\nabla\psi) + e^2A^2\psi. \qquad (4.931)\end{aligned}$$

The first term is

$$(-j\hbar\nabla)^2\psi = -\hbar^2\nabla^2\psi. \tag{4.932}$$

For the second term we use the product rule,

$$\nabla\cdot(\mathbf{A}\psi) = (\nabla\cdot\mathbf{A})\psi + \mathbf{A}\cdot\nabla\psi, \tag{4.933}$$

which gives

$$(-j\hbar\nabla)\cdot(e\mathbf{A}\psi) = -je\hbar\left[(\nabla\cdot\mathbf{A})\psi + \mathbf{A}\cdot\nabla\psi\right]. \tag{4.934}$$

The third term is

$$e\mathbf{A} \cdot (-j\hbar\nabla\psi) = -je\hbar\mathbf{A} \cdot \nabla\psi. \tag{4.935}$$

Adding the four contributions produces Eq. (4.928).

The magnetic vector potential therefore modifies the spatial phase evolution of the electron wave through terms containing both $\mathbf{A}$ and spatial derivatives of ψ.

This has a direct counterpart in the classical Hamiltonian description. The characteristic function determines the canonical momentum through Eq. (4.523),

$$\mathbf{P} = \nabla W, \tag{4.936}$$

whereas for an electron the canonical and mechanical momenta are related by Eq. (4.697),

$$\mathbf{P} = \mathbf{p} - e\mathbf{A}. \tag{4.937}$$

At the same time, the mechanical momentum is tangent to the physical trajectory,

$$\mathbf{p} = p\mathbf{t}, \tag{4.938}$$

as established in Eq. (4.746).

Consequently, when $\mathbf{A} \neq \mathbf{0}$, the gradient of the characteristic function, $\nabla W = \mathbf{P}$, is not in general parallel to the mechanical momentum $\mathbf{p}$ and hence is not necessarily parallel to the electron trajectory. The appearance of $\mathbf{A}$ in Eq. (4.928) is the wave-mechanical counterpart of this distinction between canonical and mechanical momentum.

4.8.5 Gauge transformations and the phase of the electron wave

When the electromagnetic potentials were introduced in Sec. 4.6.1, we noted that they are not uniquely determined by the physical fields. We now make this freedom explicit.

Let $\chi(\mathbf{r}, t)$ be an arbitrary sufficiently smooth scalar function. We may define a new pair of potentials by

$$\boxed{\begin{aligned} \mathbf{A}' &= \mathbf{A} + \nabla\chi, \\ \Phi' &= \Phi - \frac{\partial\chi}{\partial t}. \end{aligned}} \tag{4.939}$$

This transformation is called a *gauge transformation*. Although the potentials themselves change, the electric and magnetic fields defined in Eq. (4.676) remain unchanged.

Gauge invariance of the electromagnetic fields.

For the magnetic field,

$$\mathbf{B}' = \nabla \times \mathbf{A}' \tag{4.940}$$
$$= \nabla \times (\mathbf{A} + \nabla\chi) \tag{4.941}$$
$$= \nabla \times \mathbf{A} + \nabla \times \nabla\chi. \tag{4.942}$$

The curl of a gradient is identically zero,

$$\nabla \times \nabla\chi = \mathbf{0}, \tag{4.943}$$

and therefore

$$\boxed{\mathbf{B}' = \mathbf{B}.} \tag{4.944}$$

For the electric field, using Eq. (4.676),

$$\mathbf{E}' = -\nabla\Phi' - \frac{\partial \mathbf{A}'}{\partial t} \tag{4.945}$$
$$= -\nabla\left(\Phi - \frac{\partial\chi}{\partial t}\right) - \frac{\partial}{\partial t}(\mathbf{A} + \nabla\chi) \tag{4.946}$$
$$= -\nabla\Phi + \nabla\left(\frac{\partial\chi}{\partial t}\right) - \frac{\partial \mathbf{A}}{\partial t} - \frac{\partial}{\partial t}(\nabla\chi). \tag{4.947}$$

For a sufficiently smooth function χ, spatial and temporal derivatives commute,

$$\nabla\left(\frac{\partial\chi}{\partial t}\right) = \frac{\partial}{\partial t}(\nabla\chi), \tag{4.948}$$

so these two terms cancel. Hence,

$$\boxed{\mathbf{E}' = -\nabla\Phi - \frac{\partial \mathbf{A}}{\partial t} = \mathbf{E}.} \tag{4.949}$$

Thus different potential pairs related by Eq. (4.939) describe exactly the same electromagnetic fields.

Wave mechanics introduces one additional element: the electron wave

function must simultaneously acquire the phase

$$\boxed{\Psi'(\mathbf{r},t) = \exp\left[-j\frac{e}{\hbar}\chi(\mathbf{r},t)\right]\Psi(\mathbf{r},t).} \tag{4.950}$$

With this transformation, the combinations appearing in Eq. (4.890) transform covariantly, and the wave equation therefore describes exactly the same physics in every gauge.

Gauge covariance of the mechanical-momentum operator.

Consider first the spatial operator. Using Eq. (4.950),

$$\Psi' = e^{-je\chi/\hbar}\Psi. \tag{4.951}$$

Its gradient is

$$\nabla\Psi' = \nabla\left(e^{-je\chi/\hbar}\Psi\right) \tag{4.952}$$

$$= e^{-je\chi/\hbar}\left[\nabla\Psi - j\frac{e}{\hbar}(\nabla\chi)\Psi\right]. \tag{4.953}$$

Therefore,

$$\begin{aligned}\left(-j\hbar\nabla + e\mathbf{A}'\right)\Psi' = e^{-je\chi/\hbar}\left[-j\hbar\nabla\Psi - e(\nabla\chi)\Psi\right.\\ \left.+e\left(\mathbf{A}+\nabla\chi\right)\Psi\right].\end{aligned} \tag{4.954}$$

The terms containing $\nabla\chi$ cancel, leaving

$$\boxed{\left(-j\hbar\nabla + e\mathbf{A}'\right)\Psi' = e^{-je\chi/\hbar}\left(-j\hbar\nabla + e\mathbf{A}\right)\Psi.} \tag{4.955}$$

A completely analogous calculation for the temporal operator gives

$$\boxed{\left(j\hbar\frac{\partial}{\partial t} + e\Phi'\right)\Psi' = e^{-je\chi/\hbar}\left(j\hbar\frac{\partial}{\partial t} + e\Phi\right)\Psi.} \tag{4.956}$$

Both operators therefore acquire only the same overall phase factor under a gauge transformation. Substitution into Eq. (4.890) shows that the complete wave equation retains exactly the same form.

The vector potential and the phase of the electron wave are therefore not independent gauge-invariant quantities. A change of gauge modifies both,

while all observable predictions remain unchanged.

This result is particularly important in electron wave optics. Although the electromagnetic fields determine the classical force through Eq. (2.48), the phase accumulated by an electron wave is naturally expressed in terms of the electromagnetic potentials. The physical consequences of such phase accumulation will appear later when we discuss electron interference and phase-modifying elements.

4.9 The eikonal approximation

In Sec. 4.4.14, we showed that geometrical optics emerges from classical wave optics in the short-wavelength limit. More recently, in Sec. 4.5.2, we obtained the same connection for a free electron: the leading-order phase of the electron wave satisfies the classical Hamilton–Jacobi equation.

We now extend this result to an electron propagating through the static electric and magnetic fields of the electron microscope. This extension is important because, in a magnetic field, the phase gradient is related to the canonical momentum, whereas the physical trajectory is determined by the mechanical momentum.

The appropriate starting point is the stationary scalar relativistic electron equation derived in Eq. (4.912),

$$\left[(-j\hbar\nabla + e\mathbf{A})^2 - 2m_0 e\widehat{\Phi}\right]\psi = 0. \tag{4.957}$$

Using Eq. (4.911),

$$p^2(\mathbf{r}) = 2m_0 e\widehat{\Phi}(\mathbf{r}), \tag{4.958}$$

we may write this more compactly as

$$\boxed{(-j\hbar\nabla + e\mathbf{A})^2\,\psi = p^2(\mathbf{r})\psi.} \tag{4.959}$$

As discussed in Sec. 4.8, this is a scalar, spin-independent relativistic description of the orbital electron dynamics. It is not the complete Dirac equation in an electromagnetic field, since the latter also contains spin-dependent interactions.

4.9.1 Exact separation of amplitude and phase

Away from zeros of the wave function, we may write the stationary electron wave exactly in polar form,

$$\boxed{\psi(\mathbf{r}) = a(\mathbf{r})\exp\left[\frac{j}{\hbar}W_{\mathrm{q}}(\mathbf{r})\right],} \tag{4.960}$$

where $a(\mathbf{r}) \geq 0$ is real and $W_{\mathrm{q}}(\mathbf{r})$ is the quantum phase expressed in units of action. We use a rather than A for the amplitude in order to avoid confusion with the magnetic vector potential $\mathbf{A}$.

The notation W_{q} is deliberate. At this stage it is simply defined by the phase of the quantum wave and should not yet be identified with Hamilton's classical characteristic function W. The emergence of the classical W is precisely what the eikonal approximation will show.

It is useful to introduce the gauge-invariant phase-momentum field

$$\boxed{\boldsymbol{\pi}_{\mathrm{q}} \triangleq \nabla W_{\mathrm{q}} + e\mathbf{A}.} \tag{4.961}$$

Substitution of Eq. (4.960) into Eq. (4.959) gives two exact real equations:

$$\boxed{\|\boldsymbol{\pi}_{\mathrm{q}}\|^2 - \hbar^2 \frac{\nabla^2 a}{a} = p^2(\mathbf{r}),} \tag{4.962}$$

and

$$\boxed{\nabla \cdot \left(a^2 \boldsymbol{\pi}_{\mathrm{q}}\right) = 0.} \tag{4.963}$$

The first equation determines the phase, including a correction that depends on the spatial curvature of the amplitude. The second is a transport equation describing conservation of the stationary electron flux.

Amplitude–phase decomposition in an electromagnetic field.

Starting from Eq. (4.960),

$$\psi = a e^{jW_{\mathrm{q}}/\hbar}, \tag{4.964}$$

we first apply the covariant momentum operator

$$-j\hbar\nabla + e\mathbf{A}. \tag{4.965}$$

The gradient of ψ is

$$\nabla\psi = \left[\nabla a + \frac{j}{\hbar} a \nabla W_{\mathrm{q}}\right] e^{jW_{\mathrm{q}}/\hbar}. \tag{4.966}$$

Therefore,

$$(-j\hbar\nabla + e\mathbf{A})\,\psi = \Big[-j\hbar\nabla a + a\nabla W_{\mathrm{q}} + ae\mathbf{A}\Big]e^{jW_{\mathrm{q}}/\hbar} \tag{4.967}$$

$$= [a\boldsymbol{\pi}_{\mathrm{q}} - j\hbar\nabla a]\,e^{jW_{\mathrm{q}}/\hbar}, \tag{4.968}$$

where Eq. (4.961) has been used.

Applying the operator a second time,

$$(-j\hbar\nabla + e\mathbf{A})^2\,\psi = -j\hbar\nabla\cdot\Big[(a\boldsymbol{\pi}_{\mathrm{q}} - j\hbar\nabla a)\,e^{jW_{\mathrm{q}}/\hbar}\Big] + e\mathbf{A}\cdot[a\boldsymbol{\pi}_{\mathrm{q}} - j\hbar\nabla a]\,e^{jW_{\mathrm{q}}/\hbar}. \tag{4.969}$$

Using the product rule and collecting the terms proportional to the same phase factor gives

$$(-j\hbar\nabla + e\mathbf{A})^2\,\psi = \Big[a\,\|\boldsymbol{\pi}_{\mathrm{q}}\|^2 - \hbar^2\nabla^2 a - j\hbar\,(2\nabla a\cdot\boldsymbol{\pi}_{\mathrm{q}} + a\nabla\cdot\boldsymbol{\pi}_{\mathrm{q}})\Big]e^{jW_{\mathrm{q}}/\hbar}. \tag{4.970}$$

Substitution into Eq. (4.959) gives

$$a\,\|\boldsymbol{\pi}_{\mathrm{q}}\|^2 - \hbar^2\nabla^2 a - j\hbar\,(2\nabla a\cdot\boldsymbol{\pi}_{\mathrm{q}} + a\nabla\cdot\boldsymbol{\pi}_{\mathrm{q}}) = p^2 a. \tag{4.971}$$

Since a, W_{q}, $\mathbf{A}$, and p are real, the real and imaginary parts must vanish independently.

The real part gives

$$a\,\|\boldsymbol{\pi}_{\mathrm{q}}\|^2 - \hbar^2\nabla^2 a = p^2 a. \tag{4.972}$$

Dividing by a,

$$\|\boldsymbol{\pi}_{\mathrm{q}}\|^2 - \hbar^2\frac{\nabla^2 a}{a} = p^2, \tag{4.973}$$

which is Eq. (4.962).

The imaginary part gives

$$2\nabla a\cdot\boldsymbol{\pi}_{\mathrm{q}} + a\nabla\cdot\boldsymbol{\pi}_{\mathrm{q}} = 0. \tag{4.974}$$

Multiplying by a,

$$2a\nabla a \cdot \boldsymbol{\pi}_{\mathrm{q}} + a^2 \nabla \cdot \boldsymbol{\pi}_{\mathrm{q}} = 0. \tag{4.975}$$

Since

$$\nabla(a^2) = 2a\nabla a, \tag{4.976}$$

the product rule for the divergence gives

$$\nabla \cdot \left(a^2 \boldsymbol{\pi}_{\mathrm{q}}\right) = 0, \tag{4.977}$$

which is Eq. (4.963).

Equation (4.963),

$$\nabla \cdot \left(a^2 \boldsymbol{\pi}_{\mathrm{q}}\right) = 0, \tag{4.978}$$

has a simple physical interpretation. The quantity $a^2 = |\psi|^2$ measures the local strength of the electron wave, while $\boldsymbol{\pi}_{\mathrm{q}}$ gives the local direction and magnitude associated with its phase propagation. Their product therefore describes a local *flow* of the electron wave.

The divergence measures whether this flow accumulates or disappears from a small region of space. Thus,

$$\nabla \cdot \left(a^2 \boldsymbol{\pi}_{\mathrm{q}}\right) = 0 \tag{4.979}$$

states that, in a stationary situation, there is no net accumulation of electron-wave intensity: whatever flow enters a small volume must also leave it. This is why Eq. (4.963) is called a *stationary flux-conservation equation*.

The idea is closely analogous to the transport equation obtained for ordinary waves in Eq. (4.598). If a bundle of trajectories spreads out, the amplitude must decrease; if the bundle becomes more concentrated, the amplitude increases. In the non-relativistic Schrödinger description, the same mathematical statement is usually written as conservation of the *probability current*: $|\psi|^2$ is then interpreted as the probability density for detecting the electron, and the current describes how that probability is transported through space.

The other new term,

$$-\hbar^2 \frac{\nabla^2 a}{a}, \tag{4.980}$$

appearing in Eq. (4.962), describes something different. It measures how strongly the spatial variation of the amplitude feeds back into the equation governing the phase. The Laplacian $\nabla^2 a$ measures the local curvature of the

amplitude: it is small when the amplitude changes gently over space and becomes important when the amplitude bends or varies rapidly over short distances.

For this reason, we shall refer to

$$-\hbar^2 \frac{\nabla^2 a}{a} \tag{4.981}$$

as the *quantum amplitude-curvature correction*. It is the term that distinguishes the exact wave equation from the classical Hamilton–Jacobi equation. If this correction is negligible, the phase can evolve approximately independently of the detailed local variation of the amplitude, and the classical trajectory description emerges.

In the usual non-relativistic Schrödinger formulation, the same effect is commonly expressed in units of energy by dividing by $2m_0$:

$$\boxed{Q = -\frac{\hbar^2}{2m_0}\frac{\nabla^2 a}{a}.} \tag{4.982}$$

The quantity Q is traditionally called the *quantum potential*. Despite its name, it is not an additional external potential such as an electric potential. Rather, it is generated by the spatial shape of the wave itself: when the amplitude varies strongly, Q becomes important and the motion cannot be described accurately by the classical Hamilton–Jacobi equation alone.

4.9.2 The eikonal limit

The eikonal approximation consists of assuming that the amplitude-curvature correction in Eq. (4.962) is small compared with the squared mechanical momentum:

$$\boxed{\left|\hbar^2 \frac{\nabla^2 a}{a}\right| \ll p^2.} \tag{4.983}$$

Under this condition, Eq. (4.962) reduces to

$$\boxed{\|\nabla W_{\mathrm{q}} + e\mathbf{A}\|^2 \simeq p^2.} \tag{4.984}$$

Using Eq. (4.911),

$$\boxed{\|\nabla W_{\mathrm{q}} + e\mathbf{A}\|^2 \simeq 2m_0 e\widehat{\Phi}.} \tag{4.985}$$

But the classical Hamilton–Jacobi equation derived in Eq. (4.776) is

$$\|\nabla W + e\mathbf{A}\|^2 = 2m_0 e\widehat{\Phi}. \tag{4.986}$$

Thus the quantum phase action W_{q} and Hamilton's classical characteristic function W obey the same equation to leading order:

$$\boxed{W_{\mathrm{q}} \simeq W \qquad \text{in the eikonal regime.}} \tag{4.987}$$

This is the electromagnetic-field generalization of the free-electron result obtained in Sec. 4.5.2. The eikonal approximation and the WKB approximation are therefore two manifestations of the same short-wavelength asymptotic principle.

4.9.3 Meaning of the approximation

The local de Broglie wavenumber associated with the magnitude of the mechanical momentum is

$$k = \frac{p}{\hbar}. \tag{4.988}$$

Consequently, Eq. (4.983) may be written as

$$\boxed{\left|\frac{\nabla^2 a}{a}\right| \ll k^2.} \tag{4.989}$$

Suppose, for example, that the characteristic distance over which the amplitude changes appreciably is L_a. Dimensionally,

$$\left|\frac{\nabla^2 a}{a}\right| \sim \frac{1}{L_a^2}. \tag{4.990}$$

Equation (4.989) requires

$$kL_a \gg 1, \tag{4.991}$$

where L_a is the characteristic distance over which the amplitude changes appreciably.

The magnitude of this condition becomes particularly clear for the electrons used in a modern microscope. At an accelerating voltage of 300 kV, the relativistic electron wavelength derived in Sec. 2.4 is approximately

$$\lambda \simeq 1.97 \text{ pm}. \tag{4.992}$$

The corresponding wavenumber is therefore

$$k = \frac{2\pi}{\lambda} \simeq 3.19 \times 10^{12}\ \mathrm{m}^{-1}, \tag{4.993}$$

and hence

$$\frac{1}{k} \simeq 3.13 \times 10^{-13}\ \mathrm{m} = 0.313\ \mathrm{pm}. \tag{4.994}$$

Thus, even if the amplitude changed appreciably over a distance as short as

$$L_a = 1\ \mathrm{nm}, \tag{4.995}$$

we would have

$$kL_a \simeq 3.2 \times 10^3 \gg 1. \tag{4.996}$$

Geometrically, the same idea can be understood in terms of the phase fronts. An arbitrary smooth phase front may be curved on a large scale, but over a sufficiently small region it can be approximated by a plane, as illustrated in Fig. 4.5. Within such a region, the phase gradient is approximately constant, so the wave behaves locally like a plane wave with a well-defined propagation direction.

There is, however, an important scale requirement. The local region must be small enough that the wavefront curvature, amplitude, and electromagnetic fields change only slightly, but at the same time large enough to contain many electron wavelengths. Schematically, if L_{local} denotes the size of such a region, we require

$$\lambda \ll L_{\mathrm{local}} \ll L_{\mathrm{variation}}, \tag{4.997}$$

where $L_{\mathrm{variation}}$ represents the characteristic scale over which the amplitude, phase direction, or electromagnetic fields change appreciably.

For a 300 kV electron, for which $\lambda \simeq 1.97$ pm as discussed in Sec. 2.4, this separation of scales is enormous inside the macroscopic lenses of the microscope. A local region only a few nanometres across already contains thousands of radians of phase evolution, while the electromagnetic fields of the lenses normally vary over much larger distances.

If we estimate

$$\left| \frac{\nabla^2 a}{a} \right| \sim \frac{1}{L_a^2}, \tag{4.998}$$

then the relative magnitude of the term neglected in the eikonal approximation is approximately

$$\frac{1}{(kL_a)^2}. \tag{4.999}$$

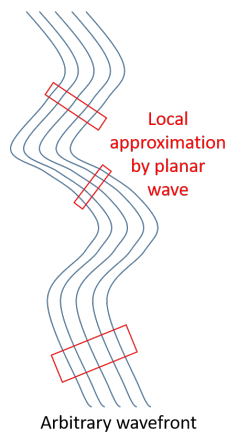


Figure 4.5: Local plane-wave approximation to a smooth, curved wavefront. Although the wavefront may vary substantially over large distances, within a sufficiently small region its neighbouring phase fronts are approximately parallel planes and the phase gradient is approximately constant. For the eikonal approximation to be valid, this local region must nevertheless remain large compared with the wavelength, while the amplitude and electromagnetic fields vary only weakly across it.

For $L_a = 1$ nm this is only

$$\frac{1}{(kL_a)^2} \simeq 10^{-7}. \tag{4.1000}$$

For variations occurring over the much larger distances characteristic of the electromagnetic lenses of the microscope, the approximation is even more extreme. For example,

L_a	kL_a	$1/(kL_a)^2$
0.1 nm	3.2×10^2	9.8×10^{-6}
1 nm	3.2×10^3	9.8×10^{-8}
1 μm	3.2×10^6	9.8×10^{-14}
1 mm	3.2×10^9	9.8×10^{-20}

(4.1001)

The enormous separation of scales explains why a 300 kV electron can normally be treated as following a classical trajectory through the macroscopic electromagnetic lenses of the microscope. The electron wavelength is of order picometres, whereas the fields produced by the lenses typically vary over micrometre-to-millimetre distances.

Notice, however, that the condition $kL_a \gg 1$ tests only the slow-variation assumption underlying the eikonal expansion. The approximation may still fail near caustics, sharp discontinuities, or in situations involving several interfering or strongly scattered wave components, even when the electron wavelength itself is extremely small.

This is the electron counterpart of the short-wavelength condition derived for ordinary waves in Sec. 4.4.14. In practice, the electromagnetic fields and the corresponding mechanical momentum must also vary sufficiently smoothly that a single local phase branch remains a meaningful description.

Once the eikonal equation has determined the phase, the remaining amplitude is governed by Eq. (4.963). In the eikonal limit, Eq. (4.775) identifies

$$\nabla W + e\mathbf{A} = \mathbf{p}, \tag{4.1002}$$

and hence the transport equation becomes

$$\boxed{\nabla \cdot \left(a^2 \mathbf{p}\right) = 0.} \tag{4.1003}$$

Thus the two equations have complementary roles: the Hamilton–Jacobi equation (Eq. 4.776) determines the electron trajectories, while the transport equation (Eq. 4.1003) determines how the wave amplitude changes as bundles of trajectories expand or contract.

4.9.4 Phase fronts and electron trajectories

There is an important difference between electrostatic and magnetic electron optics.

Surfaces of constant quantum phase are, in the eikonal limit, surfaces

$$W(\mathbf{r}) = \text{constant}. \tag{4.1004}$$

Their normal vector is

$$\nabla W. \tag{4.1005}$$

Hamilton–Jacobi theory identifies this vector with the canonical momentum,

$$\nabla W = \mathbf{P}, \tag{4.1006}$$

whereas the physical trajectory is tangent to the mechanical momentum, as established in Eq. (4.746),

$$\mathbf{p} = p\mathbf{t}. \tag{4.1007}$$

For an electron, Eq. (4.775) gives

$$\mathbf{p} = \nabla W + e\mathbf{A}. \tag{4.1008}$$

Therefore, in an electrostatic or field-free region in which $\mathbf{A}$ can be taken to vanish,

$$\nabla W = \mathbf{p}, \tag{4.1009}$$

and the electron trajectories are orthogonal to the phase fronts, exactly as light rays are orthogonal to optical wavefronts in an isotropic medium.

In a magnetic field, however,

$$\nabla W = \mathbf{p} - e\mathbf{A}, \tag{4.1010}$$

so the phase-front normal is not, in general, parallel to the physical electron trajectory. The eikonal approximation still produces the classical trajectories, but the geometrical statement that trajectories are perpendicular to surfaces of constant W no longer holds in general.

4.9.5 Simple free-space examples

The eikonal approximation is exact for several elementary free-space solutions.

For a plane wave,

$$\psi(\mathbf{r}) = a_0 \exp\left(\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}\right), \tag{4.1011}$$

the amplitude is constant,

$$\nabla^2 a_0 = 0, \tag{4.1012}$$

so the amplitude-curvature correction vanishes identically.

A spherical outgoing wave centred at $\mathbf{r}_0$ has the form

$$\psi(\mathbf{r}) = \frac{a_0}{R}\exp\left(\frac{jpR}{\hbar}\right), \qquad R = \|\mathbf{r} - \mathbf{r}_0\|. \tag{4.1013}$$

For $R \neq 0$,

$$\nabla^2\left(\frac{1}{R}\right) = 0, \tag{4.1014}$$

so the amplitude-curvature term again vanishes everywhere except at the singular source point.

The spherical example also shows that local planarity of the wavefront is a useful geometrical picture, but it is not by itself the mathematical definition of the eikonal approximation. As shown in Sec. 4.3.3, a distant spherical

wavefront becomes locally planar when its radius of curvature is large compared with the observed region. The actual eikonal condition, Eq. (4.983), concerns the magnitude of the amplitude-curvature correction relative to the local momentum.

4.9.6 Where the eikonal approximation fails

The eikonal description becomes unreliable whenever a single, slowly-varying amplitude and phase cease to provide a good local description of the wave. Important examples include:

- *Sharp edges and apertures*, where abrupt truncation of the wave generates diffraction and large spatial derivatives of the amplitude.
- *Interfaces or rapidly varying fields*, especially when they generate several reflected, refracted, or scattered branches that cannot be represented by a single local action W.
- *Caustics and focal regions.* A caustic is a region where a family of neighbouring classical trajectories becomes strongly concentrated. A focal point is the simplest example: rays that started from different nearby positions or directions converge toward the same point. More generally, the trajectories need not all meet at one point; they may form an envelope, that is, a curve or surface to which many neighbouring rays become tangent. Such an envelope is called a *caustic*.

  In a ray description, the wave amplitude increases when neighbouring trajectories are squeezed into a smaller cross-sectional area. At an ideal caustic, geometrical optics predicts that this area tends to zero and therefore that the amplitude becomes infinite. This divergence is not physical; it signals the breakdown of the ray approximation. The actual wave remains finite because diffraction and interference between the converging contributions become important. Near a caustic, the ordinary eikonal or WKB approximation must therefore be replaced by a full-wave description or by a more refined asymptotic approximation.
- *Turning points*, where the local mechanical momentum approaches zero. Since the right-hand side of Eq. (4.983) then becomes small, the semiclassical expansion necessarily fails.
- *Strong or rapidly varying specimen scattering*, where large-angle scattering or interference between multiple paths may require an explicitly wave-mechanical treatment.

For propagation through the macroscopic electromagnetic lenses of an electron microscope, the characteristic variation scales of the fields are enormously larger than the electron wavelength derived in Sec. 2.4. The eikonal approximation therefore provides an excellent basis for describing electron trajectories through the microscope column, except in special regions such as caustics.

At the specimen, the situation is more subtle. Ray tracing alone is insufficient for image formation because the recorded contrast depends on coherent scattering, diffraction, and interference. This does not mean that every eikonal approximation necessarily fails inside the specimen: for high-energy electrons and sufficiently forward scattering, eikonal or phase-object approximations can remain useful. Rather, the important point is that the complete image-formation problem must retain the wave phase and its interference, rather than replacing the electron wave solely by classical trajectories.

4.10 Paraxial approximation

The eikonal approximation of Sec. 4.9 and the paraxial approximation answer two different questions.

The *eikonal approximation* asks whether the electron wavelength is sufficiently small compared with the spatial scale over which the wave amplitude and the electromagnetic fields vary. When this is true, the phase of the electron wave obeys the classical Hamilton–Jacobi equation and a ray description becomes meaningful.

The *paraxial approximation* makes an additional geometrical assumption: the rays of interest remain close to the optical axis and make only small angles with it. If

$$\mathbf{r}(z) = \big(x(z), y(z), z\big), \tag{4.1015}$$

this means

$$\boxed{|x'| \ll 1, \qquad |y'| \ll 1,} \tag{4.1016}$$

as already introduced in Eq. (4.103). Consequently,

$$\tan\theta_x \simeq \theta_x \simeq x', \qquad \tan\theta_y \simeq \theta_y \simeq y'. \tag{4.1017}$$

Thus, the eikonal approximation allows us to speak of trajectories, whereas the paraxial approximation makes the equations governing those trajectories approximately *linear*. This linear regime is usually called *Gaussian optics*.

4.10.1 How small are paraxial angles in cryo-EM?

For electron microscopy, the paraxial condition is considerably less restrictive than it may first appear. In the angular-spectrum description of Sec. 3.6, a transverse spatial frequency R corresponds, at small angles, to

$$\theta \simeq \lambda R, \tag{4.1018}$$

as stated in Eq. (3.624).

For a 300 kV electron, Sec. 2.4 gives approximately

$$\lambda \simeq 1.97 \text{ pm}. \tag{4.1019}$$

A spatial period

$$d = 2 \text{ Å} \tag{4.1020}$$

corresponds to

$$R = \frac{1}{d}, \tag{4.1021}$$

and hence

$$\theta \simeq \frac{\lambda}{d} \simeq 9.9 \text{ mrad} \simeq 0.56^\circ. \tag{4.1022}$$

Even a spatial period of 1 Å corresponds to only

$$\theta \simeq 19.7 \text{ mrad} \simeq 1.13^\circ. \tag{4.1023}$$

Thus, the range of scattering angles that contains most of the spatial frequencies relevant to cryo-EM image formation is still small on the scale of one radian. This is why paraxial wave optics remains an excellent starting point even when discussing atomic-resolution images.

4.10.2 Why the paraxial approximation produces Gaussian optics

The origin of Gaussian optics can be understood directly from the electron characteristic function.

In Sec. 4.6.4, the exact reduced action was written as

$$W = \int \overline{M}\left(x, y, z, x', y'\right) dz, \tag{4.1024}$$

with the exact z-parameterized electron-optical Lagrangian given in Eq. (4.761).

Near the optical axis, both the trajectory and the electromagnetic fields can be expanded in powers of the transverse coordinates and slopes. For

a rotationally symmetric system, the transverse expansion has the general structure

$$\boxed{\overline{M} = \overline{M}^{(0)} + \overline{M}^{(2)} + \overline{M}^{(4)} + \cdots,} \tag{4.1025}$$

where the superscript denotes the total order in x, y, x', and y'.

The zeroth-order term describes propagation along the optical axis. The quadratic term is the first one that describes transverse focusing and propagation. Retaining only $\overline{M}^{(0)} + \overline{M}^{(2)}$ is the Gaussian, or first-order, approximation.

The physical reason for the resulting linearity is simple: differentiating a quadratic function produces something linear. Therefore, the Euler–Lagrange equations generated by the quadratic action are linear in the transverse coordinates and slopes.

Why a quadratic action gives linear ray equations.

Suppose that, in one transverse coordinate, the quadratic part of the electron-optical Lagrangian has the generic form

$$\overline{M}^{(2)} = \frac{1}{2}a(z)x'^2 + b(z)xx' + \frac{1}{2}c(z)x^2. \tag{4.1026}$$

Then

$$\frac{\partial \overline{M}^{(2)}}{\partial x'} = a(z)x' + b(z)x, \tag{4.1027}$$

and

$$\frac{\partial \overline{M}^{(2)}}{\partial x} = b(z)x' + c(z)x. \tag{4.1028}$$

Taking the total derivative of the first expression with respect to z,

$$\frac{d}{dz}\left(\frac{\partial \overline{M}^{(2)}}{\partial x'}\right) = \frac{d}{dz}\left(ax' + bx\right) \tag{4.1029}$$

$$= ax'' + a'x' + b'x + bx'. \tag{4.1030}$$

The Euler–Lagrange equation,

$$\frac{d}{dz}\left(\frac{\partial \overline{M}^{(2)}}{\partial x'}\right) - \frac{\partial \overline{M}^{(2)}}{\partial x} = 0, \tag{4.1031}$$

therefore becomes

$$ax'' + a'x' + b'x + bx' - (bx' + cx) = 0 \tag{4.1032}$$

$$ax'' + a'x' + (b' - c)\,x = 0. \tag{4.1033}$$

Thus,

$$\boxed{a(z)x'' + a'(z)x' + \left[b'(z) - c(z)\right] x = 0.} \tag{4.1034}$$

Provided $a(z) \neq 0$, this can equivalently be written as

$$\boxed{x'' + \frac{a'}{a}x' + \frac{b' - c}{a}x = 0.} \tag{4.1035}$$

Every occurrence of x, x', and x'' is to the first power. The equation is therefore a linear second-order differential equation.

The same reasoning applies independently to $y(z)$ in a rotationally symmetric system. Thus, it is not an accidental algebraic simplification that Gaussian electron optics is linear: it follows directly from truncating the electron-optical action at quadratic order.

The physical meaning is worth emphasizing. Near the optical axis, the electron experiences only the leading variation of the lens field. To this order, the restoring effect of the lens is proportional to the transverse displacement itself, just as the force of a harmonic oscillator is proportional to its displacement. Consequently, moving a ray twice as far from the axis produces twice the first-order response, and two paraxial trajectories may be added to generate another trajectory. This superposition property is precisely what makes Gaussian optics possible.

Higher-order terms in the action destroy this exact proportionality. For example, quartic terms generate cubic terms in the ray equations, so rays at different heights or angles are no longer related by simple scaling. Those nonlinear corrections are the origin of geometrical aberrations.

This observation also anticipates the origin of aberrations. Terms of fourth and higher order in Eq. (4.1025) make the ray equations nonlinear and cause different rays to depart from the ideal Gaussian mapping. We shall later interpret these departures as lens aberrations.

4.10.3 The paraxial equation of a round magnetic lens

The detailed trajectory calculation for a rotationally symmetric magnetic lens was already carried out in Sec. 4.2, so we do not repeat it here.

After removal of the Larmor rotation described in Sec. 4.2.2, the transverse coordinate $w(z)$ obeys

$$\boxed{w'' + \frac{\kappa^2(z)}{4} w = 0,} \tag{4.1036}$$

as derived in Eq. (4.167), with

$$\kappa(z) = \frac{eB_0(z)}{p}. \tag{4.1037}$$

This equation contains the essential physics of Gaussian magnetic electron optics. The quantity w measures transverse displacement from the optical axis in the rotating Larmor frame, while $\kappa^2(z)/4$ measures the local focusing strength of the magnetic field. Since the equation is linear, doubling the initial displacement or angle doubles the entire resulting trajectory, and arbitrary paraxial trajectories can be constructed by superposition.

The full magnetic-field distribution determines the detailed shape of the trajectories. In the thin-lens limit, its net focusing action can instead be summarized by a single focal length, derived previously in Eq. (4.247). This provides the physical connection between the real magnetic field and the ideal thin-lens model introduced in Sec. 3.8.

4.10.4 Two fundamental rays describe the Gaussian system

Because the paraxial trajectory equation is linear and second order, its complete behaviour can be described by only two independent solutions (Hawkes and Kasper, 2017)[Chaps. 15–17].

Choose an input plane $z = z_i$ and define two fundamental solutions, $g_p(z)$ and $h_p(z)$, by

$$\boxed{\begin{array}{rclrcl} g_p(z_i) & = & 1, & g_p'(z_i) & = & 0, \\ h_p(z_i) & = & 0, & h_p'(z_i) & = & 1. \end{array}} \tag{4.1038}$$

Their physical meanings are particularly simple (see Fig. 4.6):

- g_p describes a ray that starts at unit transverse displacement and parallel to the optical axis;

- h_p describes a ray that starts on the optical axis with unit transverse slope.

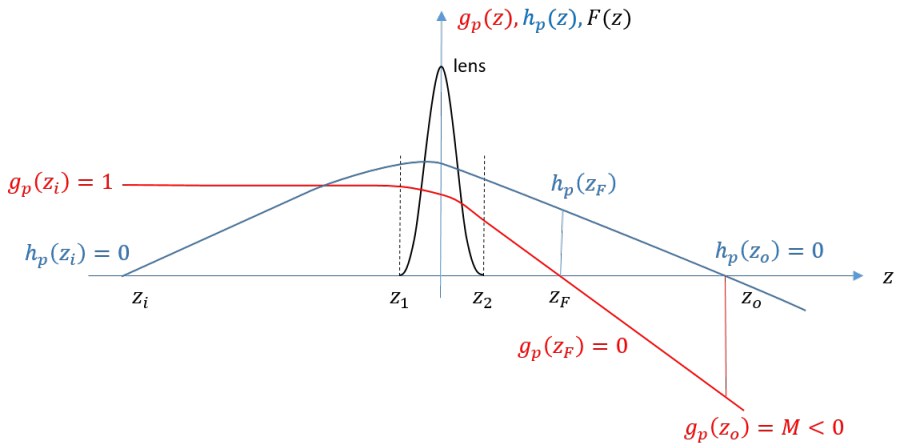


Figure 4.6: Fundamental paraxial rays $g_p(z)$ and $h_p(z)$. The solution g_p starts at unit transverse displacement and zero slope, whereas h_p starts on the optical axis with unit slope. Because the paraxial ray equation is linear, any ray is a linear combination of these two solutions. At the image plane z_o, $h_p(z_o) = 0$, so the final position is independent of the initial ray angle and $g_p(z_o) = M$ gives the transverse magnification. At the angular or Fourier plane z_F, $g_p(z_F) = 0$, so position in that plane depends only on the initial ray angle. The lens field is localized between z_1 and z_2.

Every paraxial ray is a linear combination of these two:

$$\boxed{x(z) = g_p(z)x_i + h_p(z)x_i'.} \tag{4.1039}$$

An identical relation holds for the other transverse coordinate in the Larmor frame.

Fundamental rays and the ray-transfer matrix.

Since the paraxial equation is a homogeneous second-order linear differential equation, any solution can be written as

$$x(z) = Ag_p(z) + Bh_p(z). \tag{4.1040}$$

At $z = z_i$, the boundary conditions of Eq. (4.1038) give

$$x_i = A, \qquad x_i' = B. \tag{4.1041}$$

Hence

$$x(z) = g_p(z)x_i + h_p(z)x_i'. \tag{4.1042}$$

Differentiating,

$$x'(z) = g_p'(z)x_i + h_p'(z)x_i'. \tag{4.1043}$$

Therefore,

$$\boxed{\begin{pmatrix} x(z) \\ x'(z) \end{pmatrix} = \begin{pmatrix} g_p(z) & h_p(z) \\ g_p'(z) & h_p'(z) \end{pmatrix} \begin{pmatrix} x_i \\ x_i' \end{pmatrix}.} \tag{4.1044}$$

For the magnetic-lens equation Eq. (4.167), the Wronskian

$$g_p h_p' - g_p' h_p \tag{4.1045}$$

is constant. Evaluating it at z_i using Eq. (4.1038) gives

$$\boxed{g_p h_p' - g_p' h_p = 1.} \tag{4.1046}$$

Thus the matrix in Eq. (4.1044) has unit determinant. It is precisely the ray-transfer, or ABCD, matrix introduced independently in Sec. 3.10.3, with

$$A = g_p, \qquad B = h_p, \qquad C = g_p', \qquad D = h_p'. \tag{4.1047}$$

The functions g_p and h_p should therefore not be regarded merely as a mathematical device for solving a differential equation. They answer two very physical questions: how does the microscope propagate an initial transverse *position*, and how does it propagate an initial *direction*?

4.10.5 Image planes and angular planes

Equation (4.1039) immediately reveals the two types of planes that are most important in an electron microscope.

Suppose first that, at some plane $z = z_o$,

$$\boxed{h_p(z_o) = 0.} \tag{4.1048}$$

Then

$$x_o = g_p(z_o)x_i, \tag{4.1049}$$

independently of the initial slope x_i'.

Thus, all paraxial rays leaving the same point x_i arrive at the same point x_o, even though they may have left the object in different directions. The two planes are therefore conjugate image planes. Their transverse magnification is

$$\boxed{M = g_p(z_o).} \tag{4.1050}$$

This is exactly the condition $B = 0$ obtained from the ABCD formulation in Eq. (3.669).

Consider instead a plane $z = z_F$ for which

$$\boxed{g_p(z_F) = 0.} \tag{4.1051}$$

Then

$$x_F = h_p(z_F)x_i'. \tag{4.1052}$$

The final position no longer depends on where the ray started; it depends only on its initial direction.

Consequently, all rays that leave the input plane with the same direction are brought to the same transverse point. This is an *angular plane*. In wave optics it is a Fourier plane, because propagation direction is equivalent to transverse spatial frequency.

The same result was obtained in Sec. 3.10.3: $A = 0$ identifies an angular/Fourier plane according to Eq. (3.669). Hawkes and Kasper express the same geometrical property by noting that parallel rays from the specimen plane meet at the corresponding diffraction plane (Hawkes and Kasper, 2017)[Chap. 17].

Using Eq. (3.624),

$$x_i' \simeq \theta_x \simeq \lambda R_x, \tag{4.1053}$$

and therefore

$$x_F \simeq \lambda h_p(z_F) R_x. \tag{4.1054}$$

For the familiar ideal thin-lens back focal plane, $h_p(z_F) = f$, and we recover

$$\boxed{x_F = \lambda f R_x, \qquad y_F = \lambda f R_y,} \tag{4.1055}$$

in agreement with Eq. (3.658).

The back focal plane therefore gives a physical position to every transverse spatial frequency of the electron wave.

4.10.6 The same approximation in wave optics

It is important not to confuse the paraxial approximation with replacing the electron wave by rays. Paraxial *wave* optics still retains the complex wave amplitude and therefore still describes diffraction and interference.

The characteristic mathematical feature of paraxial wave optics is the appearance of *quadratic phases*. This is not a mathematical trick. It follows directly from the geometry of propagation close to the optical axis.

The physical origin of the quadratic phase can be understood directly from the geometry of a spherical wave. Consider a point source located on the optical axis and an observation point in a plane a distance z downstream, as illustrated in Fig. 4.7. Let

$$\boldsymbol{\rho} = (x, y) \tag{4.1056}$$

denote the transverse displacement of the observation point from the optical axis. The actual propagation distance is then

$$\boxed{L(\boldsymbol{\rho}) = \sqrt{z^2 + \|\boldsymbol{\rho}\|^2}.} \tag{4.1057}$$

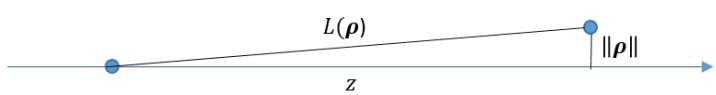


Figure 4.7: Geometry used to derive the paraxial propagation phase. A wave originates from a point on the optical axis and reaches an observation point whose transverse displacement is $\|\boldsymbol{\rho}\|$ in a plane located an axial distance z downstream. The actual propagation distance is $L = \sqrt{z^2 + \|\boldsymbol{\rho}\|^2}$. The drawing represents a two-dimensional cross-section of the full three-dimensional geometry, where $\boldsymbol{\rho} = (x, y)$.

For paraxial propagation, the transverse displacement is much smaller than the axial propagation distance,

$$\|\boldsymbol{\rho}\| \ll z. \tag{4.1058}$$

Geometrically, this means that the propagation direction makes only a small angle with the optical axis. Under this condition, Eq. (4.1057) can be expanded to give

$$\boxed{L(\boldsymbol{\rho}) \simeq z + \frac{\|\boldsymbol{\rho}\|^2}{2z}.} \tag{4.1059}$$

Thus, the first correction to the axial path length is quadratic in the transverse displacement.

Since the phase accumulated by a free electron is

$$\frac{S}{\hbar} = \frac{pL}{\hbar} = kL, \tag{4.1060}$$

using Eq. (3.163), the corresponding wave has the form

$$\boxed{\psi(\boldsymbol{\rho}, z) \propto e^{jkz} \exp\left(j\frac{k}{2z}\|\boldsymbol{\rho}\|^2\right).} \tag{4.1061}$$

The quadratic phase therefore has a direct geometrical meaning: it is the phase delay caused by the additional distance travelled by an off-axis ray relative to the axial ray.

Paraxial expansion of a spherical wave.

Starting from

$$L = \sqrt{z^2 + \|\boldsymbol{\rho}\|^2}, \tag{4.1062}$$

we factor out z,

$$L = z\sqrt{1 + \frac{\|\boldsymbol{\rho}\|^2}{z^2}}. \tag{4.1063}$$

For

$$\frac{\|\boldsymbol{\rho}\|^2}{z^2} \ll 1, \tag{4.1064}$$

the Taylor expansion

$$\sqrt{1+\epsilon} = 1 + \frac{\epsilon}{2} + \mathcal{O}(\epsilon^2) \tag{4.1065}$$

gives

$$L \simeq z + \frac{\|\boldsymbol{\rho}\|^2}{2z}. \tag{4.1066}$$

The phase of a freely propagating electron is kL, so

$$e^{jkL} \simeq \exp\left[jk\left(z + \frac{\|\boldsymbol{\rho}\|^2}{2z}\right)\right] \tag{4.1067}$$

$$= e^{jkz} \exp\left(j\frac{k}{2z}\|\boldsymbol{\rho}\|^2\right), \tag{4.1068}$$

which gives Eq. (4.1061).

Equation (4.1061) also has a simple geometrical interpretation, but two related ideas should be distinguished.

First, a spherical wavefront does not literally become a paraboloid. Rather, sufficiently close to the optical axis, the spherical surface is approximated to second order by a paraboloid. Consider a spherical wavefront of radius R,

$$\rho^2 + z^2 = R^2, \qquad \rho^2 = x^2 + y^2. \tag{4.1069}$$

Near the axis, $\rho \ll R$, and therefore

$$z = \sqrt{R^2 - \rho^2} \simeq R - \frac{\rho^2}{2R}. \tag{4.1070}$$

Thus the distance between the spherical surface and its tangent plane at the axis is

$$\boxed{\Delta z \simeq \frac{\rho^2}{2R}.} \tag{4.1071}$$

The local departure from a plane is therefore quadratic in the transverse coordinate. In this sense, the paraxial approximation replaces the spherical wavefront locally by a paraboloid of revolution.

Local parabolic approximation to a spherical wavefront.

The spherical surface satisfies

$$z = R\sqrt{1 - \frac{\rho^2}{R^2}}. \tag{4.1072}$$

Using

$$\sqrt{1-\epsilon} \simeq 1 - \frac{\epsilon}{2}, \tag{4.1073}$$

for $\epsilon \ll 1$, we obtain

$$z \simeq R\left(1 - \frac{\rho^2}{2R^2}\right) \tag{4.1074}$$

$$= R - \frac{\rho^2}{2R}. \tag{4.1075}$$

Hence

$$R - z \simeq \frac{\rho^2}{2R}, \tag{4.1076}$$

which is Eq. (4.1071).

Second, the phase is constant *on* each spherical wavefront. The quadratic phase appears when the spherical wave is sampled on a flat plane perpendicular to the optical axis. As shown in Fig. 4.7 and Eq. (4.1059), the path length to a point $\boldsymbol{\rho}$ in such a plane is

$$L(\boldsymbol{\rho}) \simeq z + \frac{\|\boldsymbol{\rho}\|^2}{2z}. \tag{4.1077}$$

The corresponding phase is therefore

$$kL \simeq kz + \frac{k}{2z}\|\boldsymbol{\rho}\|^2. \tag{4.1078}$$

After removing the spatially uniform phase kz, the transverse phase profile is

$$\boxed{\varphi(\boldsymbol{\rho}) = \frac{k}{2R}\|\boldsymbol{\rho}\|^2,} \tag{4.1079}$$

where R denotes the local radius of curvature of the spherical wave.

Thus,

$$\boxed{\begin{array}{c}\text{spherical wavefront}\\ \Downarrow \quad \text{paraxial expansion}\\ \text{locally parabolic wavefront}\\ \Downarrow\\ \text{quadratic phase across a transverse plane.}\end{array}} \tag{4.1080}$$

The coefficient of ρ^2 directly measures the wavefront curvature: a strongly curved wavefront has a small R and therefore a rapidly varying quadratic phase, whereas a nearly planar wavefront has a large R and a weak quadratic phase. In the limit

$$R \to \infty, \tag{4.1081}$$

the quadratic term vanishes and the wavefront becomes locally planar.

Thus,

$$\boxed{\text{wavefront curvature} \quad \Longleftrightarrow \quad \text{quadratic transverse phase.}} \tag{4.1082}$$

A plane wave corresponds to the limiting case $R \to \infty$, for which the quadratic phase disappears.

Free propagation

The same quadratic structure appears when propagation is expressed in spatial-frequency coordinates.

As derived in Sec. 3.7.1, a plane-wave component with transverse spatial frequencies (R_x, R_y) has longitudinal wavevector

$$k_z = k\sqrt{1 - \lambda^2 \left(R_x^2 + R_y^2\right)}. \tag{4.1083}$$

Under the paraxial condition

$$\lambda^2 \left(R_x^2 + R_y^2\right) \ll 1, \tag{4.1084}$$

this becomes

$$k_z \simeq k - \pi\lambda \left(R_x^2 + R_y^2\right). \tag{4.1085}$$

Hence propagation by a distance z multiplies each spatial-frequency component by

$$\boxed{\widehat{h}_z(R_x, R_y) \simeq e^{jkz} \exp\left[-j\pi\lambda z \left(R_x^2 + R_y^2\right)\right],} \tag{4.1086}$$

which is the Fresnel transfer function previously derived in Eq. (3.373).

Quadratic phase of free propagation in Fourier space.

The exact longitudinal component of the wavevector is

$$k_z = k\sqrt{1 - \lambda^2 R^2}, \qquad R^2 = R_x^2 + R_y^2. \tag{4.1087}$$

Using

$$\sqrt{1 - \epsilon} \simeq 1 - \frac{\epsilon}{2}, \tag{4.1088}$$

we obtain

$$k_z \simeq k\left(1 - \frac{\lambda^2 R^2}{2}\right) \tag{4.1089}$$

$$= k - \frac{k\lambda^2}{2} R^2. \tag{4.1090}$$

Since

$$k = \frac{2\pi}{\lambda}, \tag{4.1091}$$

we have

$$\frac{k\lambda^2}{2} = \pi\lambda, \tag{4.1092}$$

and therefore

$$k_z \simeq k - \pi\lambda R^2. \tag{4.1093}$$

The propagation phase is then

$$e^{jk_z z} \simeq e^{jkz} e^{-j\pi\lambda z R^2}, \tag{4.1094}$$

which gives Eq. (4.1086).

The real-space spherical-wave phase of Eq. (4.1061) and the spatial-frequency phase of Eq. (4.1086) are not two unrelated approximations. They are the real-space and Fourier-space descriptions of the same Fresnel propagation operator.

In real space, free propagation produces a quadratic phase because off-axis paths are slightly longer. In Fourier space, it produces a quadratic phase because oblique plane waves have a slightly smaller longitudinal wavevector k_z.

The quadratic phase of an ideal lens

An ideal converging lens also produces a quadratic phase, but for a different physical reason: it must compensate the different distances from different points of the lens to the focus.

Consider an incident plane wave and an ideal lens that focuses it to an axial point a distance f downstream. A point of the lens at transverse position $\boldsymbol{\rho}$ lies at a distance

$$\sqrt{f^2 + \|\boldsymbol{\rho}\|^2} \tag{4.1095}$$

from the focus. Without any phase correction, waves leaving different parts of the lens would therefore reach the focus with different phases.

For all contributions to interfere constructively at the focal point, the lens must introduce precisely the opposite phase difference. Under the paraxial approximation,

$$\sqrt{f^2 + \|\boldsymbol{\rho}\|^2} - f \simeq \frac{\|\boldsymbol{\rho}\|^2}{2f}. \tag{4.1096}$$

The required lens phase is consequently

$$\boxed{\varphi_l(\boldsymbol{\rho}) = -\frac{k}{2f}\|\boldsymbol{\rho}\|^2.} \tag{4.1097}$$

Hence the transmission function of an ideal thin lens is

$$\boxed{t_l(\boldsymbol{\rho}) = P(\boldsymbol{\rho}) \exp\left(-j\frac{k}{2f}\|\boldsymbol{\rho}\|^2\right),} \tag{4.1098}$$

in agreement with Eq. (3.483).

Why an ideal lens imposes a quadratic phase.

The geometrical distance from a point $\boldsymbol{\rho}$ in the lens plane to the axial focal point is

$$L_f(\boldsymbol{\rho}) = \sqrt{f^2 + \|\boldsymbol{\rho}\|^2}. \tag{4.1099}$$

Relative to the axial ray,

$$\Delta L_f = L_f(\boldsymbol{\rho}) - f \tag{4.1100}$$

$$\simeq \frac{\|\boldsymbol{\rho}\|^2}{2f}. \tag{4.1101}$$

Free propagation would therefore produce the additional phase

$$k\Delta L_f \simeq \frac{k}{2f}\|\boldsymbol{\rho}\|^2. \tag{4.1102}$$

For every contribution to reach the focus with the same total phase, the lens must provide the opposite phase:

$$\varphi_l = -k\Delta L_f = -\frac{k}{2f}\|\boldsymbol{\rho}\|^2. \tag{4.1103}$$

This gives Eq. (4.1097).

The lens can therefore be understood as a device that changes the curvature of the wavefront. A plane wave has zero curvature; the lens adds precisely the quadratic phase required to transform it into a converging spherical wave whose radius of curvature is f.

This is the wave-optical meaning of focusing.

Why quadratic phases produce Gaussian imaging

We can now make the connection with geometrical Gaussian optics explicit using the same input–output notation introduced in Sec. 4.10.4 and illustrated in Fig. 4.6.

Consider, for simplicity, one transverse coordinate x. Let the input object plane be located at $z = z_i$, the ideal thin lens at $z = z_\ell$, and the conjugate output image plane at $z = z_o$. We denote the corresponding transverse coordinates by

$$x_i, \qquad x_\ell, \qquad x_o, \tag{4.1104}$$

respectively.

It is convenient to introduce the positive axial propagation distances

$$\boxed{d_i \triangleq z_\ell - z_i, \qquad d_o \triangleq z_o - z_\ell.} \tag{4.1105}$$

Thus, a particular geometrical path starts from the object point x_i, crosses the lens at x_ℓ, and arrives at the output point x_o.

Under the paraxial approximation, each part of this path contributes a quadratic phase. Propagation from the input plane to the lens contributes

$$\frac{k}{2d_i}(x_\ell - x_i)^2, \tag{4.1106}$$

as follows directly from the paraxial path-length expansion of Eq. (4.1059). The ideal lens contributes the quadratic phase

$$-\frac{k}{2f}x_\ell^2, \tag{4.1107}$$

according to Eq. (4.1097), and propagation from the lens to the output plane contributes

$$\frac{k}{2d_o}(x_o - x_\ell)^2. \tag{4.1108}$$

The total phase accumulated along this path, apart from terms common to all paths, is therefore

$$\boxed{\varphi_{\mathrm{G}}(x_\ell; x_i, x_o) = \frac{k}{2d_i}(x_\ell - x_i)^2 - \frac{k}{2f}x_\ell^2 + \frac{k}{2d_o}(x_o - x_\ell)^2.} \tag{4.1109}$$

The dependence on x_ℓ is entirely quadratic. This is the wave-optical signature of Gaussian imaging.

The physical meaning of Eq. (4.1109) is particularly useful. There are many neighbouring paths connecting an input point with the output plane, distinguished by the position x_ℓ at which they cross the lens. Each path accumulates a slightly different phase because it travels a slightly different distance and receives a different phase shift from the lens. At the geometrical image point, these contributions become stationary in phase with respect to small changes of x_ℓ. This is the wave-optical counterpart of the classical stationary-action principle discussed in Sec. 4.5.3.

Requiring

$$\frac{\partial \varphi_{\mathrm{G}}}{\partial x_\ell} = 0 \tag{4.1110}$$

recovers both the thin-lens equation and the Gaussian magnification.

Gaussian imaging from the quadratic phase.

Differentiating Eq. (4.1109) with respect to the lens coordinate x_ℓ gives

$$\frac{1}{k}\frac{\partial \varphi_{\mathrm{G}}}{\partial x_\ell} = \frac{x_\ell - x_i}{d_i} - \frac{x_\ell}{f} + \frac{x_\ell - x_o}{d_o} \tag{4.1111}$$

$$= x_\ell \left(\frac{1}{d_i} + \frac{1}{d_o} - \frac{1}{f}\right) - \frac{x_i}{d_i} - \frac{x_o}{d_o}. \tag{4.1112}$$

For all neighbouring paraxial paths, corresponding to different values of x_ℓ, to connect the same pair of conjugate object and image points, the coefficient multiplying x_ℓ must vanish. Hence,

$$\boxed{\frac{1}{d_i} + \frac{1}{d_o} = \frac{1}{f}.} \tag{4.1113}$$

This is the thin-lens equation.

Once this condition is satisfied, Eq. (4.1112) reduces to

$$-\frac{x_i}{d_i} - \frac{x_o}{d_o} = 0, \tag{4.1114}$$

and therefore

$$\boxed{x_o = -\frac{d_o}{d_i} x_i.} \tag{4.1115}$$

The transverse magnification is consequently

$$\boxed{M \triangleq \frac{x_o}{x_i} = -\frac{d_o}{d_i}.} \tag{4.1116}$$

The negative sign in Eq. (4.1116) represents the inversion of a single-stage image. This is consistent with the fundamental-ray description of Sec. 4.10.5, where the image plane is characterized by

$$h_p(z_o) = 0 \tag{4.1117}$$

and the magnification is

$$M = g_p(z_o). \tag{4.1118}$$

For the single-stage imaging geometry illustrated in Fig. 4.6,

$$g_p(z_o) = M < 0, \tag{4.1119}$$

exactly as predicted by Eq. (4.1116).

This derivation is particularly revealing. The thin-lens equation is not an independent geometrical rule that must later be reconciled with wave mechanics. It is the condition under which the quadratic phases acquired during propagation to the lens, passage through the lens, and propagation to the image plane combine so that an entire family of neighbouring paths has the same phase to first order in a small change of the path.

The same quadratic structure therefore appears in three equivalent descriptions:

$$\boxed{\begin{array}{ccl} \text{quadratic reduced action} & \Longrightarrow & \text{linear paraxial ray equations,} \\ \text{quadratic wave phase} & \Longrightarrow & \text{Fresnel propagation and ideal focusing,} \\ \text{stationary quadratic phase} & \Longrightarrow & \text{Gaussian image formation.} \end{array}} \tag{4.1120}$$

This is why Gaussian optics is naturally associated with the quadratic part of the electron-optical action. At this order, ray mappings are linear and the corresponding wave phases are quadratic.

The importance of this observation will become clearer when we introduce aberrations. An ideal Gaussian lens produces precisely the quadratic phase required for first-order imaging. A real electron lens introduces additional higher-order terms. These departures from the ideal quadratic phase are the aberration phase shifts that will ultimately enter the contrast transfer function.

4.10.7 The back focal plane and the Fourier transform

One of the most important consequences of paraxial electron optics is that an ideal lens converts *propagation direction* into *position*. In wave language, the same statement becomes: an ideal lens converts *spatial frequency* into *position*.

This is the reason why the back focal plane of the objective lens is also called the *diffraction plane* or *Fourier plane*.

The connection can be understood particularly clearly by combining three results already established in the book:

1. a complex wave can be decomposed into plane-wave components by its Fourier transform, as shown in Sec. 3.6;
2. each transverse spatial frequency corresponds to a propagation direction, as expressed by Eq. (3.624);
3. a converging lens brings all rays having the same incident direction to the same point in its back focal plane, as obtained in Sec. 4.10.5.

Putting these three statements together already tells us that the back focal plane must contain the spatial-frequency decomposition of the incident electron wave.

From spatial frequency to propagation direction

Let the electron wave in the input plane be

$$\psi_i(x_i, y_i). \tag{4.1121}$$

According to the angular-spectrum representation of Sec. 3.6, it can be written as

$$\boxed{\psi_i(x_i, y_i) = \iint \widehat{\psi}_i(R_x, R_y) \exp\left[j2\pi\left(R_x x_i + R_y y_i\right)\right] dR_x\, dR_y.} \tag{4.1122}$$

This equation has a direct physical interpretation. The object wave is not being regarded as a single ray or a single plane wave. Rather, it is written as a superposition of plane waves travelling in slightly different directions. The coefficient $\widehat{\psi}_i(R_x, R_y)$ tells us how strongly the component with spatial frequency (R_x, R_y) is present.

For a plane-wave component

$$\exp\left[j2\pi\left(R_x x + R_y y\right)\right], \tag{4.1123}$$

the transverse components of the wavevector are

$$k_x = 2\pi R_x, \qquad k_y = 2\pi R_y. \tag{4.1124}$$

Since

$$k = \frac{2\pi}{\lambda}, \tag{4.1125}$$

the corresponding propagation angles satisfy, in the paraxial regime,

$$\boxed{\theta_x \simeq \lambda R_x, \qquad \theta_y \simeq \lambda R_y,} \tag{4.1126}$$

in agreement with Eq. (3.624).

Thus spatial frequency is simply another way of specifying propagation direction. Large spatial frequencies correspond to plane-wave components travelling at larger angles with respect to the optical axis.

From propagation direction to focal-plane position

Now consider what an ideal converging lens does to one of these plane-wave components.

A plane wave consists of a family of parallel rays. If those rays arrive at the lens with a small angle θ_x, the lens brings them to a point in its back focal plane whose transverse coordinate is

$$x_F \simeq f\theta_x. \tag{4.1127}$$

Similarly,

$$y_F \simeq f\theta_y. \tag{4.1128}$$

Using Eq. (4.1126),

$$\boxed{x_F = \lambda f R_x, \qquad y_F = \lambda f R_y.} \tag{4.1129}$$

Equivalently,

$$\boxed{R_x = \frac{x_F}{\lambda f}, \qquad R_y = \frac{y_F}{\lambda f}.} \tag{4.1130}$$

This is the essential geometrical property of the back focal plane.

A point in an ordinary image plane represents, approximately, a *position in the object*. A point in the back focal plane instead represents a *direction of propagation*, and therefore a *spatial frequency*.

This distinction is worth emphasizing:

image plane	:	object position $\longrightarrow$ image position,
back focal plane	:	propagation angle $\longrightarrow$ focal-plane position,
		or equivalently spatial frequency $\longrightarrow$ focal-plane position.

(4.1131)

This is precisely the distinction obtained from the fundamental rays in Sec. 4.10.5. At an image plane,

$$h_p(z_o) = 0, \tag{4.1132}$$

so the final position is independent of the initial direction. At an angular or Fourier plane,

$$g_p(z_F) = 0, \tag{4.1133}$$

so the final position is independent of the initial position and depends only on the initial direction.

Why the amplitude is the Fourier transform

The previous argument tells us *where* each spatial-frequency component appears in the back focal plane. Wave optics tells us *with what complex amplitude* it appears there.

Equation (4.1122) shows that the complex amplitude of the plane-wave component with spatial frequency (R_x, R_y) is

$$\widehat{\psi}_i(R_x, R_y). \tag{4.1134}$$

But Eq. (4.1130) tells us that the lens sends precisely this component to the point

$$\mathbf{r}_F = (x_F, y_F) = \lambda f(R_x, R_y). \tag{4.1135}$$

Therefore, apart from known scale and phase factors,

$$\boxed{\psi_F(x_F, y_F) \propto \widehat{\psi}_i\left(\frac{x_F}{\lambda f}, \frac{y_F}{\lambda f}\right).} \tag{4.1136}$$

The Fourier-transform relationship is therefore not merely a formal property of the diffraction integral. It has a very direct physical meaning:

$$\boxed{\begin{array}{c}\text{Fourier transform coefficient}\\ \widehat{\psi}_i(R_x, R_y)\\ \Updownarrow\\ \text{plane-wave component travelling at}\\ (\theta_x, \theta_y) \simeq \lambda(R_x, R_y)\\ \Downarrow \quad \text{lens}\\ \text{point in the back focal plane}\\ (x_F, y_F) = \lambda f(R_x, R_y).\end{array}} \tag{4.1137}$$

The same result from quadratic wave propagation

The same result follows directly from the quadratic phases derived in Sec. 4.10.6. This second derivation is useful because it shows explicitly how the Fourier kernel emerges from Fresnel propagation and the quadratic phase of the lens.

Consider the standard arrangement in which the input wave is placed in the front focal plane of an ideal thin lens and the field is evaluated in the back focal plane. Let the lens lie at $z = z_\ell$, so that the input plane and back focal plane are each a distance f from the lens.

Propagation from the input plane to the lens, multiplication by the quadratic lens phase, and propagation to the back focal plane together produce the Fourier-transform relation derived previously in Sec. 3.8.1.

Fourier transform generated by an ideal lens.

Let $\boldsymbol{\rho}_i = (x_i, y_i)$ denote a coordinate in the input plane, $\boldsymbol{\rho}_\ell$ a coordinate in the lens plane, and $\boldsymbol{\rho}_F = (x_F, y_F)$ a coordinate in the back focal plane.

Under Fresnel propagation over a distance f, the phase accumulated between the input and lens planes is, apart from the uniform axial phase,

$$\frac{k}{2f} \left\| \boldsymbol{\rho}_\ell - \boldsymbol{\rho}_i \right\|^2 . \tag{4.1138}$$

The ideal lens contributes

$$-\frac{k}{2f} \left\| \boldsymbol{\rho}_\ell \right\|^2 , \tag{4.1139}$$

according to Eq. (4.1097).

Propagation from the lens to its back focal plane contributes

$$\frac{k}{2f}\left\|\boldsymbol{\rho}_F-\boldsymbol{\rho}_\ell\right\|^2. \tag{4.1140}$$

The total transverse phase is therefore

$$\begin{aligned}\varphi=&\frac{k}{2f}\left\|\boldsymbol{\rho}_\ell-\boldsymbol{\rho}_i\right\|^2-\frac{k}{2f}\left\|\boldsymbol{\rho}_\ell\right\|^2\\&+\frac{k}{2f}\left\|\boldsymbol{\rho}_F-\boldsymbol{\rho}_\ell\right\|^2.\end{aligned} \tag{4.1141}$$

Expanding the squared distances and collecting terms gives

$$\begin{aligned}\varphi=&\frac{k}{2f}\left\|\boldsymbol{\rho}_i\right\|^2+\frac{k}{2f}\left\|\boldsymbol{\rho}_F\right\|^2\\&+\frac{k}{2f}\left\|\boldsymbol{\rho}_\ell\right\|^2-\frac{k}{f}\boldsymbol{\rho}_\ell\cdot\left(\boldsymbol{\rho}_i+\boldsymbol{\rho}_F\right).\end{aligned} \tag{4.1142}$$

In the standard Fourier-transform arrangement, the quadratic phase associated with propagation to the lens and the quadratic phase imposed by the lens combine so that the remaining integral can be reduced to the Fourier kernel. Equivalently, carrying out the Fresnel integrations gives the result already derived in Eq. (3.498),

$$\boxed{U_F(\mathbf{r}_F)=\frac{e^{jkf}}{j\lambda f}\exp\left(j\frac{k}{2f}\|\mathbf{r}_F\|^2\right)\widehat{PU^-}\left(\frac{\mathbf{r}_F}{\lambda f}\right).} \tag{4.1143}$$

Equation (4.1143) was derived for a generic scalar wave U. For the electron wave we simply identify

$$U\longrightarrow\psi. \tag{4.1144}$$

Thus,

$$\boxed{\psi_F(\mathbf{r}_F)=\frac{e^{jkf}}{j\lambda f}\exp\left(j\frac{k}{2f}\|\mathbf{r}_F\|^2\right)\widehat{P\psi_i}\left(\frac{\mathbf{r}_F}{\lambda f}\right).} \tag{4.1145}$$

The exponential factor outside the Fourier transform is a deterministic quadratic phase. It does not change the modulus of the Fourier coefficients at that plane,

$$\left|\exp\left(j\frac{k}{2f}\|\mathbf{r}_F\|^2\right)\right|=1, \tag{4.1146}$$

although its phase must be retained if the wave is propagated further.

What would be measured in the back focal plane?

If an ideal detector were placed directly in the back focal plane, the measured intensity would be proportional to the squared modulus of the electron wave, as discussed in Sec. 3.11. From Eq. (4.1145),

$$\boxed{I_F(\mathbf{r}_F) \propto \left|\widehat{P\psi_i}\left(\frac{\mathbf{r}_F}{\lambda f}\right)\right|^2.} \tag{4.1147}$$

The back focal plane therefore contains the diffraction pattern of the incident wave.

For a crystalline specimen, the exit wave contains particularly strong components at spatial frequencies associated with reciprocal-lattice vectors, and these appear as localized diffraction spots. For a non-periodic specimen, such as a macromolecule in cryo-EM, the spectrum is continuous rather than concentrated at a discrete set of reciprocal lattice points, but the physical principle is the same: different spatial frequencies occupy different transverse positions in the diffraction plane.

Why the back focal plane matters for image formation

The back focal plane is especially important because this is the natural plane in which the action of the objective lens on spatial frequency can be described.

The specimen produces an exit wave containing many spatial-frequency components,

$$\widehat{\psi}_{\text{exit}}(\mathbf{R}). \tag{4.1148}$$

The ideal objective lens maps each component $\mathbf{R}$ to the corresponding point

$$\mathbf{r}_F = \lambda f \mathbf{R}. \tag{4.1149}$$

A finite objective aperture therefore acts directly as a spatial-frequency filter: blocking a region of the back focal plane is equivalent to removing the corresponding spatial frequencies from the image-forming wave.

More importantly for cryo-EM, imperfections of the objective lens do not affect all spatial frequencies in exactly the same way. Different points in the back focal plane correspond to different scattering angles and therefore to different spatial frequencies. If the lens introduces an additional phase

$$\chi(\mathbf{R}), \tag{4.1150}$$

then the exit-wave spectrum is modified according to

$$\boxed{\widehat{\psi}_{\text{exit}}(\mathbf{R}) \longrightarrow P(\mathbf{R})e^{-j\chi(\mathbf{R})}\widehat{\psi}_{\text{exit}}(\mathbf{R}).} \tag{4.1151}$$

This is the central reason why the Fourier-plane description is so useful: aberrations of the objective lens become frequency-dependent phase shifts. The contrast transfer function will arise from precisely this phase modulation.

Real electron lenses

A real TEM objective lens is not an infinitesimally thin optical element. Its magnetic field occupies a finite region, and in modern objective lenses the specimen may lie inside that field. A complete treatment therefore distinguishes real and asymptotic focal planes, principal planes, and other cardinal elements (Hawkes and Kasper, 2017)[Chaps. 16–17].

For the present purpose, however, this additional geometrical machinery is not required. The first-order behaviour of the complete objective can be represented by its paraxial transfer matrix, as described in Sec. 3.10.3, or equivalently by the fundamental rays $g_p(z)$ and $h_p(z)$ introduced in Sec. 4.10.4.

The essential condition for an angular plane is

$$g_p(z_F) = 0, \tag{4.1152}$$

or, in the ABCD language,

$$A = 0, \tag{4.1153}$$

as discussed in Sec. 4.10.5. At such a plane, the final position no longer depends on the initial transverse position but only on the initial direction. Therefore the same physical interpretation survives for a thick electromagnetic lens:

$$\boxed{\text{direction} \iff \text{spatial frequency} \iff \text{position in the diffraction plane.}} \tag{4.1154}$$

This three-way correspondence is the result that we shall carry forward into the theory of aberrations and the contrast transfer function.

4.10.8 Defocus as propagation away from the Gaussian image plane

We now introduce *defocus*, which is the final ingredient needed to connect Gaussian imaging with the first form of the contrast transfer function.

In the ideal Gaussian situation, an object point located at

$$\mathbf{r}_i \tag{4.1155}$$

in the input plane

$$z = z_i \tag{4.1156}$$

is imaged sharply at the point

$$\mathbf{r}_o \tag{4.1157}$$

in the Gaussian image plane

$$z = z_o. \tag{4.1158}$$

This is the situation described geometrically in Sec. 4.10.5 and wave-optically in Sec. 4.10.6.

In practice, however, the observation plane need not coincide exactly with the Gaussian image plane. Let the actual observation plane be located at

$$z = z_{\text{obs}}. \tag{4.1159}$$

We define the signed geometrical defocus distance as

$$\boxed{\delta z \triangleq z_{\text{obs}} - z_o.} \tag{4.1160}$$

The geometrical setting is illustrated in Fig. 4.8. If the observation plane lies before the Gaussian image plane along the direction of propagation, then

$$z_{\text{obs}} < z_o, \tag{4.1161}$$

and consequently

$$\delta z < 0. \tag{4.1162}$$

This situation is called *underfocus*. Conversely,

$$z_{\text{obs}} > z_o \qquad \Longrightarrow \qquad \delta z > 0, \tag{4.1163}$$

corresponds to *overfocus*.

The physical meaning of defocus is simple. In a perfectly focused image plane, all paraxial contributions associated with one object point recombine at the corresponding image point. In a defocused plane this recombination is incomplete: the same family of trajectories now reaches slightly different transverse positions. A point object is therefore no longer imaged as an ideal point.

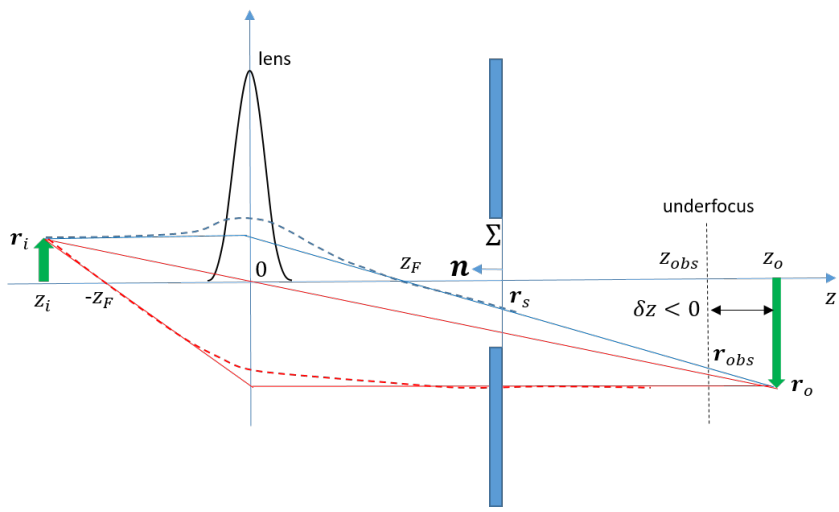


Figure 4.8: Geometrical interpretation of defocus. An object point located at $\mathbf{r}_i$ in the input plane z_i is imaged by the lens at the point $\mathbf{r}_o$ in the Gaussian image plane z_o. The back focal plane is located at z_F, and Σ denotes an aperture in that plane. If the observation plane is moved to $z_o' = z_o + \delta z$, the system is underfocused by the amount δz. In the Gaussian image plane the rays meet at the image point $\mathbf{r}_o$, whereas in the defocused plane they no longer meet at a single point. Solid lines indicate the geometrical image construction, while the colored trajectories illustrate the paths of unscattered and scattered electrons.

This can already be seen in geometrical optics. If a paraxial ray has local slope

$$\boldsymbol{\alpha} = \frac{d\mathbf{r}_o}{dz} \tag{4.1164}$$

at the Gaussian image plane, then displacement of the observation plane by δz changes its transverse intersection point by approximately

$$\boxed{\delta\mathbf{r}_o \simeq \delta z\, \boldsymbol{\alpha}.} \tag{4.1165}$$

Thus, rays that would have met at $\mathbf{r}_o$ in the Gaussian image plane are spread out in the defocused plane. If the transmitted beam is limited by an aperture to semi-angle $\alpha_{\max}$, the radius of the corresponding geometrical blur is of order

$$\boxed{c \sim |\delta z|\, \alpha_{\max}.} \tag{4.1166}$$

This estimate is particularly convenient numerically because

$$1\ \mu\text{m} \times 1\ \text{mrad} = 1\ \text{nm}. \tag{4.1167}$$

Therefore, if $|\delta z|$ is expressed in μm and $\alpha_{\max}$ in mrad, the numerical value of c is obtained directly in nm. This estimate is particularly convenient numerically because

$$1\ \mu\text{m} \times 1\ \text{mrad} = 1\ \text{nm}. \tag{4.1168}$$

Thus, for a representative semi-angle of $\alpha_{\max} = 10$ mrad and a defocus magnitude of $|\delta z| = 1\ \mu$m, Eq. (4.1166) gives

$$c \sim 10\ \text{nm}. \tag{4.1169}$$

For the same angular range,

$\lvert\delta z\rvert$	c ($\alpha_{\max} = 10$ mrad)
0.25 μm	2.5 nm
0.5 μm	5 nm
1 μm	10 nm
2 μm	20 nm
3 μm	30 nm

(4.1170)

Alternatively, for a fixed defocus magnitude of $|\delta z| = 1\ \mu$m,

$\alpha_{\max}$	c
2 mrad	2 nm
5 mrad	5 nm
10 mrad	10 nm
20 mrad	20 nm

(4.1171)

These angles are directly connected with the spatial frequencies carried by the electron wave. According to Eq. (4.1126),

$$\alpha \simeq \lambda R = \frac{\lambda}{d}, \tag{4.1172}$$

where $d = 1/R$ is the corresponding real-space period. For a 300 kV electron, Sec. 2.4 gives $\lambda \simeq 1.97$ pm. A spatial period of $d = 2$ Å therefore corresponds to

$$\alpha \simeq \frac{1.97 \text{ pm}}{200 \text{ pm}} \simeq 9.9 \text{ mrad}. \tag{4.1173}$$

Thus, for a defocus magnitude of 1 μm, rays associated with 2 Å spatial frequencies would, in a purely geometrical picture, be displaced from the Gaussian image point by roughly

$$c \sim (1\ \mu\text{m})(9.9 \text{ mrad}) \simeq 9.9 \text{ nm}. \tag{4.1174}$$

This geometrical blur should not be interpreted as the resolution of a defocused electron-microscopy image. The ray picture says only that rays with different angles no longer intersect at the same point when the observation plane is displaced from the Gaussian image plane. The electron wave remains coherent, and these angular components interfere with one another. In wave optics, defocus therefore produces an oscillatory, spatial-frequency-dependent phase shift rather than a simple geometrical smoothing. As we shall see below, this phase shift is

$$\chi_{\delta z}(R) = \pi \lambda \delta z R^2, \tag{4.1175}$$

and it is this quantity, rather than the geometrical blur radius alone, that governs contrast transfer in cryo-EM.

This geometrical picture is useful, but for image formation the wave-optical description is even more important. Defocus does not merely broaden the image; it changes the *relative phases* of the spatial frequencies composing the image wave.

Let

$$\psi_G(\mathbf{r}_o) \tag{4.1176}$$

denote the scalar electron wave in the Gaussian image plane, where

$$\mathbf{r}_o = (x_o, y_o) \tag{4.1177}$$

is the transverse coordinate. The wave in an observation plane displaced by δz from the Gaussian image plane is obtained by paraxial propagation over the distance δz.

Quadratic phase produced by defocus.

In Sec. 3.7.1, the paraxial transfer function for free propagation over a distance z was found to be

$$\widehat{h}_z(R_x, R_y) = e^{jkz} \exp\left[-j\pi\lambda z\left(R_x^2 + R_y^2\right)\right] \tag{4.1178}$$

in Eq. (3.373). Applying this result to propagation from the Gaussian image plane to the observation plane gives

$$\widehat{\psi}_{\text{obs}}(R_x, R_y) = \widehat{\psi}_G(R_x, R_y)\,\widehat{h}_{\delta z}(R_x, R_y). \tag{4.1179}$$

Therefore,

$$\widehat{\psi}_{\text{obs}}(R_x, R_y) = e^{jk\delta z}\widehat{\psi}_G(R_x, R_y) \exp\left[-j\pi\lambda\delta z\left(R_x^2 + R_y^2\right)\right]. \tag{4.1180}$$

The factor

$$e^{jk\delta z} \tag{4.1181}$$

is a spatially uniform phase and therefore does not affect the image intensity. Dropping this physically inessential global phase, we obtain

$$\boxed{\widehat{\psi}_{\text{obs}}(\mathbf{R}) = \widehat{\psi}_G(\mathbf{R}) \exp\left[-j\pi\lambda\delta z\|\mathbf{R}\|^2\right],} \tag{4.1182}$$

where

$$\mathbf{R} = (R_x, R_y). \tag{4.1183}$$

Equation (4.1182) is the basic wave-optical description of defocus. It shows that defocus leaves the spatial frequencies themselves unchanged, but multiplies each spatial frequency by a *quadratic phase factor*.

It is therefore natural to define the defocus phase by

$$\boxed{\chi_{\delta z}(\mathbf{R}) \triangleq \pi\lambda\delta z\|\mathbf{R}\|^2.} \tag{4.1184}$$

Then Eq. (4.1182) may be written more compactly as

$$\boxed{\widehat{\psi}_{\text{obs}}(\mathbf{R}) = \widehat{\psi}_G(\mathbf{R})e^{-j\chi_{\delta z}(\mathbf{R})}.} \tag{4.1185}$$

At this point, it is useful to distinguish the geometrical displacement of the observation plane from the conventional defocus parameter used in

cryo-EM. We defined

$$\delta z = z_{\text{obs}} - z_o, \tag{4.1186}$$

so that, with the positive z direction along electron propagation, underfocus corresponds to

$$\delta z < 0. \tag{4.1187}$$

In cryo-EM it is customary to report the magnitude of underfocus as a positive quantity. We therefore define

$$\boxed{\Delta f \triangleq -\delta z,} \tag{4.1188}$$

so that

$$\Delta f > 0 \qquad \text{for underfocus.} \tag{4.1189}$$

Using this convention, Eq. (4.1185) becomes

$$\widehat{\psi}_{\text{obs}}(\mathbf{R}) = \widehat{\psi}_G(\mathbf{R}) \exp\left(j\pi\lambda\Delta f \, \|\mathbf{R}\|^2\right). \tag{4.1190}$$

If the transfer factor is written as $e^{-j\chi_{\text{def}}}$, the corresponding defocus aberration phase is therefore

$$\boxed{\chi_{\text{def}}(\mathbf{R}) = -\pi\lambda\Delta f \, \|\mathbf{R}\|^2.} \tag{4.1191}$$

Defocus is therefore the simplest non-trivial image-forming phase shift of the objective system. In an ideal focused image plane, the objective transfers the Gaussian image wave without any additional frequency-dependent phase. In a defocused plane, each spatial frequency acquires the additional quadratic phase of Eq. (4.1185).

This is precisely the ingredient that will enter the first form of the contrast transfer function. Once the specimen-side approximation has been introduced, the combination of weak scattering with $e^{-j\chi_{\delta z}(\mathbf{R})}$ will lead directly to the familiar sine-type transfer law.

If we also include a finite objective aperture, represented by a pupil function

$$P(\mathbf{R}), \tag{4.1192}$$

then the paraxial transfer from the Gaussian image wave to the actual observation plane is

$$\boxed{\widehat{\psi}_{\text{obs}}(\mathbf{R}) = P(\mathbf{R})\, e^{-j\chi_{\delta z}(\mathbf{R})} \widehat{\psi}_G(\mathbf{R}),} \tag{4.1193}$$

which is the natural starting point for the later development of aberration theory and the contrast transfer function.

4.10.9 From coherent wave transfer to the detected electron image

The previous sections have described the action of the objective lens on the complex electron wave. We now connect that wave-optical description with the image recorded by the detector.

The general relation between electron-wave propagation and detection has already been established in Sec. 3.11. In particular, Eq. (3.724) showed that, for the narrow paraxial fixed-energy beam considered here, the detected electron intensity is proportional to

$$I(\mathbf{r}) \propto |\psi(\mathbf{r})|^2. \tag{4.1194}$$

Individual electrons are detected as discrete events, and their expected number within a detector region is determined by the probability current according to Eq. (3.727).

The distinction between coherent and incoherent image formation was also established in Sec. 3.11. For coherent illumination, complex amplitudes are propagated and added first, and only afterwards is their squared magnitude formed [Eq. (3.741)]. For a shift-invariant system this reduces to

$$I_o(\mathbf{r}) \propto |\psi_i * h|^2 , \tag{4.1195}$$

as shown in Eq. (3.751). By contrast, in the incoherent limit it is the intensities that are convolved with $|h|^2$, according to Eq. (3.752).

For the moment we consider the coherent limit. Partial coherence will be reintroduced later, when the corresponding temporal- and spatial-coherence envelopes are incorporated into the contrast transfer function.

We shall also postpone the physical derivation of the wave produced by the specimen. Let

$$\psi_i(\mathbf{r}) \tag{4.1196}$$

denote the electron wave entering the objective-lens imaging system. This wave is assumed to contain already the information introduced by the interaction with the specimen.

For an isoplanatic coherent imaging system, the objective acts linearly on the complex wave. In Fourier space,

$$\boxed{\widehat{\psi}_{\text{obs}}(\mathbf{R}) = H(\mathbf{R})\widehat{\psi}_i(\mathbf{R}),} \tag{4.1197}$$

where $H(\mathbf{R})$ is the coherent wave-transfer function. For the defocus-only system developed in Sec. 4.10.8,

$$\boxed{H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\delta z}(\mathbf{R})},} \tag{4.1198}$$

with

$$\chi_{\delta z}(\mathbf{R}) = \pi\lambda\delta z\|\mathbf{R}\|^2 \tag{4.1199}$$

from Eq. (4.1184). The pupil function $P(\mathbf{R})$ specifies which spatial frequencies pass through the objective aperture.

The corresponding coherent amplitude impulse response is

$$\boxed{h(\mathbf{r}) = \mathcal{F}^{-1}\left\{H(\mathbf{R})\right\}.} \tag{4.1200}$$

Thus,

$$\psi_{\text{obs}} = h * \psi_i, \tag{4.1201}$$

which is precisely the coherent propagation law already discussed in Sec. 3.11.

The point-spread interpretation provides a useful connection with the geometrical picture of defocus. Consider a contribution that, under perfect Gaussian imaging, would be localized around the image coordinate $\mathbf{r}_o$. In the actual observation plane its probability amplitude is distributed according to

$$\psi_{\text{obs}}(\mathbf{r}_{\text{obs}}) \propto h(\mathbf{r}_{\text{obs}} - \mathbf{r}_o). \tag{4.1202}$$

Consequently, repeated detections associated with this point contribution have a spatial probability distribution proportional to

$$\boxed{p(\mathbf{r}_{\text{obs}} \mid \mathbf{r}_o) \propto |h(\mathbf{r}_{\text{obs}} - \mathbf{r}_o)|^2.} \tag{4.1203}$$

This is the wave-mechanical counterpart of the geometrical defocus described in Sec. 4.10.8. In the Gaussian image plane, the contributions associated with an object point recombine at $\mathbf{r}_o$. When the observation plane is displaced from that plane, the corresponding probability amplitude becomes spatially distributed, and individual electron detections occur at positions $\mathbf{r}_{\text{obs}}$ drawn from the distribution in Eq. (4.1203).

It is important, however, not to interpret $|h|^2$ as an intensity point-spread function that can simply be convolved with the specimen intensity. That rule applies to incoherent imaging, as shown in Eq. (3.752). In the coherent regime relevant here, the electron amplitudes interfere before detection:

$$\boxed{I_{\text{obs}}(\mathbf{r}) \propto |h * \psi_i|^2.} \tag{4.1204}$$

Equation (4.1204) is simply the specialization of Eq. (3.751) to the objective-lens transfer function of Eq. (4.1198).

This distinction is the key to phase contrast. The objective lens acts linearly on the *complex wave amplitudes* through $H(\mathbf{R})$, whereas the detector responds to the squared magnitude of the resulting coherent superposition. The contrast transfer function will arise from the interference terms generated by this final operation.

4.10.10 First contrast transfer model: defocus phase contrast

We can now derive the simplest form of the contrast transfer function. The purpose at this stage is not to model the interaction between the electron and the specimen. We shall merely assume a particularly simple form for the wave entering the objective and postpone its physical derivation to the chapter devoted to specimen interaction (Chap. 5).

Let the incident wave on the imaging system be a strong uniform reference wave accompanied by a weak phase modulation,

$$\boxed{\psi_i(\mathbf{r}) \simeq 1 + j\varphi(\mathbf{r}), \qquad |\varphi(\mathbf{r})| \ll 1,} \tag{4.1205}$$

where $\varphi(\mathbf{r})$ is real. The significance and range of validity of this approximation will be discussed later. Here it serves only to isolate the mechanism by which a phase variation in the electron wave becomes an intensity variation at the detector.

Taking the Fourier transform of Eq. (4.1205) gives

$$\boxed{\widehat{\psi}_i(\mathbf{R}) = \delta(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R}).} \tag{4.1206}$$

The term $\delta(\mathbf{R})$ represents the strong unscattered reference wave. Since φ is real, its Fourier coefficients occur in complex-conjugate pairs,

$$\boxed{\widehat{\varphi}(-\mathbf{R}) = \widehat{\varphi}^*(\mathbf{R}).} \tag{4.1207}$$

This conjugate symmetry will be responsible for converting the complex phase transfer of the objective into a real sine transfer function.

For the moment, we retain only defocus and the objective aperture, so that the coherent wave transfer function is Eq. (4.1198),

$$H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\delta z}(\mathbf{R})}. \tag{4.1208}$$

We assume that the direct beam is transmitted,

$$P(\mathbf{0}) = 1, \tag{4.1209}$$

and note that

$$\chi_{\delta z}(\mathbf{0}) = 0. \tag{4.1210}$$

The reference wave is therefore unaffected by the defocus phase.

Applying Eq. (4.1197) to Eq. (4.1206) gives

$$\widehat{\psi}_{\text{obs}}(\mathbf{R}) = \delta(\mathbf{R}) + jH(\mathbf{R})\widehat{\varphi}(\mathbf{R}). \tag{4.1211}$$

Let

$$g(\mathbf{r}) = \mathcal{F}^{-1}\left\{H(\mathbf{R})\widehat{\varphi}(\mathbf{R})\right\}. \tag{4.1212}$$

Then

$$\psi_{\text{obs}}(\mathbf{r}) \simeq 1 + jg(\mathbf{r}). \tag{4.1213}$$

The objective has therefore modified the weak scattered part of the wave, but the detector still has not entered the calculation. It does so through the squared modulus.

From complex wave transfer to the sine CTF.

Using Eq. (3.724), the detected intensity is, apart from an overall normalization,

$$\begin{aligned} I_{\text{obs}}(\mathbf{r}) &= |1 + jg(\mathbf{r})|^2 \\ &= (1 + jg)\,(1 - jg^*) \\ &= 1 + jg - jg^* + |g|^2. \end{aligned} \tag{4.1214}$$

Because g is first order in the weak modulation φ, the term $|g|^2$ is second order and is neglected. Thus

$$\boxed{I_{\text{obs}}(\mathbf{r}) \simeq 1 + jg(\mathbf{r}) - jg^*(\mathbf{r}).} \tag{4.1215}$$

Define the intensity variation with respect to the uniform background by

$$\delta I(\mathbf{r}) = I_{\text{obs}}(\mathbf{r}) - 1. \tag{4.1216}$$

Taking its Fourier transform and using Eq. (4.1212),

$$\widehat{\delta I}(\mathbf{R}) = jH(\mathbf{R})\widehat{\varphi}(\mathbf{R}) - jH^*(-\mathbf{R})\widehat{\varphi}^*(-\mathbf{R}). \tag{4.1217}$$

From the Hermitian symmetry in Eq. (4.1207),

$$\widehat{\varphi}^*(-\mathbf{R}) = \widehat{\varphi}(\mathbf{R}), \tag{4.1218}$$

and consequently

$$\boxed{\widehat{\delta I}(\mathbf{R}) = j\left[H(\mathbf{R}) - H^*(-\mathbf{R})\right]\widehat{\varphi}(\mathbf{R}).} \tag{4.1219}$$

For the rotationally symmetric defocus considered here,

$$P(-\mathbf{R}) = P(\mathbf{R}), \qquad \chi_{\delta z}(-\mathbf{R}) = \chi_{\delta z}(\mathbf{R}), \tag{4.1220}$$

so

$$H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\delta z}(\mathbf{R})} \tag{4.1221}$$

and

$$H^*(-\mathbf{R}) = P(\mathbf{R})e^{+j\chi_{\delta z}(\mathbf{R})}. \tag{4.1222}$$

Therefore,

$$\begin{aligned} j\left[H(\mathbf{R}) - H^*(-\mathbf{R})\right] &= jP(\mathbf{R})\left[e^{-j\chi_{\delta z}} - e^{+j\chi_{\delta z}}\right] \\ &= 2P(\mathbf{R})\sin\left[\chi_{\delta z}(\mathbf{R})\right]. \end{aligned} \tag{4.1223}$$

Hence,

$$\boxed{\widehat{\delta I}(\mathbf{R}) = 2P(\mathbf{R})\sin\left[\chi_{\delta z}(\mathbf{R})\right]\widehat{\varphi}(\mathbf{R}).} \tag{4.1224}$$

The factor 2 in Eq. (4.1224) results from the chosen normalization of the reference wave and is normally absorbed into the definition of image contrast. The first contrast transfer function can therefore be defined, up to an overall normalization, as

$$\boxed{\mathrm{CTF}(\mathbf{R}) = P(\mathbf{R})\sin\left[\chi_{\delta z}(\mathbf{R})\right].} \tag{4.1225}$$

Using the defocus phase of Eq. (4.1184),

$$\boxed{\mathrm{CTF}(\mathbf{R}) = P(\mathbf{R})\sin\left(\pi\lambda\delta z\|\mathbf{R}\|^2\right).} \tag{4.1226}$$

For a circularly symmetric aperture we may write $R = \|\mathbf{R}\|$, giving

$$\boxed{\mathrm{CTF}(R) = P(R)\sin\left(\pi\lambda\delta z R^2\right).} \tag{4.1227}$$

If instead we use the cryo-EM convention introduced in Eq. (4.1188),

$$\Delta f = -\delta z, \qquad \Delta f > 0 \quad \text{for underfocus}, \tag{4.1228}$$

then

$$\boxed{\mathrm{CTF}(R) = -P(R)\sin\left(\pi\lambda\Delta f R^2\right).} \tag{4.1229}$$

The overall sign of the CTF depends on the conventions adopted for the Fourier transform, electron-wave phase, defocus, and image contrast. What is physically invariant at this stage is the quadratic dependence of the phase on R and the resulting oscillatory sine transfer.

The origin of the sine in Eqs. (4.1225)–(4.1229) is particularly important. The objective itself does *not* have a sine transfer function. It acts on the complex electron wave through the phase factor

$$e^{-j\chi_{\delta z}(\mathbf{R})}. \tag{4.1230}$$

The sine appears only when the detector forms the squared modulus of the wave. Because the weak phase modulation is real, its Fourier components at $\mathbf{R}$ and $-\mathbf{R}$ are complex conjugates [Eq. (4.1207)]. In the detected intensity they combine through

$$\boxed{j\left(e^{-j\chi} - e^{+j\chi}\right) = 2\sin\chi.} \tag{4.1231}$$

Thus the first CTF is a consequence of three distinct physical steps: the specimen supplies a complex electron wave; the objective modifies the phases of its spatial-frequency components; and electron detection converts the resulting coherent interference pattern into a spatial distribution of discrete electron counts.

This first model contains only defocus and aperture truncation. It already captures the essential mechanism of phase contrast. The subsequent refinements will modify the phase $\chi(\mathbf{R})$ and the amplitude of the transferred frequencies by introducing spherical aberration, astigmatism, amplitude contrast, and the different coherence envelopes.

4.11 Aberrations

The paraxial theory developed in the preceding sections describes an ideal Gaussian electron-optical system. Within that approximation, the action is retained only to quadratic order in the transverse coordinates and slopes, the ray equations are linear, and conjugate object points are mapped into conjugate image points. In wave language, the same quadratic approximation gives the ideal lens phase and the Fresnel propagation laws derived in Sec. 4.10.6.

Real electron lenses depart from this ideal behavior. Rays that should intersect at a single Gaussian image point no longer do so exactly, and

the wavefront produced by the lens differs from the wavefront predicted by Gaussian optics. These departures are collectively called *aberrations*.

Two closely related ways describe the same phenomenon. In geometrical optics, aberrations are expressed as deviations of the actual electron trajectories from the paraxial trajectories. In wave optics, they are expressed as deviations of the phase of the actual wavefront from the phase of the ideal Gaussian wavefront. The second description will be particularly useful for cryo-EM because image formation has already been formulated in terms of the Fourier-space phase transfer of the objective lens.

The connection with the analytical-mechanics formulation developed in Sec. 4.4 is deeper still. Gaussian optics results from retaining the quadratic contribution to the reduced action, as discussed in Sec. 4.10.6. Higher-order terms in the action generate departures from Gaussian imaging. Hawkes and Kasper derive the classical geometrical aberrations of round electron lenses precisely in this manner: the quadratic contribution produces the paraxial trajectories, whereas the next higher-order contribution generates the lowest-order geometrical aberrations (Hawkes and Kasper, 2017)[Chap. 24].

We shall exploit this connection systematically. Rather than treating the standard cryo-EM aberration formulas as empirical corrections to the CTF, we shall interpret them as different terms in the departure of the electron action, and therefore of the electron phase, from its ideal Gaussian value.

4.11.1 Overview and families of aberrations

Several classifications of aberrations coexist in electron optics. They are not contradictory: they emphasize different aspects of the same departure from ideal imaging. Classical electron optics usually classifies aberrations according to their dependence on object position and ray angle. Cryo-EM more commonly classifies them according to the spatial-frequency dependence and symmetry of the phase shift introduced by the objective lens. A third description expands the aberration wavefront in a complete basis, such as the Zernike polynomials.

It is useful first to distinguish these viewpoints before deriving the individual aberrations.

Classical geometrical aberrations

For a monochromatic rotationally symmetric lens, the classical lowest-order departures from Gaussian imaging are the five Seidel aberrations:

$$\boxed{\begin{array}{c}\text{spherical aberration,} \quad \text{coma,} \quad \text{astigmatism,} \\ \text{field curvature,} \quad \text{distortion.}\end{array}} \tag{4.1232}$$

Let r denote the distance of the object point from the optical axis and α a characteristic ray angle with respect to that axis. Their leading geometrical dependences may be summarized schematically as

aberration	leading ray dependence
spherical aberration	α^3
coma	$r\alpha^2$
astigmatism	$r^2\alpha$
field curvature	$r^2\alpha$
distortion	r^3

(4.1233)

The two terms proportional to $r^2\alpha$ represent different geometrical effects even though they have the same overall order.

This structure also appears naturally in the electron-optical characteristic-function treatment. Hawkes and Kasper denote the corresponding coefficients by C, K, A, F, and D for spherical aberration, coma, astigmatism, field curvature, and distortion, respectively (Hawkes and Kasper, 2017)[Chap. 24].

The dependence on r immediately provides an important physical distinction. Spherical aberration remains present for an object point on the optical axis,

$$r = 0, \tag{4.1234}$$

whereas the classical Seidel coma, astigmatism, field curvature, and distortion are field-dependent aberrations. For an objective lens imaging a very small specimen region close to the optical axis, spherical aberration is therefore especially important. Field-dependent aberrations become increasingly relevant as the object or image extends away from the axis.

Axial phase aberrations relevant to cryo-EM

The classical Seidel classification is useful for understanding the origin of lens imperfections, but it is not the most convenient classification for high-resolution cryo-EM. The quantity that enters directly into image formation

is the phase acquired by each spatial-frequency component of the electron wave.

In Sec. 4.10.8, for example, we found that displacement of the observation plane introduces the phase

$$\chi_{\delta z}(\mathbf{R}) = \pi\lambda\delta z\|\mathbf{R}\|^2, \tag{4.1235}$$

according to Eq. (4.1184). The corresponding coherent transfer function is

$$H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\delta z}(\mathbf{R})} \tag{4.1236}$$

[Eq. (4.1198)].

A real objective lens introduces additional phase terms. It is therefore natural to write

$$\boxed{H(\mathbf{R}) = P(\mathbf{R})\exp\left[-j\chi(\mathbf{R})\right],} \tag{4.1237}$$

where

$$\boxed{\chi(\mathbf{R}) = \chi_{\mathrm{def}}(\mathbf{R}) + \chi_{\mathrm{ab}}(\mathbf{R}).} \tag{4.1238}$$

Here χ_{ab} collects the phase errors produced by the imperfections of the electron-optical system.

Among the aberrations commonly relevant to cryo-EM are twofold astigmatism, spherical aberration, axial coma, trefoil or threefold astigmatism, fourfold astigmatism, and higher-order spherical and anisotropic terms. At sufficiently high resolution, several of these higher-order aberrations can be measured directly from single-particle cryo-EM data (Zivanov et al., 2020).

There is an important terminology issue. The *twofold astigmatism* routinely estimated during cryo-EM CTF determination is most naturally understood as a direction-dependent quadratic defocus. It should not be confused with the classical off-axis Seidel astigmatism listed in Eq. (4.1232). They share the name "astigmatism" because both correspond to different focusing behavior in different directions, but they arise in different expansions and describe different experimental situations.

We shall therefore start our detailed study with twofold astigmatism, because it follows directly from the defocus theory already developed in Sec. 4.10.8.

Symmetric and antisymmetric phase aberrations

A classification particularly useful in cryo-EM is obtained by comparing the phase shifts at the Friedel-related spatial frequencies $\mathbf{R}$ and $-\mathbf{R}$. Any aberration phase can be decomposed uniquely as

$$\boxed{\chi(\mathbf{R}) = \chi_{\mathrm{sym}}(\mathbf{R}) + \chi_{\mathrm{asym}}(\mathbf{R}),} \tag{4.1239}$$

where

$$\chi_{\mathrm{sym}}(\mathbf{R}) = \frac{1}{2}\left[\chi(\mathbf{R}) + \chi(-\mathbf{R})\right], \tag{4.1240}$$

$$\chi_{\mathrm{asym}}(\mathbf{R}) = \frac{1}{2}\left[\chi(\mathbf{R}) - \chi(-\mathbf{R})\right]. \tag{4.1241}$$

By construction,

$$\chi_{\mathrm{sym}}(-\mathbf{R}) = \chi_{\mathrm{sym}}(\mathbf{R}), \tag{4.1242}$$

whereas

$$\chi_{\mathrm{asym}}(-\mathbf{R}) = -\chi_{\mathrm{asym}}(\mathbf{R}). \tag{4.1243}$$

Defocus, twofold astigmatism, spherical aberration, and fourfold astigmatism belong to the symmetric family. Axial coma and trefoil provide important examples of antisymmetric aberrations.

This distinction is especially significant because the derivation of the phase-contrast CTF in Sec. 4.10.10 involved the pair of Fourier components at $\mathbf{R}$ and $-\mathbf{R}$. For a purely symmetric aberration, both acquire the same aberration phase. An antisymmetric aberration instead introduces opposite phase changes for the two members of the pair. Consequently, sufficiently general aberrations cannot always be represented by a purely real-valued CTF; the coherent transfer must then retain its complex phase explicitly (Zivanov et al., 2020).

Chromatic aberrations

The aberrations considered above describe the deterministic action of the microscope on an electron of given energy. A different family arises because real electron beams are not exactly monoenergetic and because lens excitation is not perfectly constant.

The focal properties of an electron lens depend on electron energy and lens strength. Electrons of slightly different energy therefore have slightly different Gaussian image planes. The leading axial effect is characterized by the chromatic-aberration coefficient C_c (Hawkes and Kasper, 2017)[Chap. 26].

Chromatic aberration must be distinguished conceptually from a deterministic pupil phase such as spherical aberration. Each electron of a given energy propagates coherently through the microscope with its own transfer function, but the detected image contains contributions from an ensemble of energies. The statistical framework required to combine these contributions has already been developed in Sec. 3.11.4; in particular, Eq. (3.766) expresses the detected intensity as an average over the energy distribution.

Thus chromatic aberration provides one of the physical mechanisms behind the temporal-coherence envelope that will later accompany the CTF.

Image distortions

Not every imperfection of a cryo-EM image is most naturally represented as a phase error in the objective pupil. Some imperfections alter the mapping of spatial coordinates themselves.

Classical distortion is one of the five Seidel aberrations. Modern cryo-EM also encounters effects such as anisotropic magnification, in which the magnification differs slightly between two transverse directions. Such effects modify the coordinates at which Fourier components are observed rather than merely multiplying those components by an aberration phase.

For this reason, anisotropic magnification is conveniently treated separately from symmetric and antisymmetric phase aberrations (Zivanov et al., 2020). We shall preserve this distinction throughout the chapter.

Zernike polynomials: a representation, not a new family

Once sufficiently high-order aberrations are considered, listing every possible term individually becomes cumbersome. A systematic representation of the wavefront error is then useful.

For a circular pupil, the Zernike polynomials

$$Z_n^m(\rho, \phi) \tag{4.1244}$$

provide a convenient basis in which an arbitrary sufficiently smooth aberration function can be expanded. Familiar aberrations correspond to particular low-order Zernike modes: defocus, twofold astigmatism, coma, trefoil, spherical aberration, fourfold astigmatism, and progressively higher-order terms all appear naturally within the same hierarchy.

It is important, however, not to confuse the mathematical representation with the physical origin of an aberration. A Zernike polynomial is a basis function over the pupil. It does not by itself explain why the electron lens produces that wavefront error. The physical origin will instead be connected below with the higher-order terms of the electron action. Zernike polynomials will later provide a compact and systematic language for representing the resulting aberration function.

A unified view

The different classifications introduced above can be summarized as follows:

viewpoint	quantity examined	typical examples
classical geometrical	ray displacement	$C,\ K,\ A,\ F,\ D$
cryo-EM phase	$\chi(\mathbf{R})$	defocus, C_s, astigmatism, coma
parity	$\chi(\mathbf{R})$ vs. $\chi(-\mathbf{R})$	symmetric / antisymmetric
wavefront basis	$\chi(\rho, \phi)$	Zernike modes
chromatic	energy-dependent focusing	C_c
geometrical distortion	$\mathbf{r} \mapsto \mathbf{r}'$	distortion, anisotropic magnification.

(4.1245)

These descriptions will be connected by a single underlying quantity in the next subsection: the aberration of the electron action. We shall see that the departure of the action from its Gaussian value can be interpreted simultaneously as a wavefront error, as an additional quantum-mechanical phase, and, through its derivatives, as a geometrical ray aberration. This connection will provide the basis for deriving the standard aberration phase functions used in cryo-EM rather than introducing them phenomenologically.

4.11.2 The aberration function as excess action and wavefront error

The different descriptions introduced in Sec. 4.11.1 can be connected through the characteristic function of the electron trajectory. This provides a particularly useful viewpoint because it links directly the analytical mechanics developed in Sec. 4.4, the semiclassical interpretation of the electron wave developed in Sec. 4.5, and the coherent transfer function of the objective lens introduced in Eq. (4.1237).

For a static electron-optical system, Hamilton's characteristic function is the reduced action

$$W = \int \mathbf{P} \cdot d\mathbf{r}, \tag{4.1246}$$

as established generally in Eq. (4.467) and specialized to an electron in static electromagnetic fields in Eq. (4.744). Equivalently, using the optical-axis coordinate z as trajectory parameter,

$$W = \int \overline{M}\left(x, y, z, x', y'\right) dz, \tag{4.1247}$$

according to Eq. (4.760).

Thus every physical electron trajectory carries with it a value of the reduced action. In the semiclassical regime this same quantity controls the phase of the electron wave. Indeed, Sec. 4.5.1 showed that the stationary electron wave has the WKB form

$$\psi_E(\mathbf{r}) \simeq A(\mathbf{r}) \exp\left[\frac{j}{\hbar} W(\mathbf{r})\right] \tag{4.1248}$$

[Eq. (4.618)], and the eikonal analysis extended this identification to static electromagnetic fields through Eq. (4.987).

The Gaussian action as reference

Gaussian optics provides a natural reference against which aberrations can be measured. In Sec. 4.10.2, the exact electron-optical Lagrangian was expanded as

$$\overline{M} = \overline{M}^{(0)} + \overline{M}^{(2)} + \overline{M}^{(4)} + \cdots \tag{4.1249}$$

according to Eq. (4.1025). Retaining only the zeroth- and quadratic-order terms gives Gaussian optics.

We therefore define the Gaussian reduced action as

$$\boxed{W_{\mathrm{G}} \triangleq \int \left(\overline{M}^{(0)} + \overline{M}^{(2)}\right) dz.} \tag{4.1250}$$

The actual electron lens contains the higher-order contributions omitted from Eq. (4.1250). Its characteristic function can therefore be written as

$$\boxed{W = W_{\mathrm{G}} + \Delta W_{\mathrm{ab}},} \tag{4.1251}$$

where

$$\boxed{\Delta W_{\mathrm{ab}} \triangleq W - W_{\mathrm{G}}} \tag{4.1252}$$

is the *aberration action*, or excess action relative to Gaussian optics.

The meaning of Eq. (4.1252) is geometrical as well as mechanical. Consider a point object and the family of rays emerging from the objective lens. Gaussian optics predicts a spherical wavefront whose center is the Gaussian image point. The actual lens generally produces a slightly different, aspherical wavefront. The difference between the actual and Gaussian characteristic functions measures this departure. This is the electron-optical counterpart of the classical wave-aberration construction in geometrical optics (Hawkes and Kasper, 2017)[Chap. 24].

Thus, what appears as a displacement of rays in geometrical optics is the same physical imperfection that appears as a deformation of a wavefront in wave optics.

From excess action to phase aberration

The connection with the electron wave follows immediately from Eq. (4.618). Substituting Eq. (4.1251),

$$\psi_E \simeq A \exp\left[\frac{j}{\hbar}\left(W_{\mathrm{G}} + \Delta W_{\mathrm{ab}}\right)\right] \tag{4.1253}$$

$$= A \exp\left(\frac{jW_{\mathrm{G}}}{\hbar}\right) \exp\left(\frac{j\Delta W_{\mathrm{ab}}}{\hbar}\right). \tag{4.1254}$$

The first factor contains the phase predicted by ideal Gaussian optics. The second factor is therefore the additional phase produced by the aberration:

$$\boxed{\frac{\psi_E}{\psi_{\mathrm{G}}} = \exp\left(\frac{j\Delta W_{\mathrm{ab}}}{\hbar}\right).} \tag{4.1255}$$

Our coherent-transfer convention, however, was defined in Eq. (4.1237) as

$$H(\mathbf{R}) = P(\mathbf{R}) e^{-j\chi(\mathbf{R})}. \tag{4.1256}$$

Consequently, with the sign conventions adopted in this book, the aberration phase χ_{ab} is related to the excess action by

$$\boxed{\chi_{\mathrm{ab}} = -\frac{\Delta W_{\mathrm{ab}}}{\hbar}.} \tag{4.1257}$$

The minus sign in Eq. (4.1257) is not an additional physical effect. It results entirely from the two conventions already adopted: the semiclassical electron wave was written with $\exp(jW/\hbar)$ in Eq. (4.618), whereas the microscope transfer function was written as $\exp(-j\chi)$ in Eq. (4.1237). Other conventions may therefore place the opposite sign in the definition of the aberration function.

Wavefront error in units of length

The action difference has units of action. In geometrical optics, wavefront aberrations are instead often described as a distance between the actual wavefront and an ideal reference wavefront.

This interpretation is particularly simple in a field-free region after the lens, where the momentum magnitude p is constant and $\mathbf{A} = \mathbf{0}$. From Eq. (4.744),

$$dW = \mathbf{p} \cdot d\mathbf{r}. \tag{4.1258}$$

Along the direction of propagation, Eq. (4.747) gives

$$dW = p\,ds. \tag{4.1259}$$

An action difference ΔW_{ab} may therefore be expressed as an equivalent path-length difference

$$\boxed{\Delta\ell_{\mathrm{ab}} \triangleq \frac{\Delta W_{\mathrm{ab}}}{p}.} \tag{4.1260}$$

Using the de Broglie relation

$$k = \frac{p}{\hbar} \tag{4.1261}$$

from Eq. (4.918), we obtain

$$\frac{\Delta W_{\mathrm{ab}}}{\hbar} = k\Delta\ell_{\mathrm{ab}} = \frac{2\pi}{\lambda}\Delta\ell_{\mathrm{ab}}. \tag{4.1262}$$

Hence,

$$\boxed{\chi_{\mathrm{ab}} = -k\Delta\ell_{\mathrm{ab}} = -\frac{2\pi}{\lambda}\Delta\ell_{\mathrm{ab}}.} \tag{4.1263}$$

Equation (4.1263) explains why extremely small departures of an electron wavefront can have measurable consequences. At typical cryo-EM accelerating voltages, the electron wavelength is only a few picometres, so a wavefront error much smaller than an atomic dimension can already correspond to a substantial phase shift.

The gradient of the aberration function gives the ray error

The same excess action also determines the geometrical ray aberration. For a time-independent system, Hamilton–Jacobi theory gave

$$\mathbf{P} = \nabla W \tag{4.1264}$$

in Eq. (4.523). For the Gaussian reference system,

$$\mathbf{P}_{\mathrm{G}} = \nabla W_{\mathrm{G}}. \tag{4.1265}$$

Subtracting these two relations and using Eq. (4.1252),

$$\boxed{\Delta\mathbf{P} = \mathbf{P} - \mathbf{P}_{\mathrm{G}} = \nabla\Delta W_{\mathrm{ab}}.} \tag{4.1266}$$

In the field-free region in which the wavefront is evaluated, $\mathbf{A} = \mathbf{0}$, so canonical and mechanical momenta coincide according to Eq. (4.697). The transverse part of Eq. (4.1266) is therefore

$$\Delta\mathbf{p}_{\perp} = \nabla_{\perp}\Delta W_{\mathrm{ab}}. \tag{4.1267}$$

For paraxial rays,

$$\boldsymbol{\alpha} \simeq \frac{\mathbf{p}_\perp}{p}, \tag{4.1268}$$

so the angular deviation from the Gaussian ray is

$$\boxed{\Delta\boldsymbol{\alpha} \simeq \frac{1}{p}\nabla_\perp \Delta W_{\rm ab} = \nabla_\perp \Delta\ell_{\rm ab}.} \tag{4.1269}$$

Equation (4.1269) expresses the familiar geometrical statement that rays are normal to the wavefront. If the wavefront is distorted, its local normal changes; the gradient of the wavefront error therefore determines the corresponding change of ray direction.

We have thus obtained the first important unification:

$$\boxed{\begin{array}{ccc} \Delta W_{\rm ab} & \longrightarrow & \chi_{\rm ab} = -\dfrac{\Delta W_{\rm ab}}{\hbar} \\ \text{excess action} & & \text{phase aberration} \\ \Big\downarrow \nabla_\perp & & \Big\downarrow e^{-j\chi_{\rm ab}} \\ \Delta\mathbf{p}_\perp & & \text{wave transfer} \\ \Big\downarrow /p & & \\ \Delta\boldsymbol{\alpha} & & \\ \text{ray aberration.} & & \end{array}} \tag{4.1270}$$

Why higher-order action terms generate aberrations

We can now return to the expansion of the electron-optical Lagrangian in Eq. (4.1025). Gaussian optics retains

$$\overline{M}^{(0)} + \overline{M}^{(2)}. \tag{4.1271}$$

For a rotationally symmetric lens, the first neglected contribution is $\overline{M}^{(4)}$. It therefore provides the leading correction to the Gaussian action.

A useful property of stationary action allows this correction to be calculated, to leading order, along the *Gaussian* trajectory rather than along the unknown aberrated trajectory.

Leading aberration action from the Gaussian trajectory.

Introduce a bookkeeping parameter ϵ and write

$$\overline{M}_\epsilon = \overline{M}_{\mathrm{G}} + \epsilon \overline{M}^{(4)}, \tag{4.1272}$$

where

$$\overline{M}_{\mathrm{G}} = \overline{M}^{(0)} + \overline{M}^{(2)}. \tag{4.1273}$$

Let the stationary trajectory corresponding to $\overline{M}_\epsilon$ be written as a small perturbation of the Gaussian trajectory,

$$x_\epsilon(z) = x_{\mathrm{G}}(z) + \epsilon \eta_x(z) + O(\epsilon^2), \tag{4.1274}$$

$$y_\epsilon(z) = y_{\mathrm{G}}(z) + \epsilon \eta_y(z) + O(\epsilon^2). \tag{4.1275}$$

Consequently,

$$x'_\epsilon(z) = x'_{\mathrm{G}}(z) + \epsilon \eta'_x(z) + O(\epsilon^2), \tag{4.1276}$$

$$y'_\epsilon(z) = y'_{\mathrm{G}}(z) + \epsilon \eta'_y(z) + O(\epsilon^2). \tag{4.1277}$$

The corresponding action is

$$W_\epsilon = \int_{z_0}^{z_1} \overline{M}_\epsilon \left(x_\epsilon, y_\epsilon, z, x'_\epsilon, y'_\epsilon\right) dz, \tag{4.1278}$$

which has exactly the same argument structure as Eq. (4.760).

Expanding to first order in ϵ around the Gaussian trajectory gives

$$\begin{aligned} W_\epsilon = W_{\mathrm{G}} + \epsilon \int_{z_0}^{z_1} \Bigg[& \frac{\partial \overline{M}_{\mathrm{G}}}{\partial x} \eta_x + \frac{\partial \overline{M}_{\mathrm{G}}}{\partial y} \eta_y \\ & + \frac{\partial \overline{M}_{\mathrm{G}}}{\partial x'} \eta'_x + \frac{\partial \overline{M}_{\mathrm{G}}}{\partial y'} \eta'_y + \overline{M}^{(4)} \left(x_{\mathrm{G}}, y_{\mathrm{G}}, z, x'_{\mathrm{G}}, y'_{\mathrm{G}}\right) \Bigg] dz \\ & + O(\epsilon^2), \end{aligned} \tag{4.1279}$$

where all derivatives of $\overline{M}_{\mathrm{G}}$ are evaluated along the Gaussian trajectory

$$\left(x_{\mathrm{G}}(z), y_{\mathrm{G}}(z), z, x'_{\mathrm{G}}(z), y'_{\mathrm{G}}(z)\right). \tag{4.1280}$$

We now integrate by parts the terms containing η_x' and η_y':

$$\int_{z_0}^{z_1} \frac{\partial \overline{M}_{\mathrm{G}}}{\partial x'} \eta_x' \, dz = \left[\frac{\partial \overline{M}_{\mathrm{G}}}{\partial x'} \eta_x \right]_{z_0}^{z_1} - \int_{z_0}^{z_1} \frac{d}{dz} \left(\frac{\partial \overline{M}_{\mathrm{G}}}{\partial x'} \right) \eta_x \, dz, \tag{4.1281}$$

$$\int_{z_0}^{z_1} \frac{\partial \overline{M}_{\mathrm{G}}}{\partial y'} \eta_y' \, dz = \left[\frac{\partial \overline{M}_{\mathrm{G}}}{\partial y'} \eta_y \right]_{z_0}^{z_1} - \int_{z_0}^{z_1} \frac{d}{dz} \left(\frac{\partial \overline{M}_{\mathrm{G}}}{\partial y'} \right) \eta_y \, dz. \tag{4.1282}$$

Substituting these expressions into Eq. (4.1279),

$$\begin{aligned} W_\epsilon = W_{\mathrm{G}} &+ \epsilon \int_{z_0}^{z_1} \left[\frac{\partial \overline{M}_{\mathrm{G}}}{\partial x} - \frac{d}{dz} \left(\frac{\partial \overline{M}_{\mathrm{G}}}{\partial x'} \right) \right] \eta_x \, dz \\ &+ \epsilon \int_{z_0}^{z_1} \left[\frac{\partial \overline{M}_{\mathrm{G}}}{\partial y} - \frac{d}{dz} \left(\frac{\partial \overline{M}_{\mathrm{G}}}{\partial y'} \right) \right] \eta_y \, dz \\ &+ \epsilon \left[\frac{\partial \overline{M}_{\mathrm{G}}}{\partial x'} \eta_x + \frac{\partial \overline{M}_{\mathrm{G}}}{\partial y'} \eta_y \right]_{z_0}^{z_1} \\ &+ \epsilon \int_{z_0}^{z_1} \overline{M}^{(4)} \left(x_{\mathrm{G}}, y_{\mathrm{G}}, z, x_{\mathrm{G}}', y_{\mathrm{G}}' \right) dz + O(\epsilon^2). \end{aligned} \tag{4.1283}$$

We compare the perturbed and Gaussian trajectories while keeping the boundary coordinates fixed:

$$x_\epsilon(z_0) = x_{\mathrm{G}}(z_0), \qquad x_\epsilon(z_1) = x_{\mathrm{G}}(z_1), \tag{4.1284}$$

$$y_\epsilon(z_0) = y_{\mathrm{G}}(z_0), \qquad y_\epsilon(z_1) = y_{\mathrm{G}}(z_1). \tag{4.1285}$$

Using Eqs. (4.1274) and (4.1275), this implies

$$\boxed{\eta_x(z_0) = \eta_x(z_1) = \eta_y(z_0) = \eta_y(z_1) = 0.} \tag{4.1286}$$

This is a boundary condition of the variational problem, not a consequence of the equations of motion. The boundary term in Eq. (4.1283) therefore vanishes.

Furthermore, the Gaussian trajectory is stationary with respect to the Gaussian action. Hence x_{G} and y_{G} satisfy the two Euler–Lagrange

equations

$$\frac{d}{dz}\left(\frac{\partial \overline{M}_{\mathrm{G}}}{\partial x'}\right) - \frac{\partial \overline{M}_{\mathrm{G}}}{\partial x} = 0, \tag{4.1287}$$

$$\frac{d}{dz}\left(\frac{\partial \overline{M}_{\mathrm{G}}}{\partial y'}\right) - \frac{\partial \overline{M}_{\mathrm{G}}}{\partial y} = 0. \tag{4.1288}$$

The first two integrals in Eq. (4.1283) therefore vanish.

Consequently, to leading order,

$$\boxed{\Delta W_{\mathrm{ab}}^{(4)} = \int_{z_0}^{z_1} \overline{M}^{(4)}\left(x_{\mathrm{G}}, y_{\mathrm{G}}, z, x'_{\mathrm{G}}, y'_{\mathrm{G}}\right) dz.} \tag{4.1289}$$

This result is extremely useful. The leading aberration does not require us first to solve the complete nonlinear trajectory problem. We may calculate the higher-order contribution to the action along the already known Gaussian rays. This is the basis of the eikonal or characteristic-function treatment of electron-optical aberrations (Hawkes and Kasper, 2017)[Chap. 24].

It also explains a convention that can otherwise be confusing. Since $\overline{M}^{(4)}$ is fourth order in the transverse coordinates and slopes, the corresponding excess action is quartic in the parameters describing the paraxial ray. But a geometrical ray aberration is obtained by differentiating that action, as in Eq. (4.1266). Therefore,

$$\boxed{\begin{array}{c}\text{fourth-order action or wavefront error}\\ \Downarrow\ \nabla \\ \text{third-order ray aberration.}\end{array}} \tag{4.1290}$$

Thus the familiar "third-order spherical aberration" of geometrical electron optics originates from a fourth-order contribution to the characteristic function. This distinction between the order of the wavefront error and the order of the resulting ray displacement will remain important when higher-order aberrations are introduced.

From the pupil aberration function to Fourier space

For image formation we are particularly interested in how the excess action varies with propagation direction. Let

$$\boldsymbol{\alpha} = (\alpha_x, \alpha_y) \tag{4.1291}$$

denote the paraxial direction of a ray or plane-wave component. The same direction can be labeled by the transverse spatial frequency $\mathbf{R} = (R_x, R_y)$ because, according to Eq. (4.1126),

$$\boxed{\boldsymbol{\alpha} \simeq \lambda \mathbf{R}.} \tag{4.1292}$$

For a fixed object point, the aberration action may therefore be regarded as a function either of pupil direction,

$$\Delta W_{\mathrm{ab}} = \Delta W_{\mathrm{ab}}(\boldsymbol{\alpha}), \tag{4.1293}$$

or of spatial frequency,

$$\Delta W_{\mathrm{ab}} = \Delta W_{\mathrm{ab}}(\lambda \mathbf{R}). \tag{4.1294}$$

Using Eq. (4.1257), the coherent transfer through the aberrated objective becomes

$$\boxed{H(\mathbf{R}) = P(\mathbf{R}) \exp\left[\frac{j}{\hbar} \Delta W_{\mathrm{ab}}(\lambda \mathbf{R})\right] = P(\mathbf{R}) e^{-j\chi_{\mathrm{ab}}(\mathbf{R})}.} \tag{4.1295}$$

Equation (4.1295) is the central connection needed for cryo-EM. The same function that describes the departure of the electron trajectories from Gaussian optics becomes, in the wave description, a spatial-frequency-dependent phase shift.

Defocus fits naturally into this framework as a quadratic wavefront mismatch, although we shall continue to separate the deliberately chosen defocus from the intrinsic lens aberrations, as in Eq. (4.1238). The aberration function $\chi_{\mathrm{ab}}(\mathbf{R})$ will now be built progressively, beginning with the simplest anisotropic quadratic term: twofold astigmatism.

4.11.3 Symmetry, pupil coordinates, and aberration order

In Sec. 4.11.2, we showed that the aberration of an electron lens can be described equivalently by an excess action, a wavefront error, or a phase aberration. For image formation, it is particularly convenient to regard this aberration as a function of the direction with which a ray, or equivalently a plane-wave component, crosses the objective pupil.

This change of variables is useful because the symmetry of the electron lens then places strong restrictions on the functional form of the aberration. Many of the familiar aberration families can in fact be identified before solving any detailed electron-trajectory equation.

Coordinates in the objective pupil

Let the small transverse angle of a paraxial ray with respect to the optical axis be

$$\boxed{\boldsymbol{\alpha} = (\alpha_x, \alpha_y).} \tag{4.1296}$$

Its magnitude and azimuthal angle are

$$\boxed{\alpha = \sqrt{\alpha_x^2 + \alpha_y^2}, \qquad \phi = \operatorname{atan2}(\alpha_y, \alpha_x),} \tag{4.1297}$$

so that

$$\alpha_x = \alpha \cos\phi, \qquad \alpha_y = \alpha \sin\phi. \tag{4.1298}$$

These angular coordinates have already appeared in another language. The connection between propagation angle and spatial frequency was established in Eq. (4.1126),

$$\boxed{\alpha_x \simeq \lambda R_x, \qquad \alpha_y \simeq \lambda R_y.} \tag{4.1299}$$

Consequently,

$$\boxed{\alpha \simeq \lambda R, \qquad R = \sqrt{R_x^2 + R_y^2},} \tag{4.1300}$$

and the azimuthal angle ϕ is the same whether the transverse direction is represented by $\boldsymbol{\alpha}$ or by $\mathbf{R} = (R_x, R_y)$.

The same variables can also be interpreted as physical coordinates in the back focal plane. From Eq. (4.1129),

$$\mathbf{r}_F = \lambda f \mathbf{R} \simeq f \boldsymbol{\alpha}. \tag{4.1301}$$

Thus the following three quantities label the same paraxial component:

$$\boxed{\begin{array}{ccccc} \text{ray direction} & \longleftrightarrow & \text{spatial frequency} & \longleftrightarrow & \text{back-focal-plane position} \\ \boldsymbol{\alpha} & \longleftrightarrow & \mathbf{R} & \longleftrightarrow & \mathbf{r}_F \end{array}} \tag{4.1302}$$

with

$$\boxed{\boldsymbol{\alpha} \simeq \lambda \mathbf{R}, \qquad \mathbf{r}_F \simeq f\boldsymbol{\alpha} = \lambda f \mathbf{R}.} \tag{4.1303}$$

For a circular objective aperture with maximum accepted semi-angle $\alpha_{\max}$, it will later be useful to introduce a normalized pupil radius

$$\boxed{\rho \triangleq \frac{\alpha}{\alpha_{\max}} = \frac{R}{R_{\max}} = \frac{\|\mathbf{r}_F\|}{r_{F,\max}}, \qquad 0 \le \rho \le 1.} \tag{4.1304}$$

The pair (ρ, ϕ) will be the natural coordinate system when we later introduce Zernike polynomials.

Expanding a smooth aberration around the optical axis

For a fixed object point close to the optical axis, let

$$\Delta\ell_{\mathrm{ab}}(\alpha_x, \alpha_y) \tag{4.1305}$$

denote the wavefront error introduced in Eq. (4.1260). The variables α_x and α_y describe the transverse direction of the ray, as defined in Eq. (4.1296).

The axial ray corresponds to

$$\alpha_x = \alpha_y = 0. \tag{4.1306}$$

If the wavefront error varies smoothly with ray direction in a neighborhood of this ray, it can be expanded in a two-dimensional Taylor series. This expansion provides a natural definition of the *order* of an aberration.

Expansion of the wavefront error by polynomial order.

The Taylor expansion of $\Delta\ell_{\mathrm{ab}}(\alpha_x, \alpha_y)$ around the axial ray is

$$\Delta\ell_{\mathrm{ab}}(\alpha_x, \alpha_y) = \sum_{p=0}^{\infty}\sum_{q=0}^{\infty} \frac{1}{p!\,q!} \left.\frac{\partial^{p+q}\Delta\ell_{\mathrm{ab}}}{\partial\alpha_x^p \partial\alpha_y^q}\right|_{(0,0)} \alpha_x^p \alpha_y^q. \tag{4.1307}$$

Each monomial

$$\alpha_x^p \alpha_y^q \tag{4.1308}$$

has the *total degree*

$$N = p + q. \tag{4.1309}$$

We can therefore rearrange the Taylor series by collecting all monomials with the same total degree N. This gives

$$\boxed{\Delta\ell_{\mathrm{ab}} = \Delta\ell^{(0)} + \Delta\ell^{(1)} + \Delta\ell^{(2)} + \Delta\ell^{(3)} + \cdots,} \tag{4.1310}$$

where the contribution of degree N is

$$\boxed{\Delta\ell^{(N)}(\alpha_x, \alpha_y) = \sum_{p=0}^{N} c_{p,N-p}\, \alpha_x^p \alpha_y^{N-p},} \tag{4.1311}$$

with

$$c_{p,q} = \frac{1}{p!\,q!}\,\frac{\partial^{p+q}\Delta\ell_{\rm ab}}{\partial\alpha_x^p\partial\alpha_y^q}\bigg|_{(0,0)}. \tag{4.1312}$$

Equation (4.1311) immediately shows why each fixed-order contribution is a *homogeneous polynomial*. Every monomial in $\Delta\ell^{(N)}$ contains exactly N factors of the transverse angular coordinates.

To see this explicitly, multiply both angular coordinates by the same factor κ. Then

$$\Delta\ell^{(N)}(\kappa\alpha_x,\kappa\alpha_y) = \sum_{p=0}^{N} c_{p,N-p}(\kappa\alpha_x)^p(\kappa\alpha_y)^{N-p} \tag{4.1313}$$

$$= \sum_{p=0}^{N} c_{p,N-p}\kappa^p\kappa^{N-p}\alpha_x^p\alpha_y^{N-p} \tag{4.1314}$$

$$= \kappa^N\sum_{p=0}^{N} c_{p,N-p}\alpha_x^p\alpha_y^{N-p}. \tag{4.1315}$$

Hence,

$$\boxed{\Delta\ell^{(N)}(\kappa\alpha_x,\kappa\alpha_y) = \kappa^N\Delta\ell^{(N)}(\alpha_x,\alpha_y).} \tag{4.1316}$$

Thus homogeneity is not an additional assumption about the aberration: it follows automatically from grouping the Taylor expansion according to total polynomial degree.

The scaling relation in Eq. (4.1316) has a simple geometrical interpretation. For $\kappa > 0$, multiplying both transverse angular components by κ changes the magnitude of the ray angle but not its azimuthal direction:

$$\alpha \longrightarrow \kappa\alpha, \qquad \phi \longrightarrow \phi, \tag{4.1317}$$

where α and ϕ are defined in Eq. (4.1297).

Consequently, the entire dependence of an Nth-degree term on the magnitude of the ray angle can be factored out.

Separation of radial and angular dependence.

Using

$$\alpha_x = \alpha\cos\phi, \qquad \alpha_y = \alpha\sin\phi, \tag{4.1318}$$

in Eq. (4.1311) gives

$$\Delta\ell^{(N)}(\alpha,\phi) = \sum_{p=0}^{N} c_{p,N-p}(\alpha\cos\phi)^p(\alpha\sin\phi)^{N-p} \tag{4.1319}$$

$$= \alpha^N \sum_{p=0}^{N} c_{p,N-p}\cos^p\phi\,\sin^{N-p}\phi. \tag{4.1320}$$

Defining

$$F_N(\phi) \triangleq \sum_{p=0}^{N} c_{p,N-p}\cos^p\phi\,\sin^{N-p}\phi, \tag{4.1321}$$

we obtain

$$\boxed{\Delta\ell^{(N)}(\alpha,\phi) = \alpha^N F_N(\phi).} \tag{4.1322}$$

Equation (4.1322) separates two different properties of an aberration. The integer N determines how its magnitude scales with the distance from the optical axis in the pupil, whereas the function $F_N(\phi)$ describes how that aberration varies with azimuthal direction.

For example, a rotationally symmetric aberration has no dependence on ϕ, whereas an anisotropic aberration has a nonconstant $F_N(\phi)$. We shall encounter specific examples of these angular dependences in the following subsections: twofold astigmatism, spherical aberration, coma, and higher-order aberrations.

A systematic classification of the possible angular dependences $F_N(\phi)$ will be postponed until the introduction of Zernike polynomials, where radial order and azimuthal symmetry can be treated within a single framework.

Because the phase aberration is proportional to the wavefront error according to Eq. (4.1263),

$$\chi_{\mathrm{ab}} = -\frac{2\pi}{\lambda}\Delta\ell_{\mathrm{ab}}, \tag{4.1323}$$

the same polynomial degree N applies to the action, the wavefront error, and the phase aberration.

The first few orders already have a direct physical interpretation. A constant term changes only the global phase of the wave and therefore has no observable effect on the image intensity. A linear term in the pupil coordinates produces a transverse displacement of the coherent image; in the Fourier-conjugate description, a linear phase across a real-space plane represents a tilted wavefront. Quadratic terms describe focusing, while their anisotropic part produces twofold astigmatism. These statements follow

directly from the Fourier and paraxial relations already established in the preceding sections.

Linear phase: translation and wavefront tilt.

Consider the most general first-degree wavefront error in pupil-angle coordinates,

$$\Delta\ell^{(1)}(\alpha_x, \alpha_y) = a_x\alpha_x + a_y\alpha_y, \tag{4.1324}$$

where a_x and a_y have dimensions of length.

Using the paraxial relation

$$\boldsymbol{\alpha} \simeq \lambda\mathbf{R} \tag{4.1325}$$

from Eq. (4.1299), Eq. (4.1324) becomes

$$\Delta\ell^{(1)}(\mathbf{R}) = \lambda\left(a_xR_x + a_yR_y\right). \tag{4.1326}$$

From Eq. (4.1263),

$$\chi^{(1)}(\mathbf{R}) = -2\pi\left(a_xR_x + a_yR_y\right). \tag{4.1327}$$

The corresponding coherent transfer factor is therefore

$$\boxed{H^{(1)}(\mathbf{R}) = e^{-j\chi^{(1)}(\mathbf{R})} = \exp\left[j2\pi\left(a_xR_x + a_yR_y\right)\right].} \tag{4.1328}$$

Let $\widehat{\psi}_i(\mathbf{R})$ be the wave before this phase term is applied. The output wave is

$$\psi_o(\mathbf{r}) = \iint \widehat{\psi}_i(\mathbf{R})H^{(1)}(\mathbf{R})e^{j2\pi\mathbf{R}\cdot\mathbf{r}}\,d\mathbf{R} \tag{4.1329}$$

$$= \iint \widehat{\psi}_i(\mathbf{R})e^{j2\pi\mathbf{R}\cdot(\mathbf{r}+\mathbf{a})}\,d\mathbf{R}, \tag{4.1330}$$

where

$$\mathbf{a} = (a_x, a_y). \tag{4.1331}$$

By the inverse Fourier-transform definition of Sec. 3.1,

$$\boxed{\psi_o(\mathbf{r}) = \psi_i(\mathbf{r}+\mathbf{a}).} \tag{4.1332}$$

Thus, under the Fourier and phase conventions adopted here, a linear phase in pupil coordinates produces a rigid transverse translation of the coherent image. It does not blur the image or change its internal structure.

The preceding calculation also clarifies the geometrical meaning of a first-order wavefront error. From Eq. (4.1269),

$$\Delta\boldsymbol{\alpha} = \nabla_{\boldsymbol{\alpha}} \Delta\ell_{\mathrm{ab}}^{(1)} \tag{4.1333}$$

up to the corresponding choice of transverse pupil coordinates. Since $\Delta\ell_{\mathrm{ab}}^{(1)}$ is linear, its gradient is constant. Consequently, every ray receives the same transverse correction. The ray bundle is therefore displaced as a whole rather than distorted, in agreement with the Fourier-space result of Eq. (4.1332).

For this reason, a first-order term is usually not regarded as an intrinsic image aberration: it can be absorbed into the definition of the chief ray or of the image origin.

A closely related situation occurs if the actual incident wave is tilted with respect to the optical axis used to define the Gaussian reference system. Suppose that the reference wave is taken to be untilted, whereas the actual incident wave contains the factor

$$\exp\left(j\mathbf{k}_\perp \cdot \mathbf{r}\right). \tag{4.1334}$$

Its phase differs from that of the reference wave by

$$\Delta\varphi(\mathbf{r}) = \mathbf{k}_\perp \cdot \mathbf{r}, \tag{4.1335}$$

which is linear in the transverse coordinates. Since

$$\mathbf{k}_\perp \simeq k\boldsymbol{\alpha}_0, \tag{4.1336}$$

the corresponding propagation direction is

$$\boxed{\boldsymbol{\alpha}_0 \simeq \frac{\mathbf{k}_\perp}{k}.} \tag{4.1337}$$

Thus a tilted illumination also produces a first-order difference with respect to an untilted Gaussian reference. The physical origin is different, however. A linear term generated by the lens description corresponds to

a common displacement of the ray bundle, whereas a linear term already present in the incident wave indicates that the actual chief ray is tilted relative to the one chosen as the Gaussian reference.

In both cases the essential geometrical fact is the same: a first-order wavefront error has a constant gradient and therefore changes the direction or position of the ray bundle without deforming it. Genuine image aberrations begin when the gradient varies across the pupil, as occurs for quadratic and higher-order terms.

Quadratic phase: focusing and twofold astigmatism.

The most general real quadratic wavefront error in the pupil coordinates is

$$\boxed{\Delta\ell^{(2)} = \frac{1}{2}\begin{pmatrix}\alpha_x & \alpha_y\end{pmatrix}\mathbf{Q}\begin{pmatrix}\alpha_x\\ \alpha_y\end{pmatrix},} \tag{4.1338}$$

where

$$\mathbf{Q} = \begin{pmatrix} q_{xx} & q_{xy} \\ q_{xy} & q_{yy}\end{pmatrix} \tag{4.1339}$$

is a real symmetric matrix.

Because $\mathbf{Q}$ is symmetric, its principal axes can be found by a rotation in the transverse plane. Let q_1 and q_2 be its two eigenvalues, and let ϕ_{ast} denote the orientation of the first principal axis. In coordinates aligned with these axes,

$$\Delta\ell^{(2)} = \frac{1}{2}\left(q_1\alpha_1^2 + q_2\alpha_2^2\right). \tag{4.1340}$$

Writing

$$\alpha_1 = \alpha\cos(\phi - \phi_{\text{ast}}), \qquad \alpha_2 = \alpha\sin(\phi - \phi_{\text{ast}}), \tag{4.1341}$$

Eq. (4.1340) becomes

$$\Delta\ell^{(2)} = \frac{\alpha^2}{2}\left[q_1\cos^2(\phi - \phi_{\text{ast}}) + q_2\sin^2(\phi - \phi_{\text{ast}})\right]. \tag{4.1342}$$

Using

$$\cos^2\theta = \frac{1+\cos 2\theta}{2}, \qquad \sin^2\theta = \frac{1-\cos 2\theta}{2}, \tag{4.1343}$$

we obtain

$$\boxed{\Delta\ell^{(2)} = \frac{\alpha^2}{2}\left[q_{\rm av} + q_{\rm ast}\cos 2(\phi-\phi_{\rm ast})\right],} \tag{4.1344}$$

where

$$\boxed{q_{\rm av} = \frac{q_1+q_2}{2}, \qquad q_{\rm ast} = \frac{q_1-q_2}{2}.} \tag{4.1345}$$

The first term in Eq. (4.1344) is rotationally symmetric:

$$\Delta\ell^{(2)}_{\rm iso} = \frac{q_{\rm av}}{2}\alpha^2. \tag{4.1346}$$

Using $\alpha \simeq \lambda R$ and Eq. (4.1263),

$$\chi^{(2)}_{\rm iso}(R) = -\frac{2\pi}{\lambda}\frac{q_{\rm av}}{2}\lambda^2 R^2 \tag{4.1347}$$

$$= -\pi\lambda q_{\rm av} R^2. \tag{4.1348}$$

But the defocus phase was obtained previously as

$$\chi_{\rm def}(R) = -\pi\lambda\Delta f R^2 \tag{4.1349}$$

in Eq. (4.1191). Therefore,

$$\boxed{q_{\rm av} = \Delta f.} \tag{4.1350}$$

The rotationally symmetric quadratic wavefront term is therefore exactly a focusing or defocusing term.

The second contribution,

$$\Delta\ell^{(2)}_{\rm ast} = \frac{q_{\rm ast}}{2}\alpha^2\cos 2(\phi-\phi_{\rm ast}), \tag{4.1351}$$

has a twofold angular dependence. The corresponding phase is

$$\chi^{(2)}_{\rm ast}(R,\phi) = -\pi\lambda q_{\rm ast} R^2\cos 2(\phi-\phi_{\rm ast}). \tag{4.1352}$$

Thus the complete quadratic phase may be written as

$$\boxed{\chi^{(2)}(R,\phi) = -\pi\lambda R^2\left[q_{\mathrm{av}} + q_{\mathrm{ast}}\cos 2(\phi-\phi_{\mathrm{ast}})\right].} \tag{4.1353}$$

The quadratic wavefront therefore separates uniquely into an isotropic part, which changes the focus equally in every direction, and an anisotropic part, which gives different focusing powers along two orthogonal principal directions. The latter is precisely the structure of twofold astigmatism.

Angular symmetry of a polynomial aberration

The Cartesian expansion in α_x and α_y is useful for derivations, but polar coordinates reveal the symmetry much more clearly. A homogeneous polynomial of degree N cannot have an arbitrary angular dependence: only certain angular harmonics are permitted.

Allowed angular symmetries at a given order.

Introduce the complex combinations

$$\alpha_+ \triangleq \alpha_x + j\alpha_y, \qquad \alpha_- \triangleq \alpha_x - j\alpha_y. \tag{4.1354}$$

Using Eq. (4.1297),

$$\alpha_+ = \alpha e^{j\phi}, \qquad \alpha_- = \alpha e^{-j\phi}. \tag{4.1355}$$

Any homogeneous polynomial of degree N in α_x, α_y can equivalently be written as a linear combination of products

$$\alpha_+^p \alpha_-^q, \qquad p+q=N. \tag{4.1356}$$

Such a product has the polar form

$$\alpha_+^p \alpha_-^q = \alpha^{p+q} e^{j(p-q)\phi} \tag{4.1357}$$

$$= \alpha^N e^{jm\phi}, \tag{4.1358}$$

where

$$m = p - q. \tag{4.1359}$$

Since p and q are non-negative integers and $p + q = N$, the possible absolute values of m are

$$\boxed{m = N,\, N-2,\, N-4,\, \ldots} \tag{4.1360}$$

ending at $m = 0$ for even N and at $m = 1$ for odd N.

Because the physical wavefront error is real, the complex exponentials combine into real sine and cosine terms. A generic contribution may therefore be written as

$$\boxed{\Delta\ell_{N,m}(\alpha,\phi) \propto \alpha^N \cos\left[m(\phi - \phi_{N,m})\right],} \tag{4.1361}$$

where $\phi_{N,m}$ specifies the orientation of the aberration.

Equation (4.1360) is a powerful symmetry result. For example,

$$N = 2 \quad \Longrightarrow \quad m = 0, 2, \tag{4.1362}$$
$$N = 3 \quad \Longrightarrow \quad m = 1, 3, \tag{4.1363}$$
$$N = 4 \quad \Longrightarrow \quad m = 0, 2, 4. \tag{4.1364}$$

The integer m measures the azimuthal symmetry of the wavefront error. A term with $m = 2$ has twofold angular symmetry, one with $m = 3$ has threefold symmetry, and one with $m = 4$ has fourfold symmetry. Importantly, m alone does not specify the aberration: the same angular symmetry may occur at different radial orders N.

Consequences of rotational symmetry

Consider now a perfectly rotationally symmetric objective lens and an object point on its optical axis. Rotating the entire ray configuration around the optical axis cannot change its action. The aberration wavefront must therefore satisfy

$$\boxed{\Delta\ell_{\mathrm{ab}}(\alpha,\phi+\theta) = \Delta\ell_{\mathrm{ab}}(\alpha,\phi) \qquad \text{for every } \theta.} \tag{4.1365}$$

The aberration can consequently have no dependence on ϕ, which means that only

$$m = 0 \tag{4.1366}$$

is allowed.

From Eq. (4.1360), $m = 0$ occurs only for even N. Hence an axially symmetric wavefront expansion contains terms of the form

$$\boxed{\Delta\ell_{\text{axial}}(\alpha) = a_0 + a_2\alpha^2 + a_4\alpha^4 + a_6\alpha^6 + \cdots .} \tag{4.1367}$$

The constant a_0 is an irrelevant common phase. The quadratic term belongs to Gaussian focusing; a residual quadratic term appears when the observation plane is displaced from the Gaussian image plane, producing the defocus phase derived in Sec. 4.10.8. Once the Gaussian contribution has been removed, the first non-Gaussian rotationally symmetric term is therefore proportional to

$$\alpha^4. \tag{4.1368}$$

This is the term that will give rise to spherical aberration. The next axially symmetric contribution is proportional to α^6 and leads to the higher-order spherical aberration conventionally associated with C_5.

Departures from perfect rotational symmetry permit $m \neq 0$. The first examples are already familiar from electron microscopy:

$$\begin{aligned}
N = 2,\ m = 2 \quad &\longrightarrow \quad \text{twofold astigmatism,} \\
N = 3,\ m = 1 \quad &\longrightarrow \quad \text{axial coma,} \\
N = 3,\ m = 3 \quad &\longrightarrow \quad \text{trefoil or threefold astigmatism,} \\
N = 4,\ m = 4 \quad &\longrightarrow \quad \text{fourfold astigmatism.}
\end{aligned} \tag{4.1369}$$

Higher orders continue the same hierarchy (Hawkes and Kasper, 2017; Zivanov et al., 2020).

Symmetric and antisymmetric aberrations from parity

The same expansion immediately explains the symmetric and antisymmetric classification introduced in Eqs. (4.1240) and (4.1241).

Changing from $\mathbf{R}$ to $-\mathbf{R}$ reverses the transverse propagation direction. In polar coordinates,

$$\mathbf{R} \longrightarrow -\mathbf{R} \qquad \Longleftrightarrow \qquad \phi \longrightarrow \phi + \pi, \tag{4.1370}$$

while R, and therefore α, remains unchanged.

Applying this transformation to Eq. (4.1361),

$$\cos\left[m(\phi + \pi - \phi_{N,m})\right] = \cos\left[m(\phi - \phi_{N,m}) + m\pi\right] \tag{4.1371}$$

$$= (-1)^m \cos\left[m(\phi - \phi_{N,m})\right]. \tag{4.1372}$$

Hence,

$$\boxed{\chi_{N,m}(-\mathbf{R}) = (-1)^m \chi_{N,m}(\mathbf{R}).} \tag{4.1373}$$

But Eq. (4.1360) shows that m and N always have the same parity. Therefore,

$$\boxed{\chi^{(N)}(-\mathbf{R}) = (-1)^N \chi^{(N)}(\mathbf{R}).} \tag{4.1374}$$

We therefore obtain the simple rule

$$\boxed{\begin{array}{lcll} N \text{ even} & \Longrightarrow & \chi(-\mathbf{R}) = \chi(\mathbf{R}) & \text{symmetric,} \\ N \text{ odd} & \Longrightarrow & \chi(-\mathbf{R}) = -\chi(\mathbf{R}) & \text{antisymmetric.} \end{array}} \tag{4.1375}$$

Thus defocus and twofold astigmatism, both quadratic, are symmetric; axial coma and trefoil, both cubic in the wavefront coordinates, are antisymmetric; and spherical aberration and fourfold astigmatism, whose wavefront terms are quartic, are again symmetric. This is the origin of the symmetric/antisymmetric separation used in modern high-resolution cryo-EM aberration analysis (Zivanov et al., 2020).

Wavefront order versus ray-aberration order

At this point an important terminology issue must be made explicit. The integer N used above is the polynomial degree of the *action, wavefront error, or phase aberration*. It is not generally the order quoted in the traditional names of geometrical aberrations.

As shown in Eq. (4.1290), the ray aberration is obtained by differentiating the aberration action. A wavefront term of degree N therefore produces a ray deviation of degree

$$\boxed{n = N - 1.} \tag{4.1376}$$

The relation between wavefront order and ray-aberration order is particularly clear for spherical aberration. For an axial object point in a rotationally symmetric lens, the wavefront error cannot depend on the azimuthal angle. After the quadratic Gaussian term has been removed, the first rotationally symmetric contribution is of fourth degree in the pupil angle.

From the spherical wavefront error to the ray aberration.

The most general fourth-degree rotationally symmetric wavefront

term is

$$\boxed{\Delta\ell_{\text{sph}}^{(4)} = A\alpha^4 = A\left(\alpha_x^2 + \alpha_y^2\right)^2,} \tag{4.1377}$$

where

$$\alpha^2 = \alpha_x^2 + \alpha_y^2. \tag{4.1378}$$

As discussed in Sec. 4.11.2, geometrical ray aberrations are obtained from derivatives of the aberration function. Therefore, the dependence of the transverse ray error on pupil angle is determined by the gradient

$$\nabla_{\boldsymbol{\alpha}}\Delta\ell_{\text{sph}}^{(4)}. \tag{4.1379}$$

Taking the derivatives,

$$\frac{\partial\Delta\ell_{\text{sph}}^{(4)}}{\partial\alpha_x} = 4A\alpha_x\left(\alpha_x^2 + \alpha_y^2\right), \tag{4.1380}$$

$$\frac{\partial\Delta\ell_{\text{sph}}^{(4)}}{\partial\alpha_y} = 4A\alpha_y\left(\alpha_x^2 + \alpha_y^2\right). \tag{4.1381}$$

Hence,

$$\boxed{\nabla_{\boldsymbol{\alpha}}\Delta\ell_{\text{sph}}^{(4)} = 4A\alpha^2\boldsymbol{\alpha}.} \tag{4.1382}$$

Since

$$\|\boldsymbol{\alpha}\| = \alpha, \tag{4.1383}$$

the magnitude of the corresponding geometrical ray aberration varies as

$$\boxed{\Delta r_{\text{sph}} \propto 4A\alpha^3.} \tag{4.1384}$$

Thus a fourth-degree wavefront error produces a third-degree ray aberration. The reduction of the order by one is simply the consequence of differentiating the wavefront function.

The coefficient of this cubic ray aberration is conventionally written as

$$\boxed{\Delta r_{\text{sph}} = C_3\alpha^3} \tag{4.1385}$$

up to the sign convention used to define the transverse ray error. The subscript in C_3 therefore refers to the *third-order geometrical ray aberration*, even though the corresponding wavefront error is fourth degree.

In electron microscopy, this same coefficient is traditionally called the

spherical-aberration coefficient and denoted by C_s. Accordingly, throughout this book we shall identify

$$\boxed{C_s \equiv C_3.} \tag{4.1386}$$

Comparing Eqs. (4.1382) and (4.1385), the corresponding fourth-degree wavefront term may therefore be written, apart from the sign convention used for the wavefront error, as

$$\boxed{\Delta\ell^{(4)}_{\text{sph}} = \pm\frac{C_s}{4}\alpha^4 = \pm\frac{C_3}{4}\alpha^4.} \tag{4.1387}$$

The relation between the two orders can therefore be summarized as

$$\boxed{\underbrace{\Delta\ell^{(4)}_{\text{sph}} \propto C_s\alpha^4}_{\text{fourth-degree wavefront error}} \xrightarrow{\text{differentiate}} \underbrace{\Delta r_{\text{sph}} \propto C_s\alpha^3}_{\text{third-order ray aberration}}.} \tag{4.1388}$$

Using the paraxial pupil relation Eq. (4.1300), $\alpha \simeq \lambda R$, the same fourth-degree wavefront term produces a phase aberration proportional to

$$\chi_{\text{sph}}(R) \propto C_3\lambda^3 R^4, \tag{4.1389}$$

where one power of λ is removed when the wavefront error is converted into phase through Eq. (4.1263). The exact numerical factor and sign will be derived when spherical aberration is treated explicitly.

The first few orders may therefore be organized as follows:

This distinction will be maintained throughout the following derivations: we shall use N when referring to the degree of the wavefront or phase polynomial, while traditional symbols such as C_3 and C_5 will retain their conventional electron-optical meaning.

Connection with the Zernike representation

The structure derived above also anticipates the Zernike representation that will be introduced later. On the normalized circular pupil (ρ, ϕ) of Eq. (4.1304), Zernike functions are organized by a radial order and an azimuthal order m. Their allowed indices obey the same parity constraint that emerged in Eq. (4.1360): the difference between radial and azimuthal orders is even.

This is not accidental. Both descriptions reflect the rotational symmetry of functions defined over a circular pupil. The Zernike polynomials provide an orthogonal basis on that pupil, whereas the homogeneous expansion used

N	Ray order	Allowed m	Parity	Examples
0	–	0	symmetric	global phase
1	0	1	antisymmetric	linear phase / displacement
2	1	0,2	symmetric	defocus, twofold astigmatism
3	2	1,3	antisymmetric	axial coma, trefoil
4	3	0,2,4	symmetric	spherical aberration, fourfold terms
5	4	1,3,5	antisymmetric	higher-order antisymmetric terms
6	5	0,2,4,6	symmetric	C_5 and higher-order symmetric terms

Table 4.1: Relationship between wavefront degree N, geometrical ray order, allowed azimuthal harmonics, and parity under $\mathbf{R} \to -\mathbf{R}$.

here follows directly from the Taylor expansion of the aberration action. A Zernike mode should therefore not be identified blindly with a single monomial $\alpha^N \cos(m\phi)$: because of their orthogonality construction, Zernike radial polynomials generally contain combinations of several powers of ρ.

We shall postpone the detailed Zernike formalism until the principal physical aberrations have been derived. For the moment, the essential result is that symmetry already organizes the aberration function into a small number of allowed angular families at every order:

$$\boxed{\begin{array}{c} \text{smooth excess action} \\ \Downarrow \\ \text{homogeneous order } N \\ \Downarrow \\ m = N, N-2, N-4, \ldots \\ \Downarrow \\ \begin{cases} N \text{ even}: & \text{symmetric aberration,} \\ N \text{ odd}: & \text{antisymmetric aberration,} \end{cases} \\ \Downarrow \\ \text{specific aberrations: astigmatism, coma, spherical, trefoil, } \ldots \end{array}} \tag{4.1390}$$

The simplest departure from rotationally symmetric focusing is the quadratic term with $N = 2$ and $m = 2$. We shall therefore begin the individual aberrations with twofold astigmatism, which can be understood directly as a

direction-dependent defocus.

4.11.4 Twofold astigmatism as anisotropic defocus

The astigmatism routinely estimated during cryo-EM CTF determination is most naturally understood as a direction-dependent defocus. Instead of having the same quadratic wavefront curvature in every transverse direction, the objective lens has two orthogonal principal directions with slightly different focal conditions.

This form of astigmatism should not be confused with the classical field-dependent Seidel astigmatism introduced in Sec. 4.11.1. Here we are describing an axial, quadratic anisotropy of the objective-lens wavefront.

We shall use a notation close to that commonly employed in cryo-EM. Let

$$\Delta f_M \tag{4.1391}$$

and

$$\Delta f_m \tag{4.1392}$$

denote the two principal defocus values, with

$$\Delta f_M \geq \Delta f_m, \tag{4.1393}$$

and let h denote the orientation of the direction associated with Δf_M, measured from the R_x axis.

In accordance with the convention introduced in Eq. (4.1191), positive values of Δf denote underfocus. In the present subsection, Δf should be understood as the equivalent objective-lens defocus entering the pupil phase, rather than as a literal displacement of the final detector plane.

The quadratic wavefront in its principal directions

For ordinary isotropic defocus, Eq. (4.1191) gives

$$\chi_{\mathrm{def}}(\mathbf{R}) = -\pi\lambda\Delta f\, R^2. \tag{4.1394}$$

Using the relation between wavefront error and phase, Eq. (4.1263), this corresponds to a quadratic wavefront error.

Directional defocus from a quadratic wavefront.

From Eq. (4.1263),

$$\chi_{\text{def}} = -\frac{2\pi}{\lambda}\Delta\ell_{\text{def}}. \tag{4.1395}$$

Combining this with Eq. (4.1394),

$$-\frac{2\pi}{\lambda}\Delta\ell_{\text{def}} = -\pi\lambda\Delta f\, R^2, \tag{4.1396}$$

and therefore

$$\Delta\ell_{\text{def}} = \frac{\Delta f}{2}\lambda^2 R^2. \tag{4.1397}$$

Using the paraxial relation

$$\alpha \simeq \lambda R \tag{4.1398}$$

from Eq. (4.1126), we obtain

$$\boxed{\Delta\ell_{\text{def}} = \frac{\Delta f}{2}\alpha^2.} \tag{4.1399}$$

Now suppose that the quadratic wavefront curvature is different along two orthogonal principal directions. Let these directions be denoted by M and m, with corresponding angular coordinates α_M and α_m. The natural generalization of Eq. (4.1399) is

$$\boxed{\Delta\ell^{(2)}_{\text{ast}} = \frac{1}{2}\left(\Delta f_M \alpha_M^2 + \Delta f_m \alpha_m^2\right).} \tag{4.1400}$$

This expression has the required limiting behavior. Along the M principal direction, $\alpha_m = 0$, and the wavefront is that of ordinary defocus Δf_M. Along the orthogonal direction, $\alpha_M = 0$, and the corresponding defocus is Δf_m.

Thus twofold astigmatism does not require a new type of phase law. It is the most general quadratic focusing term once rotational symmetry has been broken: ordinary defocus is recovered when

$$\Delta f_M = \Delta f_m. \tag{4.1401}$$

Directional defocus

Let

$$\phi = \angle\mathbf{R} \tag{4.1402}$$

be the azimuthal angle of the spatial-frequency vector. Since the principal direction M is oriented at the angle h, the components of the pupil angle along the two principal axes are

$$\alpha_M = \alpha \cos(\phi - h), \tag{4.1403}$$

$$\alpha_m = \alpha \sin(\phi - h). \tag{4.1404}$$

Substitution into Eq. (4.1400) gives

$$\Delta\ell_{\text{ast}}^{(2)} = \frac{\alpha^2}{2}\left[\Delta f_M \cos^2(\phi - h) + \Delta f_m \sin^2(\phi - h)\right]. \tag{4.1405}$$

This motivates the definition of the *directional defocus*

$$\boxed{\Delta f(\phi) = \Delta f_M \cos^2(\phi - h) + \Delta f_m \sin^2(\phi - h).} \tag{4.1406}$$

Equation (4.1406) is the standard astigmatic-defocus expression used in cryo-EM.

Its geometrical meaning is immediate:

$$\Delta f(h) = \Delta f_M, \tag{4.1407}$$

$$\Delta f\left(h + \frac{\pi}{2}\right) = \Delta f_m. \tag{4.1408}$$

Thus the two principal defoci occur along orthogonal directions.

Astigmatic phase in Fourier space

Using Eq. (4.1263) together with Eq. (4.1405),

$$\chi_2(R, \phi) = -\frac{2\pi}{\lambda}\Delta\ell_{\text{ast}}^{(2)} \tag{4.1409}$$

$$= -\frac{\pi}{\lambda}\alpha^2\left[\Delta f_M \cos^2(\phi - h) + \Delta f_m \sin^2(\phi - h)\right]. \tag{4.1410}$$

Since $\alpha \simeq \lambda R$,

$$\boxed{\chi_2(R, \phi) = -\pi\lambda R^2\left[\Delta f_M \cos^2(\phi - h) + \Delta f_m \sin^2(\phi - h)\right].} \tag{4.1411}$$

Equivalently, using Eq. (4.1406),

$$\boxed{\chi_2(R, \phi) = -\pi\lambda\Delta f(\phi) R^2.} \tag{4.1412}$$

The usual isotropic-defocus phase of Eq. (4.1191) is therefore recovered simply by setting

$$\Delta f_M = \Delta f_m = \Delta f. \tag{4.1413}$$

Mean defocus and astigmatic defocus

The physical content of Eq. (4.1406) becomes still clearer if we separate its rotationally symmetric and anisotropic parts. Define

$$\boxed{\Delta f_{\text{av}} \triangleq \frac{\Delta f_M + \Delta f_m}{2}, \qquad \Delta f_{\text{ast}} \triangleq \frac{\Delta f_M - \Delta f_m}{2}.} \tag{4.1414}$$

Using

$$\cos^2\theta = \frac{1+\cos 2\theta}{2}, \qquad \sin^2\theta = \frac{1-\cos 2\theta}{2}, \tag{4.1415}$$

Eq. (4.1406) becomes

$$\boxed{\Delta f(\phi) = \Delta f_{\text{av}} + \Delta f_{\text{ast}} \cos 2(\phi - h).} \tag{4.1416}$$

Consequently, Eq. (4.1412) separates as

$$\boxed{\begin{aligned}\chi_2(R,\phi) = & -\pi\lambda\Delta f_{\text{av}} R^2 \\ & -\pi\lambda\Delta f_{\text{ast}} R^2 \cos 2(\phi - h).\end{aligned}} \tag{4.1417}$$

The first term is ordinary rotationally symmetric defocus. The second is the pure twofold-astigmatism phase:

$$\boxed{\chi_{\text{ast}}(R,\phi) = -\pi\lambda\Delta f_{\text{ast}} R^2 \cos 2(\phi - h).} \tag{4.1418}$$

The name *twofold* follows directly from

$$\cos 2(\phi + \pi - h) = \cos 2(\phi - h), \tag{4.1419}$$

so the aberration repeats after a rotation by π. The maximum-defocus axis is h, while the minimum-defocus axis is $h + \pi/2$.

Care is needed when quoting the "amount of astigmatism". With the definition in Eq. (4.1414), Δf_{ast} is the amplitude of the cosine modulation and is therefore *half* the difference between the two principal defoci:

$$\boxed{\Delta f_M - \Delta f_m = 2\Delta f_{\text{ast}}.} \tag{4.1420}$$

Cryo-EM software packages often report the full difference $\Delta f_M - \Delta f_m$ instead.

Why the directional defocus is not an ellipse

The directional defocus derived in Eq. (4.1406),

$$\Delta f(\phi) = \Delta f_M \cos^2(\phi - h) + \Delta f_m \sin^2(\phi - h), \tag{4.1421}$$

should not be interpreted as the radial equation of an ellipse.

This distinction is important. The quantity $\Delta f(\phi)$ is a *scalar*: for every direction ϕ in the pupil, it gives the effective defocus experienced by that Fourier component. It is obtained by evaluating the quadratic defocus form along a unit direction.

This is precisely the point corrected in the corrigendum to Sorzano et al.: although the maximum and minimum values are Δf_M and Δf_m, the intermediate directional defocus values do not trace an ellipse.

What does have a conic geometry in Fourier space is the set of points that acquire the *same phase*. To see this, introduce spatial-frequency coordinates aligned with the two principal astigmatic directions,

$$R_M = R\cos(\phi - h), \tag{4.1422}$$

$$R_m = R\sin(\phi - h). \tag{4.1423}$$

The quadratic phase of Eq. (4.1411) can then be written as

$$\boxed{\chi_2(R_M, R_m) = -\pi\lambda\left(\Delta f_M R_M^2 + \Delta f_m R_m^2\right).} \tag{4.1424}$$

Thus the geometry observed in Fourier space is determined not by the graph of $\Delta f(\phi)$ itself, but by the level sets of the quadratic phase in Eq. (4.1424). As we shall now see, these level sets can be ellipses or hyperbolas depending on the signs of the two principal defoci.

Equal-phase contours and the sign of the principal defoci

Let us consider the set of spatial frequencies for which the quadratic astigmatic phase has a fixed value,

$$\chi_2(R_M, R_m) = \chi_0. \tag{4.1425}$$

Using Eq. (4.1424), this condition is

$$\boxed{\Delta f_M R_M^2 + \Delta f_m R_m^2 = -\frac{\chi_0}{\pi\lambda}.} \tag{4.1426}$$

The shape of this contour is governed by the signs of Δf_M and Δf_m.

Classification of the equal-phase contours.

For convenience, define

$$c_0 \triangleq -\frac{\chi_0}{\pi\lambda}. \tag{4.1427}$$

Then Eq. (4.1426) becomes

$$\Delta f_M R_M^2 + \Delta f_m R_m^2 = c_0. \tag{4.1428}$$

First suppose that the two principal defoci have the same sign,

$$\Delta f_M \Delta f_m > 0. \tag{4.1429}$$

The quadratic form is then definite. For a value of χ_0 for which the contour exists, Eq. (4.1428) can be written as

$$\boxed{\frac{R_M^2}{|c_0|/|\Delta f_M|} + \frac{R_m^2}{|c_0|/|\Delta f_m|} = 1.} \tag{4.1430}$$

This is an ellipse whose principal axes coincide with the astigmatic directions M and m.

Now suppose that the principal defoci have opposite signs. Without loss of generality, let

$$\Delta f_M > 0, \qquad \Delta f_m < 0. \tag{4.1431}$$

Equation (4.1428) then becomes

$$\Delta f_M R_M^2 - |\Delta f_m| R_m^2 = c_0. \tag{4.1432}$$

If $c_0 > 0$, division by c_0 gives

$$\boxed{\frac{R_M^2}{c_0/\Delta f_M} - \frac{R_m^2}{c_0/|\Delta f_m|} = 1,} \tag{4.1433}$$

which is a hyperbola opening along the R_M direction.

If $c_0 < 0$, Eq. (4.1432) instead becomes

$$\boxed{\frac{R_m^2}{|c_0|/|\Delta f_m|} - \frac{R_M^2}{|c_0|/\Delta f_M} = 1,} \tag{4.1434}$$

which opens along the R_m direction.

The most revealing contour is obtained for zero phase,

$$\chi_0 = 0. \tag{4.1435}$$

Equation (4.1426) then reduces to

$$\Delta f_M R_M^2 + \Delta f_m R_m^2 = 0. \tag{4.1436}$$

For opposite-sign principal defoci,

$$\Delta f_M R_M^2 = |\Delta f_m| R_m^2, \tag{4.1437}$$

and therefore

$$\boxed{R_m = \pm\sqrt{-\frac{\Delta f_M}{\Delta f_m}} R_M.} \tag{4.1438}$$

Thus the zero-phase contour is not a hyperbola but its degenerate limit: a pair of straight lines crossing at the origin.

The three physically distinct situations are illustrated in Fig. 4.9.

If both principal directions are underfocused,

$$\Delta f_M > 0, \qquad \Delta f_m > 0, \tag{4.1439}$$

then from Eq. (4.1424)

$$\chi_2(R_M, R_m) < 0 \tag{4.1440}$$

for every nonzero spatial frequency. The phase surface is locally bowl-shaped, with negative sign according to our convention, and its equal-phase contours are ellipses. The zero-phase contour consists only of the origin.

If both principal directions are overfocused,

$$\Delta f_M < 0, \qquad \Delta f_m < 0, \tag{4.1441}$$

then

$$\chi_2(R_M, R_m) > 0 \tag{4.1442}$$

for every nonzero spatial frequency. The geometry of the contours is unchanged: they are again ellipses. Only the sign of the phase has been reversed.

The qualitatively different case occurs when the two principal defoci have opposite signs. One principal direction is then underfocused while the perpendicular direction is overfocused. The quadratic phase is no longer bowl-shaped but saddle-shaped: moving away from the origin along one principal axis makes χ_2 negative, whereas moving along the other makes it positive.

The two straight lines of Eq. (4.1438) are the directions along which these two contributions exactly cancel. They separate the Fourier plane into regions of positive and negative phase. Nonzero equal-phase contours on either side of these lines are hyperbolas.

The classification can therefore be summarized as

$$\boxed{\begin{array}{rcl} \Delta f_M \Delta f_m > 0 & \Longrightarrow & \text{definite quadratic phase} \Longrightarrow \text{elliptical contours,} \\ \Delta f_M \Delta f_m < 0 & \Longrightarrow & \text{indefinite quadratic phase} \Longrightarrow \text{hyperbolic contours,} \\ \Delta f_M \Delta f_m < 0,\ \chi_2 = 0 & \Longrightarrow & \text{two intersecting straight lines.} \end{array}} \tag{4.1443}$$

There is also a degenerate boundary case in which one of the principal defoci is exactly zero. The phase then has curvature along only one principal direction, and the equal-phase conics degenerate accordingly.

Consequences for CTF zeroes. For a pure weak-phase object, and considering only this quadratic defocus contribution, the CTF is proportional to

$$\mathrm{CTF}(R_M, R_m) \propto \sin \chi_2(R_M, R_m). \tag{4.1444}$$

Its zeroes therefore satisfy

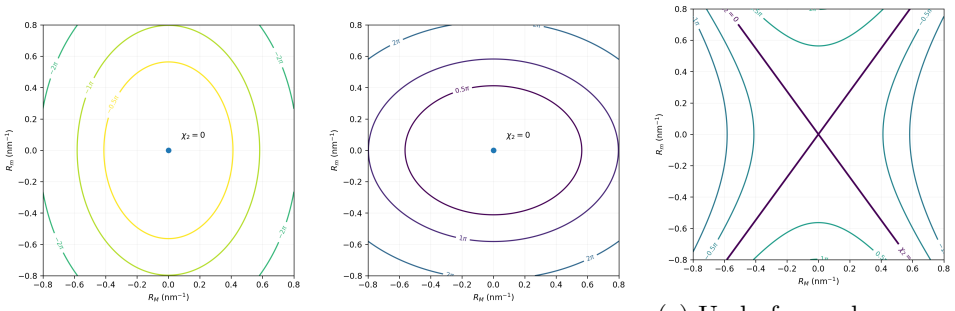
$$\boxed{\chi_2(R_M, R_m) = n\pi, \qquad n \in \mathbb{Z}.} \tag{4.1445}$$

When the two principal defoci have the same sign, the nontrivial $n \neq 0$ CTF zeroes are therefore ellipses. This is the usual situation responsible for the familiar elliptical astigmatic Thon rings.

When the two principal defoci have opposite signs, however, the $n \neq 0$ zeroes are hyperbolas. The $n = 0$ zero is the special pair of straight lines given by Eq. (4.1438).

Thus the common statement that astigmatic CTF zeroes form ellipses implicitly assumes that both principal directions lie on the same side of

focus. If one direction is underfocused and the perpendicular direction is overfocused, the topology of the CTF-zero pattern changes from elliptical to hyperbolic.



(a) Underfocus in both principal directions.

(b) Overfocus in both principal directions.

(c) Underfocus along one principal direction and overfocus along the other.

Figure 4.9: Equal-phase contours of the quadratic astigmatic phase of Eq. (4.1424). (a) When both principal directions are underfocused, $\Delta f_M > 0$ and $\Delta f_m > 0$, the phase is negative away from the origin and the equal-phase contours are ellipses. (b) When both directions are overfocused, $\Delta f_M < 0$ and $\Delta f_m < 0$, the contours remain ellipses but the phase changes sign. (c) When the principal defoci have opposite signs, the quadratic phase has a saddle structure. The nonzero equal-phase contours are hyperbolas, while the zero-phase contour consists of the two straight lines given by Eq. (4.1438).

Physical interpretation

The physical origin of twofold astigmatism is a departure from the rotational symmetry assumed for an ideal round electron lens. In the ideal case, the magnetic field of the objective lens is cylindrically symmetric about the optical axis. Consequently, rotating a ray around that axis does not change the electromagnetic field experienced by the electron, and the focusing strength is the same in every transverse direction.

A real objective lens is never perfectly rotationally symmetric. Small imperfections in the pole pieces, mechanical misalignments, contamination or charging of components, residual magnetization, and other departures from the ideal geometry can introduce an azimuthal dependence into the magnetic field. The lowest-order anisotropic perturbation has twofold, or quadrupolar, symmetry. The lens then focuses electrons slightly more strongly in one transverse direction than in the perpendicular one.

In wavefront language, this means that the quadratic curvature produced by the lens is no longer the same in every azimuthal direction. There are two orthogonal principal directions along which the quadratic form is diagonal. Along one of them the wave behaves as if the microscope had defocus Δf_M, whereas along the perpendicular direction it behaves as if the defocus were Δf_m. Intermediate directions contain the weighted combination given by Eq. (4.1406).

Thus,

$$\Delta f_M - \Delta f_m \tag{4.1446}$$

measures the strength of the twofold anisotropy, while the angle h specifies the orientation of its principal axes. If

$$\Delta f_M = \Delta f_m, \tag{4.1447}$$

the magnetic lens is rotationally symmetric at quadratic order, the anisotropic term disappears, and the phase becomes circularly symmetric.

The connection between magnetic-field asymmetry and the quadratic wavefront can be represented schematically as

$$\boxed{\begin{array}{c}\text{departure from cylindrical symmetry of the magnetic field}\\ \Downarrow\\ \text{direction-dependent focusing strength}\\ \Downarrow\\ \Delta f_M \neq \Delta f_m\\ \Downarrow\\ \text{twofold quadratic wavefront error}\\ \Downarrow\\ \text{twofold astigmatism.}\end{array}} \tag{4.1448}$$

The same interpretation explains how astigmatism is corrected in the microscope. Objective-lens stigmators generate a controllable quadrupolar field whose magnitude and orientation can be adjusted to oppose the unwanted quadrupolar component of the lens field. Ideally, the correction restores equal focusing in the two principal directions, so that

$$\Delta f_M \simeq \Delta f_m. \tag{4.1449}$$

As the difference $\Delta f_M - \Delta f_m$ increases, the quadratic phase becomes progressively more anisotropic, producing the characteristic elliptical constant-phase contours in Fourier space derived in Eq. (4.1430).

Twofold astigmatism is therefore the simplest anisotropic aberration. It does not introduce a higher polynomial order than ordinary defocus: both are quadratic wavefront errors. The difference is physical symmetry. Ordinary defocus corresponds to an isotropic quadratic curvature, whereas twofold astigmatism corresponds to a quadratic curvature with two different principal values, ultimately reflecting a twofold departure from rotational symmetry in the electron-lens field.

4.11.5 Spherical aberration

Twofold astigmatism, discussed in Sec. 4.11.4, arises because the focusing power of the objective lens is slightly different in different azimuthal directions. Spherical aberration has a fundamentally different origin. It is present even in a perfectly rotationally symmetric lens.

The paraxial approximation assumes that the transverse ray angles are small and retains only the lowest-order dependence of the electron trajectory on these angles. In a real magnetic lens, rays passing farther from the optical axis sample regions of the magnetic field for which this linear approximation is no longer exact. Consequently, the focusing strength depends slightly on the aperture angle α.

For the conventional uncorrected round magnetic lens, rays with larger aperture angle are focused more strongly than rays close to the optical axis. They therefore cross the axis closer to the lens. The Gaussian image plane is the common focus only in the paraxial limit $\alpha \to 0$. This is the physical origin of spherical aberration (Hawkes and Kasper, 2017)[Chap. 24].

The physical meaning of spherical aberration is illustrated in Fig. 4.10. In the Gaussian approximation, rays passing through different regions of the objective lens are assumed to converge to the same image point. In a real round magnetic lens with positive C_s, this is only true in the paraxial limit. Rays passing farther from the optical axis experience a larger nonlinear focusing contribution and intersect the axis before the paraxial focus.

Importantly, spherical aberration does not break rotational symmetry. For an axial object point, its magnitude depends only on

$$\alpha = \sqrt{\alpha_x^2 + \alpha_y^2}, \tag{4.1450}$$

and not on the azimuthal angle ϕ.

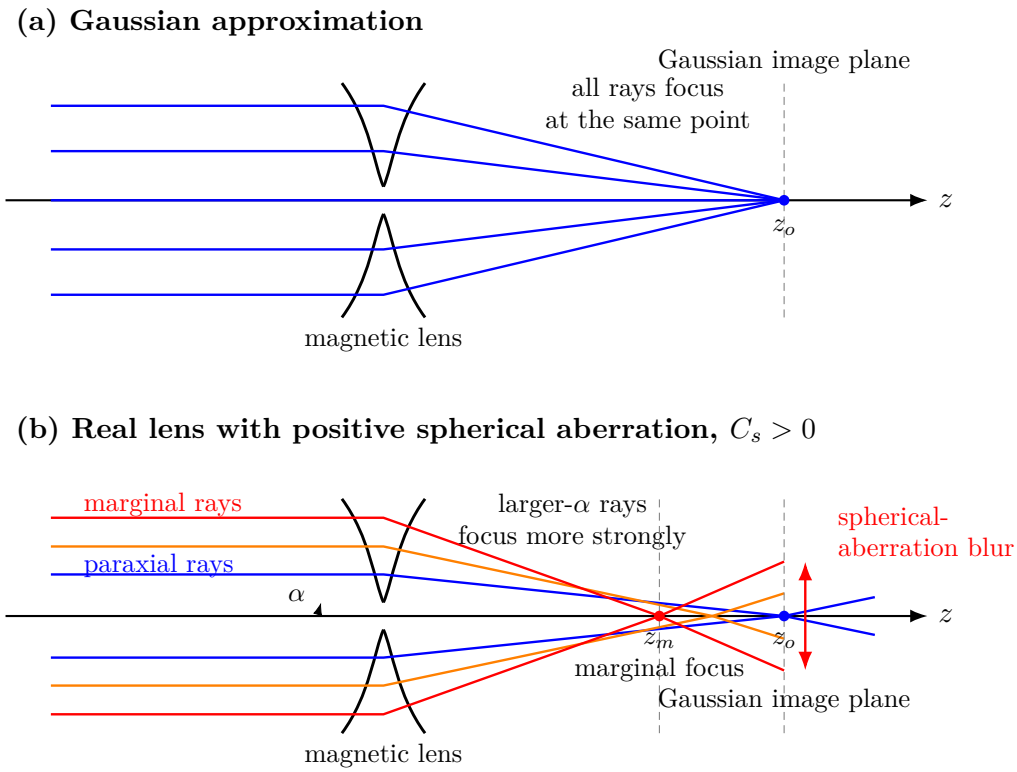


Figure 4.10: Geometrical interpretation of positive spherical aberration. (a) In Gaussian optics, all rays emerging at different aperture angles are assumed to focus at the same image point. (b) In a real rotationally symmetric magnetic lens with $C_s > 0$, marginal rays at larger aperture angles are focused more strongly and cross the optical axis before the paraxial rays. Consequently, at the Gaussian image plane they produce a finite transverse blur. The leading ray displacement scales as $\Delta r_{\text{sph}} \propto C_s \alpha^3$.

From the ray aberration to the wavefront error

The relation between the order of the spherical wavefront error and the corresponding geometrical ray aberration has already been derived above. In particular, Eq. (4.1388) showed that a fourth-degree rotationally symmetric wavefront term produces a third-order transverse ray aberration,

$$\Delta r_{\text{sph}} \propto C_s \alpha^3. \tag{4.1451}$$

The conventional spherical-aberration coefficient is therefore also denoted by C_3, as introduced in Eq. (4.1386),

$$C_s \equiv C_3. \tag{4.1452}$$

We shall not repeat that derivation here. What remains to be established is the exact coefficient and sign of the corresponding phase term under the conventions adopted in this book.

For a positive spherical-aberration coefficient C_s, the aberrated wavefront in image space is more strongly curved than the Gaussian reference wavefront. With the definition of wavefront error $\Delta\ell_{\text{ab}}$ introduced in Eq. (4.1260), this corresponds to

$$\boxed{\Delta\ell_{\text{sph}} = -\frac{C_s}{4}\alpha^4.} \tag{4.1453}$$

The sign in Eq. (4.1453) fixes the previously unspecified sign in Eq. (4.1387). Geometrically, it means that away from the axis the actual wavefront lies closer to the Gaussian image plane than the reference sphere. Its curvature is consequently larger, and the corresponding marginal rays are focused before the Gaussian focus.

Spherical-aberration phase

The phase shift corresponding to Eq. (4.1453) follows directly from the wavefront–phase relation derived in Eq. (4.1263).

Phase produced by spherical aberration.

The aberration phase and wavefront error are related by

$$\chi_{\text{ab}} = -\frac{2\pi}{\lambda}\Delta\ell_{\text{ab}}. \tag{4.1454}$$

Substituting Eq. (4.1453),

$$\chi_{\text{sph}} = -\frac{2\pi}{\lambda}\left(-\frac{C_s}{4}\alpha^4\right) \tag{4.1455}$$

$$= \frac{\pi C_s}{2\lambda}\alpha^4. \tag{4.1456}$$

The paraxial relation between aperture angle and spatial frequency is

$$\alpha \simeq \lambda R \tag{4.1457}$$

[Eq. (4.1300)]. Hence,

$$\chi_{\text{sph}}(R) = \frac{\pi C_s}{2\lambda}(\lambda R)^4 \tag{4.1458}$$

$$= \frac{\pi}{2}C_s\lambda^3 R^4. \tag{4.1459}$$

Therefore,

$$\boxed{\chi_{\text{sph}}(R) = \frac{\pi}{2}C_s\lambda^3 R^4.} \tag{4.1460}$$

Equation (4.1460) is the wave-optical expression of the same phenomenon that, in geometrical optics, produces the cubic ray displacement

$$\Delta r_{\text{sph}} \propto C_s\alpha^3. \tag{4.1461}$$

The different powers are not contradictory: the ray aberration is obtained by differentiating the fourth-degree wavefront error, as explained in Eq. (4.1388).

Why spherical aberration becomes important at high resolution

The dependence

$$\chi_{\text{sph}}(R) \propto R^4 \tag{4.1462}$$

is particularly important in electron microscopy. Close to the origin of Fourier space the spherical-aberration phase is very small, but it grows rapidly with spatial frequency. Thus a microscope may behave almost ideally at low resolution while accumulating a substantial spherical-aberration phase at high resolution.

The corresponding geometrical picture is equally direct. If the objective aperture admits rays up to the semi-angle $\alpha_{\max}$, the radius of the spherical-aberration disc in the Gaussian image plane scales as

$$\boxed{r_{\text{sph}} \propto C_s\alpha_{\max}^3.} \tag{4.1463}$$

Opening the aperture therefore increases the range of transmitted spatial frequencies but simultaneously increases the geometrical effect of spherical aberration very rapidly. This competition is one of the fundamental limitations of an uncorrected objective lens.

Combination with defocus and astigmatism

Spherical aberration and defocus act simultaneously on the same electron wave. Since both are phase contributions to the pupil transfer function, their phases add.

For the astigmatic objective considered in Sec. 4.11.4, the direction-dependent defocus is

$$\Delta f(\phi) = \Delta f_M \cos^2(\phi - h) + \Delta f_m \sin^2(\phi - h) \tag{4.1464}$$

according to Eq. (4.1406). The corresponding quadratic phase is

$$\chi_{\text{def}}(R, \phi) = -\pi\lambda\Delta f(\phi) R^2 \tag{4.1465}$$

[Eq. (4.1412)].

Adding the spherical-aberration phase of Eq. (4.1460) gives

$$\boxed{\chi(R, \phi) = -\pi\lambda\Delta f(\phi) R^2 + \frac{\pi}{2} C_s \lambda^3 R^4.} \tag{4.1466}$$

Equivalently,

$$\boxed{\chi(R, \phi) = -\pi\lambda R^2 \left[\Delta f(\phi) - \frac{1}{2} C_s \lambda^2 R^2\right].} \tag{4.1467}$$

For a non-astigmatic image,

$$\Delta f_M = \Delta f_m = \Delta f, \tag{4.1468}$$

and Eq. (4.1467) reduces to

$$\boxed{\chi(R) = -\pi\lambda R^2 \left(\Delta f - \frac{1}{2} C_s \lambda^2 R^2\right).} \tag{4.1469}$$

This equation shows an important feature of conventional cryo-EM imaging. With the positive-underfocus convention used here, $\Delta f > 0$, the defocus and spherical-aberration contributions enter the phase with opposite signs. Defocus can therefore partially compensate the phase distortion produced by spherical aberration over a finite range of spatial frequencies, although it cannot cancel the quartic aberration at all frequencies simultaneously.

Coherent transfer and the phase-contrast CTF

The corresponding coherent wave-transfer function is obtained by substituting Eq. (4.1466) into Eq. (4.1237):

$$\boxed{H(\mathbf{R}) = P(\mathbf{R}) \exp\left\{-j\left[-\pi\lambda\Delta f(\phi)R^2 + \frac{\pi}{2}C_s\lambda^3R^4\right]\right\}.} \tag{4.1470}$$

As derived previously in Sec. 4.10.10, for a weak phase object and a symmetric aberration phase, the detected phase contrast is proportional to the sine of the coherent phase. Therefore,

$$\mathrm{CTF}(R,\phi) = P(R,\phi)\sin\left[\chi(R,\phi)\right]. \tag{4.1471}$$

Using Eq. (4.1467),

$$\boxed{\mathrm{CTF}(R,\phi) = -P(R,\phi)\sin\left\{\pi\lambda R^2\left[\Delta f(\phi) - \frac{1}{2}C_s\lambda^2R^2\right]\right\}.} \tag{4.1472}$$

Equation (4.1472) is the defocus–astigmatism–spherical-aberration CTF in the pure phase-contrast limit. Amplitude contrast, phase plates, and coherence envelopes have not yet been included and will be introduced later.

For zero astigmatism, Eq. (4.1472) becomes

$$\boxed{\mathrm{CTF}(R) = -P(R)\sin\left[\pi\lambda R^2\left(\Delta f - \frac{1}{2}C_s\lambda^2R^2\right)\right].} \tag{4.1473}$$

Scherzer defocus

Equation (4.1469) shows that defocus and spherical aberration have different dependences on spatial frequency:

$$\boxed{\chi(R) = -\pi\lambda\Delta f\, R^2 + \frac{\pi}{2}C_s\lambda^3R^4.} \tag{4.1474}$$

For positive underfocus, $\Delta f > 0$, the quadratic contribution initially drives the phase towards negative values, whereas the positive quartic spherical-aberration term eventually dominates and drives it back towards positive values.

This competition can be exploited. Rather than attempting to remove spherical aberration, which is impossible by defocus alone because the two terms have different powers of R, one may choose the defocus so that the phase-contrast transfer remains large and of the same sign over the widest useful low-frequency interval. This is the idea behind *Scherzer defocus*.

Before deriving this condition, it is useful to examine the dimensions and physical magnitude of the spherical-aberration coefficient.

Units and physical scale of C_s. The spherical-aberration coefficient has dimensions of length. This can be seen directly from the geometrical ray aberration derived in Eq. (4.1463),

$$\Delta r_{\text{sph}} \propto C_s \alpha^3. \tag{4.1475}$$

Since an angle in radians is dimensionless,

$$\boxed{[C_s] = \text{m}.} \tag{4.1476}$$

The same conclusion follows independently from the phase expression Eq. (4.1460),

$$\chi_{\text{sph}} = \frac{\pi}{2} C_s \lambda^3 R^4. \tag{4.1477}$$

Since χ_{sph} is dimensionless,

$$[C_s]\,[\lambda]^3 [R]^4 = [C_s]\,\text{m}^3\text{m}^{-4} = 1, \tag{4.1478}$$

and therefore again

$$\boxed{[C_s] = \text{m}.} \tag{4.1479}$$

Values of C_s for conventional objective lenses are consequently often quoted in millimetres. For example,

$$C_s = 2.7 \text{ mm} = 2.7 \times 10^{-3} \text{ m} \tag{4.1480}$$

may at first appear enormous compared with the nanometre and Ångström length scales relevant to cryo-EM. This interpretation would be misleading, however: C_s is not itself a displacement of an electron or an image feature. It is the coefficient multiplying a high power of the aperture angle.

For example, at

$$\alpha = 10 \text{ mrad} = 10^{-2}, \tag{4.1481}$$

the characteristic geometrical spherical-aberration scale is only

$$C_s \alpha^3 = (2.7 \times 10^{-3})(10^{-2})^3 = 2.7 \times 10^{-9} \text{ m} = 2.7 \text{ nm}. \tag{4.1482}$$

More revealing for wave optics is the corresponding wavefront error, which from Eq. (4.1387) has magnitude

$$|\Delta \ell_{\text{sph}}| = \frac{C_s}{4} \alpha^4. \tag{4.1483}$$

For the same angle,

$$\boxed{|\Delta \ell_{\text{sph}}| = 6.75 \text{ pm}.} \tag{4.1484}$$

This is an extremely small physical distance. Nevertheless, the electron wavelength at the accelerating voltages used in cryo-EM is itself only a few picometres, as discussed in Sec. 2.4. Consequently, a picometre-scale wavefront error can correspond to several radians of phase. This is why a coefficient measured in millimetres can have a profound effect on atomic-resolution image formation.

Derivation of the Scherzer condition. For a weak phase object, the phase-contrast transfer is governed by the sine of the phase in Eq. (4.1474). The Scherzer condition chooses the defocus so that this sine remains large and of the same sign over a broad interval of spatial frequencies.

Scherzer defocus and Scherzer point resolution.

Starting from Eq. (4.1474),

$$\chi(R) = -\pi\lambda\Delta f R^2 + \frac{\pi}{2} C_s \lambda^3 R^4, \tag{4.1485}$$

the nonzero extremum of the phase is found from

$$\frac{d\chi}{dR} = -2\pi\lambda\Delta f R + 2\pi C_s \lambda^3 R^3 \tag{4.1486}$$

$$= 2\pi\lambda R\left(-\Delta f + C_s\lambda^2 R^2\right) = 0. \tag{4.1487}$$

Besides $R = 0$, the extremum therefore occurs at

$$\boxed{R_{\min}^2 = \frac{\Delta f}{C_s\lambda^2}.} \tag{4.1488}$$

Substituting this frequency back into the phase gives

$$\chi_{\min} = -\pi\lambda\Delta f \frac{\Delta f}{C_s\lambda^2} + \frac{\pi}{2} C_s \lambda^3 \left(\frac{\Delta f}{C_s\lambda^2}\right)^2 \tag{4.1489}$$

$$= -\frac{\pi\Delta f^2}{C_s\lambda} + \frac{\pi\Delta f^2}{2C_s\lambda}, \tag{4.1490}$$

so that

$$\boxed{\chi_{\min} = -\frac{\pi\Delta f^2}{2C_s\lambda}.} \tag{4.1491}$$

The classical Scherzer condition chooses this minimum near

$$\chi_{\min} = -\frac{2\pi}{3}. \tag{4.1492}$$

This makes $\sin\chi$ remain negative and close to its maximum magnitude over a broad frequency interval: on the way to the minimum it passes through $\chi = -\pi/2$, where $\sin\chi = -1$, reaches only $\sin(-2\pi/3) = -\sqrt{3}/2$, and then passes through $-\pi/2$ again before returning to zero.

Combining Eqs. (4.1491) and (4.1492),

$$\frac{\pi \Delta f_{\mathrm{Sch}}^2}{2C_s\lambda} = \frac{2\pi}{3}, \tag{4.1493}$$

and therefore

$$\boxed{\Delta f_{\mathrm{Sch}} = \sqrt{\frac{4}{3}C_s\lambda} \simeq 1.155\sqrt{C_s\lambda}.} \tag{4.1494}$$

With the convention used in this book, $\Delta f_{\mathrm{Sch}} > 0$ corresponds to underfocus.

Because

$$-\pi < \chi_{\min} < 0, \tag{4.1495}$$

the first nontrivial zero of the phase-contrast CTF occurs when the phase returns to zero. From Eq. (4.1474),

$$-\Delta f_{\mathrm{Sch}} + \frac{1}{2}C_s\lambda^2 R_{\mathrm{Sch}}^2 = 0, \tag{4.1496}$$

and therefore

$$R_{\mathrm{Sch}}^2 = \frac{2\Delta f_{\mathrm{Sch}}}{C_s\lambda^2}. \tag{4.1497}$$

Defining the Scherzer point resolution by

$$d_{\mathrm{Sch}} = \frac{1}{R_{\mathrm{Sch}}}, \tag{4.1498}$$

and inserting Eq. (4.1494),

$$d_{\mathrm{Sch}} = \left(\frac{C_s\lambda^2}{2\Delta f_{\mathrm{Sch}}}\right)^{1/2} \tag{4.1499}$$

$$= \left(\frac{3}{16}C_s\lambda^3\right)^{1/4}. \tag{4.1500}$$

Hence,

$$\boxed{d_{\mathrm{Sch}} = \left(\frac{3}{16}C_s\lambda^3\right)^{1/4} \simeq 0.66\left(C_s\lambda^3\right)^{1/4}.} \tag{4.1501}$$

Equations (4.1494) and (4.1501) reveal two important scalings. The optimum defocus increases only as $C_s^{1/2}$, whereas the corresponding point resolution deteriorates as $C_s^{1/4}$. The electron wavelength enters more strongly in the resolution, $d_{\mathrm{Sch}} \propto \lambda^{3/4}$.

Example: a 300 kV cryo-EM microscope. Consider a conventional 300 kV objective lens with

$$C_s = 2.7 \text{ mm}. \tag{4.1502}$$

The relativistic electron wavelength is approximately

$$\lambda \simeq 1.97 \text{ pm}. \tag{4.1503}$$

Equation (4.1494) then gives

$$\boxed{\Delta f_{\mathrm{Sch}} \simeq 84 \text{ nm},} \tag{4.1504}$$

while Eq. (4.1501) gives

$$\boxed{d_{\mathrm{Sch}} \simeq 0.249 \text{ nm} = 2.49 \text{ Å}.} \tag{4.1505}$$

Figure 4.11 shows the corresponding ideal phase-contrast CTF. No amplitude contrast, finite aperture, or coherence envelope has been included, so that the figure isolates the competition between defocus and spherical aberration.

It is striking that the Scherzer point resolution is more than two orders of magnitude larger than the electron wavelength itself. The limitation does not arise because the electron wavelength is too large, but because the objective lens cannot preserve the relative phase of all scattered waves. Spherical aberration converts the very short electron wavelength into high sensitivity to minute wavefront errors.

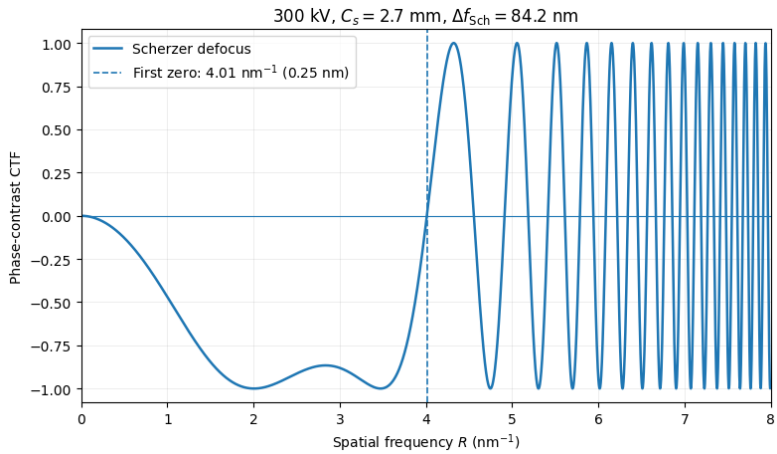


Figure 4.11: Ideal phase-contrast CTF at Scherzer defocus for a 300 kV electron microscope with $C_s = 2.7$ mm. The Scherzer defocus is approximately 84 nm and the first nontrivial zero occurs at $R \simeq 4.01$ nm^{-1}, corresponding to a point resolution of approximately 0.249 nm (2.49 Å). Beyond this first zero the CTF oscillates rapidly and the image no longer transfers all spatial frequencies with the same contrast sign.

How is spherical aberration corrected?

The discussion above naturally raises the question of what is meant by a "C_s-corrected microscope". Is the correction performed numerically after recording the image, or is the electron-optical system itself modified?

In a true C_s-corrected microscope, the correction is primarily *hardware-based*. An additional electron-optical system is inserted into the column and deliberately produces a spherical-aberration coefficient of opposite sign to that of the objective lens.

If

$$C_{s,\mathrm{obj}} > 0 \tag{4.1506}$$

is the spherical-aberration coefficient of the objective lens and

$$C_{s,\mathrm{corr}} < 0 \tag{4.1507}$$

is that introduced by the corrector, the effective coefficient is

$$\boxed{C_s^{\mathrm{eff}} = C_{s,\mathrm{obj}} + C_{s,\mathrm{corr}}.} \tag{4.1508}$$

The corrector is adjusted so that

$$\boxed{C_s^{\text{eff}} \simeq 0,} \tag{4.1509}$$

or, in some imaging modes, to a deliberately chosen small positive or negative residual value.

According to Eq. (4.1460), the corresponding phase becomes

$$\chi_{\text{sph}}(R) = \frac{\pi}{2} C_s^{\text{eff}} \lambda^3 R^4. \tag{4.1510}$$

Thus cancellation of the two C_s coefficients directly suppresses the quartic phase error responsible for spherical aberration.

Why another round lens cannot simply cancel C_s. At first sight one might imagine placing a second rotationally symmetric magnetic lens after the objective and choosing its excitation so that its spherical aberration cancels that of the first lens. This does not work for conventional static round electron lenses. Under the usual electron-optical conditions, their third-order spherical-aberration coefficient has the same sign. This limitation is one of the consequences of the Scherzer theorem.

Therefore, correcting C_s requires additional electron-optical elements that break the rotational symmetry of an ordinary round lens. In practice, these are *multipole lenses*: quadrupoles, hexapoles, octupoles, or suitable combinations of them (Haider et al., 1998).

The apparent paradox is worth emphasizing. Spherical aberration itself is rotationally symmetric, but its correction is achieved using non-rotationally symmetric fields. Several multipole elements are combined so that their unwanted anisotropic aberrations cancel one another while their desired rotationally symmetric contribution to C_s remains.

Hexapole correction. One widely used architecture employs two hexapole elements separated by appropriate transfer optics. Schematically,

$$\boxed{\begin{array}{c} \text{objective lens} \\ \Downarrow \\ +C_s \\ \text{first hexapole} \\ \Downarrow \\ -C_s \;+\; \text{additional anisotropic aberrations} \\ \text{transfer optics} \\ \Downarrow \\ \text{second hexapole} \\ \Downarrow \\ \text{cancellation of parasitic terms} \\ \Downarrow \\ C_s^{\text{eff}} \simeq 0. \end{array}} \tag{4.1511}$$

The first hexapole generates the nonlinear focusing required to produce a negative effective spherical aberration. Because a hexapole field is not rotationally symmetric, it also generates unwanted anisotropic aberrations. The transfer optics and the second hexapole are arranged so that these parasitic contributions largely cancel while the desired negative spherical-aberration contribution adds (Haider et al., 1998).

Alternative corrector designs use combinations of quadrupoles and octupoles. The underlying principle is the same: additional multipole degrees of freedom allow the electron-optical action to acquire a fourth-order term of the sign required to cancel that of the objective lens.

The role of software. Although the correction itself is produced by electromagnetic hardware, modern aberration correctors are strongly software assisted. The currents and alignments of several multipole elements must be adjusted with high precision, and the residual aberrations of the complete system must be measured repeatedly.

The control system therefore follows, schematically, a loop such as

$$\boxed{\begin{array}{c}\text{measure residual aberration}\\ \Downarrow\\ \text{estimate aberration coefficients}\\ \Downarrow\\ \text{change multipole excitations}\\ \Downarrow\\ \text{remeasure}\\ \Downarrow\\ C_s^{\text{eff}}, A_1, B_2, \ldots \text{ minimized.}\end{array}} \tag{4.1512}$$

Software therefore determines how the physical corrector should be excited, but it is the electromagnetic fields produced by the corrector that modify the wavefront before the electron reaches the detector.

Can spherical aberration instead be corrected computationally? There is an important distinction between *correcting the microscope* and *correcting the recorded data*. In conventional single-particle cryo-EM, the microscope frequently has a substantial physical value of C_s, for example several millimetres. This value is included explicitly in the CTF model through Eq. (4.1466).

If C_s and the other CTF parameters are known, their deterministic phase effect can therefore be taken into account computationally during image processing and three-dimensional reconstruction. More general residual symmetric and antisymmetric aberrations can likewise be estimated from a cryo-EM data set and corrected a posteriori (Zivanov et al., 2020).

This computational correction is extremely powerful, but it is not equivalent to hardware C_s correction. Once an image has been detected, information that was strongly attenuated by partial coherence cannot be recreated by multiplying Fourier coefficients by an inverse phase factor. Likewise, a spatial frequency lying exactly at a zero of the CTF contains no linear image contrast in that particular exposure.

In single-particle cryo-EM this problem is alleviated by recording particles at different defocus values. The CTF zeroes then occur at different spatial frequencies in different images, so that information missing from one image may be supplied by others. Nevertheless, hardware correction has the fundamentally different advantage of modifying the electron wave *before* detection.

Other strategies. Several approaches can reduce the consequences of spherical aberration without using a conventional multipole C_s corrector.

The Scherzer defocus derived in in this section uses defocus to oppose the spherical- aberration phase over a finite range of spatial frequencies. It does not remove C_s, because an R^2 term cannot cancel an R^4 term at all frequencies simultaneously.

Reducing the objective aperture also reduces spherical aberration because the geometrical ray error scales as

$$\Delta r_{\text{sph}} \propto C_s \alpha^3 \tag{4.1513}$$

and the wavefront error as

$$\Delta \ell_{\text{sph}} \propto C_s \alpha^4. \tag{4.1514}$$

However, reducing α simultaneously removes the high-angle scattered electrons that carry the highest spatial frequencies, so this strategy trades aberration against resolution.

Objective lenses can also be designed to have a smaller intrinsic C_s, but an ordinary static rotationally symmetric lens cannot in general make the coefficient vanish. True cancellation therefore requires additional degrees of freedom beyond a simple round lens.

Finally, more experimental approaches have demonstrated that a compensating phase can be introduced using nanofabricated electron phase elements, holographic elements, or shaped electromagnetic or optical fields. These methods have the same conceptual objective as the multipole corrector:

$$\boxed{\chi_{\text{corr}}(R) \simeq -\chi_{\text{sph}}(R),} \tag{4.1515}$$

but conventional electromagnetic multipole correctors remain the established approach for aberration-corrected TEM and STEM.

4.11.6 Coma and odd/antisymmetric aberrations

Defocus, twofold astigmatism, and spherical aberration all have one important property in common: their phase is unchanged when the spatial-frequency vector is reversed,

$$\mathbf{R} \longrightarrow -\mathbf{R}. \tag{4.1516}$$

They are therefore *symmetric* aberrations in the sense introduced in Sec. 4.11.1.

Coma introduces a qualitatively different behavior. Its phase changes sign when the spatial frequency is reversed,

$$\boxed{\chi_{\text{coma}}(-\mathbf{R}) = -\chi_{\text{coma}}(\mathbf{R}).} \tag{4.1517}$$

It is consequently an *antisymmetric* aberration. This distinction is much more than a classification of shapes in the pupil. Symmetric and antisymmetric phases affect image formation in different ways. Symmetric aberrations modify the real phase-contrast CTF, whereas antisymmetric aberrations introduce an additional phase rotation of the Fourier coefficients themselves. We shall derive this result below.

Physical origin of axial coma

Classical off-axis coma appears when an object point lies away from the optical axis of a rotationally symmetric lens. In cryo-EM, however, an especially important form is *axial coma*, which can be present even for a nominally axial specimen point if the illuminating beam is tilted with respect to the coma-free axis of the objective lens.

This situation is particularly relevant during microscope alignment and during beam-image-shift data collection. If the incident beam does not travel along the optical axis of the objective lens, the unscattered reference beam and the scattered Fourier components pass through the objective at different positions and angles relative to the lens symmetry axis. The spherical-aberration phase then acts differently on the two members of this system, producing an additional phase term that is odd in spatial frequency.

Figure 4.12 illustrates the two complementary ways of viewing antisymmetric aberrations. In geometrical optics, coma produces an asymmetric point-spread function because different regions of the tilted ray bundle are focused differently. In Fourier optics, the same effect appears as an odd phase function: opposite points of the objective pupil acquire phase shifts of equal magnitude and opposite sign.

Let

$$\mathbf{b} = (b_x, b_y) \tag{4.1518}$$

denote the small angular tilt of the incident beam with respect to the optical axis. More precisely, if $\mathbf{k} = (k_x, k_y, k_z)$ is the wavevector of the incident plane wave, then in the paraxial approximation

$$\boxed{\mathbf{b} \simeq \left(\frac{k_x}{k}, \frac{k_y}{k}\right) \simeq \left(\frac{k_x}{k_z}, \frac{k_y}{k_z}\right).} \tag{4.1519}$$

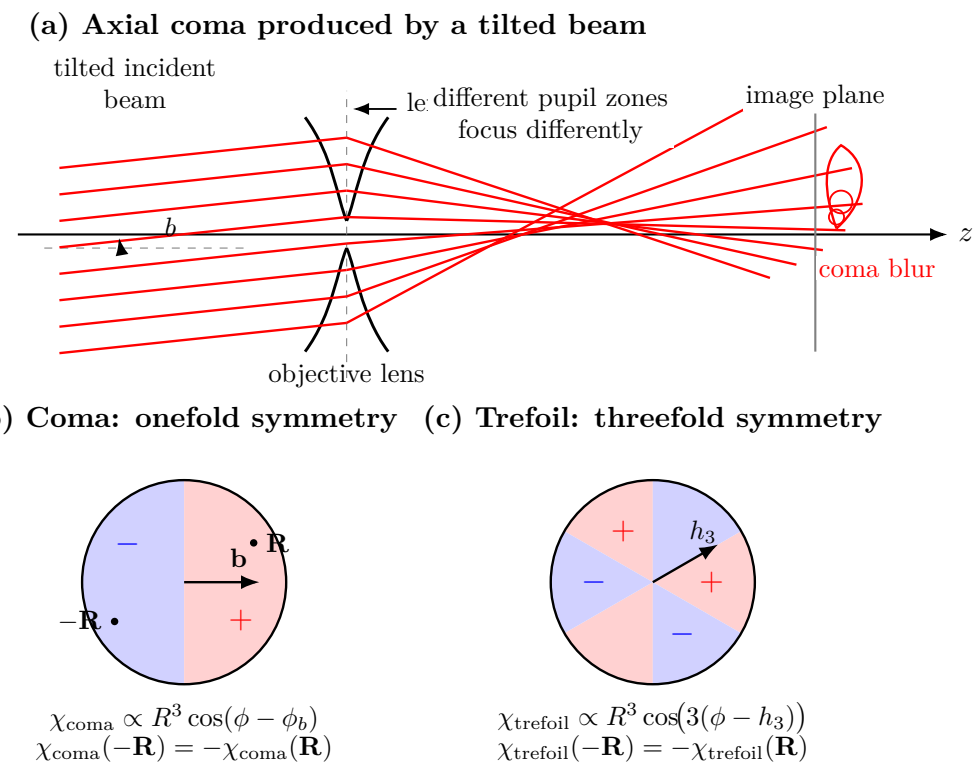


Figure 4.12: Geometrical and pupil-phase interpretation of important antisymmetric aberrations. (a) When the illuminating beam is tilted with respect to the symmetry axis of a lens having spherical aberration, rays passing through different regions of the pupil no longer focus equivalently, producing the characteristic asymmetric coma blur. To first order in beam tilt, the phase varies as $\chi_{\text{coma}} \propto C_s \lambda^2 R^2 (\mathbf{b} \cdot \mathbf{R})$ [Eq. (4.1539)]. (b) Coma has onefold angular symmetry: opposite Fourier components acquire opposite phase shifts. (c) Trefoil has the same cubic radial dependence but a threefold angular dependence. Both are antisymmetric under $\mathbf{R} \to -\mathbf{R}$.

The components b_x and b_y are therefore small angles, numerically equal to the corresponding ray slopes. Their magnitude,

$$\boxed{b = \|\mathbf{b}\| = \sqrt{b_x^2 + b_y^2},} \tag{4.1520}$$

is the total beam-tilt angle. Since angles expressed in radians are dimensionless, $\mathbf{b}$ is dimensionless.

Let

$$\boldsymbol{\alpha} = \lambda \mathbf{R} \tag{4.1521}$$

be the scattering angle associated with the spatial frequency $\mathbf{R}$, according to Eq. (4.1299).

The unscattered beam passes through the lens at the angle $\mathbf{b}$, whereas the scattered component passes through it approximately at

$$\mathbf{b} + \boldsymbol{\alpha}. \tag{4.1522}$$

The additional phase produced by this displacement from the optical axis can therefore be obtained directly from the spherical-aberration phase already derived in Eq. (4.1460).

Beam tilt generates a cubic coma phase

Coma as the interaction between beam tilt and spherical aberration.

In angular coordinates, Eq. (4.1460) may be written as

$$\chi_{\text{sph}}(\boldsymbol{\alpha}) = \frac{\pi C_s}{2\lambda} \|\boldsymbol{\alpha}\|^4, \tag{4.1523}$$

because $\boldsymbol{\alpha} \simeq \lambda \mathbf{R}$.

For an untilted beam, the unscattered reference component corresponds to $\boldsymbol{\alpha} = \mathbf{0}$. If the entire illumination is tilted by $\mathbf{b}$, the unscattered beam instead passes through the objective at $\mathbf{b}$, while a scattered component at relative angle $\boldsymbol{\alpha}$ passes through at $\mathbf{b} + \boldsymbol{\alpha}$.

The spherical-aberration phase difference between the scattered and unscattered components is therefore

$$\Delta\chi_{\text{sph}} = \frac{\pi C_s}{2\lambda} \left[\|\mathbf{b} + \boldsymbol{\alpha}\|^4 - \|\mathbf{b}\|^4 \right]. \tag{4.1524}$$

Now,

$$\|\mathbf{b}+\boldsymbol{\alpha}\|^2 = b^2 + \alpha^2 + 2\mathbf{b}\cdot\boldsymbol{\alpha}, \tag{4.1525}$$

and hence

$$\begin{aligned}\|\mathbf{b}+\boldsymbol{\alpha}\|^4 - b^4 = \alpha^4 &+ 4\alpha^2\mathbf{b}\cdot\boldsymbol{\alpha} \\ &+ 2b^2\alpha^2 + 4\left(\mathbf{b}\cdot\boldsymbol{\alpha}\right)^2 \\ &+ 4b^2\mathbf{b}\cdot\boldsymbol{\alpha}.\end{aligned} \tag{4.1526}$$

The leading term that is new and linear in the beam tilt is

$$4\alpha^2\mathbf{b}\cdot\boldsymbol{\alpha}. \tag{4.1527}$$

The meaning of Eq. (4.1526) becomes clearer if we classify its terms by their order in the small beam tilt $\mathbf{b}$.

For this purpose, $\boldsymbol{\alpha}$ is held fixed: it specifies the scattering angle of the Fourier component whose phase we are evaluating. The small parameter in the present expansion is the illumination tilt b.

Since

$$\mathbf{b}\cdot\boldsymbol{\alpha} \propto b, \tag{4.1528}$$

the different terms have the following dependence on b:

$$\alpha^4 \sim O(b^0), \tag{4.1529}$$

$$4\alpha^2\mathbf{b}\cdot\boldsymbol{\alpha} \sim O(b), \tag{4.1530}$$

$$2b^2\alpha^2 \sim O(b^2), \tag{4.1531}$$

$$4\left(\mathbf{b}\cdot\boldsymbol{\alpha}\right)^2 \sim O(b^2), \tag{4.1532}$$

$$4b^2\mathbf{b}\cdot\boldsymbol{\alpha} \sim O(b^3). \tag{4.1533}$$

Consequently, to first order in the beam tilt, Eq. (4.1526) reduces to

$$\boxed{\|\mathbf{b}+\boldsymbol{\alpha}\|^4 - b^4 = \alpha^4 + 4\alpha^2\mathbf{b}\cdot\boldsymbol{\alpha} + O(b^2).} \tag{4.1534}$$

The first term, α^4, is the spherical-aberration contribution already present for an untilted beam. The second term is the leading change produced by beam tilt:

$$4\alpha^2\mathbf{b}\cdot\boldsymbol{\alpha}. \tag{4.1535}$$

Because it is linear in $\mathbf{b}$ and cubic in the pupil angle $\boldsymbol{\alpha}$, this is the axial-coma term.

Substituting the term linear in $\mathbf{b}$ into Eq. (4.1524), the coma phase is

$$\chi_{\text{coma}} = \frac{\pi C_s}{2\lambda}\left(4\alpha^2 \mathbf{b}\cdot\boldsymbol{\alpha}\right) \tag{4.1536}$$

$$= \boxed{\frac{2\pi C_s}{\lambda}\alpha^2\left(\mathbf{b}\cdot\boldsymbol{\alpha}\right).} \tag{4.1537}$$

Using

$$\boldsymbol{\alpha} = \lambda \mathbf{R}, \tag{4.1538}$$

we obtain

$$\boxed{\chi_{\text{coma}}(\mathbf{R}) = 2\pi C_s \lambda^2 R^2 \left(\mathbf{b}\cdot\mathbf{R}\right).} \tag{4.1539}$$

If the beam tilt has magnitude b and azimuth ϕ_b, then

$$\mathbf{b}\cdot\mathbf{R} = bR\cos(\phi - \phi_b), \tag{4.1540}$$

so Eq. (4.1539) becomes

$$\boxed{\chi_{\text{coma}}(R,\phi) = 2\pi C_s \lambda^2 b R^3 \cos(\phi - \phi_b).} \tag{4.1541}$$

Equation (4.1541) reveals the characteristic structure of axial coma. Its magnitude grows as

$$R^3, \tag{4.1542}$$

and it has a onefold angular dependence

$$\cos(\phi - \phi_b). \tag{4.1543}$$

The phase is maximal along the beam-tilt direction, changes sign in the opposite direction, and vanishes along the perpendicular direction.

Why coma is antisymmetric

The antisymmetry of coma follows immediately from Eq. (4.1539). Under

$$\mathbf{R} \longrightarrow -\mathbf{R}, \tag{4.1544}$$

the factor R^2 is unchanged, while

$$\mathbf{b}\cdot(-\mathbf{R}) = -\mathbf{b}\cdot\mathbf{R}. \tag{4.1545}$$

Therefore,

$$\boxed{\chi_{\text{coma}}(-\mathbf{R}) = -\chi_{\text{coma}}(\mathbf{R}),} \tag{4.1546}$$

as stated in Eq. (4.1517).

In polar coordinates the same result follows from

$$\phi \longrightarrow \phi + \pi, \tag{4.1547}$$

because

$$\cos\left[\phi + \pi - \phi_b\right] = -\cos(\phi - \phi_b). \tag{4.1548}$$

Coma therefore gives opposite phase shifts to the two Fourier components at $\mathbf{R}$ and $-\mathbf{R}$. This is the defining property of an antisymmetric aberration.

The associated wavefront and ray aberration

The coma phase of Eq. (4.1537) corresponds, through Eq. (4.1263), to the wavefront error

$$\boxed{\Delta\ell_{\text{coma}} = -C_s\alpha^2\left(\mathbf{b}\cdot\boldsymbol{\alpha}\right).} \tag{4.1549}$$

Thus coma is a *cubic* wavefront error in the pupil coordinates.

Taking its gradient gives

$$\nabla_{\boldsymbol{\alpha}}\Delta\ell_{\text{coma}} = -C_s\left[2(\mathbf{b}\cdot\boldsymbol{\alpha})\boldsymbol{\alpha} + \alpha^2\mathbf{b}\right]. \tag{4.1550}$$

Hence the associated geometrical ray error is quadratic in aperture angle. Unlike spherical aberration, which is rotationally symmetric, the ray displacement now has a preferred transverse direction determined by $\mathbf{b}$. A point source is therefore spread asymmetrically, producing the characteristic comet-like blur from which coma takes its name.

Care is required with the word "order". Modern cryo-EM literature often refers to coma and trefoil as *third-order aberrations* because their pupil phase varies as R^3. This should not be confused with the notation $C_3 \equiv C_s$ introduced in Eq. (4.1386), where the index 3 refers to the cubic *ray* aberration generated by the quartic spherical-aberration wavefront. Different aberration conventions count different quantities.

Beam tilt also generates symmetric terms

Equation (4.1526) contains, in addition to the spherical-aberration and coma terms, contributions that are second order in the beam tilt:

$$2b^2\alpha^2 + 4\left(\mathbf{b}\cdot\boldsymbol{\alpha}\right)^2. \tag{4.1551}$$

Unlike the coma term, these contributions are unchanged under $\boldsymbol{\alpha} \to -\boldsymbol{\alpha}$ and are therefore symmetric aberrations.

Their quadratic dependence on the pupil coordinates suggests a connection with defocus and twofold astigmatism. To make this explicit, we choose one transverse axis parallel to the beam-tilt vector $\mathbf{b}$ and decompose Eq. (4.1551) into its isotropic and anisotropic parts.

Defocus and twofold astigmatism induced by beam tilt.

Choose the $\parallel$ axis along the beam-tilt direction. Then

$$\mathbf{b} = (b, 0), \qquad \boldsymbol{\alpha} = (\alpha_\parallel, \alpha_\perp), \tag{4.1552}$$

so that

$$\mathbf{b} \cdot \boldsymbol{\alpha} = b\alpha_\parallel, \qquad \alpha^2 = \alpha_\parallel^2 + \alpha_\perp^2. \tag{4.1553}$$

Substituting these expressions into Eq. (4.1551),

$$2b^2\alpha^2 + 4(\mathbf{b} \cdot \boldsymbol{\alpha})^2 = 2b^2\left(\alpha_\parallel^2 + \alpha_\perp^2\right) + 4b^2\alpha_\parallel^2 \tag{4.1554}$$

$$= 6b^2\alpha_\parallel^2 + 2b^2\alpha_\perp^2. \tag{4.1555}$$

We now separate this quadratic form into an isotropic part and an anisotropic part:

$$6b^2\alpha_\parallel^2 + 2b^2\alpha_\perp^2 = 4b^2\left(\alpha_\parallel^2 + \alpha_\perp^2\right) + 2b^2\left(\alpha_\parallel^2 - \alpha_\perp^2\right). \tag{4.1556}$$

The first term is

$$4b^2\left(\alpha_\parallel^2 + \alpha_\perp^2\right) = 4b^2\alpha^2, \tag{4.1557}$$

which is rotationally symmetric and quadratic in pupil angle. It therefore has the form of defocus.

For the second term, let θ be the angle of $\boldsymbol{\alpha}$ with respect to the beam-tilt direction. Then

$$\alpha_\parallel = \alpha\cos\theta, \qquad \alpha_\perp = \alpha\sin\theta, \tag{4.1558}$$

and therefore

$$\alpha_{\parallel}^2 - \alpha_{\perp}^2 = \alpha^2 \left(\cos^2\theta - \sin^2\theta\right) \tag{4.1559}$$

$$= \alpha^2 \cos 2\theta. \tag{4.1560}$$

Hence

$$\boxed{2b^2\alpha^2 + 4(\mathbf{b}\cdot\boldsymbol{\alpha})^2 = 4b^2\alpha^2 + 2b^2\alpha^2\cos 2\theta.} \tag{4.1561}$$

The first term is isotropic defocus, whereas the second has the $\cos 2\theta$ angular dependence characteristic of twofold astigmatism, as derived in Sec. 4.11.4.

To determine the corresponding defocus coefficients, insert the second-order contribution into the spherical-aberration phase:

$$\chi_{b^2} = \frac{\pi C_s}{2\lambda}\left(6b^2\alpha_{\parallel}^2 + 2b^2\alpha_{\perp}^2\right) \tag{4.1562}$$

$$= \frac{\pi C_s b^2}{\lambda}\left(3\alpha_{\parallel}^2 + \alpha_{\perp}^2\right). \tag{4.1563}$$

From Eq. (4.1424), using $\boldsymbol{\alpha} = \lambda\mathbf{R}$, the general quadratic astigmatic-defocus phase can be written as

$$\chi_2 = -\frac{\pi}{\lambda}\left(\Delta f_{\parallel}\alpha_{\parallel}^2 + \Delta f_{\perp}\alpha_{\perp}^2\right). \tag{4.1564}$$

Comparing coefficients with Eq. (4.1563) gives

$$\boxed{\delta\Delta f_{\parallel} = -3C_s b^2, \qquad \delta\Delta f_{\perp} = -C_s b^2.} \tag{4.1565}$$

Using the mean and astigmatic defocus definitions of Eq. (4.1414),

$$\delta\Delta f_{\text{av}} = \frac{\delta\Delta f_{\parallel} + \delta\Delta f_{\perp}}{2} = -2C_s b^2, \tag{4.1566}$$

$$\delta\Delta f_{\text{ast}} = \frac{\delta\Delta f_{\parallel} - \delta\Delta f_{\perp}}{2} = -C_s b^2. \tag{4.1567}$$

Thus

$$\boxed{\delta\Delta f_{\text{av}} = -2C_s b^2, \qquad \delta\Delta f_{\text{ast}} = -C_s b^2.} \tag{4.1568}$$

Thus beam tilt produces not only the coma term that is linear in b.

At second order in the tilt, spherical aberration also generates an isotropic defocus shift and a twofold astigmatic component whose principal axes are parallel and perpendicular to the beam-tilt direction.

This is consistent with the interpretation of astigmatism developed in Sec. 4.11.4. For small beam tilts, however, the cubic coma term is particularly important at high resolution because it grows as R^3.

Even and odd parts of the aberration phase

The preceding example suggests a general decomposition according to the behavior of the aberration phase under inversion of the spatial-frequency vector,

$$\mathbf{R} \longrightarrow -\mathbf{R}. \tag{4.1569}$$

For an arbitrary aberration phase $\chi(\mathbf{R})$, define its even and odd parts as

$$\boxed{\chi_{\text{even}}(\mathbf{R}) = \frac{1}{2}\left[\chi(\mathbf{R}) + \chi(-\mathbf{R})\right],} \tag{4.1570}$$

$$\boxed{\chi_{\text{odd}}(\mathbf{R}) = \frac{1}{2}\left[\chi(\mathbf{R}) - \chi(-\mathbf{R})\right].} \tag{4.1571}$$

Adding Eqs. (4.1570) and (4.1571) immediately gives

$$\boxed{\chi(\mathbf{R}) = \chi_{\text{even}}(\mathbf{R}) + \chi_{\text{odd}}(\mathbf{R}).} \tag{4.1572}$$

The terminology follows directly from their behavior under inversion. From Eq. (4.1570),

$$\chi_{\text{even}}(-\mathbf{R}) = \frac{1}{2}\left[\chi(-\mathbf{R}) + \chi(\mathbf{R})\right] \tag{4.1573}$$

$$= \chi_{\text{even}}(\mathbf{R}), \tag{4.1574}$$

whereas Eq. (4.1571) gives

$$\chi_{\text{odd}}(-\mathbf{R}) = \frac{1}{2}\left[\chi(-\mathbf{R}) - \chi(\mathbf{R})\right] \tag{4.1575}$$

$$= -\chi_{\text{odd}}(\mathbf{R}). \tag{4.1576}$$

Thus

$$\boxed{\begin{aligned} \chi_{\text{even}}(-\mathbf{R}) &= \chi_{\text{even}}(\mathbf{R}), \\ \chi_{\text{odd}}(-\mathbf{R}) &= -\chi_{\text{odd}}(\mathbf{R}). \end{aligned}} \tag{4.1577}$$

The decomposition is unique. Indeed, the values of $\chi(\mathbf{R})$ and $\chi(-\mathbf{R})$ determine $\chi_{\text{even}}(\mathbf{R})$ and $\chi_{\text{odd}}(\mathbf{R})$ uniquely through Eqs. (4.1570) and (4.1571).

Defocus, twofold astigmatism, and spherical aberration belong to the even part, whereas coma and trefoil belong to the odd part. The distinction is particularly important for image formation because, as shown below, the even and odd parts enter the weak-phase transfer function in qualitatively different ways.

Why antisymmetric aberrations make the transfer complex

For a weak pure-phase specimen, the general first-order intensity transfer was already derived in Eq. (4.1219):

$$\boxed{\widehat{\delta I}(\mathbf{R}) = j\left[H(\mathbf{R}) - H^*(-\mathbf{R})\right]\widehat{\varphi}(\mathbf{R}).} \tag{4.1578}$$

That derivation followed from coherent wave propagation, linearization of the detected intensity, and the Hermitian symmetry of the Fourier transform of the real object phase. We now reuse this general result to determine how even and odd aberrations enter the image transfer.

Image transfer in the presence of even and odd aberrations.

Write the coherent transfer function as

$$H(\mathbf{R}) = P(\mathbf{R})\exp\left\{-j\left[\chi_{\text{even}}(\mathbf{R}) + \chi_{\text{odd}}(\mathbf{R})\right]\right\}, \tag{4.1579}$$

and assume, as before, that the pupil is real and even,

$$P(-\mathbf{R}) = P(\mathbf{R}). \tag{4.1580}$$

Using the parity relations of Eq. (4.1577),

$$\chi_{\text{even}}(-\mathbf{R}) = \chi_{\text{even}}(\mathbf{R}), \qquad \chi_{\text{odd}}(-\mathbf{R}) = -\chi_{\text{odd}}(\mathbf{R}), \tag{4.1581}$$

we obtain

$$H^*(-\mathbf{R}) = P(\mathbf{R})\exp\left\{+j\left[\chi_{\text{even}}(\mathbf{R}) - \chi_{\text{odd}}(\mathbf{R})\right]\right\} \tag{4.1582}$$

$$= P(\mathbf{R})e^{-j\chi_{\text{odd}}(\mathbf{R})}e^{+j\chi_{\text{even}}(\mathbf{R})}. \tag{4.1583}$$

Similarly,

$$H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\text{odd}}(\mathbf{R})}e^{-j\chi_{\text{even}}(\mathbf{R})}. \tag{4.1584}$$

Substituting these expressions into Eq. (4.1219),

$$\widehat{\delta I}(\mathbf{R}) = jP(\mathbf{R})e^{-j\chi_{\text{odd}}(\mathbf{R})}\left[e^{-j\chi_{\text{even}}(\mathbf{R})} - e^{+j\chi_{\text{even}}(\mathbf{R})}\right]\widehat{\varphi}(\mathbf{R}) \tag{4.1585}$$

$$= 2P(\mathbf{R})\sin\chi_{\text{even}}(\mathbf{R})e^{-j\chi_{\text{odd}}(\mathbf{R})}\widehat{\varphi}(\mathbf{R}). \tag{4.1586}$$

Therefore,

$$\boxed{\widehat{\delta I}(\mathbf{R}) = 2P(\mathbf{R})\sin\chi_{\text{even}}(\mathbf{R})e^{-j\chi_{\text{odd}}(\mathbf{R})}\widehat{\varphi}(\mathbf{R}).} \tag{4.1587}$$

Equivalently, omitting the same conventional factor of two used when defining the CTF from Eq. (4.1224), the effective complex CTF is

$$\boxed{\text{CTF}_{\text{complex}}(\mathbf{R}) = P(\mathbf{R})\sin\chi_{\text{even}}(\mathbf{R})e^{-j\chi_{\text{odd}}(\mathbf{R})}.} \tag{4.1588}$$

Equation (4.1588) shows why the distinction between symmetric and antisymmetric aberrations is physically important. The even aberration phase determines the familiar oscillating amplitude factor

$$P(\mathbf{R})\sin\chi_{\text{even}}(\mathbf{R}), \tag{4.1589}$$

whereas the odd aberration contributes the unit-modulus factor

$$e^{-j\chi_{\text{odd}}(\mathbf{R})}. \tag{4.1590}$$

Thus an antisymmetric aberration rotates the phase of the transferred Fourier coefficient rather than changing the magnitude of the weak-phase CTF.

In particular,

$$\boxed{|\text{CTF}_{\text{complex}}(\mathbf{R})|^2 = P^2(\mathbf{R})\sin^2\chi_{\text{even}}(\mathbf{R}),} \tag{4.1591}$$

which is independent of χ_{odd}. This explains why aberrations such as coma can substantially corrupt high-resolution Fourier phases while being much less conspicuous in the ordinary power spectrum than defocus or twofold astigmatism.

Connection with the cryo-EM treatment of antisymmetric aberrations

The decomposition in Eq. (4.1588) is closely related to the formulation used by Zivanov et al. (Zivanov et al., 2020). Their image model writes the

effect of antisymmetric aberrations as an additional complex phase factor multiplying the otherwise real CTF. In their notation,

$$X_{p,\mathbf{R}} \simeq e^{j\varphi(\mathbf{R})}\,\mathrm{CTF}_{p,\mathbf{R}}\,V_{p,\mathbf{R}}, \tag{4.1592}$$

where $\varphi(\mathbf{R})$ represents the antisymmetric aberration. The sign of φ relative to χ_{odd} depends on the Fourier and transfer-function convention, but the physical content is identical: the antisymmetric component rotates the phase of each Fourier coefficient without changing its modulus.

This also explains why such aberrations can be estimated from single-particle data even when they are difficult to recognize in an individual power spectrum. Once a preliminary three-dimensional reconstruction is available, the predicted Fourier phases of each particle can be compared with the observed phases, allowing the common antisymmetric phase error to be estimated across many particles (Zivanov et al., 2020).

Translation can partially masquerade as coma

There is an additional practical complication. As shown earlier, a linear phase in Fourier space corresponds to a translation of the image. Particle-alignment algorithms are free to change the estimated particle position and can therefore absorb part of an antisymmetric phase error into an apparent image shift.

Coma varies as R^3, whereas a translation produces a phase linear in R. The two cannot cancel one another at all spatial frequencies. Nevertheless, an alignment procedure may choose a translation that partially compensates coma over the frequency range that dominates the alignment. The residual phase then contains both a cubic coma component and a linear translation component.

For this reason, the antisymmetric-aberration treatment of Zivanov et al. (Zivanov et al., 2020) includes the first-order translation terms together with the higher-order antisymmetric modes. This prevents a systematic particle-position error from being mistaken for part of the optical aberration.

Why coma becomes rapidly important at high resolution

From Eq. (4.1541),

$$|\chi_{\mathrm{coma}}| \propto C_s \lambda^2 b R^3. \tag{4.1593}$$

Thus even a small residual beam tilt can generate a large phase error at high spatial frequency.

For example, for a 300 kV microscope with

$$C_s = 2.7 \text{ mm}, \qquad \lambda \simeq 1.97 \text{ pm}, \tag{4.1594}$$

a beam tilt of only

$$b = 1 \text{ mrad} \tag{4.1595}$$

produces, along the direction of maximum coma, a phase error at $d = 2$ Å of approximately

$$\boxed{|\chi_{\text{coma}}| \simeq 8.2 \text{ rad.}} \tag{4.1596}$$

This strong R^3 dependence explains why apparently small beam-tilt misalignments may be almost irrelevant at moderate resolution but become critical when approaching atomic resolution.

Experimentally, axial coma can be reduced by coma-free beam alignment or by applying compensating beam tilt during beam-image-shift acquisition. It can also be estimated and corrected computationally during single-particle refinement, provided that the data contain sufficient high-resolution signal (Zivanov et al., 2020).

Trefoil and the physical origin of threefold astigmatism

Axial coma is not the only cubic antisymmetric aberration. Another important example is *trefoil*, also called threefold astigmatism.

Its physical origin is different from that of the beam-tilt coma derived above. Coma can arise even in an otherwise rotationally symmetric lens when the beam is tilted away from the coma-free axis. Trefoil, in contrast, requires a departure from continuous rotational symmetry in the electron-optical field itself.

A general departure from axial symmetry can be decomposed into azimuthal harmonics. A threefold component is invariant under a rotation of $2\pi/3$,

$$\phi \longrightarrow \phi + \frac{2\pi}{3}, \tag{4.1597}$$

but not under an arbitrary rotation. In electron optics, such a component is naturally associated with a hexapole, or sextupole, field: although the device has six alternating poles, its potential has threefold angular periodicity.

Threefold components may arise as residual imperfections of the electron-optical system, and they are also deliberately generated by multipole elements in aberration correctors. Trefoil should therefore be regarded as an *aberration mode*, rather than as evidence of one unique microscopic defect.

We can derive its pupil phase directly from the symmetry and smoothness of the aberration function.

Cubic wavefront error produced by threefold symmetry.

From the homogeneous expansion of the wavefront error introduced in Eq. (4.1310), the most general third-degree contribution is

$$\Delta\ell^{(3)} = c_{30}\alpha_x^3 + c_{21}\alpha_x^2\alpha_y + c_{12}\alpha_x\alpha_y^2 + c_{03}\alpha_y^3. \tag{4.1598}$$

Using the pupil polar coordinates of Eq. (4.1297),

$$\alpha_x = \alpha\cos\phi, \qquad \alpha_y = \alpha\sin\phi, \tag{4.1599}$$

every term in Eq. (4.1598) contains the common factor α^3. Hence

$$\Delta\ell^{(3)} = \alpha^3 F_3(\phi), \tag{4.1600}$$

as expected from Eq. (4.1322).

The angular dependence can be made explicit using

$$\cos^3\phi = \frac{3\cos\phi + \cos 3\phi}{4}, \tag{4.1601}$$

$$\sin^3\phi = \frac{3\sin\phi - \sin 3\phi}{4}, \tag{4.1602}$$

$$\cos^2\phi\sin\phi = \frac{\sin\phi + \sin 3\phi}{4}, \tag{4.1603}$$

$$\cos\phi\sin^2\phi = \frac{\cos\phi - \cos 3\phi}{4}. \tag{4.1604}$$

Therefore the most general cubic wavefront can contain only first and third azimuthal harmonics,

$$\Delta\ell^{(3)} = \alpha^3\left[a_1\cos\phi + b_1\sin\phi + a_3\cos 3\phi + b_3\sin 3\phi\right]. \tag{4.1605}$$

The first-harmonic terms,

$$\alpha^3\left(a_1\cos\phi + b_1\sin\phi\right), \tag{4.1606}$$

have onefold angular symmetry and correspond to coma.

For a pure threefold perturbation, however, the wavefront must satisfy

$$\Delta\ell^{(3)}\left(\alpha, \phi + \frac{2\pi}{3}\right) = \Delta\ell^{(3)}(\alpha, \phi). \tag{4.1607}$$

The first-harmonic terms do not satisfy this condition and must therefore vanish,

$$a_1 = b_1 = 0. \tag{4.1608}$$

Thus

$$\Delta\ell_{\text{trefoil}} = \alpha^3 \left[a_3 \cos 3\phi + b_3 \sin 3\phi\right]. \tag{4.1609}$$

Any combination of the sine and cosine terms can be written as a single rotated cosine. We therefore define the threefold-astigmatism coefficient A_3 and its orientation h_3 through

$$\boxed{\Delta\ell_{\text{trefoil}} = \frac{A_3}{3}\alpha^3 \cos\left[3(\phi - h_3)\right].} \tag{4.1610}$$

The factor 1/3 is a conventional normalization of the coefficient A_3.

Using the relation between wavefront error and aberration phase derived in Eq. (4.1263),

$$\chi_{\text{ab}} = -\frac{2\pi}{\lambda}\Delta\ell_{\text{ab}}, \tag{4.1611}$$

we obtain

$$\chi_{\text{trefoil}}(\alpha, \phi) = -\frac{2\pi A_3}{3\lambda}\alpha^3 \cos\left[3(\phi - h_3)\right]. \tag{4.1612}$$

Finally, using the paraxial pupil relation

$$\alpha \simeq \lambda R \tag{4.1613}$$

from Eq. (4.1300), the phase becomes

$$\boxed{\chi_{\text{trefoil}}(R, \phi) = -\frac{2\pi}{3}A_3\lambda^2 R^3 \cos\left[3(\phi - h_3)\right].} \tag{4.1614}$$

4.11.7 General higher-order aberrations and Zernike polynomials

Up to this point, we have introduced the most important aberrations one at a time. Defocus was derived in Sec. 4.10.8, where it was shown to produce the quadratic phase of Eq. (4.1191). We subsequently introduced twofold astigmatism, spherical aberration, and coma and examined the characteristic radial and angular dependence of each of them.

At this stage, however, continuing to introduce aberrations one by one

becomes increasingly inconvenient. A real objective lens may contain a superposition of many weak departures from the ideal wavefront, and there is no reason for these departures to consist of only the few aberrations that have historically been given individual names.

We therefore need a more systematic language:

> *We now need a systematic basis for describing smooth wavefront errors over a circular pupil.*

A particularly convenient answer is provided by the *Zernike polynomials*. They form a complete basis on a circular pupil and separate two distinct properties of an aberration: its radial dependence and its angular symmetry.

The purpose of introducing Zernike polynomials here is not to replace the physical interpretation developed in the preceding sections. Quite the opposite: because we already understand defocus, astigmatism, spherical aberration, and coma physically, the first Zernike modes will now appear as familiar objects rather than as an abstract collection of polynomials.

The aberration phase over the pupil

Recall from Eq. (4.1193) that an aberrated objective system can be represented in Fourier space by a pupil function multiplying the electron wave. We now generalize that expression and write

$$\boxed{H_{\mathrm{obj}}(\mathbf{R}) = P(\mathbf{R}) \exp\left[-j\chi(\mathbf{R})\right],} \tag{4.1615}$$

where

$$\mathbf{R} = (R_x, R_y) \tag{4.1616}$$

is the spatial-frequency vector, $P(\mathbf{R})$ describes the finite objective aperture, and $\chi(\mathbf{R})$ is the total aberration phase.

For pure defocus, Eq. (4.1191) gave

$$\chi_{\mathrm{def}}(\mathbf{R}) = -\pi\lambda\Delta f\, R^2, \qquad R = \|\mathbf{R}\|. \tag{4.1617}$$

The more general function $\chi(\mathbf{R})$ may depend not only on the magnitude R, but also on the azimuthal direction of $\mathbf{R}$.

It is therefore natural to introduce polar coordinates in Fourier space,

$$R_x = R\cos\varphi, \qquad R_y = R\sin\varphi. \tag{4.1618}$$

Let $R_{\max}$ denote the largest spatial frequency admitted by the objective pupil. We define the normalized radial coordinate

$$\boxed{\rho \triangleq \frac{R}{R_{\max}}, \qquad 0 \le \rho \le 1.} \tag{4.1619}$$

According to the spatial-frequency–scattering-angle relation derived in Eq. (4.1126),

$$\alpha \simeq \lambda R. \tag{4.1620}$$

Consequently,

$$\boxed{\rho = \frac{R}{R_{\max}} \simeq \frac{\alpha}{\alpha_{\max}}.} \tag{4.1621}$$

Thus, the unit disk

$$0 \le \rho \le 1, \qquad 0 \le \varphi < 2\pi \tag{4.1622}$$

represents the complete circular objective pupil.

Sometimes it is useful to express the same error as a *wave aberration* $W(\rho, \varphi)$ with dimensions of length. A path-length error W produces the phase error

$$\boxed{\chi(\rho, \varphi) = \frac{2\pi}{\lambda} W(\rho, \varphi).} \tag{4.1623}$$

The two descriptions therefore contain exactly the same information. For image formation, however, we will mostly work directly with the dimensionless phase χ, because it enters immediately in Eq. (4.1615).

The real Zernike basis

Zernike polynomials are indexed by two integers,

$$n = 0, 1, 2, \ldots \tag{4.1624}$$

and

$$m = -n, -n+2, \ldots, n-2, n. \tag{4.1625}$$

Thus,

$$\boxed{|m| \le n, \qquad n - |m| \text{ is even.}} \tag{4.1626}$$

The integer n is called the *radial order*, whereas $|m|$ determines the angular symmetry.

For the purposes of this book, we will use real, unnormalized Zernike polynomials. Their radial part is

$$\boxed{\mathcal{R}_n^{|m|}(\rho) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!\,\left(\frac{n+|m|}{2} - s\right)!\,\left(\frac{n-|m|}{2} - s\right)!} \rho^{n-2s}.} \tag{4.1627}$$

The complete real Zernike modes are then defined as

$$\boxed{Z_n^m(\rho,\varphi) = \begin{cases} \mathcal{R}_n^m(\rho)\cos(m\varphi), & m > 0, \\ \mathcal{R}_n^0(\rho), & m = 0, \\ \mathcal{R}_n^{|m|}(\rho)\sin(|m|\varphi), & m < 0. \end{cases}} \tag{4.1628}$$

This convention is deliberately chosen without additional normalization constants. Other conventions multiply the modes by factors depending on n and m, and several one-dimensional indexing schemes are also in common use. Consequently, numerical Zernike coefficients should never be compared between programs or publications without first checking the normalization and indexing convention.

The essential geometrical structure, however, is independent of these choices.

On the complete unit disk, with uniform weighting, two distinct Zernike modes are orthogonal:

$$\int_0^1 \int_0^{2\pi} Z_n^m(\rho,\varphi) Z_{n'}^{m'}(\rho,\varphi)\,\rho\,d\varphi\,d\rho = 0, \qquad (n,m) \ne (n',m'). \tag{4.1629}$$

The factor ρ appears because the area element in polar coordinates is

$$dA = \rho\,d\rho\,d\varphi. \tag{4.1630}$$

A sufficiently regular aberration function over the pupil may therefore be expanded as

$$\boxed{\chi(\rho,\varphi) = \sum_{n=0}^{\infty} \sum_{\substack{m=-n \\ n-|m| \text{ even}}}^{n} c_n^m Z_n^m(\rho,\varphi),} \tag{4.1631}$$

where the coefficients c_n^m have units of radians.
Equivalently, if the wave aberration itself is expanded,

$$W(\rho,\varphi) = \sum_{n,m} a_n^m Z_n^m(\rho,\varphi), \tag{4.1632}$$

then Eq. (4.1623) gives

$$\boxed{c_n^m = \frac{2\pi}{\lambda} a_n^m.} \tag{4.1633}$$

Angular symmetry and orientation

The usefulness of the index m becomes particularly transparent if we consider the pair Z_n^m and Z_n^{-m} for $m > 0$.

Amplitude and orientation of a Zernike aberration.

Consider an arbitrary linear combination of the two modes,

$$a\, Z_n^m(\rho,\varphi) + b\, Z_n^{-m}(\rho,\varphi). \tag{4.1634}$$

Using Eq. (4.1628),

$$= \mathcal{R}_n^m(\rho)\left[a\cos(m\varphi) + b\sin(m\varphi)\right]. \tag{4.1635}$$

Define

$$A = \sqrt{a^2 + b^2} \tag{4.1636}$$

and an orientation angle φ_0 through

$$\varphi_0 = \frac{1}{m}\,\mathrm{atan2}(b,a). \tag{4.1637}$$

Then

$$a = A\cos(m\varphi_0), \qquad b = A\sin(m\varphi_0), \tag{4.1638}$$

and therefore

$$\boxed{a\,Z_n^m + b\,Z_n^{-m} = A\,\mathcal{R}_n^m(\rho)\cos\left[m(\varphi-\varphi_0)\right].} \tag{4.1639}$$

Equation (4.1639) gives a simple physical interpretation of the two Zernike indices.

The radial index n controls how the aberration changes between the centre and the edge of the pupil, whereas $|m|$ counts the angular periodicity. Thus,

$$|m| = 1 \tag{4.1640}$$

produces a onefold angular dependence,

$$|m| = 2 \tag{4.1641}$$

a twofold dependence,

$$|m| = 3 \tag{4.1642}$$

a threefold dependence, and so forth.

The pair of coefficients corresponding to $+m$ and $-m$ does not describe two fundamentally different aberrations. Rather, the two coefficients describe the magnitude and orientation of the same m-fold aberration.

The familiar aberrations reappear

The first Zernike modes are

$$\begin{aligned}
&Z_0^0 = 1, && \\
&Z_1^1 = \rho\cos\varphi, && Z_1^{-1} = \rho\sin\varphi, \\
&Z_2^0 = 2\rho^2 - 1, && \\
&Z_2^2 = \rho^2\cos(2\varphi), && Z_2^{-2} = \rho^2\sin(2\varphi), \\
&Z_3^1 = (3\rho^3 - 2\rho)\cos\varphi, && Z_3^{-1} = (3\rho^3 - 2\rho)\sin\varphi, \\
&Z_3^3 = \rho^3\cos(3\varphi), && Z_3^{-3} = \rho^3\sin(3\varphi), \\
&Z_4^0 = 6\rho^4 - 6\rho^2 + 1, && \\
&Z_4^2 = (4\rho^4 - 3\rho^2)\cos(2\varphi), && Z_4^{-2} = (4\rho^4 - 3\rho^2)\sin(2\varphi), \\
&Z_4^4 = \rho^4\cos(4\varphi), && Z_4^{-4} = \rho^4\sin(4\varphi).
\end{aligned} \tag{4.1643}$$

Table 4.2: Low-order real Zernike modes and their usual optical interpretation. The correspondence should be understood modulo normalization conventions and, for some modes, the addition of lower-order terms such as piston, tilt, or defocus.

Mode	Angular symmetry	Typical interpretation
Z_0^0	rotational	piston
$Z_1^{\pm 1}$	onefold	tilt
Z_2^0	rotational	defocus
$Z_2^{\pm 2}$	twofold	twofold astigmatism
$Z_3^{\pm 1}$	onefold	coma
$Z_3^{\pm 3}$	threefold	trefoil / threefold astigmatism
Z_4^0	rotational	spherical-type aberration
$Z_4^{\pm 2}$	twofold	secondary twofold astigmatism
$Z_4^{\pm 4}$	fourfold	fourfold astigmatism

Their familiar optical interpretations are summarized in Table 4.2.

A graphical representation such as Fig. 4.13 makes the meaning of the angular index particularly intuitive.

There is, however, an important subtlety hidden in Table 4.2. The physical aberrations derived in the preceding sections were naturally obtained as powers of the aperture coordinate, whereas the Zernike modes have been modified by lower-order terms so that they become orthogonal on the unit disk. Consequently, the correspondence between a physical aberration and an individual Zernike mode is not always exact.

How the familiar aberrations are represented by Zernike modes.

Consider first the radial dependence of defocus. From Eq. (4.1643),

$$Z_2^0 = 2\rho^2 - 1. \tag{4.1644}$$

Therefore,

$$\boxed{\rho^2 = \frac{1}{2} Z_2^0 + \frac{1}{2} Z_0^0.} \tag{4.1645}$$

Using Eq. (4.1619), the defocus phase derived in Eq. (4.1191) can be written as

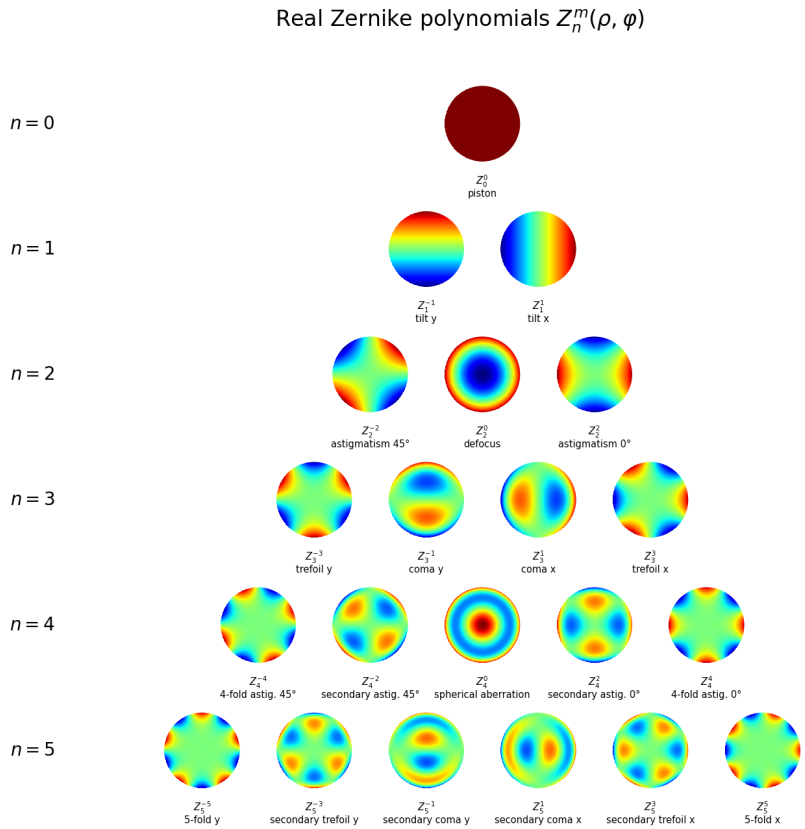


Figure 4.13: Representative low-order Zernike modes over the unit circular pupil. The radial order n increases from top to bottom, while the azimuthal index m determines the angular symmetry. The colour represents the relative sign and magnitude of the wavefront or phase perturbation; its absolute scale is arbitrary. Notice the immediate appearance of familiar patterns: rotationally symmetric defocus and spherical-type terms, twofold astigmatism, onefold coma, threefold trefoil, and fourfold astigmatism.

$$\chi_{\text{def}}(\rho) = -\pi\lambda\Delta f\, R_{\text{max}}^2 \rho^2. \tag{4.1646}$$

From Eq. (4.1643),

$$Z_2^0 = 2\rho^2 - 1, \qquad Z_0^0 = 1, \tag{4.1647}$$

and therefore

$$\boxed{\rho^2 = \frac{1}{2}\left(Z_2^0 + Z_0^0\right).} \tag{4.1648}$$

Substituting Eq. (4.1648) into the defocus phase gives

$$\boxed{\chi_{\text{def}}(\rho,\varphi) = -\frac{\pi\lambda\Delta f\, R_{\text{max}}^2}{2} Z_2^0(\rho,\varphi) - \frac{\pi\lambda\Delta f\, R_{\text{max}}^2}{2} Z_0^0(\rho,\varphi).} \tag{4.1649}$$

The second term in Eq. (4.1649) is independent of ρ and φ. It is therefore a spatially uniform phase, usually called *piston*. Multiplication of the complete wave by this constant phase factor does not change the detected intensity. Consequently, as far as image formation is concerned,

$$\boxed{\chi_{\text{def}}(\rho,\varphi) \equiv -\frac{\pi\lambda\Delta f\, R_{\text{max}}^2}{2} Z_2^0(\rho,\varphi), \qquad \text{up to piston.}} \tag{4.1650}$$

Thus, the familiar quadratic defocus aberration derived in Eq. (4.1191) corresponds to the Z_2^0 Zernike mode, apart from an optically irrelevant constant phase.

Twofold astigmatism has the angular form

$$\rho^2 \cos\left[2(\varphi - \varphi_A)\right]. \tag{4.1651}$$

Using the trigonometric identity

$$\cos\left[2(\varphi - \varphi_A)\right] = \cos(2\varphi)\cos(2\varphi_A) + \sin(2\varphi)\sin(2\varphi_A), \tag{4.1652}$$

and the definitions in Eq. (4.1643),

$$Z_2^2 = \rho^2\cos(2\varphi), \qquad Z_2^{-2} = \rho^2\sin(2\varphi), \tag{4.1653}$$

we obtain

$$\boxed{\rho^2 \cos\left[2(\varphi - \varphi_A)\right] = \cos(2\varphi_A)\, Z_2^2 + \sin(2\varphi_A)\, Z_2^{-2}.} \tag{4.1654}$$

Thus, the two coefficients of Z_2^2 and Z_2^{-2} encode the magnitude and orientation of the twofold astigmatism. In particular, Eq. (4.1654) is the $m = 2$ case of the general amplitude–orientation relation derived in Eq. (4.1639).

For coma, consider one orientation,

$$\rho^3 \cos\varphi. \tag{4.1655}$$

Since

$$Z_3^1 = (3\rho^3 - 2\rho)\cos\varphi \tag{4.1656}$$

and

$$Z_1^1 = \rho\cos\varphi, \tag{4.1657}$$

we obtain

$$\boxed{\rho^3 \cos\varphi = \frac{1}{3} Z_3^1 + \frac{2}{3} Z_1^1.} \tag{4.1658}$$

Thus, a pure cubic onefold phase contains a Zernike coma term together with a lower-order tilt term. The corresponding expression for an arbitrary orientation follows by combining the $m = +1$ and $m = -1$ components.

Finally, the radial dependence of primary spherical aberration is proportional to ρ^4. From

$$Z_4^0 = 6\rho^4 - 6\rho^2 + 1, \tag{4.1659}$$

and the expression for Z_2^0, we find

$$\boxed{\rho^4 = \frac{1}{6} Z_4^0 + \frac{1}{2} Z_2^0 + \frac{1}{3} Z_0^0.} \tag{4.1660}$$

Thus, a pure fourth-order radial phase is represented by a spherical Zernike term together with defocus and piston.

By contrast,

$$\boxed{\rho^3 \cos(3\varphi) = Z_3^3} \tag{4.1661}$$

and

$$\boxed{\rho^4 \cos(4\varphi) = Z_4^4,} \tag{4.1662}$$

with analogous equations for the sine components. These are the familiar trefoil and fourfold-astigmatism patterns.

Equations (4.1648)– (4.1660) illustrate an important concept. A Zernike mode is not identical to the physical mechanism that produced the aberration. Rather, the Zernikes provide coordinates for representing the resulting wavefront error.

For example, a physical spherical-aberration term proportional to ρ^4 does not disappear merely because part of its Zernike representation lies in Z_2^0. The decomposition simply reflects the choice of an orthogonal basis. Similarly, the lower-order tilt present in Eq. (4.1658) reflects the orthogonalization of the cubic coma-like wavefront against the linear tilt mode.

Symmetric and antisymmetric aberrations

The Zernike representation also makes another distinction that is particularly useful in cryo-EM remarkably transparent.

Consider reversing a spatial-frequency vector,

$$\mathbf{R} \longrightarrow -\mathbf{R}. \tag{4.1663}$$

In polar coordinates this transformation leaves ρ unchanged and replaces

$$\varphi \longrightarrow \varphi + \pi. \tag{4.1664}$$

Because

$$\cos[m(\varphi + \pi)] = (-1)^m \cos(m\varphi) \tag{4.1665}$$

and

$$\sin[m(\varphi + \pi)] = (-1)^m \sin(m\varphi), \tag{4.1666}$$

each Zernike mode has a well-defined parity.

Parity of the Zernike modes.

From the index condition in Eq. (4.1626), $n - |m|$ is even. Therefore, n and m have the same parity, and hence

$$(-1)^m = (-1)^n. \tag{4.1667}$$

It follows immediately that

$$\boxed{Z_n^m(\rho, \varphi + \pi) = (-1)^n Z_n^m(\rho, \varphi).} \tag{4.1668}$$

Equivalently,

$$\boxed{Z_n^m(-\mathbf{R}) = (-1)^n Z_n^m(\mathbf{R}).} \tag{4.1669}$$

Thus, all modes with even radial order satisfy

$$Z_n^m(-\mathbf{R}) = Z_n^m(\mathbf{R}), \qquad n \text{ even}, \tag{4.1670}$$

whereas all modes with odd radial order satisfy

$$Z_n^m(-\mathbf{R}) = -Z_n^m(\mathbf{R}), \qquad n \text{ odd}. \tag{4.1671}$$

This naturally separates an arbitrary aberration phase into a centrosymmetric part

$$\boxed{\chi_{\text{even}}(\mathbf{R}) = \frac{\chi(\mathbf{R}) + \chi(-\mathbf{R})}{2},} \tag{4.1672}$$

and an antisymmetric part

$$\boxed{\chi_{\text{odd}}(\mathbf{R}) = \frac{\chi(\mathbf{R}) - \chi(-\mathbf{R})}{2}.} \tag{4.1673}$$

Therefore,

$$\boxed{\chi(\mathbf{R}) = \chi_{\text{even}}(\mathbf{R}) + \chi_{\text{odd}}(\mathbf{R}).} \tag{4.1674}$$

In the Zernike basis,

$$\chi_{\text{even}} = \sum_{\substack{n \text{ even} \\ m}} c_n^m Z_n^m, \tag{4.1675}$$

whereas

$$\chi_{\text{odd}} = \sum_{\substack{n \text{ odd} \\ m}} c_n^m Z_n^m. \tag{4.1676}$$

The familiar aberrations immediately fall into these two classes. Defocus and twofold astigmatism have $n = 2$ and are symmetric. Primary spherical

aberration and fourfold astigmatism have $n = 4$ and are also symmetric. Coma and trefoil have $n = 3$ and are antisymmetric.

This distinction is especially important in cryo-EM because symmetric and antisymmetric aberrations affect the Fourier representation of the image in different ways and can therefore be estimated by different strategies.

Higher-order aberrations in modern cryo-EM

The Zernike representation is now used in high-resolution cryo-EM to describe aberrations that are not included in the traditional defocus–astigmatism–spherical-aberration model.

Experimental data analysed in that work showed clear threefold patterns associated with trefoil, sometimes combined with axial coma, as well as fourfold patterns associated with fourfold astigmatism. These effects can be sufficiently large to limit high-resolution reconstruction if they are not included in the image-formation model.

The general phase model may therefore be written in the compact form

$$\boxed{\chi(\rho,\varphi) = \sum_{n=0}^{n_{\max}} \sum_{\substack{m=-n \\ n-|m| \text{ even}}}^{n} c_n^m Z_n^m(\rho,\varphi),} \tag{4.1677}$$

where $n_{\max}$ determines the highest spatial complexity that we allow the aberration field to have.

Substitution of Eq. (4.1677) into Eq. (4.1615) gives the general coherent objective-lens transfer function

$$\boxed{H_{\text{obj}}(\rho,\varphi) = P(\rho,\varphi)\exp\left[-j\sum_{n=0}^{n_{\max}} \sum_{\substack{m=-n \\ n-|m| \text{ even}}}^{n} c_n^m Z_n^m(\rho,\varphi)\right].} \tag{4.1678}$$

Equation (4.1678) is the direct generalization of the defocus-only pupil transfer function derived in Eq. (4.1193).

There is an important conceptual point concerning Eq. (4.1677). Zernike polynomials are a *basis*; they are not the physical origin of the aberrations. A particular imperfection of the electromagnetic lenses, a misalignment, beam tilt, charging, or another instrumental effect generates some wavefront error $\chi(\rho,\varphi)$. The Zernike coefficients merely describe that resulting function in a convenient coordinate system.

This distinction is also important mathematically. Although the Zernike functions are orthogonal under the uniform pupil integral of Eq. (4.1629), experimental cryo-EM data do not provide equally reliable information at every point of Fourier space. CTF zeros, finite signal-to-noise ratio, envelope attenuation, masks, and the finite resolution of the reference all produce strongly nonuniform weighting. Consequently, the effective basis functions need not be orthogonal under the weighting imposed by the experiment.

Another sufficiently flexible set of smooth functions could in principle be used instead. Zernike polynomials are particularly convenient because the domain is circular, the modes are mathematically well behaved, and their low-order terms have immediate physical interpretations.

There is also a practical reason not to include arbitrarily many modes. A low-order truncation of Eq. (4.1677) implicitly assumes that the aberration phase varies smoothly over the pupil. Increasing $n_{\max}$ allows increasingly rapid radial and angular variations to be represented, but it also introduces more parameters that must be estimated from the experimental data. Very localized or discontinuous pupil perturbations may therefore require many Zernike modes and are not efficiently described by a low-order expansion.

The main advantage of the Zernike representation is thus not that it introduces a new physical model of the microscope. Rather, it provides a systematic extension of the physical aberration model already developed in the preceding sections:

$$\begin{aligned}
Z_2^0 &\longleftrightarrow \text{defocus}, \\
Z_2^{\pm 2} &\longleftrightarrow \text{twofold astigmatism}, \\
Z_3^{\pm 1} &\longleftrightarrow \text{coma}, \\
Z_3^{\pm 3} &\longleftrightarrow \text{trefoil / threefold astigmatism}, \\
Z_4^0 &\longleftrightarrow \text{spherical-type aberration}, \\
Z_4^{\pm 4} &\longleftrightarrow \text{fourfold astigmatism}.
\end{aligned} \tag{4.1679}$$

The arrows in Eq. (4.1679) deliberately denote a correspondence rather than a strict equality. As demonstrated by Eqs. (4.1648)– (4.1660), the exact coefficients depend on normalization conventions and on how lower-order terms such as piston, tilt, and defocus are separated from the higher-order aberrations.

With this representation, the aberration theory is no longer restricted to a short list of named defects. Any sufficiently smooth phase error over the objective pupil can be incorporated into the same Fourier-space image-formation model through Eq. (4.1678). The traditional aberrations are simply the first physically important members of this more general description.

Connection with conventional electron-optical aberration coefficients

The Zernike representation developed above provides a systematic basis for describing the aberration wavefront over a circular pupil. Classical electron optics, however, often uses a different notation in which particular physical aberrations are described by coefficients such as C_3, C_5, and the coefficients associated with the different anisotropic aberrations. It is therefore useful to make the connection between these two descriptions explicit.

The connection is already largely contained in the symmetry analysis of Sec. 4.11.2. There we showed that a homogeneous wavefront term of degree N can contain only the azimuthal orders

$$m = N, N-2, N-4, \ldots, \tag{4.1680}$$

according to Eq. (4.1360), and that each contribution can be written in the form

$$\Delta\ell_{N,m}(\alpha,\varphi) \propto \alpha^N \cos\left[m(\varphi - \varphi_{N,m})\right] \tag{4.1681}$$

[Eq. (4.1361)]. The same parity restriction appeared independently in the Zernike indices, Eq. (4.1626). This agreement is not accidental: both descriptions are consequences of the same rotational structure of smooth functions over a circular pupil.

The difference is one of purpose. The expansion in powers of α_x and α_y follows from the Taylor expansion of the electron action and is therefore naturally connected with the physical aberration coefficients of electron optics. The Zernike expansion, Eq. (4.1631), instead provides a convenient basis in which the resulting wavefront can be represented.

Higher-order spherical aberration and C_5. The distinction between the two descriptions is particularly clear for higher-order spherical aberration. The ordinary spherical-aberration coefficient was introduced in Eq. (4.1385) through the geometrical ray aberration

$$\Delta r_{\text{sph}} = C_3\alpha^3, \tag{4.1682}$$

with $C_s \equiv C_3$ according to Eq. (4.1386). The corresponding wavefront term is fourth degree in the aperture angle, Eq. (4.1387).

The next rotationally symmetric term is conventionally described by the fifth-order spherical-aberration coefficient C_5. Its name refers to the order of the corresponding *ray* aberration,

$$\Delta r_{C_5} \propto C_5\alpha^5, \tag{4.1683}$$

and therefore its wavefront error is sixth degree in α.

Wavefront term associated with C_5.

For a rotationally symmetric aberration, the transverse ray error is obtained from the radial derivative of the wavefront error, as derived in Sec. 4.11.2. Thus, up to the sign convention used for the wavefront error,

$$\frac{d\Delta\ell_{C_5}}{d\alpha} = \pm C_5\alpha^5. \tag{4.1684}$$

Integrating with respect to α gives

$$\boxed{\Delta\ell_{C_5} = \pm\frac{C_5}{6}\alpha^6} \tag{4.1685}$$

apart from an irrelevant integration constant.

Together with the C_3 contribution of Eq. (4.1387), the beginning of the rotationally symmetric spherical-aberration expansion is therefore

$$\boxed{\Delta\ell_{\mathrm{sph}}(\alpha) = \pm\left(\frac{C_3}{4}\alpha^4 + \frac{C_5}{6}\alpha^6 + \cdots\right).} \tag{4.1686}$$

This illustrates a terminology that can otherwise be confusing: C_5 is called a *fifth-order* spherical-aberration coefficient because the geometrical ray aberration varies as α^5, whereas the associated wavefront and phase functions contain a sixth-degree radial term. Thus C_5 should not be associated with Zernike radial order $n = 5$.

Using the normalized pupil coordinate introduced in Eq. (4.1304),

$$\alpha = \alpha_{\max}\rho, \tag{4.1687}$$

the radial dependence of the C_5 wavefront is simply proportional to ρ^6. From the radial-polynomial definition in Eq. (4.1627), the rotationally symmetric sixth-order Zernike polynomial is

$$Z_6^0 = 20\rho^6 - 30\rho^4 + 12\rho^2 - 1. \tag{4.1688}$$

Using the previous decompositions of ρ^2 and ρ^4 in Eqs. (4.1648) and (4.1660), respectively, we obtain

$$\boxed{\rho^6 = \frac{1}{20}Z_6^0 + \frac{1}{4}Z_4^0 + \frac{9}{20}Z_2^0 + \frac{1}{4}Z_0^0.} \tag{4.1689}$$

Consequently, the C_5 wavefront term of Eq. (4.1685) can be written as

$$\boxed{\Delta\ell_{C_5} = \pm\frac{C_5\alpha_{\max}^6}{6}\left(\frac{1}{20}Z_6^0 + \frac{1}{4}Z_4^0 + \frac{9}{20}Z_2^0 + \frac{1}{4}Z_0^0\right).} \tag{4.1690}$$

The corresponding phase follows immediately from Eq. (4.1263); there is no need to rewrite it explicitly here.

Equation (4.1689) reinforces an important point made throughout this section. The physical C_5 aberration is proportional to ρ^6; it is *not* identical to the single Zernike mode Z_6^0. The orthogonal Zernike representation of ρ^6 necessarily contains lower radial orders. In a practical fit, those lower-order components may partly overlap with piston, defocus, and the ordinary C_3 spherical-aberration contribution.

Anisotropic higher-order terms. The same reasoning applies to anisotropic aberrations. From Eq. (4.1361), a physical term of wavefront degree N and azimuthal order m has the generic form

$$\boxed{\Delta\ell_{N,m} = A_{N,m}\alpha^N \cos\left[m(\varphi - \varphi_{N,m})\right],} \tag{4.1691}$$

where $A_{N,m}$ is the corresponding aberration coefficient and the allowed values of m are those of Eq. (4.1360). In normalized pupil coordinates,

$$\Delta\ell_{N,m} \propto \rho^N \cos\left[m(\varphi - \varphi_{N,m})\right]. \tag{4.1692}$$

A particularly simple case occurs when

$$|m| = N. \tag{4.1693}$$

From Eq. (4.1627),

$$\boxed{\mathcal{R}_N^N(\rho) = \rho^N.} \tag{4.1694}$$

Thus these physical aberrations correspond directly to a pair of Zernike modes of the same order, with no admixture of lower radial powers.

For example, an arbitrarily oriented trefoil term can be written as

$$\boxed{\rho^3 \cos\left[3(\varphi - \varphi_T)\right] = \cos(3\varphi_T)\, Z_3^3 + \sin(3\varphi_T)\, Z_3^{-3},} \tag{4.1695}$$

which is the $m = 3$ case of the general amplitude–orientation relation of Eq. (4.1639). Similarly, fourfold astigmatism has

$$\boxed{\rho^4 \cos\left[4(\varphi - \varphi_4)\right] = \cos(4\varphi_4)\, Z_4^4 + \sin(4\varphi_4)\, Z_4^{-4}.} \tag{4.1696}$$

For $|m| < N$, the connection is generally less direct because the orthogonal Zernike radial polynomial contains several powers of ρ. Coma provides the first example: the physical cubic term contains both $Z_3^{\pm 1}$ and a lower-order tilt component, as shown in Eq. (4.1658). Rotationally symmetric spherical aberration behaves similarly, as illustrated by Eqs. (4.1660) and (4.1689).

The relation between the two descriptions can therefore be summarized schematically as

physical wavefront term		Zernike representation
ρ^2	$\longrightarrow$	$Z_2^0 + Z_0^0$
$\rho^2 \cos 2(\varphi - \varphi_A)$	$\longrightarrow$	$Z_2^{\pm 2}$
$\rho^3 \cos(\varphi - \varphi_C)$	$\longrightarrow$	$Z_3^{\pm 1} + Z_1^{\pm 1}$
$\rho^3 \cos 3(\varphi - \varphi_T)$	$\longrightarrow$	$Z_3^{\pm 3}$
ρ^4	$\longrightarrow$	$Z_4^0 + Z_2^0 + Z_0^0$
$\rho^4 \cos 4(\varphi - \varphi_4)$	$\longrightarrow$	$Z_4^{\pm 4}$
ρ^6	$\longrightarrow$	$Z_6^0 + Z_4^0 + Z_2^0 + Z_0^0.$

(4.1697)

The arrows in Eq. (4.1697) indicate the Zernike modes that are required, not their numerical coefficients; the exact coefficients were derived above and in Eqs. (4.1648), (4.1658), (4.1660), and (4.1689).

A complete fifth-order electron-optical theory contains many additional isotropic and anisotropic coefficients. Such a catalogue is important when deriving aberrations directly from the detailed electromagnetic field of a lens, but it is not required for the image-formation model developed here. The symmetry result of Eq. (4.1360), together with the Zernike expansion of Eq. (4.1631), already provides a systematic description of which smooth aberration terms may appear, their radial dependence, and their angular symmetry. The individual electron-optical coefficients specify the physical mechanism and magnitude of those terms; the Zernike coefficients specify their representation over the pupil.

4.11.8 The classical field-dependent Seidel aberrations

In Sec. 4.11.1 we introduced the five classical Seidel aberrations: spherical aberration, coma, astigmatism, field curvature, distortion. Most of the discussion since then has concentrated on aberrations that can be described as a phase error over the objective pupil for an object point on, or very close to, the optical axis. We now return briefly to the complete classical picture.

The essential new ingredient is the position of the object point within the field of view. In the full geometrical theory, the aberration function depends simultaneously on the object coordinate and on the coordinate of the ray within the aperture. It is this joint dependence that generates the different Seidel families (Hawkes and Kasper, 2017, Chap. 24).

In the notation used throughout this book, let

$$\mathbf{r}_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}, \qquad r_i = \|\mathbf{r}_i\|, \tag{4.1698}$$

denote the transverse position of the object point in the input plane. In Gaussian imaging, its corresponding image point is

$$\mathbf{r}_o = M\mathbf{r}_i, \tag{4.1699}$$

where the same transverse magnification M applies in both directions, as in the one-dimensional result of Eq. (4.1115).

Let

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}, \qquad \alpha = \|\boldsymbol{\alpha}\|, \tag{4.1700}$$

denote the paraxial angular coordinate of a ray within the pupil. If ϕ_i is the azimuth of $\mathbf{r}_i$ and ϕ is the azimuth of $\boldsymbol{\alpha}$, we define

$$\Delta\phi \triangleq \phi - \phi_i. \tag{4.1701}$$

Their scalar product is therefore

$$\boxed{\mathbf{r}_i \cdot \boldsymbol{\alpha} = r_i \alpha \cos \Delta\phi.} \tag{4.1702}$$

Why the five Seidel families appear

The detailed electron-optical coefficients depend on integrals through the electromagnetic lens field. Their possible functional forms, however, follow largely from symmetry. A round lens is invariant under a simultaneous rotation of the object coordinate $\mathbf{r}_i$ and the aperture coordinate $\boldsymbol{\alpha}$. The aberration wavefront must therefore be constructed from rotational scalars involving these two vectors.

The five isotropic Seidel terms from rotational symmetry.

Let

$$\mathbf{r}_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}. \tag{4.1703}$$

If both vectors are rotated by the same angle around the optical axis, their Cartesian components change, but three geometrical quantities do not: the length of $\mathbf{r}_i$, the length of $\boldsymbol{\alpha}$, and the angle between them. We may therefore introduce the three quadratic rotational invariants

$$\boxed{A \triangleq r_i^2 = x_i^2 + y_i^2, \qquad B \triangleq \alpha^2 = \alpha_x^2 + \alpha_y^2, \qquad C \triangleq \mathbf{r}_i \cdot \boldsymbol{\alpha} = x_i\alpha_x + y_i\alpha_y.} \tag{4.1704}$$

The invariance of A and B follows because a rotation does not change the length of a vector. The invariance of C follows because

$$C = r_i\alpha\cos(\phi - \phi_i), \tag{4.1705}$$

and a simultaneous rotation changes ϕ and ϕ_i by the same amount, leaving their difference unchanged.

There is one additional quadratic quantity that is invariant under proper rotations,

$$S \triangleq x_i\alpha_y - y_i\alpha_x = r_i\alpha\sin(\phi - \phi_i). \tag{4.1706}$$

Unlike A, B, and C, however, S changes sign under a mirror reflection. It therefore generates the anisotropic partners that occur in the complete magnetic-lens theory and will be left aside here. Even at fourth order its square introduces no new reflection-even structure, because

$$S^2 = r_i^2\alpha^2 - C^2 = AB - C^2. \tag{4.1707}$$

The lowest non-Gaussian contribution to the characteristic function is fourth order, as discussed in Sec. 4.11.2. Since each of A, B, and C is already of total degree two in $(x_i, y_i, \alpha_x, \alpha_y)$, every fourth-order reflection-even scalar is obtained by multiplying two of them. The six possible products are therefore

$$\boxed{A^2, \qquad AB, \qquad AC, \qquad B^2, \qquad BC, \qquad C^2.} \tag{4.1708}$$

Their degree-four character is explicit:

term	degree in $\mathbf{r}_i$	degree in $\boldsymbol{\alpha}$	total degree
$A^2 = r_i^4$	4	0	4
$AB = r_i^2\alpha^2$	2	2	4
$AC = r_i^2(\mathbf{r}_i \cdot \boldsymbol{\alpha})$	3	1	4
$B^2 = \alpha^4$	0	4	4
$BC = \alpha^2(\mathbf{r}_i \cdot \boldsymbol{\alpha})$	1	3	4
$C^2 = (\mathbf{r}_i \cdot \boldsymbol{\alpha})^2$	2	2	4

(4.1709)

The first term,

$$A^2 = r_i^4, \tag{4.1710}$$

contains no pupil coordinate. For a fixed object point it therefore adds only a constant to the wavefront and cannot change the relative phase between rays crossing different parts of the pupil. It can be discarded from the image-forming aberration function.

Five independent pupil-dependent fourth-order structures remain. Rather than using C^2 and AB separately, it is convenient to replace C^2 by the combination $2C^2 - AB$, because this separates the angularly varying astigmatic contribution from the rotationally symmetric field-curvature contribution. We therefore choose

$$B^2 = \alpha^4, \tag{4.1711}$$

$$BC = \alpha^2 \left(\mathbf{r}_i \cdot \boldsymbol{\alpha}\right), \tag{4.1712}$$

$$2C^2 - AB = 2\left(\mathbf{r}_i \cdot \boldsymbol{\alpha}\right)^2 - r_i^2\alpha^2, \tag{4.1713}$$

$$AB = r_i^2\alpha^2, \tag{4.1714}$$

$$AC = r_i^2 \left(\mathbf{r}_i \cdot \boldsymbol{\alpha}\right). \tag{4.1715}$$

Using Eq. (4.1702),

$$C = r_i\alpha\cos\Delta\phi, \tag{4.1716}$$

these become

$$B^2 = \alpha^4, \tag{4.1717}$$

$$BC = r_i \alpha^3 \cos \Delta\phi, \tag{4.1718}$$

$$2C^2 - AB = r_i^2 \alpha^2 \left(2\cos^2 \Delta\phi - 1\right) = r_i^2 \alpha^2 \cos 2\Delta\phi, \tag{4.1719}$$

$$AB = r_i^2 \alpha^2, \tag{4.1720}$$

$$AC = r_i^3 \alpha \cos \Delta\phi. \tag{4.1721}$$

These five terms have exactly the field and pupil dependences of, respectively, spherical aberration, coma, astigmatism, field curvature, and distortion.

Up to the constant conversion between a transverse coordinate in a particular aperture plane and the corresponding paraxial angle, the isotropic part of the fourth-order Seidel wavefront may therefore be written as

$$\boxed{\begin{aligned}\Delta\ell_{\text{Seidel}}(\mathbf{r}_i, \boldsymbol{\alpha}) = {} & \frac{\mathcal{C}}{4}\alpha^4 + \mathcal{K}\, r_i \alpha^3 \cos \Delta\phi \\ & + \frac{\mathcal{A}}{2}\, r_i^2 \alpha^2 \cos 2\Delta\phi + \frac{\mathcal{F}}{2}\, r_i^2 \alpha^2 \\ & + \mathcal{D}\, r_i^3 \alpha \cos \Delta\phi.\end{aligned}} \tag{4.1722}$$

Here $\mathcal{C}$, $\mathcal{K}$, $\mathcal{A}$, $\mathcal{F}$, and $\mathcal{D}$ denote the spherical-aberration, coma, astigmatism, field-curvature, and distortion coefficients, respectively. Hawkes and Kasper use the corresponding notation C, K, A, F, and D (Hawkes and Kasper, 2017, Chap. 24). We use calligraphic symbols here to avoid confusion with coefficients such as C_3 and C_5 introduced in the preceding sections.

The complete theory of round magnetic lenses also contains anisotropic partners of several of these coefficients, conventionally denoted by k, a, and d for coma, astigmatism, and distortion, respectively (Hawkes and Kasper, 2017, Chap. 24). They introduce the corresponding sine harmonics. We shall not reproduce that complete catalogue here because the symmetry and Zernike descriptions developed in Sec. 4.11.7 already provide the language needed to represent them.

The terminology "third-order Seidel aberrations" refers to the geometrical ray error, not to the degree of the wavefront polynomial. The wavefront in Eq. (4.1722) is fourth order in field and aperture coordinates. Since the ray error is obtained from derivatives of the wavefront, as discussed in

Eq. (4.1269), its leading dependence is one order lower. Consequently,

$$\boxed{\begin{array}{ccc} \text{spherical aberration} & : & \alpha^3, \\ \text{coma} & : & r_i\alpha^2, \\ \text{astigmatism} & : & r_i^2\alpha, \\ \text{field curvature} & : & r_i^2\alpha, \\ \text{distortion} & : & r_i^3, \end{array}} \tag{4.1723}$$

which recovers the scaling introduced earlier in Eq. (4.1233).

Classical Seidel coma

The classical coma contribution in Eq. (4.1722) is

$$\boxed{\Delta\ell_{\text{coma}}^{\text{Seidel}} = \mathcal{K}\, r_i\alpha^3 \cos(\phi - \phi_i).} \tag{4.1724}$$

Its magnitude is proportional to the distance of the object point from the optical axis. Consequently,

$$r_i = 0 \qquad \Longrightarrow \qquad \Delta\ell_{\text{coma}}^{\text{Seidel}} = 0. \tag{4.1725}$$

An ideal rotationally symmetric lens therefore has no classical Seidel coma for an exactly axial object point.

This should be distinguished from the axial coma studied in Sec. 4.11.6. The pupil dependence is closely related: both contain a cubic radial dependence and a first azimuthal harmonic. Their physical origins, however, are different. Classical Seidel coma is generated by an *off-axis object point* imaged through an otherwise centred rotationally symmetric system. Axial coma can instead occur for an axial object point when the illumination or the electron-optical system is displaced or tilted with respect to the coma-free axis.

For a fixed object position $\mathbf{r}_i$, the quantities r_i and ϕ_i in Eq. (4.1724) are constants. The aberration therefore appears locally as an ordinary coma-like pupil phase. Its decomposition into $Z_3^{\pm 1}$ together with the lower-order tilt contribution is the one derived in Eq. (4.1658). What identifies it as *Seidel* coma is that its amplitude is proportional to r_i and its orientation follows the azimuth ϕ_i of the object point.

Seidel astigmatism and field curvature

The astigmatism and field-curvature contributions are particularly instructive because they have the same powers of r_i and α but different angular symmetries. From Eq. (4.1722), their combined wavefront error is

$$\boxed{\Delta\ell_{\mathrm{A+F}} = \frac{r_i^2\alpha^2}{2}\left[\mathcal{F} + \mathcal{A}\cos 2(\phi - \phi_i)\right].} \tag{4.1726}$$

The field-curvature term is rotationally symmetric in the pupil. For a fixed object point it has the same quadratic pupil dependence as ordinary defocus. Its coefficient, however, grows as r_i^2. Consequently, the best-focus plane changes as one moves away from the optical axis. The locus of best focus is therefore generally a curved surface rather than the single plane predicted by Gaussian optics.

The astigmatic term has instead a twofold angular dependence. Along the radial direction of the object point,

$$\phi = \phi_i, \tag{4.1727}$$

the quadratic coefficient is proportional to

$$\mathcal{F} + \mathcal{A}, \tag{4.1728}$$

whereas in the perpendicular direction,

$$\phi = \phi_i + \frac{\pi}{2}, \tag{4.1729}$$

it is proportional to

$$\mathcal{F} - \mathcal{A}. \tag{4.1730}$$

The two orthogonal meridians therefore have different wavefront curvatures and different best-focus positions. These give the classical tangential, or meridional, and sagittal focal surfaces; their axial separation is the *astigmatic difference* (Hawkes and Kasper, 2017, Sec. 24.5).

This classical off-axis astigmatism should not be confused with the twofold astigmatism used routinely in cryo-EM and derived in Sec. 4.11.4. The latter contains the pupil term

$$\alpha^2\cos\left[2(\phi - \phi_A)\right], \tag{4.1731}$$

whose amplitude and orientation characterize the optical system and need not vanish on the optical axis. Classical Seidel astigmatism instead contains

$$r_i^2\alpha^2\cos\left[2(\phi - \phi_i)\right], \tag{4.1732}$$

so it vanishes for $r_i = 0$, grows quadratically with field position, and has a natural orientation determined by the direction of $\mathbf{r}_i$.

The Zernike interpretation is immediate. For a fixed field position, Seidel astigmatism is a linear combination of Z_2^2 and Z_2^{-2}, exactly as in Eq. (4.1654). Field curvature is a field-dependent defocus and therefore has the decomposition

$$\rho^2 = \frac{1}{2}\left(Z_2^0 + Z_0^0\right) \tag{4.1733}$$

derived in Eq. (4.1648).

Distortion

The distortion contribution is

$$\boxed{\Delta\ell_{\text{dist}} = \mathcal{D}\, r_i^3 \alpha \cos(\phi - \phi_i).} \tag{4.1734}$$

Using Eq. (4.1702), this may equivalently be written as

$$\Delta\ell_{\text{dist}} = \mathcal{D}\, r_i^2 \left(\mathbf{r}_i \cdot \boldsymbol{\alpha}\right). \tag{4.1735}$$

Why distortion shifts an image point rather than blurring it.

Apart from the constant proportionality between aperture position and $\boldsymbol{\alpha}$, differentiation of Eq. (4.1735) with respect to the pupil coordinate gives

$$\nabla_{\boldsymbol{\alpha}} \Delta\ell_{\text{dist}} = \mathcal{D}\, r_i^2 \mathbf{r}_i. \tag{4.1736}$$

The result does not depend on $\boldsymbol{\alpha}$. All rays belonging to the same object point therefore acquire the same additional transverse displacement. They remain focused to a point, but that point is shifted from its Gaussian position.

If $\mathbf{r}_o = M\mathbf{r}_i$ is the Gaussian image position defined in Eq. (4.1699), distortion changes the actual image position to

$$\mathbf{r}_o + \delta\mathbf{r}_o, \tag{4.1737}$$

with leading field dependence

$$\boxed{\|\delta\mathbf{r}_o\| \propto r_i^3.} \tag{4.1738}$$

Thus distortion changes the mapping

$$\mathbf{r}_i \longmapsto \mathbf{r}_o \tag{4.1739}$$

without, to this order, broadening the image of an individual object point.

Its Zernike interpretation is equally simple. Since its pupil dependence is linear in α, distortion appears locally as a combination of the two tilt modes Z_1^1 and Z_1^{-1}. A linear pupil phase produces a translation rather than a change in the intrinsic shape of the point-spread function.

A field-dependent pupil description

The relation between the Seidel and Zernike descriptions can now be summarized compactly. At any fixed object position $\mathbf{r}_i$, the Seidel aberrations are ordinary pupil aberrations:

$$\boxed{\begin{array}{ccl} \text{spherical aberration} & : & \alpha^4, \\ \text{coma} & : & r_i\alpha^3\cos(\phi-\phi_i), \\ \text{astigmatism} & : & r_i^2\alpha^2\cos 2(\phi-\phi_i), \\ \text{field curvature} & : & r_i^2\alpha^2, \\ \text{distortion} & : & r_i^3\alpha\cos(\phi-\phi_i). \end{array}} \tag{4.1740}$$

The special feature of the Seidel description is therefore not a new set of pupil shapes. It is the systematic dependence of their coefficients on the field coordinate $\mathbf{r}_i$.

This gives a useful interpretation of the relation between the two theories developed in the preceding sections:

> Zernike polynomials describe *what the aberration wavefront looks like over the pupil.* Seidel theory additionally describes *how that pupil wavefront changes as the object point moves through the field.*

Why field-dependent Seidel aberrations are less central in single-particle cryo-EM

The coherent pupil transfer function introduced in Eq. (4.1237) was written as

$$H(\mathbf{R}) = P(\mathbf{R})\exp\left[-j\chi(\mathbf{R})\right]. \tag{4.1741}$$

This notation implicitly assumes that the same pupil phase applies throughout the region being imaged. Appreciable Seidel field dependence requires the more general expression

$$\boxed{H(\mathbf{R};\mathbf{r}_i) = P(\mathbf{R};\mathbf{r}_i)\exp\left[-j\chi(\mathbf{R};\mathbf{r}_i)\right].} \tag{4.1742}$$

Using Eq. (4.1699), the same dependence could equivalently be parametrized by the Gaussian image coordinate $\mathbf{r}_o$.

A single-particle cryo-EM image normally uses only a small region around one particle. Suppose that the particle is centred at the field position $\mathbf{r}_{i0}$ and that

$$\mathbf{r}_i = \mathbf{r}_{i0} + \delta\mathbf{r}_i \tag{4.1743}$$

within the particle image. If the field-dependent aberrations vary only weakly across that region, then

$$\boxed{\chi\left(\mathbf{R};\mathbf{r}_{i0} + \delta\mathbf{r}_i\right) \simeq \chi\left(\mathbf{R};\mathbf{r}_{i0}\right).} \tag{4.1744}$$

The optical system may therefore be treated as approximately shift-invariant *locally*, and the usual pupil-phase CTF can be applied to that particle.

Equation (4.1744) does not imply that $\mathbf{r}_{i0}$ itself must be small. A particle may lie far from the centre of the recorded field. The approximation only requires the aberration function to change negligibly over the spatial extent of the individual particle. Different particles at different field positions may, in principle, experience slightly different local aberration parameters.

This is why the image-formation model developed in the following sections concentrates primarily on the pupil-phase aberrations that act locally on each particle: defocus, twofold astigmatism, spherical aberration, coma, and their higher-order generalizations. The classical Seidel aberrations remain part of the complete electron-optical description, but reproducing the detailed lens-field integrals for all of their coefficients is not necessary for the cryo-EM image formation model developed here.

4.11.9 Chromatic aberration

The aberrations considered in the preceding sections were deterministic: for a given electron energy and a fixed state of the microscope, each spatial frequency acquires a well-defined phase. Chromatic aberration is different. The focal strength of an electron lens depends on the electron energy, so electrons with slightly different energies are focused at slightly different axial positions. Thus a beam with a distribution of energies does not have one

unique defocus. Instead, chromatic aberration converts the energy distribution into a distribution of focal positions, as illustrated in Fig. 4.14. Each energy component propagates coherently, but the detector records the statistical combination described by Eq. (3.766). Chromatic aberration therefore appears in phase-contrast imaging primarily as a loss of *temporal coherence*.

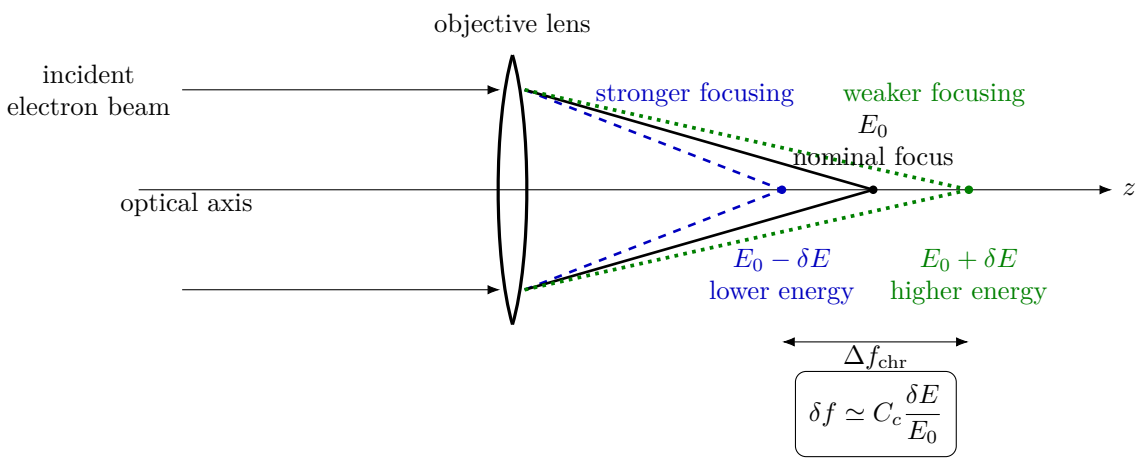


Figure 4.14: Illustration of chromatic aberration in an electron objective lens. Electrons with slightly different energies experience different focusing strengths and therefore come to focus at different axial positions. Lower-energy electrons are focused more strongly, whereas higher-energy electrons are focused more weakly. A finite energy distribution is thus converted into a distribution of effective defocus values. The sign of δf depends on the defocus convention adopted in the text.

This effect was introduced qualitatively in Sec. 3.11.4. In particular, Eq. (3.767) defined the chromatic-aberration coefficient C_c through the first-order relation

$$\boxed{\delta f \simeq C_c \frac{\delta E}{E_0},} \tag{4.1745}$$

up to the sign convention adopted for defocus. Here E_0 is the mean electron energy and δE is a small deviation from it.

Equation (4.1745) should be understood as an operational definition of the chromatic coefficient used in the present image-formation model. More detailed electron-optical definitions may introduce relativistic correction factors in the relation between energy, accelerating voltage, and focal displacement; these factors can be absorbed into the experimentally relevant derivative of focal position with respect to energy. What matters here is that,

to first order, an energy fluctuation produces a proportional fluctuation of defocus.

From chromatic defocus spread to a coherence envelope

The connection with the defocus theory developed earlier is particularly simple. According to Eq. (4.1191), the defocus part of the aberration phase is

$$\chi_{\text{def}}(\mathbf{R}) = -\pi\lambda\Delta f\, R^2, \qquad R = \|\mathbf{R}\|. \tag{4.1746}$$

A fluctuation δf of the effective defocus therefore produces the additional phase

$$\boxed{\delta\chi(\mathbf{R}) = -\pi\lambda\delta f\, R^2.} \tag{4.1747}$$

Temporal-coherence envelope produced by focus spread.

Let the effective defocus fluctuate around its mean value with standard deviation σ_f. From Eq. (4.1747),

$$\frac{\partial\chi}{\partial f} = -\pi\lambda R^2. \tag{4.1748}$$

The generic Gaussian temporal-coherence result derived in Eq. (3.776) therefore gives

$$\boxed{E_c(R) = \exp\left[-\frac{1}{2}\left(\pi\lambda\sigma_f\right)^2 R^4\right].} \tag{4.1749}$$

If the focus spread is specified by its full width at half maximum, Eq. (3.773) gives

$$\Delta f_{\text{FWHM}} = 2\sqrt{2\ln 2}\,\sigma_f. \tag{4.1750}$$

Substitution into Eq. (4.1749) yields

$$\boxed{E_c(R) = \exp\left[-\frac{1}{\ln 2}\left(\frac{\pi}{4}\lambda\Delta f_{\text{FWHM}}\right)^2 R^4\right].} \tag{4.1751}$$

Equation (4.1751) makes several properties of chromatic damping immediately apparent. First,

$$E_c(0) = 1, \tag{4.1752}$$

so the low spatial frequencies are essentially unaffected. Second, the exponent increases as R^4, so the damping becomes rapidly stronger at high resolution. Third, the damping depends quadratically on both the electron wavelength and the focus spread.

Other sources of temporal focus spread

Not every fluctuation of focus is caused by a difference in electron energy. During an exposure, fluctuations of the accelerating voltage and of the objective-lens current can also change the focal position. These effects are usually grouped with chromatic aberration because their consequence for the image is the same: the recorded image averages over a distribution of defocus values.

It is useful to formulate this more generally. Let small fluctuations of energy, accelerating voltage, and objective-lens current be denoted by δE, δV, and δI, respectively. To first order, the resulting focus fluctuation can be written as

$$\boxed{\delta f = a_E \delta E + a_V \delta V + a_I \delta I,} \tag{4.1753}$$

where

$$a_E = \left.\frac{\partial f}{\partial E}\right|_0, \qquad a_V = \left.\frac{\partial f}{\partial V}\right|_0, \qquad a_I = \left.\frac{\partial f}{\partial I}\right|_0. \tag{4.1754}$$

If these fluctuations are statistically independent and have zero mean, their variances add:

$$\boxed{\sigma_f^2 = a_E^2 \sigma_E^2 + a_V^2 \sigma_V^2 + a_I^2 \sigma_I^2.} \tag{4.1755}$$

This equation is more fundamental than any particular parametrization of the microscope instabilities.

In the conventional simplified model, the different sensitivities are often expressed through the chromatic coefficient itself, giving approximately

$$\boxed{\sigma_f^2 \simeq C_c^2 \left[\left(\frac{\sigma_E}{E_0}\right)^2 + \left(\frac{\sigma_V}{V_0}\right)^2 + 4\left(\frac{\sigma_I}{I_0}\right)^2 \right].} \tag{4.1756}$$

The numerical factors in Eq. (4.1756) depend somewhat on the electron-optical convention and on the response of the particular objective lens. For this reason, the experimentally relevant quantity for image formation is ultimately the effective focus spread σ_f, rather than the individual instability parameters.

Why current fluctuations contribute with a factor of 4.

The factor of 4 associated with the objective-lens current in Eq. (4.1756) may initially appear arbitrary, but it follows directly from the fundamental physics of magnetic electron lenses and standard statistical error propagation.

For a standard round magnetic lens, the focal length f is determined by the accelerating voltage V (which dictates the electron kinetic energy E) and the current I passing through the lens coils. To a first approximation, Busch's theorem dictates that the focal length scales as

$$f = K\frac{V}{I^2}, \tag{4.1757}$$

where K is a constant depending on the specific lens geometry and the charge-to-mass ratio of the electron.

Taking the natural logarithm of both sides separates the variables:

$$\ln f = \ln K + \ln V - 2\ln I. \tag{4.1758}$$

Differentiating this expression relates small fluctuations in the operational parameters to the resulting fluctuation in focal length:

$$\boxed{\frac{\delta f}{f_0} = \frac{\delta V}{V_0} - 2\frac{\delta I}{I_0},} \tag{4.1759}$$

where f_0, V_0, and I_0 are the nominal operating values. Because the kinetic energy E is directly proportional to the accelerating voltage ($E = eV$), an independent energy spread δE in the beam behaves identically to a voltage fluctuation, giving an equivalent fractional term $\delta E/E_0$.

The chromatic aberration coefficient C_c acts as the scaling constant that converts these fractional changes into an absolute focal shift. The total instantaneous focus fluctuation is therefore

$$\delta f = C_c\frac{\delta E}{E_0} + C_c\frac{\delta V}{V_0} - 2C_c\frac{\delta I}{I_0}. \tag{4.1760}$$

During an exposure, the fluctuations in beam energy, accelerating voltage, and lens current can be modeled as statistically independent random processes with zero mean. Consequently, their variances add.

The variance of the focus spread, $\sigma_f^2 = \langle(\delta f)^2\rangle$, becomes

$$\sigma_f^2 = C_c^2 \left[\left(\frac{\sigma_E}{E_0}\right)^2 + \left(\frac{\sigma_V}{V_0}\right)^2 + (-2)^2 \left(\frac{\sigma_I}{I_0}\right)^2 \right]. \tag{4.1761}$$

Squaring the prefactor of -2 for the current fluctuation naturally generates the factor of 4:

$$\boxed{\sigma_f^2 = C_c^2 \left[\left(\frac{\sigma_E}{E_0}\right)^2 + \left(\frac{\sigma_V}{V_0}\right)^2 + 4 \left(\frac{\sigma_I}{I_0}\right)^2 \right].} \tag{4.1762}$$

Thus, the factor of 4 is not a generic convention, but a direct consequence of the inverse-square dependence of the focal length on the magnetic coil current.

If all three distributions are Gaussian, the same quadrature relation can be written directly in terms of their FWHM values,

$$\boxed{\Delta f_{\mathrm{FWHM}} \simeq C_c \sqrt{\left(\frac{\Delta E_{\mathrm{FWHM}}}{E_0}\right)^2 + \left(\frac{\Delta V_{\mathrm{FWHM}}}{V_0}\right)^2 + 4 \left(\frac{\Delta I_{\mathrm{FWHM}}}{I_0}\right)^2}.} \tag{4.1763}$$

Once this effective FWHM focus spread is known, all of these mechanisms enter the transfer function through the single envelope Eq. (4.1751).

What the envelope means for the CTF

The chromatic envelope does not introduce a new deterministic aberration phase at the mean electron energy. Instead, it attenuates the coherent transfer produced by that phase. If $H_{\mathrm{coh}}(\mathbf{R})$ denotes the coherent transfer function of Eq. (4.1237), the temporally averaged transfer is, under the Gaussian approximation,

$$\boxed{H(\mathbf{R}) = E_c(R)\, H_{\mathrm{coh}}(\mathbf{R}).} \tag{4.1764}$$

The corresponding power spectrum contains

$$|H(\mathbf{R})|^2 = E_c^2(R)\, |H_{\mathrm{coh}}(\mathbf{R})|^2 . \tag{4.1765}$$

Thus an envelope quoted for the CTF amplitude must be squared when its effect on the power spectrum is considered.

The Gaussian form is not fundamental. From Eq. (3.771), the temporal-coherence factor is the characteristic function of the actual distribution of

energy or focus fluctuations. A non-Gaussian or asymmetric energy distribution will therefore produce a different envelope and may, in general, also introduce a residual phase shift. Equation (4.1749) is the particularly useful result obtained when the effective focus distribution is narrow and Gaussian.

For cryo-EM this distinction is important because the energy spread of the source, the chromatic coefficient of the objective lens, and the stability of the microscope determine how rapidly high-resolution information is attenuated. Chromatic aberration is therefore not merely another item in the catalogue of geometrical aberrations: together with the finite energy spread of the illumination, it sets one of the fundamental temporal-coherence limits of the microscope.

4.12 Amplitude contrast

Amplitude contrast is conceptually different from the aberrations studied in Sec. 4.11. An aberration modifies the phase of the electron wave as it propagates through the microscope and is represented by the aberration phase $\chi(\mathbf{R})$. Amplitude contrast instead arises because the magnitude of the electron wave emerging from the specimen need not be spatially uniform.

For example, some electrons may be removed from the coherent image-forming beam by inelastic scattering or by elastic scattering to angles that are not transmitted by the objective aperture. The microscopic origin of these processes will be discussed in Chap. 5. Here we only need their consequence for image formation: in addition to a weak phase modulation, the electron wave entering the objective may contain a weak real amplitude modulation.

This distinction is important. Amplitude contrast does *not* correspond to an additional term in the aberration phase $\chi(\mathbf{R})$ and should therefore not be classified as another lens aberration.

4.12.1 A weak object can modify both amplitude and phase

The first CTF model of Sec. 4.10.10 assumed the pure weak-phase wave

$$\psi_i(\mathbf{r}) \simeq 1 + j\varphi(\mathbf{r}) \tag{4.1766}$$

[Eq. (4.1205)]. We now allow the specimen also to produce a small real modulation $a(\mathbf{r})$ of the wave amplitude:

$$\boxed{\psi_i(\mathbf{r}) \simeq 1 + a(\mathbf{r}) + j\varphi(\mathbf{r}), \qquad |a| \ll 1, \qquad |\varphi| \ll 1.} \tag{4.1767}$$

The function $a(\mathbf{r})$ is a *signed fractional amplitude modulation*. Thus an attenuation of the electron wave corresponds to $a < 0$ with the convention of Eq. (4.1767).

The approximation in Eq. (4.1767) follows directly from writing the specimen transmission function in amplitude–phase form.

Weak amplitude–phase approximation.

Let the specimen transmission function be

$$t(\mathbf{r}) = A(\mathbf{r})e^{j\varphi(\mathbf{r})}, \tag{4.1768}$$

where $A(\mathbf{r})$ is real. For a weak object, write

$$A(\mathbf{r}) = 1 + a(\mathbf{r}), \qquad |a| \ll 1. \tag{4.1769}$$

Expanding the exponential to first order,

$$e^{j\varphi} = 1 + j\varphi + \mathcal{O}(\varphi^2). \tag{4.1770}$$

Therefore,

$$\begin{aligned} t(\mathbf{r}) &= (1+a)(1+j\varphi) + \mathcal{O}(\varphi^2) \\ &= 1 + a + j\varphi + ja\varphi + \mathcal{O}(\varphi^2). \end{aligned} \tag{4.1771}$$

The product $a\varphi$ is second order in the weak-object perturbation and can be neglected together with a^2 and φ^2. Hence,

$$\boxed{t(\mathbf{r}) \simeq 1 + a(\mathbf{r}) + j\varphi(\mathbf{r}).} \tag{4.1772}$$

Equation (4.1772) makes the distinction particularly transparent. The real perturbation a changes the *magnitude* of the electron wave, whereas the imaginary perturbation $j\varphi$ changes its *phase*.

4.12.2 Transfer of amplitude and phase contrast

Taking the Fourier transform of Eq. (4.1767),

$$\widehat{\psi}_i(\mathbf{R}) = \delta(\mathbf{R}) + \widehat{a}(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R}). \tag{4.1773}$$

Since both a and φ are real,

$$\widehat{a}(-\mathbf{R}) = \widehat{a}^*(\mathbf{R}), \qquad \widehat{\varphi}(-\mathbf{R}) = \widehat{\varphi}^*(\mathbf{R}). \tag{4.1774}$$

Consider first a symmetric aberration phase, so that

$$\chi(-\mathbf{R}) = \chi(\mathbf{R}), \tag{4.1775}$$

and a centrosymmetric pupil,

$$P(-\mathbf{R}) = P(\mathbf{R}). \tag{4.1776}$$

The coherent wave-transfer function is then Eq. (4.1237),

$$H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi(\mathbf{R})}. \tag{4.1777}$$

The phase and amplitude components of the weak object are transferred differently when the detector forms the image intensity.

The sine and cosine contrast-transfer terms.

Applying the coherent transfer function to Eq. (4.1773) gives

$$\widehat{\psi}_{\text{obs}}(\mathbf{R}) = \delta(\mathbf{R}) + H(\mathbf{R})\left[\widehat{a}(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R})\right]. \tag{4.1778}$$

Let the weak part of the output wave be $g(\mathbf{r})$, so that

$$\psi_{\text{obs}}(\mathbf{r}) = 1 + g(\mathbf{r}). \tag{4.1779}$$

To first order in the weak object,

$$I_{\text{obs}}(\mathbf{r}) = |1+g|^2 \simeq 1 + g + g^*. \tag{4.1780}$$

Thus, defining

$$\delta I(\mathbf{r}) = I_{\text{obs}}(\mathbf{r}) - 1, \tag{4.1781}$$

its Fourier transform is

$$\begin{aligned}\widehat{\delta I}(\mathbf{R}) = {} & H(\mathbf{R})\left[\widehat{a}(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R})\right] \\ & + H^*(-\mathbf{R})\left[\widehat{a}(\mathbf{R}) - j\widehat{\varphi}(\mathbf{R})\right],\end{aligned} \tag{4.1782}$$

where Eq. (4.1774) has been used. Collecting the two contributions,

$$\widehat{\delta I}(\mathbf{R}) = \left[H(\mathbf{R}) + H^*(-\mathbf{R})\right]\widehat{a}(\mathbf{R}) + j\left[H(\mathbf{R}) - H^*(-\mathbf{R})\right]\widehat{\varphi}(\mathbf{R}). \tag{4.1783}$$

For a symmetric aberration,

$$H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi(\mathbf{R})}, \qquad H^*(-\mathbf{R}) = P(\mathbf{R})e^{+j\chi(\mathbf{R})}. \tag{4.1784}$$

Consequently,

$$H(\mathbf{R}) + H^*(-\mathbf{R}) = 2P(\mathbf{R})\cos\chi(\mathbf{R}), \tag{4.1785}$$

whereas, as already derived in Eq. (4.1231),

$$j\left[H(\mathbf{R}) - H^*(-\mathbf{R})\right] = 2P(\mathbf{R})\sin\chi(\mathbf{R}). \tag{4.1786}$$

Therefore,

$$\boxed{\widehat{\delta I}(\mathbf{R}) = 2P(\mathbf{R})\left[\cos\chi(\mathbf{R})\,\widehat{a}(\mathbf{R}) + \sin\chi(\mathbf{R})\,\widehat{\varphi}(\mathbf{R})\right].} \tag{4.1787}$$

Equation (4.1787) is the central result of this section. It shows that the two kinds of object modulation are transferred in *quadrature*:

$$\boxed{\begin{array}{ccc} \text{phase modulation} & \longrightarrow & \sin\chi, \\ \text{amplitude modulation} & \longrightarrow & \cos\chi. \end{array}} \tag{4.1788}$$

The sine transfer derived in Sec. 4.10.10 is therefore not the complete weak-object result; it is the special case $a(\mathbf{r}) = 0$.

4.12.3 Why amplitude contrast survives at focus

The quadrature relation of Eq. (4.1788) has an immediate physical consequence. At a frequency for which the aberration phase vanishes,

$$\chi(\mathbf{R}) = 0, \tag{4.1789}$$

we have

$$\sin\chi = 0, \qquad \cos\chi = 1. \tag{4.1790}$$

Equation (4.1787) therefore reduces to

$$\boxed{\widehat{\delta I}(\mathbf{R}) = 2P(\mathbf{R})\widehat{a}(\mathbf{R}), \qquad \chi(\mathbf{R}) = 0.} \tag{4.1791}$$

Thus a pure weak-phase object produces no first-order intensity contrast when the phase shift between the reference and scattered waves vanishes, whereas an amplitude-modulating object remains visible.

This explains why amplitude contrast contributes especially to the low-frequency image contrast. For an ordinary objective without a phase plate,

$$\chi(\mathbf{0}) = 0, \tag{4.1792}$$

so the pure phase-contrast CTF vanishes at the origin, whereas the amplitude component does not.

4.12.4 The conventional one-parameter amplitude-contrast model

Equation (4.1787) contains two independent specimen functions, $\widehat{a}(\mathbf{R})$ and $\widehat{\varphi}(\mathbf{R})$. In general, there is therefore no reason for amplitude and phase contrast to be described by a *single* scalar CTF multiplying one object function.

The conventional cryo-EM CTF makes an additional approximation: the amplitude and phase modulations are assumed to have sufficiently similar spatial dependence that, over the frequencies of interest, one may write

$$\boxed{\widehat{a}(\mathbf{R}) \simeq Q(\mathbf{R})\widehat{\varphi}(\mathbf{R}).} \tag{4.1793}$$

Substitution into Eq. (4.1787) gives, apart from the overall factor 2,

$$\boxed{\mathrm{CTF}_Q(\mathbf{R}) = P(\mathbf{R})\left[\sin\chi(\mathbf{R}) + Q(\mathbf{R})\cos\chi(\mathbf{R})\right].} \tag{4.1794}$$

There is an important interpretation issue here. The parameter $Q(\mathbf{R})$ in Eq. (4.1794) is most safely understood as the *relative weight of amplitude and phase contrast in the effective image-formation model.* It should not in general be identified literally with the fraction of electrons removed from the beam. The latter is a microscopic scattering quantity, whereas Eq. (4.1793) already contains the additional assumption that the amplitude and phase object functions have the same spatial dependence.

4.12.5 Normalized amplitude contrast

Another common convention uses a normalized amplitude-contrast coefficient A, defined so that the squared weights of the phase and amplitude components sum to one:

$$\boxed{A \triangleq \frac{Q}{\sqrt{1+Q^2}}, \qquad \sqrt{1-A^2} = \frac{1}{\sqrt{1+Q^2}}.} \tag{4.1795}$$

Dividing Eq. (4.1794) by the irrelevant global factor $\sqrt{1+Q^2}$ gives

$$\boxed{\mathrm{CTF}_A(\mathbf{R}) = P(\mathbf{R})\left[\sqrt{1-A^2}\,\sin\chi(\mathbf{R}) + A\cos\chi(\mathbf{R})\right].} \tag{4.1796}$$

The distinction between Eqs. (4.1794) and (4.1796) is therefore only one of normalization when Q is constant.

4.12.6 Amplitude contrast as an effective CTF phase shift

The combination of sine and cosine in Eq. (4.1796) can be written as a single shifted sine. This observation is frequently useful in CTF fitting.

Amplitude contrast as a constant phase shift.

Define an angle ϕ_A by

$$\sin\phi_A = A, \qquad \cos\phi_A = \sqrt{1-A^2}. \tag{4.1797}$$

Thus,

$$\boxed{\phi_A = \arcsin A = \arctan Q.} \tag{4.1798}$$

Using

$$\sin(\chi+\phi_A) = \sin\chi\cos\phi_A + \cos\chi\sin\phi_A, \tag{4.1799}$$

we obtain

$$\boxed{\sqrt{1-A^2}\sin\chi + A\cos\chi = \sin(\chi+\phi_A).} \tag{4.1800}$$

Consequently,

$$\boxed{\mathrm{CTF}_A(\mathbf{R}) = P(\mathbf{R})\sin\left[\chi(\mathbf{R})+\phi_A\right].} \tag{4.1801}$$

Equivalently, using the unnormalized coefficient Q,

$$\boxed{\sin\chi + Q\cos\chi = \sqrt{1+Q^2}\,\sin\left[\chi+\arctan Q\right].} \tag{4.1802}$$

Equation (4.1801) shows that amplitude contrast shifts the CTF zeros. Without amplitude contrast,

$$\chi(\mathbf{R}_n) = n\pi. \tag{4.1803}$$

With a constant amplitude-contrast contribution, Eq. (4.1801) instead gives

$$\boxed{\chi(\mathbf{R}_n) = n\pi - \phi_A.} \tag{4.1804}$$

Thus even a frequency-independent amplitude-contrast coefficient changes the positions of the CTF zero crossings.

At zero aberration phase, Eq. (4.1796) gives

$$\boxed{\mathrm{CTF}_A = A\,P, \qquad \chi = 0.} \tag{4.1805}$$

If the direct beam is transmitted, $P(\mathbf{0}) = 1$, then

$$\mathrm{CTF}_A(\mathbf{0}) = A. \tag{4.1806}$$

This is the mathematical expression of the low-frequency contrast introduced by the amplitude component.

4.12.7 The phase shift by amplitude contrast is not an aberration

Equation (4.1801) may suggest that amplitude contrast can simply be regarded as another contribution to the aberration phase. Physically, this interpretation would be misleading.

The aberration phase $\chi(\mathbf{R})$ describes a phase accumulated by the coherent electron wave during propagation through the microscope. It is related to a wavefront error and, as shown in Sec. 4.11.2, ultimately to departures of electron trajectories from Gaussian imaging.

The quantity ϕ_A in Eq. (4.1798) has a different origin. It appears only after the real and imaginary parts of the weak object wave have been combined in the detected intensity. The representation

$$\sqrt{1 - A^2}\sin\chi + A\cos\chi = \sin(\chi + \phi_A) \tag{4.1807}$$

is therefore an *algebraic reparametrization of the CTF*, not the statement that the objective lens has acquired an additional wavefront aberration.

This distinction also separates amplitude contrast from a physical phase plate. A phase plate genuinely introduces an additional phase difference between selected components of the electron wave; amplitude contrast only produces an equivalent constant shift after the phase- and amplitude-contrast contributions have been combined in the image intensity.

4.12.8 How good is the constant amplitude-contrast approximation?

The most general weak-object result remains Eq. (4.1787). Nothing in that equation requires $\widehat{a}(\mathbf{R})$ and $\widehat{\varphi}(\mathbf{R})$ to be proportional. The introduction of a single amplitude-contrast coefficient is therefore a model reduction, not a fundamental property of electron scattering.

Indeed, the physical processes responsible for attenuation may depend on specimen composition, thickness, scattering angle, energy loss, and the apertures or energy filters used in the experiment. It is therefore natural in the more general model to allow

$$Q = Q(\mathbf{R}). \tag{4.1808}$$

This possibility was already present in Eq. (4.1794). The approximation

$$Q(\mathbf{R}) \simeq Q_0 \tag{4.1809}$$

is useful because it reduces the amplitude contribution to one scalar parameter and, through Eq. (4.1802), to a constant CTF phase offset.

For the standard weak-object CTF model this approximation is usually sufficient. When a more faithful forward model of the specimen is required, however, amplitude attenuation should be incorporated into the specimen transmission function itself rather than represented by a single empirical parameter.

4.13 CTF envelopes

The aberrations studied in Sec. 4.11 describe the coherent action of the microscope on a particular electron wave. For a fixed electron energy, illumination direction, and microscope state, the objective lens produces a definite aberration phase $\chi(\mathbf{R})$ and therefore a definite coherent transfer function.

A real exposure, however, does not generally correspond to one unique coherent realization. The electron energy fluctuates within the energy distribution of the source, different points of a finite source illuminate the specimen from slightly different directions, and the specimen or microscope may move during the exposure. Each realization propagates coherently, but the detector averages the corresponding intensities.

This statistical averaging was established generally in Sec. 3.11. Temporal and spatial coherence envelopes are therefore *not aberrations*. They

describe the loss of contrast that occurs when slightly different coherent transfer functions are combined at detection.

The same mathematical mechanism also describes image degradation caused by motion during the exposure. Motion is not partial coherence in the strict source-statistical sense, but it is convenient to discuss it here because it also produces an averaging envelope.

4.13.1 The general origin of an envelope

Suppose that ξ denotes a fluctuating experimental parameter, such as electron energy, illumination angle, or specimen position. For each fixed value of ξ, the microscope is coherent and produces a weak-object image with Fourier-domain contrast

$$\widehat{\delta I}_\xi(\mathbf{R}) = C_\xi(\mathbf{R})\widehat{o}(\mathbf{R}), \tag{4.1810}$$

where $\widehat{o}(\mathbf{R})$ denotes the relevant weak-object function. Since the detector averages intensities over the statistical ensemble,

$$\left\langle \widehat{\delta I}_\xi(\mathbf{R}) \right\rangle_\xi = \langle C_\xi(\mathbf{R}) \rangle_\xi \, \widehat{o}(\mathbf{R}). \tag{4.1811}$$

Thus, after weak-object linearization, averaging detected intensities is equivalent to averaging the corresponding contrast-transfer factors.

Suppose now that the fluctuation changes only the phase of the coherent transfer factor,

$$H_\xi(\mathbf{R}) = H_0(\mathbf{R}) \exp\left[-j\delta\chi_\xi(\mathbf{R})\right]. \tag{4.1812}$$

Then

$$\boxed{\langle H_\xi(\mathbf{R}) \rangle_\xi = H_0(\mathbf{R}) E_\xi(\mathbf{R}),} \tag{4.1813}$$

where

$$\boxed{E_\xi(\mathbf{R}) \triangleq \left\langle e^{-j\delta\chi_\xi(\mathbf{R})} \right\rangle_\xi .} \tag{4.1814}$$

The envelope is therefore the characteristic function of the distribution of the phase perturbation.

This observation is extremely general. Different physical mechanisms produce different functions $\delta\chi_\xi$ and different statistical distributions, but the averaging operation itself is always the one in Eq. (4.1814).

Gaussian fluctuations produce Gaussian envelopes.

Suppose that the fluctuating parameter ξ follows a zero-mean Gaussian distribution with standard deviation σ_ξ,

$$p(\xi) = \frac{1}{\sqrt{2\pi}\sigma_\xi} \exp\left[-\frac{\xi^2}{2\sigma_\xi^2}\right]. \tag{4.1815}$$

If the corresponding phase perturbation is linear in ξ,

$$\delta\chi_\xi(\mathbf{R}) = g(\mathbf{R})\xi, \tag{4.1816}$$

then Eq. (4.1814) becomes

$$\boxed{E_\xi(\mathbf{R}) = \int_{-\infty}^{\infty} p(\xi) e^{-jg(\mathbf{R})\xi} d\xi.} \tag{4.1817}$$

This integral has exactly the structure of a Fourier transform, although the transformed variable is now ξ, not a spatial coordinate. Recall that with the Fourier-transform convention introduced in Eq. (3.2),

$$\widehat{f}(R_\xi) = \int_{-\infty}^{\infty} f(\xi) e^{-j2\pi R_\xi \xi} d\xi. \tag{4.1818}$$

Comparing this expression with Eq. (4.1817), we identify the Fourier frequency conjugate to ξ as

$$\boxed{R_\xi = \frac{g(\mathbf{R})}{2\pi}.} \tag{4.1819}$$

Therefore,

$$E_\xi(\mathbf{R}) = \widehat{p}\left(\frac{g(\mathbf{R})}{2\pi}\right). \tag{4.1820}$$

Notice that R_ξ is not the spatial frequency $\mathbf{R}$ of the image. It is the variable Fourier-conjugate to the fluctuating physical quantity ξ. The dependence on the image spatial frequency enters through the function $g(\mathbf{R})$.

We now evaluate this Fourier transform explicitly. Substituting Eq. (4.1815) into Eq. (4.1817) gives

$$E_\xi(\mathbf{R}) = \frac{1}{\sqrt{2\pi}\sigma_\xi} \int_{-\infty}^{\infty} \exp\left[-\frac{\xi^2}{2\sigma_\xi^2}\right] \exp\left[-jg(\mathbf{R})\xi\right] d\xi. \tag{4.1821}$$

Introduce the dimensionless variable

$$u = \frac{\xi}{\sigma_\xi}, \qquad d\xi = \sigma_\xi \, du, \tag{4.1822}$$

and define

$$a \triangleq \sigma_\xi g(\mathbf{R}). \tag{4.1823}$$

Equation (4.1821) becomes

$$E_\xi(\mathbf{R}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} e^{-jau} du. \tag{4.1824}$$

To evaluate this integral without assuming beforehand the Fourier transform of a Gaussian, define

$$F(a) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} e^{-jau} du. \tag{4.1825}$$

Differentiating with respect to a,

$$\frac{dF}{da} = -\frac{j}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-u^2/2} e^{-jau} du. \tag{4.1826}$$

Since

$$u e^{-u^2/2} = -\frac{d}{du} e^{-u^2/2}, \tag{4.1827}$$

Eq. (4.1826) can be written as

$$\frac{dF}{da} = \frac{j}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d}{du}\left(e^{-u^2/2}\right) e^{-jau} du. \tag{4.1828}$$

Integrating by parts, the boundary term vanishes because $e^{-u^2/2} \to 0$ as $|u| \to \infty$, and we obtain

$$\begin{aligned} \frac{dF}{da} &= -a\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2/2}e^{-jau}du \\ &= -aF(a). \end{aligned} \tag{4.1829}$$

Thus F satisfies the differential equation

$$\frac{dF}{da} = -aF. \tag{4.1830}$$

Its solution is

$$F(a) = Ce^{-a^2/2}. \tag{4.1831}$$

At $a = 0$,

$$F(0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2/2}du = 1, \tag{4.1832}$$

because the Gaussian distribution is normalized. Hence $C = 1$, and

$$\boxed{F(a) = e^{-a^2/2}.} \tag{4.1833}$$

Finally, using Eq. (4.1823),

$$\boxed{E_\xi(\mathbf{R}) = \exp\left[-\frac{1}{2}\sigma_\xi^2 g^2(\mathbf{R})\right].} \tag{4.1834}$$

Equation (4.1834) is the origin of many of the Gaussian envelopes encountered in electron microscopy. It also shows why an envelope becomes stronger whenever the coherent phase becomes more sensitive to the fluctuating parameter.

4.13.2 Temporal coherence

Temporal partial coherence was introduced in Sec. 3.11.4. The detected intensity for an energy-incoherent source was shown in Eq. (3.766) to be an average over coherently propagated electron waves of different energies.

The most important effect of the energy distribution on the objective lens is the chromatic focus spread discussed in Sec. 4.11.9. A fluctuation δf of defocus changes the aberration phase by

$$\delta\chi = -\pi\lambda\delta f\, R^2 \tag{4.1835}$$

according to Eq. (4.1747). For a Gaussian focus distribution of standard deviation σ_f, Eq. (4.1834) immediately gives

$$\boxed{E_t(R) = \exp\left[-\frac{1}{2}\left(\pi\lambda\sigma_f\right)^2 R^4\right],} \tag{4.1836}$$

which is the temporal-coherence envelope already derived in Eq. (4.1749).

The important point here is conceptual. Chromatic aberration determines how strongly focus changes with energy; the energy distribution determines which focal positions are sampled; and the exponential damping in Eq. (4.1836) appears only after these coherent realizations are averaged.

4.13.3 Finite source size and spatial coherence

Spatial partial coherence has a closely analogous origin. As discussed in Sec. 3.11.4, a finite effective source produces an ensemble of illumination directions. Let

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_x \\ \theta_y \end{pmatrix} \tag{4.1837}$$

be a small illumination tilt relative to the optical axis.

According to the angular-spectrum relation derived in Eq. (3.367), the corresponding transverse spatial frequency of the incident plane wave is

$$\boxed{\mathbf{R}_0 \simeq \frac{\boldsymbol{\theta}}{\lambda}.} \tag{4.1838}$$

For axial illumination, an object spatial frequency $\mathbf{R}$ is transferred through the pupil at $\mathbf{R}$. With tilted illumination, the unscattered reference beam passes through the pupil at $\mathbf{R}_0$, whereas the corresponding scattered component passes through it at

$$\mathbf{R} + \mathbf{R}_0. \tag{4.1839}$$

The relevant aberration phase difference is therefore

$$\boxed{\chi_{\boldsymbol{\theta}}(\mathbf{R}) = \chi\left(\mathbf{R} + \mathbf{R}_0\right) - \chi(\mathbf{R}_0).} \tag{4.1840}$$

Phase variation produced by a small illumination tilt.

For a sufficiently small illumination angle, Eq. (4.1840) can be ex-

panded to first order in $\mathbf{R}_0$:

$$\chi(\mathbf{R}+\mathbf{R}_0) \simeq \chi(\mathbf{R}) + \mathbf{R}_0 \cdot \boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R}), \tag{4.1841}$$

$$\chi(\mathbf{R}_0) \simeq \chi(\mathbf{0}) + \mathbf{R}_0 \cdot \boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{0}). \tag{4.1842}$$

Consequently,

$$\begin{aligned}\chi_{\boldsymbol{\theta}}(\mathbf{R}) \simeq\ & \chi(\mathbf{R}) - \chi(\mathbf{0}) \\ & + \mathbf{R}_0 \cdot \left[\boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R}) - \boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{0})\right].\end{aligned} \tag{4.1843}$$

For the usual axial aberrations, $\chi(\mathbf{0}) = 0$ and $\boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{0}) = 0$. Using Eq. (4.1838), the fluctuating part of the phase is therefore

$$\boxed{\delta\chi_{\boldsymbol{\theta}}(\mathbf{R}) \simeq \frac{\boldsymbol{\theta}}{\lambda} \cdot \boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R}).} \tag{4.1844}$$

Equation (4.1844) has a particularly simple interpretation: spatial coherence is lost when a small change in illumination direction produces a large change in aberration phase.

Let the illumination directions have the circular Gaussian distribution

$$\boxed{p(\boldsymbol{\theta}) = \frac{1}{\pi\alpha_s^2} \exp\left[-\frac{\theta_x^2+\theta_y^2}{\alpha_s^2}\right].} \tag{4.1845}$$

With this convention, each Cartesian component has variance

$$\sigma_{\theta_x}^2 = \sigma_{\theta_y}^2 = \frac{\alpha_s^2}{2}. \tag{4.1846}$$

Thus α_s is an effective Gaussian angular width; it should not be confused with the sharp radius of a hard circular condenser aperture.

Spatial-coherence envelope for a Gaussian source.

Substituting Eq. (4.1844) into the general averaging expression of Eq. (4.1814), we obtain

$$\begin{aligned}E_s(\mathbf{R}) = & \iint p(\boldsymbol{\theta}) \\ & \times \exp\left[-\frac{j}{\lambda}\boldsymbol{\theta} \cdot \boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R})\right] d\theta_x\, d\theta_y.\end{aligned} \tag{4.1847}$$

Let us evaluate this integral explicitly. First, write the gradient in Cartesian components,

$$\boldsymbol{\nabla}_{\mathbf{R}}\chi = \begin{pmatrix} \dfrac{\partial\chi}{\partial R_x} \\ \dfrac{\partial\chi}{\partial R_y} \end{pmatrix}. \tag{4.1848}$$

Therefore, the scalar product in Eq. (4.1847) is

$$\boldsymbol{\theta}\cdot\boldsymbol{\nabla}_{\mathbf{R}}\chi = \theta_x\frac{\partial\chi}{\partial R_x} + \theta_y\frac{\partial\chi}{\partial R_y}. \tag{4.1849}$$

The phase factor can consequently be separated as

$$\begin{aligned} &\exp\left[-\frac{j}{\lambda}\left(\theta_x\frac{\partial\chi}{\partial R_x} + \theta_y\frac{\partial\chi}{\partial R_y}\right)\right] \\ &\quad = \exp\left[-\frac{j\theta_x}{\lambda}\frac{\partial\chi}{\partial R_x}\right]\exp\left[-\frac{j\theta_y}{\lambda}\frac{\partial\chi}{\partial R_y}\right]. \end{aligned} \tag{4.1850}$$

The Gaussian illumination distribution of Eq. (4.1845) can also be factorized:

$$\begin{aligned} p(\boldsymbol{\theta}) &= \frac{1}{\pi\alpha_s^2}\exp\left[-\frac{\theta_x^2+\theta_y^2}{\alpha_s^2}\right] \\ &= \left[\frac{1}{\sqrt{\pi}\alpha_s}\exp\left(-\frac{\theta_x^2}{\alpha_s^2}\right)\right]\left[\frac{1}{\sqrt{\pi}\alpha_s}\exp\left(-\frac{\theta_y^2}{\alpha_s^2}\right)\right]. \end{aligned} \tag{4.1851}$$

Thus,

$$p(\boldsymbol{\theta}) = p_x(\theta_x)\,p_y(\theta_y), \tag{4.1852}$$

where both p_x and p_y are one-dimensional Gaussian distributions.

To identify their standard deviations, compare

$$p_x(\theta_x) = \frac{1}{\sqrt{\pi}\alpha_s}\exp\left[-\frac{\theta_x^2}{\alpha_s^2}\right] \tag{4.1853}$$

with the standard Gaussian form

$$p(\theta_x) = \frac{1}{\sqrt{2\pi}\sigma_\theta}\exp\left[-\frac{\theta_x^2}{2\sigma_\theta^2}\right]. \tag{4.1854}$$

It follows that

$$\boxed{\sigma_\theta^2 = \frac{\alpha_s^2}{2}.} \tag{4.1855}$$

Using Eqs. (4.1850) and (4.1851), the double integral of Eq. (4.1847) becomes

$$\begin{aligned} E_s(\mathbf{R}) = &\left\{\int_{-\infty}^{\infty} p_x(\theta_x) \exp\left[-\frac{j\theta_x}{\lambda}\frac{\partial\chi}{\partial R_x}\right] d\theta_x\right\} \\ &\times \left\{\int_{-\infty}^{\infty} p_y(\theta_y) \exp\left[-\frac{j\theta_y}{\lambda}\frac{\partial\chi}{\partial R_y}\right] d\theta_y\right\}. \end{aligned} \tag{4.1856}$$

We have therefore reduced the two-dimensional problem to two independent one-dimensional Gaussian averages.

The first integral has exactly the form evaluated in Eq. (4.1834), with

$$\xi = \theta_x, \qquad \sigma_\xi^2 = \frac{\alpha_s^2}{2}, \qquad g_x(\mathbf{R}) = \frac{1}{\lambda}\frac{\partial\chi}{\partial R_x}. \tag{4.1857}$$

Hence,

$$E_{s,x}(\mathbf{R}) = \exp\left[-\frac{\alpha_s^2}{4\lambda^2}\left(\frac{\partial\chi}{\partial R_x}\right)^2\right]. \tag{4.1858}$$

Similarly,

$$E_{s,y}(\mathbf{R}) = \exp\left[-\frac{\alpha_s^2}{4\lambda^2}\left(\frac{\partial\chi}{\partial R_y}\right)^2\right]. \tag{4.1859}$$

Since the two factors multiply,

$$\begin{aligned} E_s(\mathbf{R}) &= E_{s,x}(\mathbf{R})E_{s,y}(\mathbf{R}) \\ &= \exp\left[-\frac{\alpha_s^2}{4\lambda^2}\left\{\left(\frac{\partial\chi}{\partial R_x}\right)^2 + \left(\frac{\partial\chi}{\partial R_y}\right)^2\right\}\right]. \end{aligned} \tag{4.1860}$$

Finally, by definition,

$$\|\boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R})\|^2 = \left(\frac{\partial\chi}{\partial R_x}\right)^2 + \left(\frac{\partial\chi}{\partial R_y}\right)^2, \tag{4.1861}$$

and therefore

$$\boxed{E_s(\mathbf{R}) = \exp\left[-\frac{\alpha_s^2}{4\lambda^2}\left\|\boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R})\right\|^2\right].} \tag{4.1862}$$

Equation (4.1862) is the general Gaussian spatial-coherence envelope. Unlike the temporal envelope, its frequency dependence is controlled by the *gradient* of the coherent aberration phase across the pupil.

Defocus and spherical aberration

For a rotationally symmetric objective, the defocus–spherical-aberration phase derived in Eq. (4.1466) is

$$\chi(R) = -\pi\lambda\Delta f R^2 + \frac{\pi}{2}C_s\lambda^3 R^4. \tag{4.1863}$$

Since the phase depends only on R,

$$\left\|\boldsymbol{\nabla}_{\mathbf{R}}\chi\right\| = \left|\frac{d\chi}{dR}\right|. \tag{4.1864}$$

Differentiating,

$$\frac{d\chi}{dR} = 2\pi\lambda\left(C_s\lambda^2 R^3 - \Delta f R\right). \tag{4.1865}$$

Substitution into Eq. (4.1862) gives

$$\boxed{E_s(R) = \exp\left\{-\pi^2\alpha_s^2\left[C_s\lambda^2 R^3 - \Delta f R\right]^2\right\}.} \tag{4.1866}$$

This is the familiar spatial-coherence envelope associated with finite illumination angular spread. Its precise sign inside the square follows from the defocus convention used in Eq. (4.1466). With another convention for Δf, the same result is often written with a plus sign.

The physical meaning of Eq. (4.1866) is particularly interesting. The envelope is not determined directly by the magnitude of χ, but by its slope with respect to spatial frequency. A range of illumination angles samples nearby points of the aberration function; therefore contrast is preserved when that function is locally flat and lost when it varies rapidly across the pupil.

For astigmatic or higher-order aberrations, the most general expression is still Eq. (4.1862). In polar coordinates,

$$\boxed{\|\boldsymbol{\nabla}_{\mathbf{R}}\chi\|^2 = \left(\frac{\partial\chi}{\partial R}\right)^2 + \frac{1}{R^2}\left(\frac{\partial\chi}{\partial\phi}\right)^2.} \tag{4.1867}$$

Thus an anisotropic aberration can generate an anisotropic spatial-coherence envelope. Replacing the exact gradient merely by a direction-dependent defocus term is an approximation that is adequate when the angular derivative of the aberration phase is negligible.

It is also worth emphasizing that Eq. (4.1866) assumes a *Gaussian* distribution of illumination angles. A literal uniformly illuminated circular source or a hard angular aperture produces a different characteristic function and therefore a different envelope. The Gaussian expression is widely useful because real effective source distributions are often approximately smooth and because it provides a simple one-parameter description of the illumination angular spread.

4.13.4 Averaging caused by specimen motion

Motion during an exposure gives another example of the same statistical principle, although it should not strictly be called partial coherence. A Fourier component accumulates a realization-dependent phase as the specimen moves, and the detector averages that phase factor over the exposure.

Two geometrically distinct motions are useful to consider: translation parallel to the image plane and displacement along the optical axis.

Uniform in-plane drift

Let the specimen translate during an exposure by the total vector

$$\Delta\mathbf{r}_d. \tag{4.1868}$$

We model a uniform straight-line drift centered around the mean specimen position as

$$\mathbf{u}(s) = s\,\Delta\mathbf{r}_d, \qquad -\frac{1}{2} \le s \le \frac{1}{2}. \tag{4.1869}$$

By the Fourier shift property, translation by $\mathbf{u}$ multiplies a Fourier component by

$$\exp\left[-j2\pi\mathbf{R}\cdot\mathbf{u}\right]. \tag{4.1870}$$

Envelope produced by uniform linear drift.

The trajectory parameter s in Eq. (4.1869) runs uniformly from $-1/2$ to $1/2$. Since this interval has unit length, its probability density is simply

$$p(s) = 1, \qquad -\frac{1}{2} \leq s \leq \frac{1}{2}. \tag{4.1871}$$

The motion envelope is therefore the average of the Fourier phase factor generated by all positions visited during the exposure:

$$E_{\parallel}(\mathbf{R}) = \int_{-1/2}^{1/2} \exp\left[-j2\pi s\, \mathbf{R} \cdot \Delta\mathbf{r}_d\right] ds. \tag{4.1872}$$

To simplify the notation, define the dimensionless scalar

$$\boxed{a \triangleq \mathbf{R} \cdot \Delta\mathbf{r}_d.} \tag{4.1873}$$

The quantity a is dimensionless because $\mathbf{R}$ has units of cycles per unit length and $\Delta\mathbf{r}_d$ has units of length. Equation (4.1872) then becomes

$$E_{\parallel}(a) = \int_{-1/2}^{1/2} e^{-j2\pi a s} ds. \tag{4.1874}$$

For $a \neq 0$, an antiderivative is

$$\int e^{-j2\pi a s}\, ds = \frac{e^{-j2\pi a s}}{-j2\pi a}. \tag{4.1875}$$

Evaluating it at the integration limits,

$$\begin{aligned} E_{\parallel}(a) &= \left.\frac{e^{-j2\pi a s}}{-j2\pi a}\right|_{-1/2}^{1/2} \\ &= \frac{e^{-j\pi a} - e^{+j\pi a}}{-j2\pi a}. \end{aligned} \tag{4.1876}$$

Using Euler's identity,

$$e^{-jx} - e^{jx} = -2j \sin x, \tag{4.1877}$$

with $x = \pi a$, Eq. (4.1876) reduces to

$$E_{\|}(a) = \frac{-2j\sin(\pi a)}{-j2\pi a} = \frac{\sin(\pi a)}{\pi a}. \tag{4.1878}$$

We define the normalized sinc function as

$$\boxed{\operatorname{sinc}(x) \triangleq \frac{\sin(\pi x)}{\pi x}.} \tag{4.1879}$$

Therefore,

$$\boxed{E_{\|}(\mathbf{R}) = \operatorname{sinc}\left(\mathbf{R}\cdot\Delta\mathbf{r}_d\right).} \tag{4.1880}$$

At $\mathbf{R}\cdot\Delta\mathbf{r}_d = 0$, the expression is understood by continuity. Since

$$\lim_{x\to 0}\frac{\sin(\pi x)}{\pi x} = 1, \tag{4.1881}$$

we have

$$E_{\|} = 1 \qquad \text{when } \mathbf{R}\cdot\Delta\mathbf{r}_d = 0. \tag{4.1882}$$

The argument of the sinc function contains the projection of the total drift onto the direction of the spatial frequency. Writing

$$\Delta\mathbf{r}_d = D_R\widehat{\mathbf{d}}, \tag{4.1883}$$

and denoting by γ the angle between $\mathbf{R}$ and the drift direction $\widehat{\mathbf{d}}$,

$$\mathbf{R}\cdot\Delta\mathbf{r}_d = RD_R\cos\gamma. \tag{4.1884}$$

Hence,

$$\boxed{E_{\|}(R,\gamma) = \operatorname{sinc}\left(RD_R\cos\gamma\right).} \tag{4.1885}$$

Thus a straight drift attenuates most strongly those Fourier components whose frequency vector is parallel to the motion. For frequencies perpendicular to the drift,

$$\gamma = \frac{\pi}{2}, \tag{4.1886}$$

so that

$$E_{\parallel} = 1. \tag{4.1887}$$

A pure translation along one direction therefore produces an *anisotropic* loss of information.

The exact envelope is therefore directional. If

$$\Delta \mathbf{r}_d = D_R \widehat{\mathbf{d}}, \tag{4.1888}$$

then

$$E_{\parallel}(\mathbf{R}) = \operatorname{sinc}\left(D_R \mathbf{R} \cdot \widehat{\mathbf{d}}\right). \tag{4.1889}$$

A frequently used scalar approximation replaces the projected spatial frequency by its magnitude,

$$\boxed{E_{\parallel}(R) \simeq \operatorname{sinc}(D_R R).} \tag{4.1890}$$

This radial form is convenient but should not obscure the fact that a fixed directed drift actually produces anisotropic attenuation.

Random two-dimensional Gaussian jitter

A different model assumes that the specimen does not drift systematically in one direction, but fluctuates randomly around a fixed mean position. Let the instantaneous displacement from that mean position be

$$\mathbf{u} = \begin{pmatrix} u_x \\ u_y \end{pmatrix}. \tag{4.1891}$$

We first consider an isotropic Gaussian distribution,

$$\boxed{p(\mathbf{u}) = \frac{1}{2\pi\sigma_u^2} \exp\left[-\frac{u_x^2 + u_y^2}{2\sigma_u^2}\right],} \tag{4.1892}$$

where σ_u^2 is the variance of the displacement along each Cartesian direction,

$$\left\langle u_x^2 \right\rangle = \left\langle u_y^2 \right\rangle = \sigma_u^2. \tag{4.1893}$$

This model describes stationary jitter: at any instant the specimen position is drawn from the same distribution around the same mean position.

A displacement $\mathbf{u}$ multiplies a Fourier component by the shift factor

$$e^{-j2\pi \mathbf{R}\cdot\mathbf{u}}. \tag{4.1894}$$

The motion envelope is therefore obtained by averaging this phase factor over the displacement distribution.

Envelope produced by isotropic Gaussian jitter.

Using Eq. (4.1814), with the random variable now being the two-dimensional displacement $\mathbf{u}$,

$$E_{\mathrm{G}}(\mathbf{R}) = \iint p(\mathbf{u}) e^{-j2\pi \mathbf{R}\cdot\mathbf{u}} du_x\, du_y. \tag{4.1895}$$

Writing

$$\mathbf{R} = \begin{pmatrix} R_x \\ R_y \end{pmatrix}, \tag{4.1896}$$

the scalar product is

$$\mathbf{R}\cdot\mathbf{u} = R_x u_x + R_y u_y. \tag{4.1897}$$

Consequently,

$$e^{-j2\pi \mathbf{R}\cdot\mathbf{u}} = e^{-j2\pi R_x u_x} e^{-j2\pi R_y u_y}. \tag{4.1898}$$

The probability distribution in Eq. (4.1892) also factorizes,

$$p(\mathbf{u}) = p_x(u_x) p_y(u_y), \tag{4.1899}$$

where

$$p_x(u_x) = \frac{1}{\sqrt{2\pi}\sigma_u} \exp\left[-\frac{u_x^2}{2\sigma_u^2}\right], \tag{4.1900}$$

and similarly for $p_y(u_y)$. Thus Eq. (4.1895) becomes

$$\begin{aligned} E_{\mathrm{G}}(\mathbf{R}) = & \left[\int_{-\infty}^{\infty} p_x(u_x) e^{-j2\pi R_x u_x} du_x\right] \\ & \times \left[\int_{-\infty}^{\infty} p_y(u_y) e^{-j2\pi R_y u_y} du_y\right]. \end{aligned} \tag{4.1901}$$

Each factor has exactly the Gaussian form evaluated in Eq. (4.1834). For the x direction,

$$\xi = u_x, \qquad g = 2\pi R_x, \qquad \sigma_\xi = \sigma_u, \tag{4.1902}$$

so that

$$E_{\mathrm{G},x} = \exp\left[-2\pi^2 \sigma_u^2 R_x^2\right]. \tag{4.1903}$$

Similarly,

$$E_{\mathrm{G},y} = \exp\left[-2\pi^2\sigma_u^2 R_y^2\right]. \tag{4.1904}$$

Multiplying the two contributions,

$$\begin{aligned} E_{\mathrm{G}}(\mathbf{R}) &= \exp\left[-2\pi^2\sigma_u^2\left(R_x^2 + R_y^2\right)\right] \\ &= \boxed{\exp\left[-2\pi^2\sigma_u^2 R^2\right].} \end{aligned} \tag{4.1905}$$

The Gaussian displacement model therefore produces a Gaussian decay in spatial frequency. Notice that the root-mean-square magnitude of the two-dimensional displacement is

$$u_{\mathrm{rms}} \triangleq \sqrt{\langle \|\mathbf{u}\|^2 \rangle} = \sqrt{2}\,\sigma_u. \tag{4.1906}$$

Equation (4.1905) can consequently also be written as

$$\boxed{E_{\mathrm{G}}(R) = \exp\left[-\pi^2 u_{\mathrm{rms}}^2 R^2\right].} \tag{4.1907}$$

This model differs fundamentally from the uniform straight drift of Eq. (4.1880). Straight drift generates a sinc function because the visited positions are uniformly distributed along a line segment. Gaussian jitter generates a Gaussian envelope because the visited positions follow a Gaussian distribution in the image plane.

The more general anisotropic Gaussian model is

$$\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_u\right), \tag{4.1908}$$

where $\mathbf{\Sigma}_u$ is the 2×2 displacement covariance matrix. The same characteristic-function calculation gives

$$\boxed{E_{\mathrm{G}}(\mathbf{R}) = \exp\left[-2\pi^2 \mathbf{R}^T \mathbf{\Sigma}_u \mathbf{R}\right].} \tag{4.1909}$$

Thus anisotropic mechanical vibrations naturally produce an anisotropic motion envelope.

Gaussian random walk

The preceding Gaussian-jitter model assumes that the specimen fluctuates around a fixed mean position. Another possibility is that each new position is obtained by adding a random displacement to the *current* position.

The specimen then has no restoring tendency towards a fixed position and performs a random walk.

Let the displacement after the nth step be

$$\boxed{\mathbf{u}_n = \mathbf{u}_{n-1} + \boldsymbol{\eta}_n, \qquad \mathbf{u}_0 = \mathbf{0},} \tag{4.1910}$$

where the increments $\boldsymbol{\eta}_n$ are independent two-dimensional Gaussian random vectors,

$$\boldsymbol{\eta}_n \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\text{step}}^2 \mathbf{I}\right). \tag{4.1911}$$

Thus each Cartesian component of every step has variance σ_{step}^2.

After n steps,

$$\mathbf{u}_n = \sum_{m=1}^{n} \boldsymbol{\eta}_m. \tag{4.1912}$$

Because independent Gaussian random variables add to another Gaussian random variable, $\mathbf{u}_n$ is Gaussian, but its variance increases with n.

Envelope of a discrete Gaussian random walk.

Consider first one Cartesian component,

$$u_{x,n} = \sum_{m=1}^{n} \eta_{x,m}. \tag{4.1913}$$

Since the increments are independent,

$$\text{Var}(u_{x,n}) = \sum_{m=1}^{n} \text{Var}(\eta_{x,m}) = n\sigma_{\text{step}}^2. \tag{4.1914}$$

The same result holds for $u_{y,n}$, so that

$$\boxed{\mathbf{u}_n \sim \mathcal{N}\left(\mathbf{0}, n\sigma_{\text{step}}^2 \mathbf{I}\right).} \tag{4.1915}$$

At step n, the ensemble-averaged Fourier phase factor is therefore, from Eq. (4.1905),

$$\begin{aligned} E_n(\mathbf{R}) &= \left\langle e^{-j2\pi \mathbf{R}\cdot\mathbf{u}_n} \right\rangle \\ &= \exp\left[-2\pi^2 n\sigma_{\text{step}}^2 R^2\right]. \end{aligned} \tag{4.1916}$$

Suppose that an exposure contains N equally weighted positions, $n = 0, \ldots, N-1$. The complete exposure averages the transfer factors from all these positions,

$$E_{\mathrm{RW},N}(R) = \frac{1}{N} \sum_{n=0}^{N-1} \exp\left[-2\pi^2 n \sigma_{\mathrm{step}}^2 R^2\right]. \tag{4.1917}$$

Define

$$\beta \triangleq 2\pi^2 \sigma_{\mathrm{step}}^2 R^2. \tag{4.1918}$$

Then

$$E_{\mathrm{RW},N}(R) = \frac{1}{N} \sum_{n=0}^{N-1} e^{-n\beta}. \tag{4.1919}$$

This is a finite geometric series. For a ratio

$$q = e^{-\beta}, \tag{4.1920}$$

the identity

$$\sum_{n=0}^{N-1} q^n = \frac{1-q^N}{1-q} \tag{4.1921}$$

gives

$$\boxed{E_{\mathrm{RW},N}(R) = \frac{1-e^{-N\beta}}{N\left(1-e^{-\beta}\right)}, \qquad \beta = 2\pi^2 \sigma_{\mathrm{step}}^2 R^2.} \tag{4.1922}$$

Unlike stationary Gaussian jitter, the random walk becomes progressively more dispersed as the exposure proceeds. Early parts of the exposure are therefore less strongly damped than later ones.

Continuous random walk: Brownian motion

The continuous-time limit of the Gaussian random walk is Brownian motion. Let D_m denote the two-dimensional translational diffusion coefficient, with units of length^2/time. The displacement components then satisfy

$$\boxed{\left\langle u_x^2(t) \right\rangle = \left\langle u_y^2(t) \right\rangle = 2D_m t.} \tag{4.1923}$$

Consequently,

$$\left\langle \|\mathbf{u}(t)\|^2 \right\rangle = 4D_m t. \tag{4.1924}$$

Envelope produced by Brownian motion.

At time t, Eq. (4.1923) shows that the displacement has the Gaussian covariance

$$\mathbf{\Sigma}(t) = 2D_m t\,\mathbf{I}. \tag{4.1925}$$

Using the general Gaussian displacement envelope of Eq. (4.1909),

$$\begin{aligned} E(R,t) &= \exp\left[-2\pi^2 \mathbf{R}^T \left(2D_m t\,\mathbf{I}\right)\mathbf{R}\right] \\ &= \exp\left[-4\pi^2 D_m R^2 t\right]. \end{aligned} \tag{4.1926}$$

If the exposure lasts a time T, and all times contribute equally, the detected image contains the time average

$$E_{\mathrm{B}}(R) = \frac{1}{T}\int_0^T \exp\left[-4\pi^2 D_m R^2 t\right] dt. \tag{4.1927}$$

Define

$$a \triangleq 4\pi^2 D_m R^2. \tag{4.1928}$$

Then

$$E_{\mathrm{B}}(R) = \frac{1}{T}\int_0^T e^{-at} dt. \tag{4.1929}$$

Since

$$\int e^{-at}\,dt = -\frac{1}{a}e^{-at}, \tag{4.1930}$$

we obtain

$$\begin{aligned} E_{\mathrm{B}}(R) &= \frac{1}{T}\left[-\frac{1}{a}e^{-at}\right]_0^T \\ &= \frac{1}{aT}\left(1 - e^{-aT}\right). \end{aligned} \tag{4.1931}$$

Substituting back $a = 4\pi^2 D_m R^2$ gives

$$\boxed{E_{\mathrm{B}}(R) = \frac{1 - \exp\left(-4\pi^2 D_m T R^2\right)}{4\pi^2 D_m T R^2}.} \tag{4.1932}$$

At $R = 0$, both numerator and denominator vanish, but using

$$\lim_{x\to 0} \frac{1 - e^{-x}}{x} = 1, \tag{4.1933}$$

we recover

$$E_{\mathrm{B}}(0) = 1. \tag{4.1934}$$

The Brownian envelope is neither a Gaussian nor a sinc. Its form follows from averaging Gaussian displacement distributions whose variance grows linearly with time.

For small

$$x = 4\pi^2 D_m T R^2, \tag{4.1935}$$

using

$$1 - e^{-x} = x - \frac{x^2}{2} + \mathcal{O}(x^3), \tag{4.1936}$$

Eq. (4.1932) becomes

$$E_{\mathrm{B}}(R) \simeq 1 - 2\pi^2 D_m T R^2 + \mathcal{O}(R^4). \tag{4.1937}$$

At large spatial frequencies,

$$4\pi^2 D_m T R^2 \gg 1, \tag{4.1938}$$

the exponential term is negligible and

$$\boxed{E_{\mathrm{B}}(R) \simeq \frac{1}{4\pi^2 D_m T R^2}.} \tag{4.1939}$$

Thus the ensemble-averaged Brownian-motion transfer decreases asymptotically as R^{-2} rather than exponentially.

The three in-plane motion models considered above illustrate a general principle. The Fourier-space attenuation is the characteristic function of the distribution of positions sampled during the exposure:

$$\boxed{\begin{array}{lcl} \text{uniform positions along a line} & \longrightarrow & \text{sinc}, \\ \text{stationary Gaussian positions} & \longrightarrow & \text{Gaussian envelope}, \\ \text{Gaussian random walk} & \longrightarrow & \text{time average of progressively broader Gaussians.} \end{array}} \tag{4.1940}$$

The functional form of a motion envelope is therefore not universal. It encodes the statistical model assumed for the specimen trajectory.

Axial vibration and the Bessel envelope

Displacement of the specimen along the optical axis changes its effective defocus. Let the axial displacement during the exposure be modeled as a sinusoidal vibration,

$$\boxed{\delta f(t) = D_F \cos(\omega t),} \tag{4.1941}$$

where D_F is the vibration amplitude. According to Eq. (4.1747), this produces the phase variation

$$\delta\chi(t) = -\pi\lambda D_F R^2 \cos(\omega t). \tag{4.1942}$$

Bessel envelope produced by sinusoidal axial motion.

If the exposure contains many vibration periods, the phase factor is averaged uniformly over

$$\tau = \omega t \in [0, 2\pi]. \tag{4.1943}$$

Therefore,

$$E_\perp(R) = \frac{1}{2\pi}\int_0^{2\pi} \exp\left[j\pi\lambda D_F R^2 \cos\tau\right] d\tau. \tag{4.1944}$$

The integral representation of the Bessel function of the first kind, order zero, is

$$J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{jx\cos\tau} d\tau. \tag{4.1945}$$

Consequently,

$$\boxed{E_\perp(R) = J_0\left(\pi\lambda D_F R^2\right).} \tag{4.1946}$$

The Bessel form in Eq. (4.1946) is therefore not a universal envelope for arbitrary axial motion. It follows specifically from a sinusoidal displacement sampled uniformly in phase during the exposure.

For example, if the specimen instead moved uniformly and monotonically along z through a total distance D_F, the corresponding average would be a sinc function rather than a Bessel function:

$$E_\perp^{\text{linear}}(R) = \operatorname{sinc}\left(\frac{\lambda D_F R^2}{2}\right). \tag{4.1947}$$

Thus, the functional form of a motion envelope contains information about the assumed motion trajectory.

Combined motion envelopes

There is no unique motion envelope. As shown in the preceding subsections, the Fourier-space attenuation depends on the statistical model adopted for the specimen trajectory. Uniform linear drift, stationary Gaussian jitter, a random walk, and sinusoidal vibration sample different distributions of specimen positions and therefore produce different transfer factors.

It is nevertheless useful to separate motion into components parallel and perpendicular to the image plane. Let

$$E_{\parallel}(\mathbf{R}) \tag{4.1948}$$

denote the transfer factor produced by in-plane motion, and let

$$E_{\perp}(\mathbf{R}) \tag{4.1949}$$

denote the factor produced by motion along the optical axis.

If the two motions are statistically independent, their phase factors can be averaged separately. Thus,

$$\boxed{E_{\text{motion}}(\mathbf{R}) = E_{\perp}(\mathbf{R})\, E_{\parallel}(\mathbf{R}).} \tag{4.1950}$$

The particular functional form of E_{motion} therefore depends on the models chosen for the two components of motion.

For example, if the axial motion is the sinusoidal vibration of Eq. (4.1941), then Eq. (4.1946) gives

$$E_{\perp}(R) = J_0\left(\pi\lambda D_F R^2\right). \tag{4.1951}$$

If this motion is combined with the uniform in-plane drift of Eq. (4.1869), whose envelope is given by Eq. (4.1880), the result is

$$\boxed{E_{\text{motion}}(\mathbf{R}) = J_0\left(\pi\lambda D_F R^2\right) \operatorname{sinc}\left(\mathbf{R}\cdot\Delta\mathbf{r}_d\right).} \tag{4.1952}$$

This is the particular Bessel–sinc model obtained from sinusoidal axial motion and uniform straight-line in-plane drift.

Under the radial approximation of Eq. (4.1890), this becomes

$$\boxed{E_{\text{motion}}(R) \simeq J_0\left(\pi\lambda D_F R^2\right) \operatorname{sinc}(D_R R).} \tag{4.1953}$$

The radial form is convenient, but the exact straight-drift factor is anisotropic, as shown in Eq. (4.1889).

Other in-plane motion models simply replace the sinc factor. For stationary isotropic Gaussian jitter, Eq. (4.1905) gives

$$\boxed{E_{\text{motion}}(R) = E_{\perp}(R) \exp\left[-2\pi^2 \sigma_u^2 R^2\right].} \tag{4.1954}$$

For anisotropic Gaussian jitter, Eq. (4.1909) gives

$$\boxed{E_{\text{motion}}(\mathbf{R}) = E_{\perp}(\mathbf{R}) \exp\left[-2\pi^2 \mathbf{R}^T \mathbf{\Sigma}_u \mathbf{R}\right].} \tag{4.1955}$$

For a discrete Gaussian random walk, the corresponding factor is instead given by Eq. (4.1922),

$$\boxed{E_{\text{motion}}(R) = E_{\perp}(R) \frac{1 - e^{-N\beta}}{N\left(1 - e^{-\beta}\right)}, \qquad \beta = 2\pi^2 \sigma_{\text{step}}^2 R^2.} \tag{4.1956}$$

In the continuous Brownian-motion limit, Eq. (4.1932) gives

$$\boxed{E_{\text{motion}}(R) = E_{\perp}(R) \frac{1 - \exp\left(-4\pi^2 D_m T R^2\right)}{4\pi^2 D_m T R^2}.} \tag{4.1957}$$

The axial component is equally model-dependent. Sinusoidal axial vibration gives the Bessel factor of Eq. (4.1946), whereas uniform axial motion gives the sinc-type factor of Eq. (4.1947). Therefore, neither the Bessel factor nor the sinc factor is universal: each is a consequence of a particular trajectory model.

The general lesson is

$$\boxed{\begin{aligned} \text{trajectory model} \quad &\longrightarrow \quad \text{distribution of sampled positions} \\ &\longrightarrow \quad \text{average Fourier phase factor} \\ &\longrightarrow \quad E_{\text{motion}}. \end{aligned}} \tag{4.1958}$$

Thus, when a motion envelope is introduced in an image-formation model, the assumed motion statistics should also be stated.

Finally, the term “envelope” should be used with some care. Gaussian jitter and Brownian-motion factors are positive and monotonically decreasing, but the sinc and Bessel factors can cross zero and change sign. In those cases, the complete transfer factor contains both attenuation and phase reversals. Taking its absolute value gives a positive envelope in the informal sense, but discards the sign changes that belong to the actual forward model.

4.13.5 Why independent envelopes multiply

The temporal-coherence, spatial-coherence, and motion factors are often combined by multiplication. This follows directly from statistical independence rather than being an additional optical assumption.

Product of statistically independent envelopes.

Suppose the total fluctuating phase is the sum

$$\delta\chi = \delta\chi_t + \delta\chi_s + \delta\chi_m, \tag{4.1959}$$

where the subscripts denote temporal, spatial, and motion fluctuations. Then

$$e^{-j\delta\chi} = e^{-j\delta\chi_t} e^{-j\delta\chi_s} e^{-j\delta\chi_m}. \tag{4.1960}$$

If the three random processes are statistically independent,

$$\left\langle e^{-j\delta\chi} \right\rangle = \left\langle e^{-j\delta\chi_t} \right\rangle \left\langle e^{-j\delta\chi_s} \right\rangle \left\langle e^{-j\delta\chi_m} \right\rangle \tag{4.1961}$$

$$= \boxed{E_t\, E_s\, E_{\text{motion}}.} \tag{4.1962}$$

If the fluctuations are correlated, the factorization in Eq. (4.1962) is not exact and the joint statistical distribution must be averaged instead. Thus multiplication of envelopes is justified when the underlying physical fluctuations are independent or can be treated as approximately independent.

4.13.6 A single effective Gaussian envelope

The previous subsections have shown that different physical mechanisms produce different attenuation factors. Temporal partial coherence gives the envelope of Eq. (4.1836); finite illumination angular spread gives Eq. (4.1862); and different motion models lead to sinc, Gaussian, Bessel, or random-walk envelopes.

In practice, however, cryo-EM image processing often replaces all these mechanisms by a single empirical Gaussian attenuation,

$$\boxed{E_{\text{eff}}(R) = \exp\left(-\frac{B}{4}R^2\right),} \tag{4.1963}$$

where B has units of length squared and is conventionally referred to as a *B-factor*.

This effective Gaussian should not, in general, be interpreted as a new physical source of contrast loss. Rather, it is a phenomenological model that summarizes the combined attenuation of several mechanisms over a restricted range of spatial frequencies.

When the Gaussian form is exact

There are situations in which Eq. (4.1963) follows directly from a physical model. For example, the isotropic Gaussian displacement model derived in Eq. (4.1905) gives

$$E_{\mathrm{G}}(R) = \exp\left(-2\pi^2\sigma_u^2 R^2\right). \tag{4.1964}$$

Comparison with Eq. (4.1963) gives

$$\boxed{B = 8\pi^2\sigma_u^2.} \tag{4.1965}$$

Thus, when the dominant attenuation is genuinely caused by isotropic Gaussian positional uncertainty, the effective B-factor can be given a direct physical interpretation in terms of the displacement variance.

Why a Gaussian is often a useful approximation

A broader justification can be obtained from the characteristic-function description introduced in Eq. (4.1814). Consider, for instance, an arbitrary zero-mean in-plane displacement $\mathbf{u}$ with finite second moments. Its transfer factor is

$$\boxed{E(\mathbf{R}) = \left\langle e^{-j2\pi\mathbf{R}\cdot\mathbf{u}} \right\rangle.} \tag{4.1966}$$

Quadratic approximation of a general displacement envelope.

Starting from Eq. (4.1966),

$$E(\mathbf{R}) = \left\langle e^{-j2\pi\mathbf{R}\cdot\mathbf{u}} \right\rangle. \tag{4.1967}$$

Using the expansion

$$\begin{aligned} e^{-j2\pi\mathbf{R}\cdot\mathbf{u}} \simeq 1 &- j2\pi\mathbf{R}\cdot\mathbf{u} \\ &- 2\pi^2(\mathbf{R}\cdot\mathbf{u})^2, \end{aligned} \tag{4.1968}$$

and taking the statistical average of both sides,

$$\begin{aligned} E(\mathbf{R}) &= \left\langle e^{-j2\pi\mathbf{R}\cdot\mathbf{u}} \right\rangle \\ &\simeq \left\langle 1 - j2\pi\mathbf{R}\cdot\mathbf{u} - 2\pi^2(\mathbf{R}\cdot\mathbf{u})^2 \right\rangle . \end{aligned} \tag{4.1969}$$

Using the linearity of the statistical average,

$$\boxed{\begin{aligned} E(\mathbf{R}) \simeq 1 &- j2\pi\mathbf{R}\cdot\langle\mathbf{u}\rangle \\ &- 2\pi^2 \left\langle (\mathbf{R}\cdot\mathbf{u})^2 \right\rangle . \end{aligned}} \tag{4.1970}$$

If the displacement is centered,

$$\langle\mathbf{u}\rangle = \mathbf{0}, \tag{4.1971}$$

the linear term vanishes. Introducing the covariance matrix

$$\mathbf{\Sigma}_u = \left\langle \mathbf{u}\mathbf{u}^T \right\rangle , \tag{4.1972}$$

we obtain

$$\left\langle (\mathbf{R}\cdot\mathbf{u})^2 \right\rangle = \mathbf{R}^T\mathbf{\Sigma}_u\mathbf{R}. \tag{4.1973}$$

Therefore,

$$\boxed{E(\mathbf{R}) \simeq 1 - 2\pi^2\mathbf{R}^T\mathbf{\Sigma}_u\mathbf{R}} \tag{4.1974}$$

to second order in $\mathbf{R}$.

Now use

$$e^{-x} = 1 - x + \mathcal{O}(x^2). \tag{4.1975}$$

Since

$$x = 2\pi^2\mathbf{R}^T\mathbf{\Sigma}_u\mathbf{R} \tag{4.1976}$$

is already of second order in spatial frequency, Eq. (4.1974) can be represented, to the same order, as

$$\boxed{E(\mathbf{R}) \simeq \exp\left[-2\pi^2\mathbf{R}^T\mathbf{\Sigma}_u\mathbf{R}\right] .} \tag{4.1977}$$

Thus, even when the underlying displacement distribution is not Gaussian, its low-frequency attenuation begins with a quadratic dependence on $\mathbf{R}$ provided that the displacement has finite variance. A Gaussian envelope is therefore a natural second-order approximation for many positional-

averaging processes.

For an isotropic displacement distribution,

$$\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}, \tag{4.1978}$$

and Eq. (4.1977) reduces to

$$E(R) \simeq \exp\left(-2\pi^2 \sigma_u^2 R^2\right). \tag{4.1979}$$

Why the approximation is not universal

The preceding argument should not be interpreted as saying that every physical envelope is intrinsically Gaussian in R. The exact frequency dependence depends on how the fluctuating variable enters the phase.

For example, temporal partial coherence gives Eq. (4.1836),

$$E_t(R) = \exp\left[-\frac{1}{2}\left(\pi\lambda\sigma_f\right)^2 R^4\right], \tag{4.1980}$$

whose exponent scales as R^4 rather than R^2.

Similarly, the spatial-coherence envelope of Eq. (4.1862) depends on

$$\left\|\boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R})\right\|^2, \tag{4.1981}$$

which is generally not proportional to R^2. Uniform linear drift gives the sinc factor of Eq. (4.1880), whereas sinusoidal axial motion gives the Bessel factor of Eq. (4.1946).

Therefore,

$$E_{\text{eff}}(R) = \exp\left(-\frac{B}{4}R^2\right) \tag{4.1982}$$

should usually be understood as an approximation to the combined attenuation over the frequency interval relevant to the experiment, not as an exact representation of all underlying mechanisms.

Effective B-factor for several attenuation mechanisms

If several independent mechanisms contribute to the attenuation, their exact combined envelope is, according to Eq. (4.1962),

$$E_{\text{exact}}(\mathbf{R}) = \prod_k E_k(\mathbf{R}). \tag{4.1983}$$

Taking logarithms,

$$\boxed{\log E_{\text{exact}}(\mathbf{R}) = \sum_k \log E_k(\mathbf{R}).} \tag{4.1984}$$

The effective Gaussian approximation amounts to replacing this complete frequency dependence by

$$\boxed{\log E_{\text{exact}}(R) \approx -\frac{B_{\text{eff}}}{4} R^2.} \tag{4.1985}$$

Thus B_{eff} can be viewed as the coefficient of a quadratic approximation to the logarithm of the total attenuation. Its numerical value generally depends on the spatial-frequency interval over which the approximation is fitted.

This also explains why the fitted B-factor should not normally be decomposed into unique physical contributions. Different combinations of energy spread, illumination incoherence, specimen motion, detector response, radiation damage, and other high-frequency losses may produce similar effective decays over a finite frequency interval.

Amplitude versus power-spectrum envelopes

The envelope in Eq. (4.1963) multiplies the Fourier amplitude of the image-forming transfer function. Consequently, if

$$E_{\text{eff}}(R) = \exp\left(-\frac{B}{4} R^2\right), \tag{4.1986}$$

then its contribution to the power spectrum is

$$\boxed{E_{\text{eff}}^2(R) = \exp\left(-\frac{B}{2} R^2\right).} \tag{4.1987}$$

It is therefore important to specify whether a reported Gaussian B-factor refers to an amplitude envelope or to a power-spectrum envelope, since the corresponding exponential coefficients differ by a factor of two.

The effective Gaussian model is useful precisely because it compresses a large number of poorly known attenuation mechanisms into a single parameter. Its simplicity is its principal advantage, but also its main limitation: the parameter B describes the observed decay without uniquely identifying its physical origin.

4.14 Phase plates

The previous section introduced an effective CTF phase shift ϕ_A associated with amplitude contrast. As emphasized there, ϕ_A is not a physical phase accumulated by the electron wave: it appears only after the amplitude and phase components of the weak object have been combined in the detected intensity.

A *phase plate* is fundamentally different. It is a physical electron-optical element that deliberately changes the relative phase of different components of the electron wave. Its purpose is to convert the phase modulation produced by a weak specimen into measurable image intensity without relying entirely on defocus (Malac et al., 2021; Axelrod et al., 2024).

The basic idea is closely related to the back focal plane discussed in Sec. 4.10.7. There, Eq. (4.1129) established that

$$\boxed{x_F = \lambda f R_x, \qquad y_F = \lambda f R_y.} \tag{4.1988}$$

Thus different spatial frequencies occupy different positions in the back focal plane. In particular, the unscattered or reference beam is concentrated around $\mathbf{R} = \mathbf{0}$, whereas scattered components appear at non-zero spatial frequencies. A device placed in this plane, or in a conjugate diffraction plane, can therefore change their relative phase.

This possibility is especially important for cryo-EM because biological specimens are predominantly weak phase objects. Conventional defocus phase contrast deliberately uses lens aberrations to change the relative phase of the reference and scattered waves. A phase plate attempts to do this more directly.

4.14.1 Why low spatial frequencies are poorly transferred in focus

For a pure weak-phase object, the coherent CTF derived in Sec. 4.10.10 is, for symmetric aberrations,

$$\mathrm{CTF}(\mathbf{R}) = P(\mathbf{R}) \sin \chi(\mathbf{R}). \tag{4.1989}$$

At low spatial frequency, the usual aberration phase approaches zero. For example, retaining defocus and spherical aberration,

$$\chi(R) = -\pi \lambda \Delta f R^2 + \frac{\pi}{2} C_s \lambda^3 R^4 + \cdots . \tag{4.1990}$$

Hence,

$$\chi(R) \longrightarrow 0 \qquad \text{as} \qquad R \longrightarrow 0. \tag{4.1991}$$

Since

$$\sin\chi \simeq \chi \tag{4.1992}$$

for small χ, the low-frequency transfer behaves approximately as

$$\boxed{\mathrm{CTF}(R) \simeq -\pi\lambda\Delta f R^2, \qquad R \to 0,} \tag{4.1993}$$

when defocus is the dominant term.

Thus the very spatial frequencies that describe the large-scale shape of a particle are weakly transferred in a conventional nearly focused image. Increasing the defocus generates stronger low-frequency contrast, but it also introduces the familiar oscillations and zero crossings of the CTF.

The purpose of an ideal phase plate is to change this situation by introducing approximately a quarter-wave relative phase shift,

$$|\Delta\phi_{\mathrm{pp}}| \simeq \frac{\pi}{2}, \tag{4.1994}$$

between the reference and scattered components. The corresponding phase-contrast transfer then changes, ideally, from a sine-like to a cosine-like form.

4.14.2 A phase plate as a pupil-plane transfer function

A general phase plate can be represented by a complex transmission function in the diffraction plane,

$$\boxed{Q_{\mathrm{pp}}(\mathbf{R}) = a_{\mathrm{pp}}(\mathbf{R}) \exp\left[-j\phi_{\mathrm{pp}}(\mathbf{R})\right],} \tag{4.1995}$$

where $a_{\mathrm{pp}}(\mathbf{R})$ describes any attenuation or obstruction and $\phi_{\mathrm{pp}}(\mathbf{R})$ is the phase introduced by the device.

For an ideal pure phase plate,

$$a_{\mathrm{pp}}(\mathbf{R}) = 1, \tag{4.1996}$$

and the coherent wave-transfer function becomes

$$\boxed{H_{\mathrm{pp}}(\mathbf{R}) = P(\mathbf{R}) \exp\left\{-j\left[\chi(\mathbf{R}) + \phi_{\mathrm{pp}}(\mathbf{R})\right]\right\}.} \tag{4.1997}$$

At first sight, one might therefore be tempted simply to replace χ by $\chi+\phi_{\mathrm{pp}}$. There is, however, an important subtlety. Image formation depends on the phase of the scattered wave *relative to the reference wave*. A common phase added to both is physically irrelevant.

The reference component at $\mathbf{R} = \mathbf{0}$ must therefore be retained explicitly.

4.14.3 Only the phase relative to the direct beam matters

The weak-phase-object derivation of Sec. 4.10.10 assumed implicitly that the coherent transfer of the direct beam was unity. A phase plate may modify the direct beam itself, so it is useful to repeat the derivation without making that assumption.

Weak-phase CTF with an arbitrary reference-beam phase.

For a weak phase object, Eq. (4.1205) gives

$$\psi_i(\mathbf{r}) \simeq 1 + j\varphi(\mathbf{r}), \tag{4.1998}$$

and hence

$$\widehat{\psi}_i(\mathbf{R}) = \delta(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R}). \tag{4.1999}$$

Let $H(\mathbf{R})$ now denote the complete coherent transfer function, including the phase plate. The output wave is

$$\widehat{\psi}_o(\mathbf{R}) = H(\mathbf{0})\delta(\mathbf{R}) + jH(\mathbf{R})\widehat{\varphi}(\mathbf{R}). \tag{4.2000}$$

Define

$$H_0 \triangleq H(\mathbf{0}). \tag{4.2001}$$

In real space, the output consists of the constant reference wave H_0 plus a weak scattered contribution $g(\mathbf{r})$,

$$\psi_o(\mathbf{r}) = H_0 + g(\mathbf{r}). \tag{4.2002}$$

To first order,

$$\begin{aligned} I_o &= |H_0 + g|^2 \\ &\simeq |H_0|^2 + H_0^* g + H_0 g^*. \end{aligned} \tag{4.2003}$$

Therefore the Fourier transform of the intensity modulation is

$$\boxed{\widehat{\delta I}(\mathbf{R}) = j\left[H_0^* H(\mathbf{R}) - H_0 H^*(-\mathbf{R})\right]\widehat{\varphi}(\mathbf{R}).} \tag{4.2004}$$

Equation (4.2004) generalizes Eq. (4.1219). The latter is recovered when

$$H_0 = 1. \tag{4.2005}$$

Now consider an ideal pure phase plate and a real centrosymmetric pupil. Using Eq. (4.1997),

$$H_0 = e^{-j\phi_{\mathrm{pp}}(\mathbf{0})}, \tag{4.2006}$$

where we have used

$$P(\mathbf{0}) = 1, \qquad \chi(\mathbf{0}) = 0. \tag{4.2007}$$

Define the phase introduced relative to the reference beam as

$$\boxed{\Delta\phi_{\mathrm{pp}}(\mathbf{R}) \triangleq \phi_{\mathrm{pp}}(\mathbf{R}) - \phi_{\mathrm{pp}}(\mathbf{0}).} \tag{4.2008}$$

Suppose also that the phase-plate profile is centrosymmetric,

$$\phi_{\mathrm{pp}}(-\mathbf{R}) = \phi_{\mathrm{pp}}(\mathbf{R}). \tag{4.2009}$$

Using the decomposition

$$\chi = \chi_{\mathrm{sym}} + \chi_{\mathrm{asym}}, \tag{4.2010}$$

we obtain

$$H_0^* H(\mathbf{R}) = P(\mathbf{R}) e^{-j\chi_{\mathrm{asym}}(\mathbf{R})} e^{-j[\chi_{\mathrm{sym}}(\mathbf{R}) + \Delta\phi_{\mathrm{pp}}(\mathbf{R})]}, \tag{4.2011}$$

$$H_0 H^*(-\mathbf{R}) = P(\mathbf{R}) e^{-j\chi_{\mathrm{asym}}(\mathbf{R})} e^{+j[\chi_{\mathrm{sym}}(\mathbf{R}) + \Delta\phi_{\mathrm{pp}}(\mathbf{R})]}. \tag{4.2012}$$

Substitution into Eq. (4.2004) gives, after absorbing the conventional factor of two into the contrast normalization,

$$\boxed{\mathrm{CTF}_{\mathrm{pp}}(\mathbf{R}) = P(\mathbf{R}) e^{-j\chi_{\mathrm{asym}}(\mathbf{R})} \sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \Delta\phi_{\mathrm{pp}}(\mathbf{R})\right].} \tag{4.2013}$$

Equation (4.2013) is the central result of this section. The physically relevant quantity is not the absolute phase $\phi_{\mathrm{pp}}(\mathbf{R})$ introduced by the device but the *difference*

$$\Delta\phi_{\mathrm{pp}}(\mathbf{R}) = \phi_{\mathrm{pp}}(\mathbf{R}) - \phi_{\mathrm{pp}}(\mathbf{0}) \tag{4.2014}$$

between each scattered component and the direct reference beam.

Consequently, it makes no difference whether a particular technology shifts mainly the direct beam or mainly the scattered beams. Only their relative phase enters the image-formation model.

4.14.4 The ideal quarter-wave phase plate

For an ideal phase plate, the relative phase of every non-zero scattered spatial frequency would satisfy

$$\Delta\phi_{\mathrm{pp}} = \pm\frac{\pi}{2}. \tag{4.2015}$$

The sign depends on whether the reference beam or the scattered beam is advanced or retarded and determines the polarity of the resulting image.

Choosing, for definiteness,

$$\Delta\phi_{\mathrm{pp}} = +\frac{\pi}{2}, \tag{4.2016}$$

Eq. (4.2013) becomes

$$\boxed{\mathrm{CTF}_{\mathrm{pp}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\cos\left[\chi_{\mathrm{sym}}(\mathbf{R})\right].} \tag{4.2017}$$

With the opposite sign,

$$\mathrm{CTF}_{\mathrm{pp}} = -Pe^{-j\chi_{\mathrm{asym}}}\cos\chi_{\mathrm{sym}}, \tag{4.2018}$$

which contains the same information with reversed contrast.

The reason why a quarter-wave shift is optimal follows immediately.

Why the ideal phase shift is $\pi/2$.

At very low spatial frequency, where

$$\chi_{\mathrm{sym}}(\mathbf{R}) \simeq 0, \tag{4.2019}$$

Eq. (4.2013) gives

$$\mathrm{CTF}_{\mathrm{pp}} \simeq P(\mathbf{R})\sin\Delta\phi_{\mathrm{pp}}. \tag{4.2020}$$

The magnitude of the transferred phase contrast is therefore bounded by

$$|\sin\Delta\phi_{\mathrm{pp}}| \leq 1. \tag{4.2021}$$

The maximum is attained when

$$\boxed{\Delta\phi_{\mathrm{pp}} = \frac{\pi}{2} + n\pi, \qquad n \in \mathbb{Z}.} \tag{4.2022}$$

> Thus a relative quarter-wave shift gives the largest possible first-order phase contrast for a weak phase object.

This result explains the appeal of phase-plate imaging. Conventional defocus contrast behaves, from Eq. (4.1993), as

$$\mathrm{CTF}(R) = O(R^2) \tag{4.2023}$$

near the origin, whereas an ideal quarter-wave phase plate gives

$$\boxed{|\mathrm{CTF}_{\mathrm{pp}}(R)| \simeq P(R)} \tag{4.2024}$$

for non-zero spatial frequencies above the phase-plate cut-on.

This last qualification is important. Exactly $\mathbf{R} = \mathbf{0}$ corresponds to the mean transmitted beam itself; no component can possess a phase relative to itself. The ideal statement therefore refers to arbitrarily low *non-zero* spatial frequencies in the limit of an infinitesimally narrow reference beam.

4.14.5 The cut-on spatial frequency

No real phase plate can distinguish an infinitely narrow direct beam from scattered electrons at arbitrarily small angles. There is therefore usually a finite transition region around $\mathbf{R} = \mathbf{0}$ over which the desired relative phase shift has not yet been established. This defines an effective *cut-on spatial frequency*.

Suppose that a characteristic transverse radius r_c in the back focal plane separates the region associated with the direct beam from the region receiving the desired phase shift. From Eq. (4.1988),

$$r_F = \lambda f R, \tag{4.2025}$$

and therefore

$$\boxed{R_c \simeq \frac{r_c}{\lambda f}.} \tag{4.2026}$$

The corresponding real-space periodicity is

$$\boxed{d_c = \frac{1}{R_c} \simeq \frac{\lambda f}{r_c}.} \tag{4.2027}$$

For an idealized Zernike-type step profile, for example,

$$\Delta\phi_{\mathrm{pp}}(R) = \begin{cases} 0, & R < R_c, \\ \phi_0, & R \geq R_c. \end{cases} \tag{4.2028}$$

The phase contrast below R_c is then essentially the same as in conventional bright-field imaging, whereas frequencies above R_c receive the phase-plate shift.

The cut-on frequency is especially important in cryo-EM because the lowest spatial frequencies carry much of the information used to recognize and align particles. A smaller value of R_c therefore allows phase contrast to extend to larger real-space features (Malac et al., 2021).

There is also a Fourier consequence of the shape of the transition. A sharp step in the pupil-plane phase introduces high-frequency components into the pupil function and can generate ringing around strong features in the image. Smooth phase profiles are therefore often preferable to an abrupt phase transition (Malac et al., 2021).

Equation (4.2026) also explains why additional transfer optics are sometimes used with phase plates. By enlarging the diffraction pattern at the physical location of the phase plate, a given device dimension can correspond to a smaller spatial-frequency cut-on.

4.14.6 Combination with amplitude contrast

The phase plate and the amplitude-contrast term of Sec. 4.12 enter the CTF in different ways but can be combined compactly.

For a centrosymmetric pure phase plate, the phase and amplitude object components of Eq. (4.1787) acquire the common relative phase

$$\chi_{\mathrm{sym}} + \Delta\phi_{\mathrm{pp}}. \tag{4.2029}$$

Thus, using the normalized amplitude-contrast coefficient A of Eq. (4.1795),

$$\boxed{\begin{aligned}\mathrm{CTF}_{A,\mathrm{pp}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})} \\ \times\left\{\sqrt{1-A^2}\sin\left[\chi_{\mathrm{sym}} + \Delta\phi_{\mathrm{pp}}\right]\right. \\ \left. +A\cos\left[\chi_{\mathrm{sym}} + \Delta\phi_{\mathrm{pp}}\right]\right\}.\end{aligned}} \tag{4.2030}$$

Using

$$\phi_A = \arcsin A \tag{4.2031}$$

from Eq. (4.1798), Eq. (4.2030) becomes

$$\boxed{\mathrm{CTF}_{A,\mathrm{pp}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \Delta\phi_{\mathrm{pp}}(\mathbf{R}) + \phi_A\right].} \tag{4.2032}$$

Although the two shifts appear additively in Eq. (4.2032), their physical origins remain different:

$\Delta\phi_{\mathrm{pp}}$	:	physical relative phase imposed on the electron wave,
ϕ_A	:	algebraic phase shift representing amplitude–phase mixing in the weak-object intensity model.

(4.2033)

Neither quantity should be confused with a lens aberration.

4.14.7 A phase plate does not remove the other aberrations

The phrase "in-focus phase-plate imaging" can give the misleading impression that the rest of the electron optics disappears. It does not.

Even if

$$\Delta f = 0, \tag{4.2034}$$

the symmetric aberration phase can still contain spherical aberration and higher-order terms,

$$\chi_{\mathrm{sym}}(R) = \frac{\pi}{2} C_s \lambda^3 R^4 + \chi_{\mathrm{sym}}^{\mathrm{higher}}(R) + \cdots . \tag{4.2035}$$

For an ideal quarter-wave phase plate,

$$\mathrm{CTF}_{\mathrm{pp}} \propto \cos \chi_{\mathrm{sym}}, \tag{4.2036}$$

so sufficiently large aberration phase still produces oscillations and zero crossings.

More generally, phase-plate imaging and defocus are not mutually exclusive. A finite defocus simply enters the same total phase:

$$\boxed{\mathrm{CTF}_{\mathrm{pp}} \propto \sin\left[-\pi\lambda\Delta f R^2 + \frac{\pi}{2} C_s \lambda^3 R^4 + \Delta\phi_{\mathrm{pp}}(R) + \cdots\right].} \tag{4.2037}$$

The phase plate therefore shifts and reshapes the CTF; it does not remove the need to model the microscope aberrations.

4.14.8 Real phase plates are not necessarily pure phase elements

A material phase plate may scatter electrons elastically or inelastically, while support structures may attenuate or completely remove selected regions of

the diffraction pattern. The coherent part of such effects can be represented by the general complex phase-plate transmission introduced in Eq. (4.1995),

$$Q_{\rm pp}(\mathbf{R}) = a_{\rm pp}(\mathbf{R})e^{-j\phi_{\rm pp}(\mathbf{R})}. \tag{4.2038}$$

In this case the effect of the plate is not, in general, equivalent to simply adding a phase to χ. Nevertheless, a closed-form expression for the weak-phase CTF can still be obtained.

CTF of a general complex phase plate.

Let the coherent transfer function, including the phase plate, be

$$H(\mathbf{R}) = P(\mathbf{R})a_{\rm pp}(\mathbf{R})\exp\left\{-j\left[\chi(\mathbf{R}) + \phi_{\rm pp}(\mathbf{R})\right]\right\}. \tag{4.2039}$$

For the direct reference beam,

$$H_0 \equiv H(\mathbf{0}) = a_0 e^{-j\phi_0}, \tag{4.2040}$$

where

$$a_0 \equiv a_{\rm pp}(\mathbf{0}), \qquad \phi_0 \equiv \phi_{\rm pp}(\mathbf{0}), \tag{4.2041}$$

and we have used

$$P(\mathbf{0}) = 1, \qquad \chi(\mathbf{0}) = 0. \tag{4.2042}$$

The mean intensity of the unscattered beam is therefore

$$I_0 = |H_0|^2 = a_0^2. \tag{4.2043}$$

It is useful to define the normalized image contrast as

$$c(\mathbf{r}) = \frac{I(\mathbf{r}) - I_0}{I_0}. \tag{4.2044}$$

Dividing the general weak-phase result Eq. (4.2004) by I_0 gives

$$\begin{aligned}\widehat{c}(\mathbf{R}) = \frac{j}{a_0}\Big\{&P(\mathbf{R})a_{\rm pp}(\mathbf{R})e^{-j[\chi(\mathbf{R})+\phi_{\rm pp}(\mathbf{R})-\phi_0]}\\ &-P(-\mathbf{R})a_{\rm pp}(-\mathbf{R})e^{+j[\chi(-\mathbf{R})+\phi_{\rm pp}(-\mathbf{R})-\phi_0]}\Big\}\,\widehat{\varphi}(\mathbf{R}).\end{aligned} \tag{4.2045}$$

As in the rest of the chapter, we absorb the common factor of two into the contrast normalization when defining the CTF. We now assume that the objective pupil itself is real and centrosymmetric,

$$P(-\mathbf{R}) = P(\mathbf{R}). \tag{4.2046}$$

To simplify Eq. (4.2045), decompose the phase-plate attenuation into its symmetric and antisymmetric parts,

$$a_{\mathrm{s}}(\mathbf{R}) \equiv \frac{a_{\mathrm{pp}}(\mathbf{R}) + a_{\mathrm{pp}}(-\mathbf{R})}{2}, \tag{4.2047}$$

$$a_{\mathrm{a}}(\mathbf{R}) \equiv \frac{a_{\mathrm{pp}}(\mathbf{R}) - a_{\mathrm{pp}}(-\mathbf{R})}{2}, \tag{4.2048}$$

and similarly decompose its phase into

$$\Delta\phi_{\mathrm{pp,s}}(\mathbf{R}) \equiv \frac{\phi_{\mathrm{pp}}(\mathbf{R}) + \phi_{\mathrm{pp}}(-\mathbf{R})}{2} - \phi_0, \tag{4.2049}$$

$$\phi_{\mathrm{pp,a}}(\mathbf{R}) \equiv \frac{\phi_{\mathrm{pp}}(\mathbf{R}) - \phi_{\mathrm{pp}}(-\mathbf{R})}{2}. \tag{4.2050}$$

The subtraction of ϕ_0 appears only in the symmetric part because the physically relevant phase is measured relative to the direct beam.

Using the aberration decomposition

$$\chi = \chi_{\mathrm{sym}} + \chi_{\mathrm{asym}}, \tag{4.2051}$$

define

$$\Theta(\mathbf{R}) \equiv \chi_{\mathrm{sym}}(\mathbf{R}) + \Delta\phi_{\mathrm{pp,s}}(\mathbf{R}), \tag{4.2052}$$

$$\Psi(\mathbf{R}) \equiv \chi_{\mathrm{asym}}(\mathbf{R}) + \phi_{\mathrm{pp,a}}(\mathbf{R}). \tag{4.2053}$$

Then the two terms appearing in Eq. (4.2045) become

$$a_{\mathrm{pp}}(\mathbf{R}) e^{-j[\chi(\mathbf{R}) + \phi_{\mathrm{pp}}(\mathbf{R}) - \phi_0]} = (a_{\mathrm{s}} + a_{\mathrm{a}})\, e^{-j(\Theta + \Psi)}, \tag{4.2054}$$

$$a_{\mathrm{pp}}(-\mathbf{R}) e^{+j[\chi(-\mathbf{R}) + \phi_{\mathrm{pp}}(-\mathbf{R}) - \phi_0]} = (a_{\mathrm{s}} - a_{\mathrm{a}})\, e^{+j(\Theta - \Psi)}. \tag{4.2055}$$

Their difference can be factored as

$$\begin{aligned}
&(a_\mathrm{s} + a_\mathrm{a})\, e^{-j(\Theta+\Psi)} - (a_\mathrm{s} - a_\mathrm{a})\, e^{+j(\Theta-\Psi)} \\
&\quad = e^{-j\Psi}\left[a_\mathrm{s}\left(e^{-j\Theta} - e^{+j\Theta}\right) + a_\mathrm{a}\left(e^{-j\Theta} + e^{+j\Theta}\right)\right] \\
&\quad = 2e^{-j\Psi}\left[-ja_\mathrm{s}\sin\Theta + a_\mathrm{a}\cos\Theta\right]. \qquad (4.2056)
\end{aligned}$$

After absorbing the same factor of two used in the conventional CTF definition, we finally obtain

$$\boxed{\begin{aligned}
\mathrm{CTF}_\mathrm{pp}(\mathbf{R}) = \frac{P(\mathbf{R})}{a_0} e^{-j[\chi_\mathrm{asym}(\mathbf{R})+\phi_\mathrm{pp,a}(\mathbf{R})]} \\
\times \{a_\mathrm{s}(\mathbf{R})\sin\left[\chi_\mathrm{sym}(\mathbf{R}) + \Delta\phi_\mathrm{pp,s}(\mathbf{R})\right] \\
+ja_\mathrm{a}(\mathbf{R})\cos\left[\chi_\mathrm{sym}(\mathbf{R}) + \Delta\phi_\mathrm{pp,s}(\mathbf{R})\right]\}.
\end{aligned}} \qquad (4.2057)$$

Equation (4.2057) gives a more precise meaning to the statement that a real phase plate cannot always be represented by simply adding a phase to χ. Even for a perfectly centrosymmetric device, a spatial-frequency-dependent coherent attenuation remains,

$$\frac{a_\mathrm{pp}(\mathbf{R})}{a_\mathrm{pp}(\mathbf{0})}. \qquad (4.2058)$$

Only when this ratio is approximately unity does the phase plate reduce to the pure phase-shift model of Eq. (4.2013).

If the attenuation itself is not centrosymmetric, $a_\mathrm{a} \neq 0$, the second term in Eq. (4.2057) introduces an additional imaginary contribution proportional to $\cos[\chi_\mathrm{sym} + \Delta\phi_\mathrm{pp,s}]$. The effect can then no longer be represented by any single real phase shift of the conventional sine CTF.

4.14.9 Physical implementations

Many physical mechanisms have been proposed to realize the relative phase shift required by Eq. (4.2008). Their engineering details differ substantially, but from the viewpoint of image formation they all attempt to construct a suitable complex pupil function $Q_\mathrm{pp}(\mathbf{R})$ (Malac et al., 2021; Axelrod et al., 2024).

Approach	Phase-shifting mechanism	Typical considerations
Thin-film Zernike	Mean inner potential of a thin material film; the direct beam passes through a central hole while scattered electrons traverse the film	Conceptually simple and capable of near-quarter-wave contrast, but the film introduces scattering and may suffer from charging, contamination, a fixed phase shift, and a relatively sharp cut-on
Hole-free or Volta	A beam-induced surface-potential modification in a continuous heated film changes the phase near the direct beam	Avoids a prefabricated central hole and has been successfully used in cryo-EM; the phase shift evolves during irradiation and additional high-resolution signal loss has been measured
Electrostatic	A localized electrostatic potential changes the electron action and therefore its phase	Phase shift can in principle be tuned electrically; practical issues include support structures, obstruction, charging, alignment, and fabrication
Laser	The ponderomotive interaction with a high-intensity optical field changes the phase of the direct electron beam	No material need lie in the electron beam and stable tunable phase shifts are possible; high optical intensity, resonant cavities, enlarged diffraction optics, and precise alignment are required
Other field-based or electron-optical designs	Magnetic fields, anamorphotic transformations, or other engineered electron-optical interactions	Potentially avoid some limitations of material plates, but generally require additional electron optics or more complicated alignment

The thin-film Zernike phase plate is named after the same Frits Zernike whose polynomials were introduced in Sec. 4.11.7, but the two uses of the name refer to different ideas. The phase plate implements Zernike's phase-contrast principle; the Zernike polynomials provide a basis for representing functions over a circular pupil.

Thin-film Zernike phase plates were important demonstrations of electron-optical phase contrast, but material in the diffraction plane can introduce scattering, charging, contamination, and sharp phase transitions. The broad

review of Malac et al. (2021) discusses these effects and the large family of alternative implementations.

The hole-free or Volta phase plate instead uses a continuous heated film and a beam-induced surface potential around the direct beam (Danev et al., 2014). It enabled practical in-focus cryo-EM and near-atomic-resolution single-particle analysis (Danev and Baumeister, 2016). Its phase profile is not fixed, however, and measurements have also shown a spatial-frequency-dependent loss of signal associated with the plate (Buijsse et al., 2020).

A conceptually different approach is the laser phase plate. A high-intensity optical field produces a ponderomotive potential and therefore changes the phase of electrons passing through the field. This avoids placing material in the electron beam and eliminates scattering and charging associated with a solid phase plate (Schwartz et al., 2019). Recent developments and remaining technical challenges are reviewed by Axelrod et al. (2024).

From the standpoint of the image-formation model, however, these technologies need not be treated separately. What matters is the resulting function

$$Q_{\mathrm{pp}}(\mathbf{R}) = a_{\mathrm{pp}}(\mathbf{R}) e^{-j\phi_{\mathrm{pp}}(\mathbf{R})}. \tag{4.2059}$$

4.14.10 Phase-plate fluctuations lead naturally to an envelope

For a perfectly stable phase plate, $\phi_{\mathrm{pp}}(\mathbf{R})$ is part of the deterministic coherent transfer function. If instead the phase shift changes during an exposure or varies among images, the distinction between a phase plate and an envelope becomes important.

Suppose

$$\Delta\phi_{\mathrm{pp}}(\mathbf{R}, \xi) = \overline{\Delta\phi}_{\mathrm{pp}}(\mathbf{R}) + \delta\phi_{\mathrm{pp}}(\mathbf{R}, \xi), \tag{4.2060}$$

where ξ labels different realizations. For each realization the phase plate remains a coherent optical element. If the fluctuations are subsequently averaged at detection, their characteristic function is

$$\boxed{E_{\mathrm{pp}}(\mathbf{R}) = \left\langle e^{-j\delta\phi_{\mathrm{pp}}(\mathbf{R},\xi)} \right\rangle_{\xi}.} \tag{4.2061}$$

Under the same assumptions developed in the previous section, this factor can act as an additional CTF envelope.

Thus,

$$\boxed{\begin{array}{lcl} \overline{\Delta\phi}_{\rm pp} & : & \text{deterministic phase-plate transfer,} \\ \delta\phi_{\rm pp} & : & \text{fluctuating phase-plate contribution,} \\ E_{\rm pp} & : & \text{contrast loss after statistical averaging.} \end{array}} \tag{4.2062}$$

4.14.11 What a phase plate does—and does not do

A phase plate should finally be placed in the same conceptual hierarchy as the other ingredients of the image-formation model.

It does not create new spatial frequencies and does not increase the fundamental information limit set by the electron wavelength, apertures, aberrations, coherence, specimen stability, and detector performance. Instead, it redistributes how efficiently the spatial frequencies already present in the exit wave are converted into image intensity. Its most important advantage for weak biological specimens is the stronger transfer of low spatial frequencies (Malac et al., 2021; Axelrod et al., 2024).

The deterministic part of the image-formation model can therefore be summarized as

$$\boxed{\begin{array}{c} \text{lens aberrations} \quad \longrightarrow \quad \chi(\mathbf{R}), \\ \text{phase plate} \quad \longrightarrow \quad \Delta\phi_{\rm pp}(\mathbf{R}), \\ \text{amplitude contrast} \quad \longrightarrow \quad \phi_A, \\ \Downarrow \\ \mathrm{CTF}_{\rm coh}(\mathbf{R}) = P(\mathbf{R}) e^{-j\chi_{\rm asym}(\mathbf{R})} \sin\left[\chi_{\rm sym}(\mathbf{R}) + \Delta\phi_{\rm pp}(\mathbf{R}) + \phi_A\right]. \end{array}} \tag{4.2063}$$

The envelopes developed in the following section act at a different stage: they describe how much of this coherent contrast survives the statistical averaging performed during the exposure.

4.15 The contrast transfer function

We have introduced the ingredients of the electron-optical image-formation model separately. Defocus and lens aberrations modify the phase of the coherently propagated electron wave; the objective aperture limits the range of transmitted spatial frequencies; amplitude contrast introduces a component in quadrature with phase contrast; and fluctuations of energy, illumination

direction, or specimen position lead to statistical averaging and attenuation envelopes.

We can now assemble these effects into a single contrast-transfer model.

It is useful to do this progressively. We shall start from the simplest defocus-only CTF, then introduce deterministic aberrations, distinguish symmetric and antisymmetric aberrations, add amplitude contrast, and finally incorporate the statistical attenuation produced by partial coherence and motion.

An important qualification should be stated at the outset. A scalar CTF is not the most general description of electron image formation. It arises after the weak-object linearization introduced in Sec. 4.10.10, together with the assumptions of an approximately shift-invariant imaging system and sufficiently simple illumination statistics. The more general partially coherent description is the mutual-intensity formulation of Eq. (3.739). The CTF is the particularly useful reduction of that general theory that applies under the conditions normally encountered in single-particle cryo-EM.

4.15.1 Level 1: the defocus-only CTF

The simplest model was derived in Sec. 4.10.10. For a weak phase object,

$$\psi_i(\mathbf{r}) \simeq 1 + j\varphi(\mathbf{r}), \tag{4.2064}$$

the defocused objective transfers the electron wave coherently according to

$$H_{\text{def}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\text{def}}(\mathbf{R})} \tag{4.2065}$$

[Eq. (4.1198)]. With the cryo-EM defocus convention of Eq. (4.1188),

$$\boxed{\chi_{\text{def}}(R) = -\pi\lambda\Delta f R^2.} \tag{4.2066}$$

The corresponding detected phase contrast is

$$\boxed{\text{CTF}_{\text{def}}(R) = P(R)\sin\left[\chi_{\text{def}}(R)\right] = -P(R)\sin\left(\pi\lambda\Delta f R^2\right),} \tag{4.2067}$$

which is Eq. (4.1229).

At this first level, there are already three conceptually different operations:

$$\boxed{\begin{array}{ccl} P(\mathbf{R}) & : & \text{coherent selection of transmitted frequencies,} \\ e^{-j\chi_{\text{def}}(\mathbf{R})} & : & \text{coherent phase modification,} \\ \sin\chi_{\text{def}} & : & \text{conversion of wave phase into detected intensity contrast.} \end{array}} \tag{4.2068}$$

In particular, the sine does not describe the action of the lens itself. As shown in the derivation leading to Eq. (4.1231), it appears only after the two Friedel-related Fourier components interfere with the reference wave and the detector forms the intensity.

4.15.2 Level 2: deterministic coherent aberrations

For an ideal but aberrated microscope operating at a fixed electron energy, fixed illumination direction, and fixed microscope state, the propagation is still coherent. The only change is that the quadratic defocus phase is replaced by the complete deterministic aberration phase,

$$\boxed{H_{\mathrm{coh}}(\mathbf{R}) = P(\mathbf{R}) \exp\left[-j\chi(\mathbf{R})\right],} \tag{4.2069}$$

which is the transfer function introduced in Eq. (4.1237).

The phase may contain defocus, twofold astigmatism, spherical aberration, coma, trefoil, fourfold astigmatism, and any higher-order terms considered in Sec. 4.11.1. In the Zernike representation of Eq. (4.1631), these contributions are simply different coordinates of the same pupil-phase function.

For example, retaining only defocus, twofold astigmatism, and spherical aberration gives

$$\boxed{\chi(R,\phi) = -\pi\lambda\Delta f(\phi)R^2 + \frac{\pi}{2}C_s\lambda^3R^4,} \tag{4.2070}$$

as derived in Eq. (4.1466). If this phase is centrosymmetric,

$$\chi(-\mathbf{R}) = \chi(\mathbf{R}), \tag{4.2071}$$

the phase-contrast CTF becomes

$$\boxed{\mathrm{CTF}_{\mathrm{sym}}(\mathbf{R}) = P(\mathbf{R}) \sin\left[\chi(\mathbf{R})\right],} \tag{4.2072}$$

generalizing the defocus-only result of Eq. (4.2067).

4.15.3 Level 3: symmetric and antisymmetric aberrations

The restriction

$$\chi(-\mathbf{R}) = \chi(\mathbf{R}) \tag{4.2073}$$

is not valid for all aberrations. As shown in Sec. 4.11.1, the general phase can be decomposed uniquely according to Eq. (4.1239),

$$\boxed{\chi(\mathbf{R}) = \chi_{\text{sym}}(\mathbf{R}) + \chi_{\text{asym}}(\mathbf{R}),} \tag{4.2074}$$

with

$$\chi_{\text{sym}}(-\mathbf{R}) = \chi_{\text{sym}}(\mathbf{R}), \qquad \chi_{\text{asym}}(-\mathbf{R}) = -\chi_{\text{asym}}(\mathbf{R}). \tag{4.2075}$$

Defocus, twofold astigmatism, spherical aberration, and fourfold astigmatism are symmetric. Axial coma, trefoil, and the other odd-parity Zernike families are antisymmetric, as discussed after Eq. (4.1668).

The effect of this decomposition on the CTF can be derived directly.

Phase-contrast CTF with symmetric and antisymmetric aberrations.

For a weak phase object, the general linearized intensity-transfer result derived in Sec. 4.10.10 is

$$\widehat{\delta I}(\mathbf{R}) = j\left[H(\mathbf{R}) - H^*(-\mathbf{R})\right]\widehat{\varphi}(\mathbf{R}), \tag{4.2076}$$

apart from the overall normalization factor introduced there.

For a centrosymmetric real pupil,

$$P(-\mathbf{R}) = P(\mathbf{R}), \tag{4.2077}$$

and using Eq. (4.2074),

$$H(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\text{sym}}(\mathbf{R})}e^{-j\chi_{\text{asym}}(\mathbf{R})}. \tag{4.2078}$$

At the Friedel-related frequency,

$$\begin{aligned} H^*(-\mathbf{R}) &= P(\mathbf{R})\exp\left[+j\chi(-\mathbf{R})\right] \\ &= P(\mathbf{R})\exp\left[+j\chi_{\text{sym}}(\mathbf{R}) - j\chi_{\text{asym}}(\mathbf{R})\right] \\ &= P(\mathbf{R})e^{-j\chi_{\text{asym}}(\mathbf{R})}e^{+j\chi_{\text{sym}}(\mathbf{R})}. \end{aligned} \tag{4.2079}$$

Therefore,

$$\begin{aligned} H(\mathbf{R}) - H^*(-\mathbf{R}) &= P(\mathbf{R})e^{-j\chi_{\text{asym}}(\mathbf{R})}\left[e^{-j\chi_{\text{sym}}} - e^{+j\chi_{\text{sym}}}\right] \\ &= -2jP(\mathbf{R})e^{-j\chi_{\text{asym}}(\mathbf{R})}\sin\left[\chi_{\text{sym}}(\mathbf{R})\right]. \end{aligned} \tag{4.2080}$$

Substituting into Eq. (4.2076), and absorbing the factor 2 into the contrast normalization, gives

$$\boxed{\mathrm{CTF}_{\mathrm{coh}}(\mathbf{R}) = P(\mathbf{R}) \sin\left[\chi_{\mathrm{sym}}(\mathbf{R})\right] e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}.} \tag{4.2081}$$

Equation (4.2081) reveals a particularly useful separation:

$$\boxed{\begin{array}{lcl} \chi_{\mathrm{sym}} & \longrightarrow & \text{contrast oscillations and CTF zeroes,} \\ \chi_{\mathrm{asym}} & \longrightarrow & \text{Fourier phase of the transferred image.} \end{array}} \tag{4.2082}$$

The antisymmetric aberration appears as a unit-modulus factor,

$$\left| e^{-j\chi_{\mathrm{asym}}(\mathbf{R})} \right| = 1. \tag{4.2083}$$

Consequently, within the weak-phase coherent model,

$$\boxed{\left|\mathrm{CTF}_{\mathrm{coh}}(\mathbf{R})\right| = P(\mathbf{R}) \left|\sin \chi_{\mathrm{sym}}(\mathbf{R})\right|.} \tag{4.2084}$$

Antisymmetric aberrations therefore do not change the CTF modulus or its zeroes at this level; they change the phases of Fourier coefficients.

This is one reason why antisymmetric aberrations are much less evident in a conventional power-spectrum representation than defocus or astigmatism. Their effect becomes apparent when phase relationships among individual particle images are exploited.

Although Eq. (4.2081) is complex, the predicted image remains real. Indeed,

$$\boxed{\mathrm{CTF}_{\mathrm{coh}}(-\mathbf{R}) = \mathrm{CTF}^{*}_{\mathrm{coh}}(\mathbf{R}),} \tag{4.2085}$$

because the sine factor is even and $e^{-j\chi_{\mathrm{asym}}}$ has the required conjugate symmetry. Thus the generalized complex CTF has exactly the Hermitian symmetry required for the inverse Fourier transform to be real.

4.15.4 Adding amplitude contrast

Amplitude contrast has a different physical origin from lens aberrations, as emphasized in Sec. 4.12. For the weak complex object of Eq. (4.1767), the amplitude and phase components are transferred in quadrature, according to Eq. (4.1787).

The conventional one-parameter model further assumes Eq. (4.1793), allowing amplitude and phase contrast to be represented by one effective

object function. With the normalized amplitude-contrast coefficient A introduced in Eq. (4.1795), the symmetric-only CTF was

$$P(\mathbf{R})\left[\sqrt{1-A^2}\sin\chi + A\cos\chi\right]. \tag{4.2086}$$

Repeating the symmetric–antisymmetric decomposition above gives

$$\boxed{\begin{aligned}\mathrm{CTF}_{\mathrm{coh}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})} \\ \times\left[\sqrt{1-A^2}\,\sin\chi_{\mathrm{sym}}(\mathbf{R}) + A\,\cos\chi_{\mathrm{sym}}(\mathbf{R})\right].\end{aligned}} \tag{4.2087}$$

Using the amplitude-contrast phase

$$\phi_A = \arcsin A \tag{4.2088}$$

from Eq. (4.1798), this can be written more compactly as

$$\boxed{\mathrm{CTF}_{\mathrm{coh}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \phi_A\right].} \tag{4.2089}$$

If a centrosymmetric phase plate is present, the coherent relative phase derived in Sec. 4.14 must also be included. Thus Eq. (4.2089) generalizes to

$$\boxed{\mathrm{CTF}_{\mathrm{coh}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \Delta\phi_{\mathrm{pp}}(\mathbf{R}) + \phi_A\right].} \tag{4.2090}$$

The three phase-like quantities appearing inside the sine have different physical meanings. The aberration phase χ_{sym} is accumulated through the deterministic electron optics; $\Delta\phi_{\mathrm{pp}}$ is an intentionally imposed coherent relative phase; and ϕ_A is the effective phase shift resulting from amplitude–phase mixing in the weak-object intensity model.

This equation contains all the effects considered so far that belong to a *single coherent realization* of the experiment. The pupil changes the transmitted amplitude, the deterministic aberrations change the phase of the electron wave, and amplitude contrast changes the way the object wave interferes with the reference wave at detection.

4.15.5 What does "coherence" mean here?

Before introducing the envelopes, it is worth clarifying the word *coherence*. Expressions such as "spatial coherence" and "temporal coherence" can be misleading if they are interpreted as properties that an individual electron simply either possesses or lacks.

Coherence is fundamentally a *relation between two components of a statistical wave field*. In Sec. 3.11.2, spatial coherence was defined through the mutual coherence function

$$\Gamma_s(\mathbf{r}_1, \mathbf{r}_2) = \langle \psi_s(\mathbf{r}_1)\psi_s^*(\mathbf{r}_2)\rangle \tag{4.2091}$$

[Eq. (3.732)]. This quantity asks whether the relative phase between the wave at $\mathbf{r}_1$ and the wave at $\mathbf{r}_2$ is reproducible across the ensemble.

Similarly, Eq. (3.759) introduced

$$W(E, E') = \langle a(E)a^*(E')\rangle\,, \tag{4.2092}$$

which asks whether two different energy components maintain a reproducible phase relation.

Thus the appropriate question is never simply

> "Is the electron beam coherent?"

but rather

> "Which components of the electron field are coherent with which other components, and with respect to which degree of freedom?"

In particular, perfect coherence does not require every emitted electron to possess the same absolute phase.

A random global phase does not destroy coherence.

Suppose that the nth realization of the source is

$$\psi_n(\mathbf{r}) = e^{j\phi_n}\psi_0(\mathbf{r}), \tag{4.2093}$$

where the global phase ϕ_n may vary arbitrarily from one emitted electron to another. Then

$$\begin{aligned}\psi_n(\mathbf{r}_1)\psi_n^*(\mathbf{r}_2) &= e^{j\phi_n}\psi_0(\mathbf{r}_1)e^{-j\phi_n}\psi_0^*(\mathbf{r}_2)\\ &= \psi_0(\mathbf{r}_1)\psi_0^*(\mathbf{r}_2).\end{aligned} \tag{4.2094}$$

Averaging over all realizations therefore gives

$$\boxed{\Gamma_s(\mathbf{r}_1, \mathbf{r}_2) = \psi_0(\mathbf{r}_1)\psi_0^*(\mathbf{r}_2),} \tag{4.2095}$$

independently of the distribution of the global phases ϕ_n.

The physically relevant quantity is therefore the *relative* phase within each realization and its reproducibility across the ensemble, not a common absolute phase shared by all detected electrons.

With this interpretation, the conventional terminology can be made more precise:

temporal partial coherence	:	correlation between different energies or frequencies,
spatial partial coherence	:	correlation between different transverse source points or, equivalently, illumination directions.

(4.2096)

In the standard cryo-EM source model, different energies are often treated as an incoherent mixture, as in Eq. (3.760), while different effective source points are treated as mutually incoherent illumination states. Each individual member of either ensemble nevertheless propagates *coherently* through the microscope.

This is the central distinction:

$$\boxed{\begin{array}{c}\text{one realization}\\ \Downarrow\\ \text{coherent wave propagation}\\ \text{many mutually partially coherent or incoherent realizations}\\ \Downarrow\\ \text{statistical averaging at detection.}\end{array}} \tag{4.2097}$$

4.15.6 Level 4: statistical averaging and envelope functions

The distinction just made explains why an envelope is fundamentally different from an aberration.

For a fixed realization ξ, the microscope has a definite coherent transfer function,

$$H_\xi(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_\xi(\mathbf{R})}. \tag{4.2098}$$

The variable ξ may represent energy, illumination angle, specimen position, lens-current fluctuation, or another changing experimental quantity. Each realization therefore has a perfectly well-defined phase. The attenuation appears only because the detector combines many such realizations.

For example, the detected energy-averaged intensity is given by Eq. (3.766),

$$I(\mathbf{r}) \propto \int p(E) \left| \mathcal{L}_E \psi_E^{\text{in}}(\mathbf{r}) \right|^2 dE, \tag{4.2099}$$

and the corresponding spatial average over illumination directions is Eq. (3.779).

Thus an aberration and an envelope act at two logically different stages:

$$\boxed{\begin{array}{lcc} \text{aberration} & : & \text{phase of each coherent realization,} \\ \text{envelope} & : & \text{survival of contrast after averaging realizations.} \end{array}} \tag{4.2100}$$

When statistical averaging becomes a multiplicative envelope

The replacement of the full ensemble average by a multiplicative envelope requires an additional approximation. It is useful to state this explicitly.

Multiplicative envelope from a symmetric phase distribution.

Let a nominal symmetric phase be

$$\chi_{\text{sym}}(\mathbf{R}), \tag{4.2101}$$

and suppose that a fluctuating experimental parameter produces the additional phase

$$\delta\chi_\xi(\mathbf{R}). \tag{4.2102}$$

For simplicity, consider first pure phase contrast. For realization ξ,

$$C_\xi(\mathbf{R}) = P(\mathbf{R}) e^{-j\chi_{\text{asym}}(\mathbf{R})} \sin\left[\chi_{\text{sym}}(\mathbf{R}) + \delta\chi_\xi(\mathbf{R})\right]. \tag{4.2103}$$

Using

$$\begin{aligned} \sin(\chi_{\text{sym}} + \delta\chi_\xi) = {} & \sin\chi_{\text{sym}} \cos\delta\chi_\xi \\ & + \cos\chi_{\text{sym}} \sin\delta\chi_\xi, \end{aligned} \tag{4.2104}$$

the ensemble average becomes

$$\begin{aligned} \langle C_\xi(\mathbf{R}) \rangle = {} & P(\mathbf{R}) e^{-j\chi_{\text{asym}}(\mathbf{R})} \\ & \times \left[\sin\chi_{\text{sym}} \langle\cos\delta\chi_\xi\rangle + \cos\chi_{\text{sym}} \langle\sin\delta\chi_\xi\rangle\right]. \end{aligned} \tag{4.2105}$$

If the distribution of phase fluctuations is symmetric around zero,

$$p(\delta\chi) = p(-\delta\chi), \tag{4.2106}$$

then $\sin\delta\chi$ is odd and therefore

$$\langle \sin\delta\chi_\xi \rangle = 0. \tag{4.2107}$$

Defining

$$\boxed{E_\xi(\mathbf{R}) \triangleq \langle \cos\delta\chi_\xi(\mathbf{R}) \rangle = \left\langle e^{-j\delta\chi_\xi(\mathbf{R})} \right\rangle,} \tag{4.2108}$$

where the second equality follows because the characteristic function is real for a symmetric distribution, we obtain

$$\boxed{\langle C_\xi(\mathbf{R}) \rangle = E_\xi(\mathbf{R}) C_0(\mathbf{R}).} \tag{4.2109}$$

Equation (4.2109) explains why the envelope appears multiplicatively in the usual CTF. It is not a fundamental additional optical element. It is the characteristic function of the statistical phase distribution generated by the ensemble.

If the distribution of fluctuations is asymmetric,

$$\langle \sin\delta\chi_\xi \rangle \neq 0, \tag{4.2110}$$

and the statistical average also introduces a phase shift. In that more general situation the effect cannot be represented by a purely real positive envelope.

Likewise, if the fluctuations change not only the phase but also the pupil, magnification, or other properties of the imaging system, a single multiplicative envelope need not exist. The full ensemble average of Sec. 3.11 is then the more fundamental description.

4.15.7 The physically motivated envelopes

Under the approximations derived in Sec. 4.13, several statistically independent effects can be represented by multiplicative factors.

For temporal partial coherence, the Gaussian focus-spread model gives Eq. (4.1836),

$$\boxed{E_t(R) = \exp\left[-\frac{1}{2}\left(\pi\lambda\sigma_f\right)^2 R^4\right].} \tag{4.2111}$$

Here the physical sequence is

$$\text{energy or lens-strength variation} \longrightarrow \text{focus variation} \longrightarrow \text{phase variation} \longrightarrow E_t. \tag{4.2112}$$

Chromatic aberration therefore determines the sensitivity of the coherent phase to energy, whereas the envelope appears only after the distribution of energies is averaged.

For finite illumination angular spread, the Gaussian source model gives Eq. (4.1862),

$$\boxed{E_s(\mathbf{R}) = \exp\left[-\frac{\alpha_s^2}{4\lambda^2}\left\|\boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R})\right\|^2\right].} \tag{4.2113}$$

In the standard use of this expression, χ is dominated by the symmetric axial aberrations such as defocus and spherical aberration. The envelope measures how rapidly the coherent pupil phase changes when the illumination direction is varied.

Motion introduces another statistical average. Unlike temporal and spatial partial coherence, motion is not conventionally a source coherence property, but mathematically it enters in the same way because different specimen positions are sampled during the exposure. Its transfer factor depends on the adopted motion model, as discussed in Sec. 4.13.4. We shall denote it generically by

$$E_m(\mathbf{R}). \tag{4.2114}$$

If the underlying fluctuations are statistically independent, Eq. (4.1962) gives

$$\boxed{E_{\text{tot}}(\mathbf{R}) = E_t(\mathbf{R})E_s(\mathbf{R})E_m(\mathbf{R}).} \tag{4.2115}$$

In practical applications, these factors may instead be replaced by the single effective Gaussian envelope introduced in Sec. 4.13.6,

$$\boxed{E_{\text{tot}}(R) \approx E_{\text{eff}}(R) = \exp\left(-\frac{B}{4}R^2\right).} \tag{4.2116}$$

4.15.8 The complete factorized CTF model

We can finally assemble the coherent and statistical contributions.

Under the weak-object approximation, a centrosymmetric pupil, a constant normalized amplitude-contrast coefficient A, and envelope mechanisms that can be represented by a real multiplicative factor, the complete local

CTF is

$$\boxed{\begin{aligned}\mathrm{CTF}(\mathbf{R}) = E_{\mathrm{tot}}(\mathbf{R})P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})} \\ \times \left[\sqrt{1-A^2}\,\sin\chi_{\mathrm{sym}}(\mathbf{R}) + A\,\cos\chi_{\mathrm{sym}}(\mathbf{R})\right].\end{aligned}} \tag{4.2117}$$

Equivalently, using $\phi_A = \arcsin A$ from Eq. (4.1798),

$$\boxed{\mathrm{CTF}(\mathbf{R}) = E_{\mathrm{tot}}(\mathbf{R})P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \Delta\phi_{\mathrm{pp}}(\mathbf{R}) + \phi_A\right].} \tag{4.2118}$$

The symmetric phase can contain all even-parity aberrations,

$$\chi_{\mathrm{sym}} = \chi_{\mathrm{def}} + \chi_{\mathrm{2fold}} + \chi_{C_s} + \chi_{\mathrm{4fold}} + \cdots, \tag{4.2119}$$

whereas

$$\chi_{\mathrm{asym}} = \chi_{\mathrm{coma}} + \chi_{\mathrm{trefoil}} + \cdots. \tag{4.2120}$$

For the common defocus–astigmatism–spherical-aberration model,

$$\begin{aligned}\chi_{\mathrm{sym}}(R,\phi) = & -\pi\lambda\Delta f(\phi)R^2 \\ & + \frac{\pi}{2}C_s\lambda^3R^4 + \chi_{\mathrm{sym}}^{\mathrm{higher}}(R,\phi).\end{aligned} \tag{4.2121}$$

Substitution into Eq. (4.2118) gives

$$\boxed{\begin{aligned}\mathrm{CTF}(R,\phi) = E_{\mathrm{tot}}(R,\phi)P(R,\phi)e^{-j\chi_{\mathrm{asym}}(R,\phi)} \\ \times \sin\Big[-\pi\lambda\Delta f(\phi)R^2 + \frac{\pi}{2}C_s\lambda^3R^4 \\ +\chi_{\mathrm{sym}}^{\mathrm{higher}}(R,\phi) + \Delta\phi_{\mathrm{pp}}(R,\phi) + \phi_A\Big].\end{aligned}} \tag{4.2122}$$

This expression displays the different physical layers of the model explicitly:

factor	physical origin	mathematical action
$P(\mathbf{R})$	objective aperture	coherent amplitude selection
$E_{\mathrm{tot}}(\mathbf{R})$	ensemble averaging	attenuation
$\chi_{\mathrm{sym}}(\mathbf{R})$	deterministic even aberrations	contrast oscillations
$\chi_{\mathrm{asym}}(\mathbf{R})$	deterministic odd aberrations	Fourier phase shift
ϕ_A	amplitude contrast	phase/amplitude quadrature mixing.

(4.2123)

4.15.9 What is coherent and what is incoherent in the final model?

The factorization in Eq. (4.2117) should not obscure the very different origins of its terms.

Defocus and lens aberrations are *coherent*. For a given electron energy, illumination direction, and microscope configuration, they assign a definite phase to every spatial-frequency component. They do not, by themselves, imply any averaging or loss of coherence.

The objective aperture is also part of the coherent optical system. It changes the amplitude transmitted by one realization, rather than representing statistical averaging over different realizations.

Amplitude contrast is neither an aberration nor a partial-coherence effect. It describes a property of the specimen exit wave and changes the relative weight with which amplitude and phase modulations contribute to the detected image.

Temporal and spatial partial coherence are different. They describe statistical correlations within an ensemble of incident electron states. In the common cryo-EM approximations, electrons of different effective energies or illumination directions are treated as mutually incoherent ensemble members. Each member propagates coherently, but their detected intensities are statistically combined. The resulting loss of contrast is represented by the envelope.

Specimen motion has yet another physical origin. It does not describe coherence of the source at all. Nevertheless, because the detector integrates over different specimen positions during the exposure, the mathematics again produces an average of Fourier phase factors and hence a motion envelope.

We may therefore summarize the entire construction as

$$\boxed{\begin{array}{c}\text{specimen exit wave}\\ \Downarrow\\ \text{one coherent realization}\\ P(\mathbf{R})\exp\left[-j\left(\chi_{\text{sym}}+\chi_{\text{asym}}\right)\right]\\ \Downarrow\\ \text{coherent interference and intensity detection}\\ \Downarrow\\ \text{average over energy, illumination direction, and motion}\\ \Downarrow\\ \text{CTF}(\mathbf{R})\text{ of Eq. (4.2117).}\end{array}} \tag{4.2124}$$

The word "coherence" is therefore best understood as shorthand for the statistical correlation between specified components of the electron field. It does not mean that all electrons are emitted at the same point, with the same energy, or with the same absolute phase. The labels "temporal" and "spatial" identify the degree of freedom over which the correlation is being considered.

4.15.10 What is visible in the power spectrum?

The distinction between symmetric and antisymmetric aberrations becomes especially transparent in the power spectrum. Taking the squared modulus of Eq. (4.2117) gives

$$\boxed{\begin{aligned}|\text{CTF}(\mathbf{R})|^2 &= E_{\text{tot}}^2(\mathbf{R})P^2(\mathbf{R})\\ &\quad\times\left[\sqrt{1-A^2}\sin\chi_{\text{sym}}(\mathbf{R})+A\cos\chi_{\text{sym}}(\mathbf{R})\right]^2.\end{aligned}} \tag{4.2125}$$

The factor

$$e^{-j\chi_{\text{asym}}} \tag{4.2126}$$

has disappeared because its modulus is one. Thus the conventional Thon rings are determined primarily by the symmetric aberrations, amplitude contrast, and envelopes. Antisymmetric aberrations alter Fourier phases without directly changing the ideal CTF power at first order.

This also illustrates why estimating the complete optical state of a modern cryo-EM data set requires more information than a rotationally averaged micrograph power spectrum. Symmetric aberrations are naturally visible

through the positions and shapes of CTF oscillations; attenuation mechanisms are visible through their decay; but antisymmetric aberrations must be inferred from phase-sensitive relationships among images.

4.16 The background of the image power spectrum

The contrast transfer function is commonly estimated from the power spectrum of an experimental micrograph. A typical power spectrum does not, however, consist simply of the squared CTF. The CTF oscillations are superimposed on a much larger, smoothly varying background, as illustrated schematically in Fig. 4.15. One often describes the observed spectrum by an expression of the form

$$\boxed{S_{\text{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})\,|\text{CTF}(\mathbf{R})|^2\,,} \tag{4.2127}$$

where $B(\mathbf{R})$ is a slowly varying background and $A(\mathbf{R})$ determines the amplitude of the CTF-modulated component.

Although Eq. (4.2127) is extremely useful, its physical interpretation requires some care. There is no single physical process called "the background". Electron-counting statistics, detector response, electronic noise, and other processes enter the image at different stages and are therefore transferred differently.

Indeed, early quantitative CTF models already distinguished noise added before and after the microscope transfer function (Sorzano et al., 2007). A particularly useful physical decomposition was given by Zeng et al. (2007), who distinguished structural specimen background, electron shot noise, and detector-electronic noise. More recent experimental work has clarified several of these contributions (Baxter et al., 2009; McMullan et al., 2015; Ruskin et al., 2013a).

We shall first derive the general PSD structure and then examine the physical origin of its different components.

4.16.1 Why the transfer function appears squared in the power spectrum

Let $x(\mathbf{r})$ denote a weak-object image quantity before the action of the microscope and let $\text{CTF}(\mathbf{R})$ denote the complete linear image transfer function derived in Sec. 4.15. Thus,

$$\widehat{y}(\mathbf{R}) = \text{CTF}(\mathbf{R})\widehat{x}(\mathbf{R}). \tag{4.2128}$$

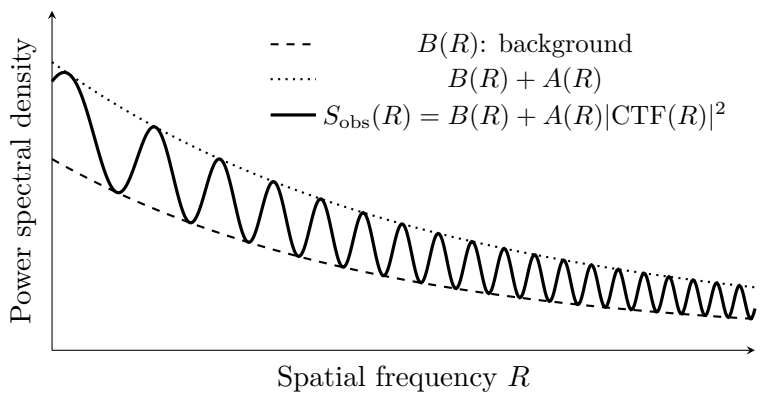


Figure 4.15: Schematic decomposition of a cryo-EM image power spectrum. The observed power spectral density can be represented approximately as a slowly varying background $B(R)$ plus a structural contribution $A(R)|\text{CTF}(R)|^2$. The function $B(R) + A(R)$ defines a smooth upper trend, whereas the squared CTF causes the power to oscillate between the upper trend and the background. Near an ideal CTF zero, the observed PSD approaches $B(R)$; near a CTF maximum, it approaches $B(R) + A(R)$. The curves are schematic and are intended to illustrate the decomposition rather than a particular experimental data set.

The power spectral density is the ensemble-average squared modulus of the Fourier coefficients. Consequently, the transfer function enters the PSD through its squared modulus.

Propagation of a power spectrum through a linear system.

Starting from Eq. (4.2128),

$$\widehat{y}(\mathbf{R}) = \mathrm{CTF}(\mathbf{R})\widehat{x}(\mathbf{R}), \tag{4.2129}$$

its squared modulus is

$$\begin{aligned} |\widehat{y}(\mathbf{R})|^2 &= \widehat{y}(\mathbf{R})\widehat{y}^*(\mathbf{R}) \\ &= \mathrm{CTF}(\mathbf{R})\widehat{x}(\mathbf{R})\,\mathrm{CTF}^*(\mathbf{R})\widehat{x}^*(\mathbf{R}) \\ &= |\mathrm{CTF}(\mathbf{R})|^2\,|\widehat{x}(\mathbf{R})|^2\,. \end{aligned} \tag{4.2130}$$

Taking the statistical average,

$$\left\langle |\widehat{y}(\mathbf{R})|^2 \right\rangle = |\mathrm{CTF}(\mathbf{R})|^2 \left\langle |\widehat{x}(\mathbf{R})|^2 \right\rangle. \tag{4.2131}$$

Defining

$$S_x(\mathbf{R}) \triangleq \left\langle |\widehat{x}(\mathbf{R})|^2 \right\rangle, \qquad S_y(\mathbf{R}) \triangleq \left\langle |\widehat{y}(\mathbf{R})|^2 \right\rangle, \tag{4.2132}$$

we obtain

$$\boxed{S_y(\mathbf{R}) = |\mathrm{CTF}(\mathbf{R})|^2\, S_x(\mathbf{R}).} \tag{4.2133}$$

$S_x(\mathbf{R})$ and $S_y(\mathbf{R})$ are called the Power Spectrum Density (PSD) of the x and y signals, respectively.

The modulus signs in Eq. (4.2133) are important. For a purely symmetric aberration the usual CTF is real, and one often writes simply CTF^2. In the general model of Sec. 4.15, however, antisymmetric aberrations make the CTF complex. The correct PSD factor is therefore

$$|\mathrm{CTF}(\mathbf{R})|^2\,, \tag{4.2134}$$

not $\mathrm{CTF}^2(\mathbf{R})$.

Using Eq. (4.2118),

$$\mathrm{CTF}(\mathbf{R}) = E_{\mathrm{tot}}(\mathbf{R})P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \phi_A\right], \tag{4.2135}$$

we find

$$\boxed{\begin{aligned}|\mathrm{CTF}(\mathbf{R})|^2 = E_{\mathrm{tot}}^2(\mathbf{R})P^2(\mathbf{R}) \\ \times \sin^2\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \phi_A\right].\end{aligned}} \tag{4.2136}$$

The antisymmetric aberration phase has disappeared because

$$\left|e^{-j\chi_{\mathrm{asym}}}\right|^2 = 1. \tag{4.2137}$$

Thus, within the weak-object linear model, the ordinary micrograph power spectrum is directly sensitive to symmetric aberrations and envelope attenuation, but not to the phase produced by antisymmetric aberrations. This is the power-spectrum counterpart of the result derived in Eq. (4.2084).

4.16.2 Why independent image components add in power

The second important property of the PSD is that statistically uncorrelated contributions add in *power*, even though their image amplitudes add before the power spectrum is calculated.

Addition of independent power spectra.

Let the image be the sum of two random fields,

$$i(\mathbf{r}) = a(\mathbf{r}) + b(\mathbf{r}). \tag{4.2138}$$

Assume that the fields are wide-sense stationary, so that their means are independent of position,

$$\mu_a = \langle a(\mathbf{r})\rangle, \qquad \mu_b = \langle b(\mathbf{r})\rangle, \tag{4.2139}$$

and define the zero-mean fluctuations

$$\delta a(\mathbf{r}) = a(\mathbf{r}) - \mu_a, \qquad \delta b(\mathbf{r}) = b(\mathbf{r}) - \mu_b. \tag{4.2140}$$

The corresponding fluctuation of the total image is

$$\delta i(\mathbf{r}) = \delta a(\mathbf{r}) + \delta b(\mathbf{r}). \tag{4.2141}$$

The autocovariance of the image is

$$C_i(\boldsymbol{\rho}) = \langle \delta i(\mathbf{r})\delta i^*(\mathbf{r}+\boldsymbol{\rho})\rangle\,, \tag{4.2142}$$

where, by stationarity, it depends only on the displacement $\boldsymbol{\rho}$ and not on the absolute position $\mathbf{r}$.

Substituting $\delta i = \delta a + \delta b$ into Eq. (4.2142),

$$\begin{aligned} C_i(\boldsymbol{\rho}) = & \langle \delta a(\mathbf{r}) \delta a^*(\mathbf{r}+\boldsymbol{\rho}) \rangle + \langle \delta b(\mathbf{r}) \delta b^*(\mathbf{r}+\boldsymbol{\rho}) \rangle \\ & + \langle \delta a(\mathbf{r}) \delta b^*(\mathbf{r}+\boldsymbol{\rho}) \rangle + \langle \delta b(\mathbf{r}) \delta a^*(\mathbf{r}+\boldsymbol{\rho}) \rangle . \end{aligned} \tag{4.2143}$$

The first two terms are the autocovariances

$$C_a(\boldsymbol{\rho}) = \langle \delta a(\mathbf{r}) \delta a^*(\mathbf{r}+\boldsymbol{\rho}) \rangle , \tag{4.2144}$$

and

$$C_b(\boldsymbol{\rho}) = \langle \delta b(\mathbf{r}) \delta b^*(\mathbf{r}+\boldsymbol{\rho}) \rangle . \tag{4.2145}$$

If a and b are statistically independent, then $\delta a(\mathbf{r})$ and $\delta b(\mathbf{r}+\boldsymbol{\rho})$ are also independent. Therefore,

$$\begin{aligned} \langle \delta a(\mathbf{r}) \delta b^*(\mathbf{r}+\boldsymbol{\rho}) \rangle &= \langle \delta a(\mathbf{r}) \rangle \langle \delta b^*(\mathbf{r}+\boldsymbol{\rho}) \rangle \\ &= 0, \end{aligned} \tag{4.2146}$$

because both centered fields have zero mean. The other cross term vanishes in the same way. Hence,

$$\boxed{C_i(\boldsymbol{\rho}) = C_a(\boldsymbol{\rho}) + C_b(\boldsymbol{\rho}).} \tag{4.2147}$$

According to the Wiener–Khinchin theorem, the power spectral density is the Fourier transform of the autocovariance,

$$S_x(\mathbf{R}) = \int C_x(\boldsymbol{\rho}) e^{-j2\pi \mathbf{R}\cdot\boldsymbol{\rho}} d\boldsymbol{\rho}. \tag{4.2148}$$

Taking the Fourier transform of Eq. (4.2147) and using the linearity of the Fourier transform gives

$$\boxed{S_i(\mathbf{R}) = S_a(\mathbf{R}) + S_b(\mathbf{R}).} \tag{4.2149}$$

The centering in the previous derivation is important. If we use the raw autocorrelation

$$R_x(\boldsymbol{\rho}) = \langle x(\mathbf{r}) x^*(\mathbf{r}+\boldsymbol{\rho}) \rangle , \tag{4.2150}$$

then

$$R_x(\boldsymbol{\rho}) = C_x(\boldsymbol{\rho}) + |\mu_x|^2. \tag{4.2151}$$

For independent a and b, the cross terms of the raw autocorrelation do not vanish in general; rather,

$$\langle a(\mathbf{r})b^*(\mathbf{r}+\boldsymbol{\rho})\rangle = \mu_a\mu_b^*. \tag{4.2152}$$

These constant mean terms contribute only at zero spatial frequency. Indeed, the full power spectrum can be written as

$$\boxed{S_i^{\mathrm{raw}}(\mathbf{R}) = S_a(\mathbf{R}) + S_b(\mathbf{R}) + |\mu_a+\mu_b|^2\delta(\mathbf{R}).} \tag{4.2153}$$

Thus, after subtraction of the mean image, or for $\mathbf{R} \neq \mathbf{0}$, the independent power spectra add exactly as in Eq. (4.2149).

4.16.3 A layered physical model

It is useful to organize the different contributions to a cryo-EM micrograph according to the stage at which they enter the image-formation chain. This physical hierarchy also explains why the much simpler background-plus-CTF model commonly used for CTF estimation works as well as it does.

Let

$$s(\mathbf{r}) \tag{4.2154}$$

denote the reproducible specimen signal, such as the projected contrast of the particle of interest, and let

$$b_{\mathrm{el}}(\mathbf{r}) \tag{4.2155}$$

denote an irreproducible but *elastic structural background* arising from the surrounding vitreous ice, support film, contaminants, buffer, and other specimen components.

Both contributions are generated in the specimen before electron-optical image formation. Under the weak-object linear imaging model developed in Sec. 4.15, they are therefore transferred by the same image-intensity contrast transfer function,

$$\mathrm{CTF}(\mathbf{R}). \tag{4.2156}$$

This is the complete CTF of Eq. (4.2117), not the coherent wave transfer function $H(\mathbf{R})$ introduced earlier.

If the particle signal and the elastic structural background are uncorrelated random fields, their PSDs add according to Eq. (4.2149),

$$\boxed{S_{\mathrm{spec}}(\mathbf{R}) = S_s(\mathbf{R}) + S_{\mathrm{el}}(\mathbf{R}).} \tag{4.2157}$$

The electron-optical contribution to the image PSD is therefore

$$\boxed{S_{\text{image,el}}(\mathbf{R}) = |\text{CTF}(\mathbf{R})|^2 \left[S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R})\right].} \tag{4.2158}$$

The modulus is required because the complete CTF may be complex when antisymmetric aberrations are present, as shown in Eq. (4.2117).

The detector then records the structural image. Let $\text{MTF}(\mathbf{R})$ denote its modulation transfer function, whose physical origin and measurement are developed in Sec. 6.5.2 of Chap. 6. The structural contribution to the recorded PSD is then

$$\boxed{\begin{aligned} S_{\text{struct,det}}(\mathbf{R}) = {} & |\,\text{MTF}(\mathbf{R})|^2 |\,\text{CTF}(\mathbf{R})|^2 \\ & \times \left[S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R})\right]. \end{aligned}} \tag{4.2159}$$

At this point the part of the PSD that can be described purely by the electron optics of this chapter has been identified. Other physical processes also contribute to the measured micrograph, but their detailed description belongs elsewhere in the book. We therefore collect all power not explicitly represented by Eq. (4.2159) into an operational background function

$$B(\mathbf{R}). \tag{4.2160}$$

Defining

$$\boxed{A(\mathbf{R}) \triangleq |\,\text{MTF}(\mathbf{R})|^2 \left[S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R})\right],} \tag{4.2161}$$

the observed PSD can be written in the familiar form

$$\boxed{S_{\text{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R}) |\,\text{CTF}(\mathbf{R})|^2.} \tag{4.2162}$$

Equation (4.2162) is the practical form used by many CTF-estimation procedures. If $A(\mathbf{R})$ and $B(\mathbf{R})$ vary slowly compared with the CTF oscillations, the PSD contains a slowly varying lower trend on top of which the structural power oscillates according to $|\,\text{CTF}\,|^2$.

Near an ideal CTF zero,

$$|\,\text{CTF}(\mathbf{R})|^2 \simeq 0, \tag{4.2163}$$

and therefore

$$\boxed{S_{\text{obs}}(\mathbf{R}) \simeq B(\mathbf{R}).} \tag{4.2164}$$

Near a CTF maximum, by contrast, the elastic structural contribution raises the PSD above this lower trend. This gives rise to the smooth lower and upper bounds often used in practical CTF estimation.

What generates the Thon rings? Equation (4.2162) also clarifies that the visible CTF oscillations need not originate primarily from the particles of interest. Any sufficiently broadband elastic structural spectrum present in the specimen can reveal the microscope transfer function.

Amorphous carbon has long been used for this purpose. More importantly for cryo-EM, the development of the vitreous-ice PSD above shows that the ice itself provides a broad elastic structural spectrum. McMullan et al. (2015) demonstrated experimentally that amorphous ice alone produces measurable Thon rings. Thus $S_{\mathrm{el}}(\mathbf{R})$ may provide a substantial fraction of the broadband structural power against which the CTF oscillations are observed.

This is advantageous for CTF estimation. The biological object does not need to possess strong Fourier components at every spatial frequency: the surrounding amorphous material itself probes the transfer function over a broad frequency range.

Structural background should therefore not be confused with detector noise. It may contain no reproducible information about the particle of interest, but it is nevertheless genuine specimen structure and is therefore modulated by the CTF.

What is contained in $B(\mathbf{R})$? The background $B(\mathbf{R})$ in Eq. (4.2162) should not be associated with one particular microscopic process. It is an operational quantity that collects several effects not represented explicitly by the elastic structural term.

Inelastic scattering is one such contribution. It changes both the energy and direction of the electrons and generates several energy-loss channels. As developed in Sec. 5.12, inelastically scattered electrons are not fundamentally equivalent to incoherent background: some retain phase information, while chromatic aberration and specimen-induced decoherence determine how much of that information survives. In conventional cryo-EM, however, a substantial fraction of their contribution is absorbed empirically into the slowly varying background.

The finite number of detected electrons introduces shot noise. Its Poisson origin and the resulting ideal white-noise PSD are developed in Sec. 6.8.1 of Chap. 6.

The detector itself modifies the electron signal and introduces additional fluctuations. The detector MTF, noise power spectrum, DQE, and electronic noise are treated in Chap. 6.

Electron-counting detectors introduce a further effect, coincidence loss,

discussed in Sec. 6.8.5. Coincidence loss is particularly important conceptually because it is *not* an independent additive source of PSD. It changes the nonlinear mapping between incident electrons and recorded counts and therefore modifies both the structural signal and the counting statistics.

Why the empirical background has no unique physical form. The practical background $B(\mathbf{R})$ is usually smooth compared with the oscillating CTF, but there is no universal theoretical expression for its shape.

In the model of Sorzano et al. (2007), the post-CTF background was represented by a flexible combination of a constant, an exponential-like term, and positive and negative Gaussian components. The purpose was to reproduce the experimentally observed two-dimensional PSD, including possible anisotropy, rather than to associate every fitted component with a particular microscopic mechanism. Subsequent work emphasized that substantial parameter cross-talk may occur between the fitted background and CTF parameters (Sorzano et al., 2009).

Modern CTF-estimation methods often retain this pragmatic approach. CTFFIND4, for example, estimates a slowly varying background of the experimental amplitude spectrum by smoothing it and subtracting the result before comparing the remaining oscillations with the theoretical CTF (Rohou and Grigorieff, 2015).

The most useful definition of the empirical background is therefore

$$\boxed{\begin{array}{c} B(\mathbf{R}) = \text{power not explained by the chosen} \\ \text{CTF-modulated structural model.} \end{array}} \tag{4.2165}$$

In practical CTF estimation one usually imposes the additional assumption that this residual power varies sufficiently slowly with $\mathbf{R}$ to be separated from the CTF oscillations.

The empirical success of such a model should not be confused with a microscopic identification of $B(\mathbf{R})$. A fitted smooth component may absorb several unrelated physical processes.

Physical hierarchy. The important distinction is therefore not simply between "signal" and "background", but between the stages at which the different contributions enter the experiment.

The part treated explicitly in this chapter is

$$\begin{gathered}\text{particle} + \text{ice} + \text{support} + \text{contaminants}\\ \Downarrow\\ \text{elastic structural fluctuations}\\ \Downarrow\\ |\,\mathrm{CTF}(\mathbf{R})|^2\\ \Downarrow\\ |\,\mathrm{MTF}(\mathbf{R})|^2\\ \Downarrow\\ \text{recorded structural PSD.}\end{gathered} \tag{4.2166}$$

Inelastic interactions modify the electron state inside the specimen and are developed in Chap. 5. Shot noise, detector transfer, electronic noise, and coincidence loss arise during the detection process and are developed in Chap. 6.

The compact expression

$$S_{\mathrm{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})|\,\mathrm{CTF}(\mathbf{R})|^2 \tag{4.2167}$$

should therefore be regarded as a useful phenomenological reduction of this richer physical hierarchy, not as a fundamental decomposition of the micrograph into two uniquely defined microscopic sources.

This distinction becomes particularly important when fitted parameters are assigned physical meaning. The positions of the CTF zeroes are strongly constrained by the deterministic electron optics developed in this chapter. By contrast, the slowly varying functions $A(\mathbf{R})$ and $B(\mathbf{R})$ can absorb several unrelated specimen, imaging, and detector effects. An excellent fit of the observed PSD does not, by itself, establish that the fitted background or envelope parameters correspond uniquely to particular microscopic mechanisms.

4.16.4 Power spectrum of the ice

A common but misleading intuition is that the vitreous ice surrounding a particle merely adds featureless noise to the micrograph. Elastic scattering from the ice is instead part of the specimen scattering potential and therefore undergoes the same electron-optical image formation as the particle itself.

This distinction was made explicitly by Zeng et al. (2007), who separated a structural image background arising from the entire specimen, including

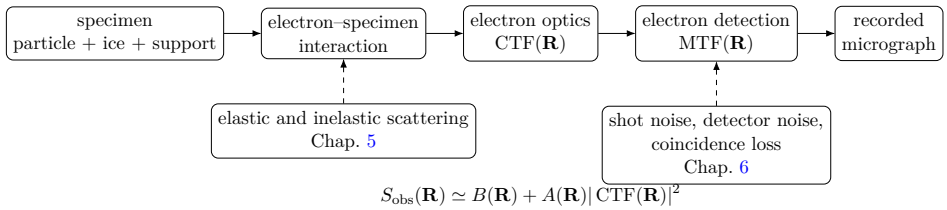


Figure 4.16: Physical hierarchy underlying the practical power-spectrum model. Elastic structural fluctuations generated in the specimen are transferred by the electron-optical CTF and subsequently by the detector. Electron–specimen interactions, including inelastic scattering, are developed in Chap. 5; shot noise, detector noise, detector transfer, and coincidence loss are developed in Chap. 6. The empirical decomposition $S_{\mathrm{obs}} \simeq B + A|\,\mathrm{CTF}\,|^2$ compresses these different physical processes into a convenient model for CTF estimation.

ice and support film, from sample-independent shot noise and detector-electronic noise. The structural component was multiplied by the CTF, microscope envelope, and detector MTF.

Experimental measurements by Baxter et al. (2009) likewise found a substantial "structural noise" contribution arising from variations in the specimen and its surrounding matrix. Its magnitude was comparable to that of the reproducible molecular signal in the data sets studied.

Perhaps the clearest demonstration is provided by McMullan et al. (2015), who recorded images of films containing essentially pure amorphous ice and observed clear Thon rings in their power spectra. The existence of these rings shows directly that amorphous ice contains elastic structural fluctuations that are modulated by the CTF. The ice spectrum is not flat: a particularly strong feature occurs near a spacing of 3.7 Å, reflecting short-range order in amorphous water.

Modern simulations reach the same conclusion from the opposite direction. Physically realistic simulation of cryo-EM images requires a non-trivial spatial model of the amorphous ice potential; atomistic and Gaussian-random-field ice models can reproduce characteristic experimental image statistics (Parkhurst et al., 2024).

Thus, the ice has an important contribution to

$$S_{\mathrm{el}}(\mathbf{R}), \tag{4.2168}$$

which is multiplied by $|\,\mathrm{CTF}(\mathbf{R})|^2$. It should not all be assigned to the additive term $B(\mathbf{R})$ of Eq. (4.2127).

The structural power spectrum of amorphous ice

Amorphous ice is not a featureless source of additive noise. Its molecular arrangement produces an elastic structural spectrum, and this spectrum is transferred by the microscope in exactly the same way as any other weak specimen contrast. The experimental results of McMullan et al. (2015) provide a particularly clear demonstration: images of films containing essentially pure amorphous ice exhibit Thon rings, proving that the ice itself provides broadband elastic structural power that is modulated by the CTF.

The situation is, however, more subtle than for a stable amorphous solid such as pre-irradiated carbon. During irradiation, the molecular configuration of the ice changes. Consequently, Fourier amplitudes recorded at different times during an exposure are not perfectly correlated. McMullan et al. (2015) modeled this evolution as a Gaussian Markov process and used it to derive the dose-dependent power spectrum of amorphous ice.

We develop the same argument here using this chapter's notation.

The spectrum of a static ice configuration. Consider first a hypothetical frozen configuration of amorphous ice that does not change during the exposure. Let

$$S_{\mathrm{ice},0}(R) \tag{4.2169}$$

denote the ensemble-averaged elastic structural power spectrum per unit squared electron dose of such configurations.

Because amorphous ice is statistically isotropic, this quantity is approximately a function only of

$$R = \|\mathbf{R}\|. \tag{4.2170}$$

It is not, however, spectrally white. In particular, McMullan et al. (2015) observed a strong broad maximum around

$$\boxed{R_{\mathrm{ice}} \simeq \frac{1}{3.7\,\text{Å}}.} \tag{4.2171}$$

They associate this feature with the next-nearest-neighbour oxygen arrangement of tetrahedrally coordinated water molecules. Thus the molecular short-range order of amorphous water leaves a characteristic signature in its power spectrum.

For a static configuration exposed with a total electron dose D, the image amplitude is proportional to D. Its structural power therefore scales as

$$S_{\mathrm{ice,static}}(R; D) = D^2 S_{\mathrm{ice},0}(R). \tag{4.2172}$$

After image formation and detection, this contribution becomes, according to Eq. (4.2158),

$$\boxed{S_{\text{ice,det}}(\mathbf{R}; D) = |\text{MTF}(\mathbf{R})|^2 \, |\text{CTF}(\mathbf{R})|^2 \, D^2 S_{\text{ice},0}(R).} \tag{4.2173}$$

Thus even pure amorphous ice is capable of generating Thon rings: $S_{\text{ice},0}(R)$ supplies the structural power, while $|\,\text{CTF}(\mathbf{R})|^2$ produces its oscillatory modulation.

Beam-induced rearrangement of the water molecules. The static model of Eq. (4.2172) is not appropriate during an electron exposure. An incident high-energy electron perturbs the local atomic configuration of the ice. Subsequent electrons therefore encounter slightly different configurations.

Following McMullan et al. (2015), let

$$\mathbf{x}_k(q) \tag{4.2174}$$

be the position of water molecule k after an accumulated dose q, measured in electrons per unit area. The simplest approximation is that an additional dose Δq produces a Gaussian random displacement,

$$\Delta \mathbf{x}_k = \mathbf{x}_k(q + \Delta q) - \mathbf{x}_k(q), \tag{4.2175}$$

with probability density

$$\boxed{p(\Delta \mathbf{x}_k \mid \Delta q) = \frac{1}{\left(2\pi\sigma_0^2 \Delta q\right)^{3/2}} \exp\left[-\frac{\|\Delta \mathbf{x}_k\|^2}{2\sigma_0^2 \Delta q}\right].} \tag{4.2176}$$

Here σ_0^2 is the mean-squared displacement in *one Cartesian direction* induced per unit electron dose. Thus

$$\left\langle \Delta x^2 \right\rangle = \left\langle \Delta y^2 \right\rangle = \left\langle \Delta z^2 \right\rangle = \sigma_0^2 \Delta q. \tag{4.2177}$$

This model is Markovian because the probability distribution of the next configuration depends only on the present configuration and on the additional accumulated dose, not on the detailed history by which the present configuration was reached.

Loss of Fourier correlation with accumulated dose. The important consequence of Eq. (4.2176) is that the Fourier amplitudes of the ice decorrelate progressively as irradiation proceeds.

Correlation of the ice Fourier amplitudes.

Consider a transverse displacement

$$\Delta\mathbf{r} = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}. \tag{4.2178}$$

At image spatial frequency $\mathbf{R}$, this displacement changes the Fourier phase of the contribution from a molecule by

$$e^{-j2\pi\mathbf{R}\cdot\Delta\mathbf{r}}. \tag{4.2179}$$

According to Eq. (4.2176), the two transverse components are independent Gaussian variables with

$$\mathrm{Var}(\Delta x) = \mathrm{Var}(\Delta y) = \sigma_0^2 \Delta q. \tag{4.2180}$$

The statistical average of Eq. (4.2179) is therefore the characteristic function of a two-dimensional Gaussian,

$$\left\langle e^{-j2\pi\mathbf{R}\cdot\Delta\mathbf{r}} \right\rangle = \exp\left[-2\pi^2\sigma_0^2\Delta q R^2\right]. \tag{4.2181}$$

Define

$$\boxed{\alpha_R \triangleq 2\pi^2\sigma_0^2 R^2.} \tag{4.2182}$$

Then

$$\boxed{\left\langle e^{-j2\pi\mathbf{R}\cdot\Delta\mathbf{r}} \right\rangle = e^{-\alpha_R \Delta q}.} \tag{4.2183}$$

Equation (4.2183) is the essential result. Two ice configurations separated by only a small accumulated dose remain strongly correlated, whereas configurations separated by a sufficiently large dose become effectively unrelated. The decorrelation is faster at higher spatial frequency because

$$\alpha_R \propto R^2. \tag{4.2184}$$

From molecular displacements to correlation between ice configurations.

Consider two configurations of the amorphous ice separated by an accumulated electron dose Δq. Let the transverse position of water molecule n in the first configuration be

$$\mathbf{r}_n, \tag{4.2185}$$

and let its position in the second configuration be

$$\mathbf{r}'_n = \mathbf{r}_n + \Delta\mathbf{r}_n. \tag{4.2186}$$

According to the Gaussian Markov model of Eq. (4.2176), the displacement $\Delta\mathbf{r}_n$ is a zero-mean Gaussian random variable whose variance in each transverse direction is

$$\sigma_0^2 \Delta q. \tag{4.2187}$$

Let $F_{\mathrm{w}}(\mathbf{R})$ denote the elastic scattering amplitude of one water molecule at spatial frequency $\mathbf{R}$. Apart from normalization constants that are common to the two configurations, the Fourier amplitude of the first ice configuration can be written as

$$f(\mathbf{R}) = F_{\mathrm{w}}(\mathbf{R}) \sum_{n=1}^{N} e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_n}. \tag{4.2188}$$

For the second configuration,

$$\begin{aligned} f'(\mathbf{R}) &= F_{\mathrm{w}}(\mathbf{R}) \sum_{n=1}^{N} e^{-j2\pi\mathbf{R}\cdot\mathbf{r}'_n} \\ &= F_{\mathrm{w}}(\mathbf{R}) \sum_{n=1}^{N} e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_n} e^{-j2\pi\mathbf{R}\cdot\Delta\mathbf{r}_n}. \end{aligned} \tag{4.2189}$$

We now keep the first ice configuration fixed and average only over the random displacements that generate the second configuration. By

the linearity of the statistical average,

$$\begin{aligned}\left\langle f'(\mathbf{R}) \mid \{\mathbf{r}_n\}\right\rangle_{\Delta\mathbf{r}} = F_{\mathrm{w}}(\mathbf{R}) \sum_{n=1}^{N} e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_n} \\ \times \left\langle e^{-j2\pi\mathbf{R}\cdot\Delta\mathbf{r}_n}\right\rangle .\end{aligned} \tag{4.2190}$$

All molecules are assumed to follow the same displacement statistics, so the last expectation is independent of n. From Eq. (4.2183),

$$\left\langle e^{-j2\pi\mathbf{R}\cdot\Delta\mathbf{r}_n}\right\rangle = e^{-\alpha_R \Delta q}, \tag{4.2191}$$

where

$$\alpha_R = 2\pi^2\sigma_0^2 R^2. \tag{4.2192}$$

Therefore,

$$\boxed{\left\langle f'(\mathbf{R}) \mid \{\mathbf{r}_n\}\right\rangle_{\Delta\mathbf{r}} = e^{-\alpha_R \Delta q} f(\mathbf{R}).} \tag{4.2193}$$

This result has an immediate interpretation: after a dose Δq, the component of the new ice Fourier amplitude that remains correlated with the original configuration is reduced by the factor $e^{-\alpha_R \Delta q}$.

To obtain the correlation between the two configurations, multiply by $f^*(\mathbf{R})$ and average also over the ensemble of possible initial ice configurations:

$$\begin{aligned}\left\langle f'(\mathbf{R}) f^*(\mathbf{R})\right\rangle &= \left\langle \left\langle f'(\mathbf{R}) \mid \{\mathbf{r}_n\}\right\rangle_{\Delta\mathbf{r}} f^*(\mathbf{R})\right\rangle_{\mathrm{ice}} \\ &= e^{-\alpha_R \Delta q} \left\langle f(\mathbf{R}) f^*(\mathbf{R})\right\rangle_{\mathrm{ice}} \\ &= e^{-\alpha_R \Delta q} \left\langle |f(\mathbf{R})|^2\right\rangle_{\mathrm{ice}} .\end{aligned} \tag{4.2194}$$

Hence,

$$\boxed{\left\langle f'(\mathbf{R}) f^*(\mathbf{R})\right\rangle = S_{\mathrm{ice},0}(R) e^{-\alpha_R \Delta q},} \tag{4.2195}$$

where

$$S_{\mathrm{ice},0}(R) \triangleq \left\langle |f(\mathbf{R})|^2\right\rangle_{\mathrm{ice}} \tag{4.2196}$$

is the structural power spectrum of an instantaneous ice configuration, up to the normalization adopted for the electron dose.

Equation (4.2195) provides the missing connection between molecular motion and image statistics. The Gaussian displacement model does not merely attenuate the Fourier amplitude of an individual molecule. Because

the Fourier amplitude of the complete ice configuration is the sum of the molecular amplitudes, the same characteristic-function factor determines how much of the later configuration remains correlated with the earlier one.

The correlation coefficient between configurations separated by dose Δq is therefore

$$\boxed{\rho_{\text{ice}}(R, \Delta q) = e^{-\alpha_R \Delta q} = \exp\left(-2\pi^2 \sigma_0^2 R^2 \Delta q\right).} \tag{4.2197}$$

Thus the ice loses memory of its previous configuration exponentially with accumulated dose, and the loss of correlation is faster at higher spatial frequency.

For two frames i and j, each receiving a dose d, the accumulated dose separating their ice configurations is

$$\Delta q = |i - j| d. \tag{4.2198}$$

Equation (4.2195) therefore gives

$$\boxed{\left\langle f_i(\mathbf{R}) f_j^*(\mathbf{R}) \right\rangle = d^2 S_{\text{ice},0}(R) e^{-\alpha_R |i-j| d},} \tag{4.2199}$$

where the factor d^2 appears because the Fourier amplitude collected in each frame is proportional to the electron dose d.

Power spectrum of a dose-fractionated exposure. We can now apply the configuration-correlation result of Eq. (4.2195) to a dose-fractionated exposure.

Suppose that the exposure is divided into M frames, each receiving a dose d, so that the total dose is

$$D = Md. \tag{4.2200}$$

The Fourier transform of the complete summed exposure is

$$f_\Sigma(\mathbf{R}) = \sum_{i=1}^{M} f_i(\mathbf{R}). \tag{4.2201}$$

Its expected structural power is therefore

$$\begin{aligned} S_{\text{ice}}^{(M)}(\mathbf{R}) &= \left\langle |f_\Sigma(\mathbf{R})|^2 \right\rangle \\ &= \sum_{i=1}^{M} \sum_{j=1}^{M} \left\langle f_i(\mathbf{R}) f_j^*(\mathbf{R}) \right\rangle. \end{aligned} \tag{4.2202}$$

The diagonal terms $i = j$ contain the powers of the individual frames. The off-diagonal terms $i \neq j$ contain their mutual correlations. Consequently, the difference between a static specimen and a dynamically rearranging specimen appears entirely through these off-diagonal terms.

Power spectrum of the summed frames.

Substituting Eq. (4.2199) into Eq. (4.2202) gives

$$S_{\text{ice}}^{(M)}(R) = d^2 S_{\text{ice},0}(R) \sum_{i=1}^{M} \sum_{j=1}^{M} e^{-\alpha_R |i-j| d}. \tag{4.2203}$$

Define the correlation between two consecutive frames as

$$\boxed{\eta_R \triangleq e^{-\alpha_R d}.} \tag{4.2204}$$

Then two frames separated by k frame intervals have correlation

$$\eta_R^k. \tag{4.2205}$$

There are M terms with $i = j$, each having $k = 0$. For a separation of one frame interval, $k = 1$, there are $M - 1$ pairs above the diagonal and $M - 1$ corresponding pairs below it. More generally, for a separation k there are $M - k$ pairs on either side of the diagonal. Therefore,

$$\boxed{\begin{aligned} S_{\text{ice}}^{(M)}(R) &= d^2 S_{\text{ice},0}(R) \\ &\quad \times \left[M + 2 \sum_{k=1}^{M-1} (M-k) \eta_R^k \right]. \end{aligned}} \tag{4.2206}$$

The finite sum can be evaluated using

$$\sum_{k=1}^{M-1} (M-k) \eta_R^k = \frac{\eta_R \left[(M-1) - M\eta_R + \eta_R^M \right]}{(1-\eta_R)^2}. \tag{4.2207}$$

Hence,

$$\boxed{S_{\rm ice}^{(M)}(R) = d^2 S_{\rm ice,0}(R) \times \left[M + \frac{2\eta_R \left[(M-1) - M\eta_R + \eta_R^M \right]}{(1-\eta_R)^2} \right].} \tag{4.2208}$$

Equation (4.2206) makes the role of molecular motion particularly transparent. If there is essentially no rearrangement during the exposure,

$$\alpha_R d \to 0, \qquad \eta_R \to 1, \tag{4.2209}$$

all frames contain essentially the same ice structure. Every pair of frames is then fully correlated and

$$S_{\rm ice}^{(M)}(R) \longrightarrow d^2 S_{\rm ice,0}(R) M^2 = D^2 S_{\rm ice,0}(R). \tag{4.2210}$$

The Fourier amplitudes add coherently before their squared modulus is taken.

At the opposite extreme, if configurations in different frames are completely decorrelated,

$$\eta_R \to 0, \tag{4.2211}$$

all off-diagonal terms disappear and

$$\boxed{S_{\rm ice}^{(M)}(R) \longrightarrow M d^2 S_{\rm ice,0}(R).} \tag{4.2212}$$

Only the powers of the individual frames survive. The transition between the M^2 and M limits is precisely what allows the beam-induced rearrangement of amorphous ice to be measured from dose-fractionated movies.

Continuous-motion limit. A particularly transparent result is obtained by allowing the dose fractionation to become arbitrarily fine. Let

$$M \to \infty, \qquad d \to 0, \qquad M d = D. \tag{4.2213}$$

The double sum in Eq. (4.2202) becomes an integral over accumulated dose.

Continuous-dose power spectrum of amorphous ice.

Using dose variables q_1 and q_2,

$$S_{\text{ice}}(R;D) = S_{\text{ice},0}(R) \times \int_0^D \int_0^D e^{-\alpha_R|q_1-q_2|} dq_1\, dq_2. \tag{4.2214}$$

The integrand is symmetric about $q_1 = q_2$, so

$$S_{\text{ice}}(R;D) = 2S_{\text{ice},0}(R) \times \int_0^D \int_0^{q_1} e^{-\alpha_R(q_1-q_2)} dq_2\, dq_1. \tag{4.2215}$$

The inner integral is

$$\int_0^{q_1} e^{-\alpha_R(q_1-q_2)} dq_2 = \frac{1-e^{-\alpha_R q_1}}{\alpha_R}. \tag{4.2216}$$

Therefore,

$$\begin{aligned} S_{\text{ice}}(R;D) &= \frac{2S_{\text{ice},0}(R)}{\alpha_R} \int_0^D \left(1-e^{-\alpha_R q}\right) dq \\ &= \frac{2S_{\text{ice},0}(R)}{\alpha_R^2} \left[\alpha_R D + e^{-\alpha_R D} - 1\right]. \end{aligned} \tag{4.2217}$$

Hence,

$$\boxed{S_{\text{ice}}(R;D) = \frac{2S_{\text{ice},0}(R)}{\alpha_R^2} \left[\alpha_R D + e^{-\alpha_R D} - 1\right].} \tag{4.2218}$$

It is useful to write Eq. (4.2218) relative to the static result of Eq. (4.2172). Define

$$x \triangleq \alpha_R D \tag{4.2219}$$

and

$$\boxed{G_{\text{ice}}(x) \triangleq \frac{2\left(x + e^{-x} - 1\right)}{x^2}.} \tag{4.2220}$$

Then

$$\boxed{S_{\text{ice}}(R;D) = D^2 S_{\text{ice},0}(R) G_{\text{ice}}\left(\alpha_R D\right).} \tag{4.2221}$$

The function G_{ice} describes the loss of power caused by the progressive decorrelation of the ice configuration during the exposure.

Low-dose and high-dose limits. For

$$x = \alpha_R D \ll 1, \tag{4.2222}$$

we have

$$e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \cdots, \tag{4.2223}$$

and therefore

$$G_{\text{ice}}(x) = 1 - \frac{x}{3} + \mathcal{O}(x^2). \tag{4.2224}$$

Consequently,

$$\boxed{S_{\text{ice}}(R; D) \simeq D^2 S_{\text{ice},0}(R), \qquad \alpha_R D \ll 1.} \tag{4.2225}$$

The ice configuration has changed little and Fourier amplitudes acquired at different times add almost coherently.

At the opposite extreme,

$$\alpha_R D \gg 1, \tag{4.2226}$$

Eq. (4.2220) gives

$$G_{\text{ice}}(x) \simeq \frac{2}{x}, \tag{4.2227}$$

and therefore

$$\boxed{S_{\text{ice}}(R; D) \simeq \frac{2D}{\alpha_R} S_{\text{ice},0}(R).} \tag{4.2228}$$

The structural power then grows only linearly with dose rather than quadratically. Increasing the exposure no longer produces coherent reinforcement of the same ice structure because the molecular configuration has decorrelated.

The detected ice spectrum. Combining Eq. (4.2221) with the microscope and detector transfer functions gives the structural contribution of the amorphous ice to the experimentally recorded PSD,

$$\boxed{\begin{aligned} S_{\text{ice,det}}(\mathbf{R}; D) = {} & |\text{MTF}(\mathbf{R})|^2 \, |\text{CTF}(\mathbf{R})|^2 \\ & \times D^2 S_{\text{ice},0}(R) G_{\text{ice}}\left(\alpha_R D\right). \end{aligned}} \tag{4.2229}$$

This expression clarifies the origin of the observed Thon rings from amorphous ice. Four conceptually different factors are present:

$$\boxed{\begin{array}{rcl} S_{\mathrm{ice},0}(R) & : & \text{molecular structure of amorphous ice,} \\ |\,\mathrm{CTF}(\mathbf{R})|^2 & : & \text{electron-optical modulation,} \\ G_{\mathrm{ice}}(\alpha_R D) & : & \text{beam-induced decorrelation during the exposure,} \\ |\,\mathrm{MTF}(\mathbf{R})|^2 & : & \text{detector modulation transfer.} \end{array}} \tag{4.2230}$$

The broad maximum of $S_{\mathrm{ice},0}(R)$ around $R \simeq 1/(3.7\,\text{Å})$ makes the Thon rings particularly visible around this resolution, whereas the R^2 dependence of α_R in Eq. (4.2182) causes beam-induced decorrelation to become progressively stronger at high spatial frequency.

Why summing frame power spectra can reveal the ice more clearly. The model also explains an initially surprising observation of McMullan et al. (2015). For pre-irradiated amorphous carbon, the atomic configuration is essentially unchanged from frame to frame. The Fourier amplitudes are therefore strongly correlated and summing the images before calculating the power spectrum preserves all cross terms in Eq. (4.2202). In the ideal static limit,

$$S_{\mathrm{sum}} \propto M^2 d^2. \tag{4.2231}$$

For amorphous ice, however, the molecular configuration changes during the exposure. The cross terms between sufficiently separated frames are attenuated by the factor

$$e^{-\alpha_R |i-j| d} \tag{4.2232}$$

of Eq. (4.2199). A long summed image therefore loses much of the high-frequency structural power.

By contrast, if the power spectrum is calculated separately from each short frame and the resulting power spectra are then added, no inter-frame phase coherence is required. The structural power present within each short exposure is retained, while averaging many power spectra reduces the statistical fluctuations of the estimated spectrum. This is why McMullan et al. (2015) observed the ice Thon rings much more clearly in the *sum of the power spectra of the individual frames* than in the *power spectrum of the summed image*. The opposite behavior was observed for stabilized amorphous carbon.

Optimum dose for observing the ice Thon rings. The same model predicts that there is an optimum exposure per image for observing the ice signal. Suppose an exposure contains M elementary frames of dose d, and groups of m successive frames are summed before their power spectra are calculated. Each grouped image then has dose

$$D_m = md. \tag{4.2233}$$

For a fixed total dose, the structural contribution in the sum of these power spectra depends on

$$z = \alpha_R md \tag{4.2234}$$

through the function

$$\boxed{g(z) = \frac{z + e^{-z} - 1}{z}.} \tag{4.2235}$$

For small z,

$$g(z) \simeq \frac{z}{2}, \tag{4.2236}$$

whereas

$$g(z) \to 1 \tag{4.2237}$$

for large z. Thus the ice signal initially increases as successive frames are combined, but eventually saturates as the ice configurations within the grouped exposure become decorrelated.

The noise in the corresponding sum of noise-whitened power spectra grows as $\sqrt{m}$ for fixed total dose. Since $z \propto m$, the Thon-ring signal-to-noise ratio is therefore proportional to

$$\frac{g(z)}{\sqrt{z}}. \tag{4.2238}$$

Its maximum occurs at

$$z \simeq 2.149, \tag{4.2239}$$

giving the optimum dose per image

$$\boxed{D_{\text{opt}}(R) = \frac{2.149}{\alpha_R} = \frac{2.149}{2\pi^2 \sigma_0^2 R^2}.} \tag{4.2240}$$

From their experimental dose-fractionated images, McMullan et al. (2015) obtained

$$\sigma_0^2 \simeq 0.37 \, \frac{\text{Å}^2}{\text{e}^-/\text{Å}^2} \tag{4.2241}$$

in each Cartesian direction. At the strong ice feature

$$R = \frac{1}{3.7\,\text{Å}}, \tag{4.2242}$$

Eq. (4.2240) gives an optimum exposure of approximately

$$\boxed{D_{\text{opt}} \simeq 4\,\text{e}^-/\text{Å}^2,} \tag{4.2243}$$

in agreement with their measurements.

Additional attenuation from finite ice thickness. A second effect must be considered when the ice is not very thin. Molecules at different depths are imaged at different effective defocus values. Consequently, even if the molecular configuration were static, the Thon-ring oscillations from different depths would not coincide exactly.

For a simple defocus-dominated CTF,

$$\text{CTF}(R, z) \simeq \sin\left[\chi_0(R) - \pi\lambda R^2 z\right], \tag{4.2244}$$

where z denotes the position within the specimen thickness.

Because the power spectrum contains the squared CTF, averaging over depth does not give the usual amplitude-coherence envelope directly.

Defocus averaging through a uniform ice layer.

For a specimen of thickness t, take

$$-\frac{t}{2} \le z \le \frac{t}{2}. \tag{4.2245}$$

Using

$$\sin^2 x = \frac{1}{2}\left[1 - \cos(2x)\right], \tag{4.2246}$$

the depth-averaged CTF power is

$$\begin{aligned}\left\langle \text{CTF}^2(R)\right\rangle_z = \frac{1}{2} \\ - \frac{1}{2t}\int_{-t/2}^{t/2} \cos\left[2\chi_0(R) - 2\pi\lambda R^2 z\right] dz.\end{aligned} \tag{4.2247}$$

The sine part of the integral vanishes by symmetry, and therefore

$$\boxed{\left\langle \mathrm{CTF}^2(R) \right\rangle_z = \frac{1}{2}\left[1 - \frac{\sin\xi}{\xi}\cos 2\chi_0(R)\right],} \tag{4.2248}$$

where

$$\boxed{\xi = \pi\lambda R^2 t.} \tag{4.2249}$$

Thus the amplitude of the Thon-ring modulation in the *power spectrum* is multiplied by

$$\boxed{E_{\mathrm{thick}}^{\mathrm{PSD}}(R) = \frac{\sin\left(\pi\lambda R^2 t\right)}{\pi\lambda R^2 t}.} \tag{4.2250}$$

This is the factor discussed by McMullan et al. (2015). Notice that it can change sign and therefore is more precisely a modulation-transfer factor than a positive envelope.

The first zero occurs when

$$\lambda R^2 t = 1. \tag{4.2251}$$

Consequently, sufficiently thick ice causes the Thon-ring modulation to vanish and reverse at high spatial frequency. In addition to this geometrical defocus averaging, McMullan et al. (2015) note that thicker ice also increases multiple elastic and inelastic scattering, further reducing the observable high-resolution signal.

4.17 Summary

This chapter has followed a long path, beginning with the motion of a single electron in electromagnetic fields and ending with the frequency-domain model used to describe cryo-EM images. The intermediate developments—classical trajectories, analytical mechanics, wave mechanics, paraxial optics, aberration theory, partial coherence, and power-spectrum analysis—are not independent topics. They form a chain of successive reductions of the same physical problem.

The central logical flow of the chapter can be summarized as

$$
\boxed{
\begin{array}{c}
\text{electromagnetic fields} \\
\Downarrow \\
\text{electron trajectories and action} \\
\Downarrow \\
\text{electron-wave phase} \\
\Downarrow \\
\text{paraxial Gaussian imaging} \\
\Downarrow \\
\text{Fourier plane and aberration phase } \chi(\mathbf{R}) \\
\Downarrow \\
\text{coherent wave transfer} \\
\Downarrow \\
\text{weak-object intensity transfer} \\
\Downarrow \\
\mathrm{CTF}(\mathbf{R}) \\
\Downarrow \\
\text{statistical envelopes} \\
\Downarrow \\
\text{observed image and power spectrum.}
\end{array}
}
\tag{4.2252}
$$

The final objective is not the detailed trajectory through every individual lens of the microscope. Rather, all of that electron-optical physics is ultimately compressed into a small set of functions, principally

$$
P(\mathbf{R}), \qquad \chi(\mathbf{R}), \qquad E_{\mathrm{tot}}(\mathbf{R}), \qquad A, \tag{4.2253}
$$

which determine the contrast transfer function and therefore the relationship between the specimen exit wave and the recorded image.

From the Lorentz force to an electron lens

The chapter began with the simplest magnetic problem: an electron in a uniform axial field,

$$
\mathbf{B} = B_0 \hat{\mathbf{z}}. \tag{4.2254}
$$

The Lorentz force causes the transverse momentum to rotate at the relativistic cyclotron angular frequency

$$\boxed{\omega_c = \frac{eB_0}{\gamma m_0},} \tag{4.2255}$$

as derived in Eq. (4.8). The trajectory is therefore a helix. This simple problem establishes how a magnetic field bends the electron trajectory, but it does not yet describe a lens.

A real round magnetic lens has an axial field $B_0(z)$ that varies with z. Maxwell's equations then require transverse field components. Near the optical axis,

$$\boxed{B_x \simeq -\frac{x}{2}B_0'(z), \qquad B_y \simeq -\frac{y}{2}B_0'(z), \qquad B_z \simeq B_0(z),} \tag{4.2256}$$

according to Eq. (4.76).

Introducing

$$\boxed{\kappa(z) = \frac{eB_0(z)}{p},} \tag{4.2257}$$

as in Eq. (4.159), the apparently coupled transverse motion can be simplified by moving to the Larmor rotating frame. After the rotation has been removed, the transverse coordinate w satisfies

$$\boxed{w'' + \frac{\kappa^2(z)}{4}w = 0,} \tag{4.2258}$$

which is Eq. (4.167).

This equation contains one of the important physical results of the chapter: a rotationally symmetric magnetic field produces, after the Larmor rotation has been separated out, an isotropic focusing action whose strength is proportional to B_0^2.

In the thin-lens approximation, the corresponding focal length is

$$\boxed{\frac{1}{f} = \frac{1}{4}\int_{-\infty}^{\infty} \kappa^2(z)\,dz = \frac{e^2}{4p^2}\int_{-\infty}^{\infty} B_0^2(z)\,dz,} \tag{4.2259}$$

as obtained in Eq. (4.247).

Thus the first part of the chapter establishes the first link in Eq. (4.2252):

$$\boxed{B_0(z) \longrightarrow \text{Lorentz force} \longrightarrow \text{electron trajectories} \longrightarrow \text{focusing.}} \tag{4.2260}$$

From trajectories to wavefronts: the action

Individual trajectories are useful, but image formation is more naturally described in terms of whole families of trajectories. This motivated the transition from geometrical optics to analytical mechanics in Secs. 4.3 and 4.4.

In ordinary geometrical optics, Fermat's principle states that the optical path is stationary,

$$\delta \int n(\mathbf{r})\, ds = 0, \tag{4.2261}$$

as derived in Eq. (4.347). The corresponding scalar eikonal W satisfies

$$\boxed{\nabla W \cdot \nabla W = n^2(\mathbf{r}),} \tag{4.2262}$$

according to Eq. (4.398). Its gradient determines the local ray direction, so rays and wavefronts are dual descriptions of the same geometrical-optics solution.

Analytical mechanics generalizes this idea. Starting from the action

$$S = \int L\, dt, \tag{4.2263}$$

the principle of stationary action gives the Euler–Lagrange equations. The Legendre transformation introduces the Hamiltonian, and Hamilton's principal function satisfies

$$\boxed{\frac{\partial S}{\partial t} + H\left(\mathbf{r}, \nabla S, t\right) = 0,} \tag{4.2264}$$

which is the Hamilton–Jacobi equation of Eq. (4.505).

This formulation is especially important for the remainder of the chapter because the action is simultaneously connected with

$$\boxed{\begin{array}{ccc} \text{trajectory} & \longleftrightarrow & \nabla S \\ \text{wavefront} & \longleftrightarrow & S = \text{constant} \\ \text{quantum phase} & \longleftrightarrow & S/\hbar. \end{array}} \tag{4.2265}$$

The action therefore provides the conceptual bridge between geometrical electron optics and electron-wave optics.

Electrons in electromagnetic fields

For an electron, the natural description involves the electrostatic potential Φ and magnetic vector potential $\mathbf{A}$. The exact relativistic Hamiltonian developed in Sec. 4.6 is

$$\boxed{H(\mathbf{r},\mathbf{P},t) = c\sqrt{(m_0c)^2 + \|\mathbf{P}+e\mathbf{A}\|^2} - m_0c^2 - e\Phi,} \tag{4.2266}$$

which is Eq. (4.718). Here $\mathbf{P}$ is the canonical momentum. The mechanical momentum is

$$\mathbf{p} = \mathbf{P} + e\mathbf{A}. \tag{4.2267}$$

For static fields, Hamilton's characteristic function W obeys

$$\boxed{\|\nabla W + e\mathbf{A}\|^2 = 2m_0e\widehat{\Phi},} \tag{4.2268}$$

as derived in Eq. (4.776). The relativistically corrected acceleration potential $\widehat{\Phi}$ determines the local mechanical momentum through

$$\boxed{p^2 = 2m_0e\widehat{\Phi}.} \tag{4.2269}$$

These equations are particularly important because exactly the same combination

$$\nabla W + e\mathbf{A} \tag{4.2270}$$

will reappear in the wave description.

The discussion of round lenses, cylindrical lenses, dipoles, quadrupoles, hexapoles, and octupoles in Sec. 4.7 gives the physical interpretation of the potentials entering these equations. The detailed hardware of the microscope can be complicated, but for image formation its deterministic effect is ultimately reduced to

$$\boxed{\{\text{round lenses, multipoles, alignments}\} \longrightarrow \{\text{Gaussian imaging}, \chi(\mathbf{R})\}.} \tag{4.2271}$$

Dipoles mainly steer the beam, quadrupoles participate directly in first-order anisotropic focusing, and hexapoles and octupoles provide nonlinear actions that can be arranged to compensate lens aberrations. The multipole symmetry of a physical field should not, however, be identified directly with the symmetry of the residual aberration of the complete microscope.

The electron wave and the semiclassical limit

The trajectory description was then reconnected to wave mechanics in Sec. 4.8. In the scalar, spin-independent relativistic approximation, an electron in electromagnetic potentials satisfies

$$\boxed{\left[\left(j\hbar\frac{\partial}{\partial t}+e\Phi\right)^2-c^2\left(-j\hbar\nabla+e\mathbf{A}\right)^2-m_0^2c^4\right]\Psi=0,} \tag{4.2272}$$

which is Eq. (4.890).

For stationary fields, this reduces to

$$\boxed{\left(-j\hbar\nabla+e\mathbf{A}\right)^2\psi=p^2(\mathbf{r})\psi,} \tag{4.2273}$$

according to Eq. (4.904), or equivalently

$$\boxed{\left[\left(-j\hbar\nabla+e\mathbf{A}\right)^2-2m_0e\widehat{\Phi}\right]\psi=0,} \tag{4.2274}$$

as in Eq. (4.912).

The formal correspondence with the classical Hamilton–Jacobi equation is already evident:

$$\boxed{\nabla W \quad\longleftrightarrow\quad -j\hbar\nabla.} \tag{4.2275}$$

The eikonal approximation in Sec. 4.9 makes this correspondence precise. Writing the wave exactly as

$$\psi(\mathbf{r})=a(\mathbf{r})\exp\left[\frac{j}{\hbar}W_{\mathrm{q}}(\mathbf{r})\right], \tag{4.2276}$$

the exact phase equation is

$$\boxed{\left\|\nabla W_{\mathrm{q}}+e\mathbf{A}\right\|^2-\hbar^2\frac{\nabla^2 a}{a}=p^2.} \tag{4.2277}$$

This is Eq. (4.962).

If the amplitude varies slowly enough that

$$\boxed{\left|\hbar^2\frac{\nabla^2 a}{a}\right|\ll p^2,} \tag{4.2278}$$

as required by Eq. (4.983), the amplitude-curvature term can be neglected and

$$\boxed{\left\|\nabla W_{\mathrm{q}}+e\mathbf{A}\right\|^2\simeq 2m_0e\widehat{\Phi}.} \tag{4.2279}$$

Comparison with Eq. (4.776) gives

$$\boxed{W_{\mathrm{q}} \simeq W.} \tag{4.2280}$$

This is Eq. (4.987).

Thus classical electron trajectories are not a different physical theory. They emerge as the short-wavelength limit of the electron wave.

The paraxial approximation and Gaussian optics

The eikonal approximation establishes when rays are meaningful. The *paraxial* approximation introduces an additional assumption: electron trajectories make only small angles with the optical axis.

For a free propagation distance z, the exact geometrical path length

$$L = \sqrt{z^2 + \|\boldsymbol{\rho}\|^2} \tag{4.2281}$$

becomes, according to Eq. (4.1059),

$$\boxed{L \simeq z + \frac{\|\boldsymbol{\rho}\|^2}{2z}.} \tag{4.2282}$$

Consequently, a spherical electron wave acquires the quadratic phase

$$\boxed{\psi(\boldsymbol{\rho}, z) \propto e^{jkz} \exp\left[j\frac{k}{2z}\|\boldsymbol{\rho}\|^2\right],} \tag{4.2283}$$

as derived in Eq. (4.1061).

This quadratic structure is the mathematical origin of Gaussian optics. In the ray description, every paraxial trajectory can be written as a linear combination of two fundamental rays,

$$x(z) = g_p(z)x_i + h_p(z)x_i', \tag{4.2284}$$

or

$$\boxed{\begin{pmatrix} x(z) \\ x'(z) \end{pmatrix} = \begin{pmatrix} g_p(z) & h_p(z) \\ g_p'(z) & h_p'(z) \end{pmatrix} \begin{pmatrix} x_i \\ x_i' \end{pmatrix},} \tag{4.2285}$$

as shown in Eq. (4.1044).

Two particular planes are essential for image formation.

If

$$\boxed{h_p(z_o) = 0,} \tag{4.2286}$$

then the output position is independent of the input angle:

$$x_o = g_p(z_o)x_i. \tag{4.2287}$$

The plane is therefore an image plane, as established in Eq. (4.1048).

If instead

$$\boxed{g_p(z_F) = 0,} \tag{4.2288}$$

the output position is independent of the initial position and depends only on propagation direction. This is an angular plane, and in wave optics it is a Fourier plane [Eq. (4.1051)].

The back focal plane: where spatial frequency becomes position

The connection between direction and spatial frequency is

$$\theta_x \simeq \lambda R_x, \qquad \theta_y \simeq \lambda R_y. \tag{4.2289}$$

For an ideal thin lens, the back focal plane therefore satisfies

$$\boxed{x_F = \lambda f R_x, \qquad y_F = \lambda f R_y.} \tag{4.2290}$$

This is the central result of Sec. 4.10.7.

Equation (4.2290) gives the back focal plane its special role in image formation:

$$\boxed{\text{propagation direction} \iff \text{spatial frequency} \iff \text{position in the back focal plane.}} \tag{4.2291}$$

The objective aperture can therefore select spatial frequencies directly. More importantly, any phase error introduced by the objective lens at a particular location in the back focal plane is equivalently a spatial-frequency-dependent phase error.

This is the point at which the detailed electron-optical column can be reduced to a Fourier-domain transfer function:

$$\boxed{\widehat{\psi}_{\text{out}}(\mathbf{R}) = H(\mathbf{R})\widehat{\psi}_{\text{in}}(\mathbf{R}),} \tag{4.2292}$$

with

$$\boxed{H_{\text{coh}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi(\mathbf{R})}.} \tag{4.2293}$$

This is the essential content of Eqs. (4.1151) and (4.2069).

Defocus: the first image-forming phase

The simplest non-trivial contribution to χ is defocus. With the cryo-EM convention that

$$\Delta f > 0 \tag{4.2294}$$

corresponds to underfocus, the defocus phase is

$$\boxed{\chi_{\text{def}}(\mathbf{R}) = -\pi\lambda\Delta f R^2, \qquad R = \|\mathbf{R}\|,} \tag{4.2295}$$

as derived in Eq. (4.1191).

The important conceptual result is that defocus does not simply "blur" the wave. Each spatial frequency remains coherent but acquires a different phase:

$$H_{\text{def}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi_{\text{def}}(\mathbf{R})}. \tag{4.2296}$$

A geometrical displacement from the Gaussian image plane has therefore become a quadratic phase modulation in Fourier space.

Why the contrast transfer function is a sine

The coherent wave transfer function Eq. (4.2293) is *not* yet the CTF measured in an image. The objective lens acts on complex wave amplitudes, whereas the detector records intensity.

For a weak pure-phase object,

$$\boxed{\psi_i(\mathbf{r}) \simeq 1 + j\varphi(\mathbf{r}), \qquad |\varphi| \ll 1,} \tag{4.2297}$$

as assumed in Eq. (4.1205). After coherent propagation and linearization of the detected intensity, Eq. (4.1219) gives

$$\boxed{\widehat{\delta I}(\mathbf{R}) = j\left[H(\mathbf{R}) - H^*(-\mathbf{R})\right]\widehat{\varphi}(\mathbf{R}).} \tag{4.2298}$$

For a real centrosymmetric pupil and an even aberration phase, Eq. (4.2298) reduces to

$$\boxed{\widehat{\delta I}(\mathbf{R}) = 2P(\mathbf{R})\sin\chi(\mathbf{R})\widehat{\varphi}(\mathbf{R}),} \tag{4.2299}$$

as derived in Eq. (4.1224).

After absorbing the conventional factor of two into the contrast normalization,

$$\boxed{\text{CTF}(\mathbf{R}) = P(\mathbf{R})\sin\chi(\mathbf{R}).} \tag{4.2300}$$

This is one of the chapter's central results. The sine does not describe the action of the lens on the electron wave. The lens contributes the complex phase factor

$$e^{-j\chi}. \tag{4.2301}$$

The sine appears only after the strong reference wave and the weak scattered wave interfere, and the detector forms the intensity.

Aberrations are errors in accumulated action

The section on aberrations, Sec. 4.11, generalized the defocus phase. The fundamental interpretation is that an aberration corresponds to an excess action, or equivalently to a wavefront error.

If W denotes a wavefront error expressed as a length, the associated phase error is

$$\boxed{\chi = \frac{2\pi}{\lambda} W,} \tag{4.2302}$$

as stated in Eq. (4.1623). Equivalently, the aberration phase can be derived directly from the excess action through Eq. (4.1257).

The gradient of the aberration function determines the corresponding ray error. Thus geometrical aberrations and wave aberrations are not different defects: they are different representations of the same departure from Gaussian imaging.

For the aberrations most important in conventional cryo-EM, the symmetric phase contains the quadratic defocus contribution and the quartic spherical-aberration contribution,

$$\boxed{\chi_{\text{sym}}(R,\phi) = -\pi\lambda\Delta f(\phi)R^2 + \frac{\pi}{2}C_s\lambda^3 R^4 + \chi_{\text{sym}}^{\text{higher}}(R,\phi).} \tag{4.2303}$$

This is Eq. (4.2121).

Twofold astigmatism is represented through the directional defocus

$$\boxed{\Delta f(\phi) = \Delta f_M \cos^2(\phi - \phi_{\text{ast}}) + \Delta f_m \sin^2(\phi - \phi_{\text{ast}}),} \tag{4.2304}$$

from Eq. (4.1406), or equivalently

$$\boxed{\Delta f(\phi) = \Delta f_{\text{av}} + \Delta f_{\text{ast}} \cos 2(\phi - h).} \tag{4.2305}$$

Thus astigmatism is simply anisotropic quadratic phase curvature.

Spherical aberration contributes

$$\boxed{\chi_{\text{sph}}(R) = \frac{\pi}{2} C_s \lambda^3 R^4,} \tag{4.2306}$$

according to Eq. (4.1460). Its R^4 dependence explains why it becomes increasingly important at high resolution.

Symmetric and antisymmetric aberrations

One of the most useful classifications developed in this chapter is the decomposition

$$\boxed{\chi(\mathbf{R}) = \chi_{\text{sym}}(\mathbf{R}) + \chi_{\text{asym}}(\mathbf{R}),} \tag{4.2307}$$

with

$$\boxed{\begin{aligned} \chi_{\text{sym}}(-\mathbf{R}) &= \chi_{\text{sym}}(\mathbf{R}), \\ \chi_{\text{asym}}(-\mathbf{R}) &= -\chi_{\text{asym}}(\mathbf{R}). \end{aligned}} \tag{4.2308}$$

This is the parity decomposition derived in Eqs. (4.1570)–(4.1577).

Defocus, twofold astigmatism, and spherical aberration are symmetric. Coma and trefoil are antisymmetric.

For a weak phase object, the resulting coherent intensity transfer is

$$\boxed{\text{CTF}_{\text{coh}}(\mathbf{R}) = P(\mathbf{R}) \sin \chi_{\text{sym}}(\mathbf{R}) e^{-j\chi_{\text{asym}}(\mathbf{R})},} \tag{4.2309}$$

as obtained in Eq. (4.1588).

This formula is conceptually important. Symmetric aberrations determine the oscillating CTF amplitude and therefore its zeroes, whereas antisymmetric aberrations generate a Fourier phase shift of unit modulus.

Consequently,

$$\left| e^{-j\chi_{\text{asym}}} \right| = 1, \tag{4.2310}$$

so antisymmetric aberrations may seriously corrupt an image while remaining much less visible in its ordinary power spectrum.

Zernike polynomials provide a general aberration basis

The named aberrations do not exhaust all possible departures from the ideal wavefront. A convenient general representation over the circular objective pupil is provided by the real Zernike polynomials:

$$\boxed{\chi(\rho, \varphi) = \sum_{n=0}^{\infty} \sum_{\substack{m=-n \\ n-|m| \text{ even}}}^{n} c_n^m Z_n^m(\rho, \varphi),} \tag{4.2311}$$

as introduced in Eq. (4.1631).

The Zernike expansion is a *representation* of the final pupil aberration, not a catalogue of physical lens elements. This distinction is important in view of Sec. 4.7: the multipole order of an electrode or pole-piece configuration should not be identified directly with the Zernike order of the residual wavefront error produced by the complete optical system.

For image formation, the utility of the Zernike description is that all deterministic aberrations, however they arise physically, can ultimately be absorbed into the single function

$$\chi(\mathbf{R}). \tag{4.2312}$$

Amplitude contrast

The pure-phase model is not exact because the specimen can modify both the amplitude and phase of the exit wave. In Sec. 4.12, this was represented by a normalized amplitude-contrast fraction A.

The weak-object CTF becomes

$$\boxed{\mathrm{CTF}_A(\mathbf{R}) = P(\mathbf{R})\left[\sqrt{1-A^2}\,\sin\chi(\mathbf{R}) + A\cos\chi(\mathbf{R})\right],} \tag{4.2313}$$

according to Eq. (4.1796).

Defining

$$\boxed{\phi_A = \arcsin A,} \tag{4.2314}$$

the same expression can be written

$$\boxed{\mathrm{CTF}_A(\mathbf{R}) = P(\mathbf{R})\sin\left[\chi(\mathbf{R}) + \phi_A\right],} \tag{4.2315}$$

as shown in Eq. (4.1801).

This compact representation should not obscure the physical distinction: ϕ_A is an effective phase shift that summarizes the mixing of amplitude and phase contrast. It is *not* a lens aberration.

Chromatic aberration and statistical averaging

All aberrations considered so far assign a definite phase to each spatial frequency for a fixed state of the microscope. Chromatic aberration introduces another idea: the electron energy is distributed, so the effective focus is also distributed.

To first order,

$$\boxed{\delta f \simeq C_c \frac{\delta E}{E_0},} \tag{4.2316}$$

according to Eq. (4.1745). Since defocus produces the phase

$$\chi_{\text{def}} = -\pi\lambda\Delta f R^2, \tag{4.2317}$$

a focus fluctuation produces

$$\boxed{\delta\chi = -\pi\lambda\delta f R^2.} \tag{4.2318}$$

For a Gaussian distribution of focus with standard deviation σ_f, averaging these coherent realizations gives

$$\boxed{E_t(R) = \exp\left[-\frac{1}{2}\left(\pi\lambda\sigma_f\right)^2 R^4\right],} \tag{4.2319}$$

as derived in Eq. (4.1749) and repeated in Eq. (4.1836).

Chromatic aberration is therefore an optical property of the lens, but the temporal-coherence envelope is a *statistical* consequence of combining that aberration with an energy or focus distribution.

The general origin of CTF envelopes

The preceding example was generalized in Sec. 4.13. Suppose that a fluctuating variable ξ changes the coherent phase by $\delta\chi_\xi(\mathbf{R})$. The corresponding envelope is

$$\boxed{E_\xi(\mathbf{R}) = \left\langle e^{-j\delta\chi_\xi(\mathbf{R})}\right\rangle_\xi.} \tag{4.2320}$$

This is Eq. (4.1814).

Equation (4.2320) is one of the most general results of the chapter. An envelope is the characteristic function of the distribution of a random phase perturbation.

For a Gaussian variable, this immediately gives a Gaussian damping factor. Temporal coherence, finite source size, and Gaussian specimen motion are different physical effects, but they all follow from this same averaging principle.

For an isotropic Gaussian angular source, Eq. (4.1862) gives

$$\boxed{E_s(\mathbf{R}) = \exp\left[-\frac{\alpha_s^2}{4\lambda^2}\left\|\boldsymbol{\nabla}_{\mathbf{R}}\chi(\mathbf{R})\right\|^2\right].} \tag{4.2321}$$

Spatial-coherence damping is therefore governed by the gradient of the aberration phase.

Specimen motion gives the same characteristic-function mathematics. For a uniform in-plane drift $\Delta\mathbf{r}_d$,

$$\boxed{E_{\|}(\mathbf{R}) = \operatorname{sinc}\left(\mathbf{R}\cdot\Delta\mathbf{r}_d\right),} \tag{4.2322}$$

from Eq. (4.1880), whereas isotropic Gaussian jitter gives

$$\boxed{E_G(R) = \exp\left[-2\pi^2\sigma_u^2 R^2\right],} \tag{4.2323}$$

according to Eq. (4.1905).

If the underlying fluctuations are statistically independent, their envelopes multiply:

$$\boxed{E_{\text{tot}} = E_t E_s E_{\text{motion}} \cdots,} \tag{4.2324}$$

as proved in Eq. (4.1962).

For practical work, several attenuation mechanisms are often summarized phenomenologically by

$$\boxed{E_{\text{eff}}(R) = \exp\left(-\frac{B}{4}R^2\right),} \tag{4.2325}$$

as introduced in Eq. (4.1963). This B-factor representation is convenient but is not universal: chromatic damping is proportional to e^{-cR^4}, uniform drift produces a sinc, axial vibration may produce a Bessel function, and Brownian motion has yet another functional form.

The complete contrast transfer function

All the ingredients of the image-formation model were finally assembled in Sec. 4.15.

Under the weak-object approximation, for a real centrosymmetric pupil, a constant normalized amplitude-contrast coefficient A, and envelope mechanisms that can be represented multiplicatively, the complete CTF is

$$\boxed{\begin{aligned}\mathrm{CTF}(\mathbf{R}) = {} & E_{\text{tot}}(\mathbf{R})P(\mathbf{R})e^{-j\chi_{\text{asym}}(\mathbf{R})} \\ & \times\left[\sqrt{1-A^2}\,\sin\chi_{\text{sym}}(\mathbf{R}) + A\cos\chi_{\text{sym}}(\mathbf{R})\right].\end{aligned}} \tag{4.2326}$$

This is Eq. (4.2117).

Equivalently,

$$\boxed{\mathrm{CTF}(\mathbf{R}) = E_{\mathrm{tot}}(\mathbf{R})P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \phi_A\right].} \tag{4.2327}$$

This is Eq. (4.2118).

For the conventional cryo-EM model,

$$\boxed{\begin{aligned}\mathrm{CTF}(R,\phi) = E_{\mathrm{tot}}(R,\phi)P(R,\phi)e^{-j\chi_{\mathrm{asym}}(R,\phi)} \\ \times \sin\Big[-\pi\lambda\Delta f(\phi)R^2 + \frac{\pi}{2}C_s\lambda^3R^4 \\ +\chi_{\mathrm{sym}}^{\mathrm{higher}}(R,\phi) + \phi_A\Big],\end{aligned}} \tag{4.2328}$$

which is Eq. (4.2122).

This equation is the principal result toward which most of the chapter has been directed. Each factor has a distinct origin:

$$\boxed{\begin{array}{ccl} P(\mathbf{R}) & : & \text{objective aperture,} \\ \chi_{\mathrm{sym}} & : & \text{deterministic symmetric aberrations,} \\ \chi_{\mathrm{asym}} & : & \text{deterministic antisymmetric aberrations,} \\ \phi_A & : & \text{amplitude–phase contrast mixing,} \\ E_{\mathrm{tot}} & : & \text{statistical averaging.} \end{array}} \tag{4.2329}$$

It is particularly important not to merge these concepts physically. Aberrations are deterministic phase errors. The aperture is coherent amplitude selection. Amplitude contrast originates in the specimen. Envelopes arise from averaging different realizations. They appear multiplicatively in the final equation only because the approximations used in deriving the weak-object model permit this useful factorization.

From the CTF to the image power spectrum

The final part of the chapter connected the image-formation model with what is actually inspected experimentally when estimating the CTF: the micrograph power spectrum.

If a random image component x is transferred linearly according to

$$\widehat{y}(\mathbf{R}) = H(\mathbf{R})\widehat{x}(\mathbf{R}), \tag{4.2330}$$

then its power spectral density satisfies

$$\boxed{S_y(\mathbf{R}) = |H(\mathbf{R})|^2 S_x(\mathbf{R}),} \tag{4.2331}$$

as proved in Eq. (4.2133). Hence the CTF appears in a power spectrum as

$$\boxed{|\,\mathrm{CTF}(\mathbf{R})|^2,} \tag{4.2332}$$

not as CTF^2.

This distinction matters because the complete CTF may be complex. Using Eq. (4.2327),

$$\begin{aligned} |\,\mathrm{CTF}(\mathbf{R})|^2 = E_{\mathrm{tot}}^2(\mathbf{R})P^2(\mathbf{R}) \\ \times \sin^2\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \phi_A\right]. \end{aligned} \tag{4.2333}$$

The antisymmetric phase has disappeared because its modulus is one. This is the result derived explicitly in Eq. (4.2136).

The second statistical result needed for the PSD is that independent, or more generally zero-cross-covariance, image components add in power:

$$\boxed{S_{a+b}(\mathbf{R}) = S_a(\mathbf{R}) + S_b(\mathbf{R}),} \tag{4.2334}$$

apart from the zero-frequency contribution of non-zero means. This was derived from the autocovariance and the Wiener–Khinchin theorem in Sec. 4.16.2.

These results lead to the practical background-plus-CTF description

$$\boxed{S_{\mathrm{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})|\,\mathrm{CTF}(\mathbf{R})|^2,} \tag{4.2335}$$

which is Eq. (4.2162).

Here $A(\mathbf{R})$ represents the slowly varying structural power that is modulated by the CTF, including contributions from the particle, vitreous ice, support, and other elastic specimen structure. The function $B(\mathbf{R})$ is operational: it represents the remaining slowly varying power not described by the chosen CTF-modulated structural model.

Near an ideal CTF zero,

$$|\,\mathrm{CTF}\,|^2 \simeq 0, \tag{4.2336}$$

so

$$S_{\mathrm{obs}} \simeq B. \tag{4.2337}$$

Between the zeroes, structural power rises above this background. This produces the oscillating Thon-ring pattern used for CTF estimation.

Why vitreous ice itself reveals the CTF

An important application of the PSD framework is the structural spectrum of vitreous ice, developed in Sec. 4.16.4. The ice is not merely featureless noise: its molecular structure supplies broadband elastic scattering that is modulated by the microscope CTF.

For a static ice configuration exposed to total dose D, the structural power would scale as

$$D^2 S_{\mathrm{ice},0}(R). \tag{4.2338}$$

Beam-induced molecular motion progressively destroys the correlation between configurations. For the Gaussian Markov model developed in Sec. 4.16.4, the Fourier correlation after an accumulated dose difference Δq is

$$\boxed{\rho_{\mathrm{ice}}(R, \Delta q) = e^{-\alpha_R \Delta q}, \qquad \alpha_R = 2\pi^2 \sigma_0^2 R^2.} \tag{4.2339}$$

In the continuous-dose limit, the resulting structural ice PSD is

$$\boxed{S_{\mathrm{ice}}(R; D) = D^2 S_{\mathrm{ice},0}(R) G_{\mathrm{ice}}\left(\alpha_R D\right),} \tag{4.2340}$$

where, from Eq. (4.2220),

$$\boxed{G_{\mathrm{ice}}(x) = \frac{2(x + e^{-x} - 1)}{x^2}.} \tag{4.2341}$$

After microscope and detector transfer,

$$\boxed{\begin{aligned} S_{\mathrm{ice,det}}(R; D) = {} & |\,\mathrm{MTF}(R)|^2 |\,\mathrm{CTF}(R)|^2 \\ & \times D^2 S_{\mathrm{ice},0}(R) G_{\mathrm{ice}}\left(\alpha_R D\right), \end{aligned}} \tag{4.2342}$$

which is Eq. (4.2229).

This example closes the logical circle of the chapter. A microscopic structural fluctuation in the specimen is converted into an exit-wave component, transferred by the electron optics, attenuated by physical and statistical effects, recorded by the detector, and finally observed as oscillating power in the micrograph PSD.

The minimal image-formation model to retain

The detailed derivations of this chapter are needed to understand where the image model comes from and where its approximations enter. Once

those derivations have been established, however, the core model can be remembered through a small number of equations.

First, the objective acts coherently on the electron wave:

$$\boxed{H_{\mathrm{coh}}(\mathbf{R}) = P(\mathbf{R})e^{-j\chi(\mathbf{R})}.} \tag{4.2343}$$

Second, the dominant symmetric phase of a conventional cryo-EM objective is

$$\boxed{\chi_{\mathrm{sym}}(R,\phi) \simeq -\pi\lambda\Delta f(\phi)R^2 + \frac{\pi}{2}C_s\lambda^3R^4 + \cdots.} \tag{4.2344}$$

Third, weak-object detection converts the coherent phase modulation into an intensity CTF:

$$\boxed{\mathrm{CTF}(\mathbf{R}) = E_{\mathrm{tot}}(\mathbf{R})P(\mathbf{R})e^{-j\chi_{\mathrm{asym}}(\mathbf{R})}\sin\left[\chi_{\mathrm{sym}}(\mathbf{R}) + \phi_A\right].} \tag{4.2345}$$

Fourth, statistical phase fluctuations generate envelopes through

$$\boxed{E(\mathbf{R}) = \left\langle e^{-j\delta\chi(\mathbf{R})}\right\rangle.} \tag{4.2346}$$

Finally, the experimental power spectrum sees the squared modulus:

$$\boxed{S_{\mathrm{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})|\,\mathrm{CTF}(\mathbf{R})|^2.} \tag{4.2347}$$

These five equations summarize the image-forming content of the chapter. They also identify the successive approximations that have been made:

$$\boxed{\begin{array}{ccl} \text{short wavelength} & \Rightarrow & \text{eikonal/geometrical optics,} \\ \text{small propagation angles} & \Rightarrow & \text{paraxial Gaussian optics,} \\ \text{small specimen modulation} & \Rightarrow & \text{weak-object linearization,} \\ \text{shift-invariant local imaging} & \Rightarrow & \text{multiplicative Fourier transfer,} \\ \text{appropriate statistical factorization} & \Rightarrow & \text{multiplicative envelopes.} \end{array}} \tag{4.2348}$$

The contrast transfer function is therefore not a fundamental law introduced independently of the preceding physics. It is the final reduced description obtained after progressively simplifying the electron's electromagnetic dynamics, wave propagation, lens action, aberrations, specimen contrast, and statistical averaging.

Final perspective

The many physical elements of an electron microscope do not appear individually in the final cryo-EM image model. The detailed magnetic and electrostatic fields determine trajectories; the trajectories determine the accumulated action; the action determines the electron-wave phase; the paraxial approximation reduces the first-order system to Gaussian imaging; the objective back focal plane maps propagation angle to spatial frequency; deviations from Gaussian imaging become the aberration phase $\chi(\mathbf{R})$; weak-object interference converts this phase into intensity contrast; and averaging over different experimental realizations produces the CTF envelopes.

The entire chapter can therefore be compressed conceptually into

$$\boxed{\begin{aligned} \{\Phi, \mathbf{A}\} &\longrightarrow \text{trajectories} \\ &\longrightarrow W \\ &\longrightarrow \psi \\ &\longrightarrow H_{\text{coh}} = Pe^{-j\chi} \\ &\longrightarrow \text{CTF} \\ &\longrightarrow S_{\text{obs}}. \end{aligned}} \tag{4.2349}$$

The principal practical result is Eq. (4.2117), or its compact form Eq. (4.2118). Everything preceding it explains the physical origin of its factors; everything following it explains how those factors manifest themselves in an experimental cryo-EM micrograph.

Chapter 5

Interaction with the specimen

5.1 From the microscope to the specimen

In Chap. 4, we followed the electron from the source through the electromagnetic fields of the microscope and described how the objective lens transforms the wave emerging from the specimen. In particular, the coherent action of the objective was represented by the transfer function

$$H(\mathbf{R}) = P(\mathbf{R}) \exp[-j\chi(\mathbf{R})] \tag{5.1}$$

[Eq. (4.1237)], and the subsequent sections developed the contrast transfer function, amplitude contrast, and the different mechanisms that attenuate the transferred signal.

There was, however, an essential quantity that we treated as already known: the electron wave *leaving the specimen*. We now turn to the physics that creates this wave.

Let the entrance and exit surfaces of the specimen be at $z = z_{\text{in}}$ and $z = z_{\text{out}}$, respectively. We define

$$\psi_{\text{in}}(x, y) \triangleq \psi(x, y, z_{\text{in}}), \qquad \psi_{\text{exit}}(x, y) \triangleq \psi(x, y, z_{\text{out}}). \tag{5.2}$$

The central task of this chapter is to understand the transformation

$$\boxed{\psi_{\text{in}} \xrightarrow{\text{three-dimensional specimen}} \psi_{\text{exit}}.} \tag{5.3}$$

We may represent this transformation abstractly by a specimen operator $\mathcal{S}$,

$$\boxed{\psi_{\text{exit}} = \mathcal{S}\, \psi_{\text{in}}.} \tag{5.4}$$

Much of this chapter can be understood as progressively constructing more realistic approximations to $\mathcal{S}$. For a sufficiently thin specimen it will reduce to multiplication by a two-dimensional transmission function. For a three-dimensional specimen it will become a sequence of interactions and free-space propagations, leading to the multislice description. Once inelastic scattering and radiation damage are included, even the apparently simple notation of Eq. (5.4) will require some qualification.

The important conceptual separation is

$$\boxed{\begin{array}{lcl} \text{electron source} \longrightarrow \psi_{\text{in}} & \xrightarrow{\mathcal{S}} & \psi_{\text{exit}} \\ & \text{specimen interaction} & \downarrow \\ & & \text{electron optics} \longrightarrow \text{detected image.} \end{array}} \tag{5.5}$$

The contrast transfer function developed in Sec. 4.15 therefore acts on information already encoded in the exit wave. The microscope does not image the atomic coordinates directly: the specimen must first convert its three-dimensional structure into an electron wave.

Three physically different consequences of electron–matter interaction

When a high-energy electron traverses a biological specimen, several physical processes are possible. For image formation it is useful to separate them first into *elastic* and *inelastic* interactions.

In an *elastic scattering* event, the internal state of the specimen is unchanged. To an excellent approximation, the energy of the imaging electron therefore remains unchanged,

$$\boxed{E_f = E_i \qquad \text{(elastic scattering).}} \tag{5.6}$$

The momentum of the electron may nevertheless change direction. In wave language, the specimen changes the amplitude and, particularly importantly for cryo-EM, the phase and propagation direction of the electron wave.

Because the specimen is left in the same internal state, different elastic scattering paths can remain mutually coherent and interfere. It is this coherent elastic scattering that carries most of the high-resolution structural information used in single-particle analysis. A central objective of the first part of this chapter will therefore be to calculate the elastic component of ψ_{exit} from the three-dimensional electrostatic potential of the specimen.

In an *inelastic scattering* event, by contrast, part of the electron energy is transferred to the specimen. If the transferred energy is denoted by $\hbar\omega$,

$$\boxed{E_f = E_i - \hbar\omega, \qquad \hbar\omega > 0 \qquad \text{(inelastic scattering)}.} \tag{5.7}$$

The specimen is then left in a different internal state. In a biological specimen, the energy transfer may produce electronic excitation, ionization, collective electronic excitation, molecular vibration, or other excitations of the material.

This has two distinct consequences for imaging. First, the outgoing electron has a different energy, and therefore a slightly different wavelength and different electron-optical behavior. In particular, Sec. 4.11.9 showed that energy differences are converted by chromatic aberration into focus differences. Second, the specimen has acquired information about the interaction through its changed internal state. When those internal states are not measured, different inelastic channels cannot in general be treated as one coherent electron wave interfering with the elastic channel.

This point will require a small amount of quantum mechanics later in the chapter. For the moment it is sufficient to regard inelastic scattering as producing a collection of possible outgoing channels rather than a single elastic exit wave.

Finally, some interactions do more than modify the electron that has just traversed the specimen. They modify the specimen itself. Excitation and ionization can initiate bond breaking, radical formation, charge redistribution, molecular motion, and eventually changes in chemical composition and structure. These processes constitute *radiation damage*.

Radiation damage is therefore conceptually different from the elastic–inelastic distinction. It is not simply a third possible outgoing electron channel. Instead, it means that the object being imaged is no longer static. If D denotes the accumulated electron dose, the electrostatic potential should more generally be written as

$$\boxed{V(\mathbf{r}) \quad \longrightarrow \quad V(\mathbf{r}; D).} \tag{5.8}$$

An electron arriving later in the exposure may therefore interact with a different specimen from an electron arriving earlier.

This distinction is particularly important in single-particle cryo-EM. A movie is not merely a repeated observation of a fixed object disturbed by mechanical motion and shot noise. During the exposure, the underlying molecular structure itself progressively changes. The dose-dependent attenuation of high-resolution information encountered experimentally ultimately has its microscopic origin in these electron–matter interactions.

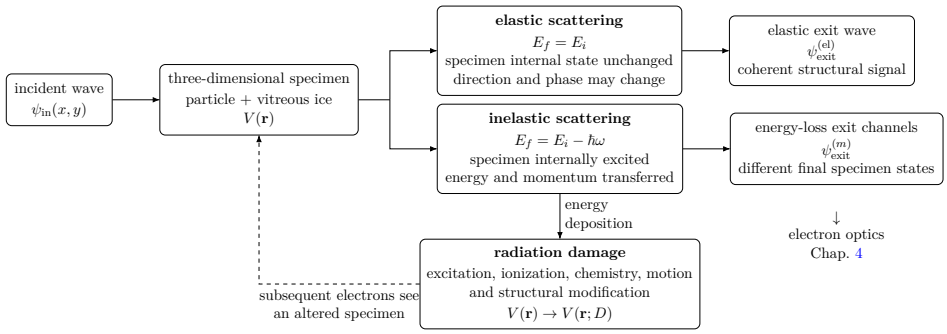


Figure 5.1: Conceptual map of electron–specimen interaction. The incident electron wave enters a three-dimensional cryogenic specimen and can undergo elastic or inelastic scattering. Elastic scattering leaves the internal state of the specimen unchanged and produces the coherent exit wave that carries the principal structural information used in single-particle cryo-EM. Inelastic scattering transfers energy to the specimen and produces different energy-loss channels associated with different final states of the specimen. Energy deposition can in turn initiate radiation damage, so that subsequent electrons interact with a dose-dependent specimen $V(\mathbf{r}; D)$. The diagram emphasizes the dominant connection between inelastic energy deposition and radiation damage in biological specimens; energetic elastic collisions can also transfer energy to atomic nuclei and will be discussed separately when relevant.

What the interaction model must ultimately provide

Figure 5.1 gives the physical map for the rest of this chapter. It also shows that there is no single level of specimen description appropriate to every question.

For the coherent elastic image-formation problem, the central quantity is the elastic exit wave,

$$\psi_{\text{exit}}^{(\text{el})}(x, y). \tag{5.9}$$

For inelastic scattering, we must additionally know how the outgoing electrons are distributed among different energy-loss and specimen excitation channels. For radiation damage, we must understand how those interactions progressively modify the object itself.

Schematically, the specimen model must therefore provide

$$\boxed{\begin{array}{rcl} \text{elastic scattering} & \longrightarrow & \psi_{\text{exit}}^{(\text{el})}, \\ \text{inelastic scattering} & \longrightarrow & \text{probabilities of the different energy-loss channels}, \\ \text{radiation damage} & \longrightarrow & V(\mathbf{r}; D). \end{array}} \tag{5.10}$$

At first we shall concentrate on the elastic branch. This allows us to replace the microscopic many-particle specimen by an electrostatic potential $V(\mathbf{r})$ and to ask how a fast electron propagates through that potential. This route leads naturally from atomic charge distributions to atomic potentials, from atomic potentials to electron scattering factors, and finally from the three-dimensional specimen potential to the multislice description.

Only after this elastic framework is established will we return to the inelastic branch. At that point a potential acting on a single electron will no longer be sufficient: we will need a minimal quantum-mechanical description in which the specimen itself can change state.

The logic of the chapter is therefore

$$\boxed{\begin{array}{c} \text{atomic charge} \longrightarrow V(\mathbf{r}) \longrightarrow \text{elastic scattering} \longrightarrow \psi_{\text{exit}} \\ \longrightarrow \text{multislice and three-dimensional propagation}, \\ \text{followed by} \\ \text{inelastic transitions} \longrightarrow \text{energy loss and decoherence} \longrightarrow \text{radiation damage}. \end{array}} \tag{5.11}$$

The first step is therefore to determine the electrostatic potential generated by matter and, in particular, by the atoms that constitute a biological specimen.

5.2 The specimen as a quantum scatterer

The quantum-mechanical foundations of the electron–specimen interaction were developed in Appendix A. We shall not repeat those derivations here. Instead, we take them as the starting point for constructing the physical models of matter used in electron microscopy.

Let $|\Xi_0\rangle_{\rm s}$ denote the initial internal state of the specimen and let $\hat{H}_{\rm int}$ denote the interaction Hamiltonian coupling the imaging electron to the specimen. As derived in Sec. A.12, if the specimen is finally found in the state $|\Xi_m\rangle_{\rm s}$, the corresponding effective interaction acting on the imaging electron is

$$\boxed{\hat{U}_{m0} = {}_{\rm s}\langle\Xi_m|\hat{H}_{\rm int}|\Xi_0\rangle_{\rm s}.} \tag{5.12}$$

The bra and ket in Eq. (5.12) contract only the specimen degrees of freedom. Consequently, $\hat{U}_{m0}$ is still an operator acting on the imaging electron.

Two cases will organize much of this chapter:

$$\boxed{\begin{array}{lcl} m = 0 & : & \text{elastic channel,} \\ m \neq 0 & : & \text{inelastic channel.} \end{array}} \tag{5.13}$$

The physical meaning of this distinction, including energy transfer, specimen recoil, matrix elements, and the role of orthogonal unobserved specimen states, was developed in Secs. A.9.2, A.12, and A.13.

Our purpose here is different. We now ask what Eq. (5.12) becomes for matter whose dominant interaction with the imaging electron is electromagnetic. This will lead naturally from the microscopic Coulomb interaction to the electrostatic specimen potential used throughout electron microscopy.

5.2.1 The charge-density description of the specimen

Consider the charge contained in the specimen. In a microscopic description, the nuclei have charges $+Z_A e$ and the bound electrons have charge $-e$. It is convenient to collect all of these charges into a single charge-density operator,

$$\boxed{\hat{\rho}_{\rm s}(\mathbf{r}') = e\sum_A Z_A \delta\left(\mathbf{r}' - \hat{\mathbf{R}}_A\right) - e\sum_j \delta\left(\mathbf{r}' - \hat{\mathbf{r}}_j\right).} \tag{5.14}$$

Here $\hat{\mathbf{R}}_A$ and $\hat{\mathbf{r}}_j$ denote the position operators of the specimen nuclei and bound electrons, respectively.

For most elastic image-formation calculations, nuclear motion during the passage of the fast electron is neglected and the nuclear positions are subsequently replaced by fixed coordinates $\mathbf{R}_A$. We retain the more general notation for the moment because it also naturally describes inelastic channels.

If the imaging electron is at position $\mathbf{r}$, its charge is $-e$. In Appendix A, the microscopic electron–specimen Coulomb interaction was written explicitly as a sum over nuclei and bound electrons; see Eq. (A.459). Using the charge-density operator of Eq. (5.14), the same interaction can be written much more compactly.

From pairwise Coulomb interactions to the charge-density form.

Starting from Eq. (5.14),

$$\hat{\rho}_{\mathrm{s}}(\mathbf{r}') = e\sum_A Z_A \delta\left(\mathbf{r}' - \hat{\mathbf{R}}_A\right) - e\sum_j \delta\left(\mathbf{r}' - \hat{\mathbf{r}}_j\right), \tag{5.15}$$

consider

$$-\frac{e}{4\pi\varepsilon_0}\int \frac{\hat{\rho}_{\mathrm{s}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'. \tag{5.16}$$

Substituting the charge-density operator gives

$$\begin{aligned} &-\frac{e}{4\pi\varepsilon_0}\int \frac{e\sum_A Z_A\delta\left(\mathbf{r}' - \hat{\mathbf{R}}_A\right) - e\sum_j \delta\left(\mathbf{r}' - \hat{\mathbf{r}}_j\right)}{|\mathbf{r}-\mathbf{r}'|}\, d^3r' \\ &= -\sum_A \frac{Z_A e^2}{4\pi\varepsilon_0|\mathbf{r}-\hat{\mathbf{R}}_A|} + \sum_j \frac{e^2}{4\pi\varepsilon_0|\mathbf{r}-\hat{\mathbf{r}}_j|}, \end{aligned} \tag{5.17}$$

where the Dirac delta has selected the position of each charge.

This is precisely the microscopic Coulomb interaction derived in Eq. (A.459). Therefore,

$$\boxed{\hat{H}_{\mathrm{int}}(\mathbf{r}) = -\frac{e}{4\pi\varepsilon_0}\int \frac{\hat{\rho}_{\mathrm{s}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'.} \tag{5.18}$$

Equation (5.18) is therefore not a new physical assumption. It is simply the pairwise Coulomb interaction of Eq. (A.459) written in terms of the total

charge-density operator of the specimen.

Positive nuclear charge gives an attractive contribution to the imaging-electron potential energy, whereas negative bound-electron charge gives a repulsive contribution.

5.2.2 Elastic and inelastic charge-density channels

The quantum channel decomposition of Eq. (5.12) can now be applied directly to the charge-density operator. Define

$$\boxed{\rho_{m0}(\mathbf{r}) \equiv {}_{\mathrm{s}}\langle \Xi_m | \hat{\rho}_{\mathrm{s}}(\mathbf{r}) | \Xi_0 \rangle_{\mathrm{s}}.} \tag{5.19}$$

The function $\rho_{m0}(\mathbf{r})$ describes the charge-density component through which the specimen couples the initial state $|\Xi_0\rangle_{\mathrm{s}}$ to the final state $|\Xi_m\rangle_{\mathrm{s}}$.

Substituting Eq. (5.18) into Eq. (5.12) gives

$$\boxed{U_{m0}(\mathbf{r}) = -\frac{e}{4\pi\varepsilon_0} \int \frac{\rho_{m0}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'.} \tag{5.20}$$

It is convenient to define the corresponding channel potential

$$\boxed{V_{m0}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho_{m0}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r',} \tag{5.21}$$

so that

$$\boxed{U_{m0}(\mathbf{r}) = -eV_{m0}(\mathbf{r}).} \tag{5.22}$$

The elastic channel

For $m = 0$,

$$\boxed{\rho_{00}(\mathbf{r}) = {}_{\mathrm{s}}\langle \Xi_0 | \hat{\rho}_{\mathrm{s}}(\mathbf{r}) | \Xi_0 \rangle_{\mathrm{s}}.} \tag{5.23}$$

This is the expectation value of the charge-density operator in the unchanged specimen state. We shall denote it simply by

$$\rho(\mathbf{r}) \equiv \rho_{00}(\mathbf{r}). \tag{5.24}$$

The corresponding potential is

$$\boxed{V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r',} \tag{5.25}$$

and the elastic interaction experienced by the imaging electron is

$$\boxed{U_{\mathrm{el}}(\mathbf{r}) = -eV(\mathbf{r}).} \tag{5.26}$$

This is the origin of the scalar electrostatic specimen potential used throughout conventional elastic image-formation theory. The many-particle quantum state of the specimen has not been replaced by classical point electrons. Rather, its relevant diagonal charge-density matrix element has been compressed into the ordinary function $\rho(\mathbf{r})$ and hence into $V(\mathbf{r})$.

Inelastic channels

For $m \neq 0$,

$$\boxed{\rho_{m0}(\mathbf{r}) = {}_{\mathrm{s}}\langle \Xi_m | \hat{\rho}_{\mathrm{s}}(\mathbf{r}) | \Xi_0 \rangle_{\mathrm{s}}} \tag{5.27}$$

is an off-diagonal matrix element of the charge-density operator.

It is not an ordinary static charge distribution. It is a *transition charge density*: it describes the spatial variation of charge through which the Coulomb interaction couples two different internal states of the specimen.

Its associated transition potential is

$$\boxed{V_{m0}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho_{m0}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d^3r', \qquad m \neq 0.} \tag{5.28}$$

Unlike the static elastic density, an off-diagonal transition density need not be a real-valued function. It is a matrix element associated with a transition rather than an ordinary measurable charge distribution.

Thus elastic and inelastic Coulomb scattering share the same kernel. What changes is the specimen matrix element that supplies its source:

$$\boxed{\begin{aligned} \rho_{00} &= \langle \Xi_0 | \hat{\rho}_{\mathrm{s}} | \Xi_0 \rangle \quad \longrightarrow \quad \text{static elastic potential,} \\ \rho_{m0} &= \langle \Xi_m | \hat{\rho}_{\mathrm{s}} | \Xi_0 \rangle \quad \longrightarrow \quad \text{inelastic transition potential.} \end{aligned}} \tag{5.29}$$

This unified viewpoint will be useful later when we return to energy loss, plasmons, electronic excitations, and other inelastic processes. For the moment, however, we restrict attention to the elastic channel and ask how the ordinary potential $V(\mathbf{r})$ is represented and used in electron microscopy.

5.3 The electrostatic potential of matter

From the quantum-mechanical viewpoint, the elastic specimen is now described by the scalar function $V(\mathbf{r})$ of Eq. (5.25). From the viewpoint of electron microscopy, however, this is only the beginning.

Electron microscopists encounter the same interaction in several closely related forms:

$$\boxed{\begin{array}{ccccc} \text{charge density} & \longleftrightarrow & \text{electrostatic potential} & \longleftrightarrow & \text{electron scattering factor} \\ \rho(\mathbf{r}) & & V(\mathbf{r}) & & f_e(\mathbf{R}_3). \end{array}} \tag{5.30}$$

The purpose of this section is to establish the connection among these representations, first for an isolated atom and then for a complete biological macromolecule. This will also clarify what is meant by atomic screening, why electron scattering is particularly sensitive to charge redistribution, how tabulated atomic scattering factors are related to real-space potentials, and what approximations are made when a protein is represented as a sum of isolated atoms.

5.3.1 Charge density, Coulomb potential, and Poisson's equation

Equation (5.25) expresses the potential as a nonlocal Coulomb integral over the elastic charge density:

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d^3r'. \tag{5.31}$$

The same relation may be expressed locally through Poisson's equation,

$$\boxed{\nabla^2 V(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\varepsilon_0}.} \tag{5.32}$$

Here V is measured in volts and ρ in $\mathrm{C\,m^{-3}}$.

The Coulomb integral emphasizes that the potential at $\mathbf{r}$ contains contributions from charge throughout the specimen, whereas Poisson's equation relates the local curvature of the potential to the local charge density. They are two equivalent descriptions of the same electrostatic field.

From Poisson's equation to the Coulomb potential.

The Green function of the three-dimensional Laplace operator satisfies

$$\boxed{\nabla^2 \frac{1}{|\mathbf{r}-\mathbf{r}'|} = -4\pi\delta(\mathbf{r}-\mathbf{r}').} \tag{5.33}$$

Assuming that the potential vanishes sufficiently far from a localized charge distribution, define

$$V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'. \tag{5.34}$$

Applying the Laplacian with respect to $\mathbf{r}$ gives

$$\nabla^2 V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \rho(\mathbf{r}')\nabla^2 \frac{1}{|\mathbf{r}-\mathbf{r}'|}\, d^3r' \tag{5.35}$$

$$= -\frac{1}{\varepsilon_0} \int \rho(\mathbf{r}')\delta(\mathbf{r}-\mathbf{r}')\, d^3r' \tag{5.36}$$

$$= -\frac{\rho(\mathbf{r})}{\varepsilon_0}. \tag{5.37}$$

Thus the Coulomb integral of Eq. (5.25) is the solution of Poisson's equation, Eq. (5.32).

5.3.2 The screened potential of an atom

The simplest useful example is an isolated atom of atomic number Z. Place its nucleus at the origin and denote the electronic number density by $n_a(\mathbf{r})$. Within the usual frozen-nucleus description, its elastic charge density is

$$\boxed{\rho_a(\mathbf{r}) = Ze\,\delta(\mathbf{r}) - e\,n_a(\mathbf{r}).} \tag{5.38}$$

For a neutral atom,

$$\int n_a(\mathbf{r})\, d^3r = Z. \tag{5.39}$$

Substitution of Eq. (5.38) into Eq. (5.25) gives

$$\boxed{V_a(\mathbf{r}) = \frac{e}{4\pi\varepsilon_0}\left[\frac{Z}{r} - \int \frac{n_a(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'\right].} \tag{5.40}$$

The two terms have very different physical origins. The first is the positive Coulomb potential of the nucleus. The second is the negative potential generated by the bound electron cloud. Their competition is traditionally described as *screening* of the nuclear charge.

Electrostatic potential of a spherically symmetric electron cloud.

Starting from Eq. (5.40),

$$V_a(\mathbf{r}) = \frac{e}{4\pi\varepsilon_0}\left[\frac{Z}{r} - \int \frac{n_a(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'\right], \tag{5.41}$$

we need to evaluate the contribution of the electron cloud,

$$I(r) \triangleq \int \frac{n_a(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'. \tag{5.42}$$

If the atomic density is spherically symmetric,

$$n_a(\mathbf{r}') = n_a(r'), \tag{5.43}$$

the result can depend only on

$$r = \|\mathbf{r}\|. \tag{5.44}$$

Choose the polar axis of the spherical coordinates of $\mathbf{r}'$ along $\mathbf{r}$. If θ is the angle between $\mathbf{r}$ and $\mathbf{r}'$, then

$$|\mathbf{r}-\mathbf{r}'| = \sqrt{r^2 + r'^2 - 2rr'\cos\theta}. \tag{5.45}$$

Since

$$d^3r' = r'^2 \sin\theta\, dr'\, d\theta\, d\varphi, \tag{5.46}$$

Eq. (5.42) becomes

$$I(r) = \int_0^\infty n_a(r') r'^2\, dr' \int_0^{2\pi} d\varphi \int_0^\pi \frac{\sin\theta\, d\theta}{\sqrt{r^2 + r'^2 - 2rr'\cos\theta}}. \tag{5.47}$$

The azimuthal integral gives 2π. For the polar integral, define

$$u = \cos\theta, \qquad du = -\sin\theta\, d\theta. \tag{5.48}$$

Then

$$\int_0^\pi \frac{\sin\theta\, d\theta}{\sqrt{r^2 + r'^2 - 2rr'\cos\theta}} = \int_{-1}^1 \frac{du}{\sqrt{r^2 + r'^2 - 2rr'u}} \tag{5.49}$$

$$= \frac{1}{rr'}\left[(r + r') - |r - r'|\right]. \tag{5.50}$$

There are now two cases.

If

$$r' < r, \tag{5.51}$$

then $|r - r'| = r - r'$, and therefore

$$\int d\Omega' \frac{1}{|\mathbf{r} - \mathbf{r}'|} = \frac{4\pi}{r}. \tag{5.52}$$

If instead

$$r' > r, \tag{5.53}$$

then $|r - r'| = r' - r$, giving

$$\int d\Omega' \frac{1}{|\mathbf{r} - \mathbf{r}'|} = \frac{4\pi}{r'}. \tag{5.54}$$

Thus,

$$\boxed{\int d\Omega' \frac{1}{|\mathbf{r} - \mathbf{r}'|} = \frac{4\pi}{\max(r, r')}.} \tag{5.55}$$

Substituting this result into Eq. (5.47) and splitting the radial integral at $r' = r$ gives

$$I(r) = \int_0^r n_a(r') r'^2 \frac{4\pi}{r}\, dr' + \int_r^\infty n_a(r') r'^2 \frac{4\pi}{r'}\, dr' \tag{5.56}$$

$$= \frac{1}{r}\int_0^r 4\pi r'^2 n_a(r')\, dr' + \int_r^\infty 4\pi r' n_a(r')\, dr'. \tag{5.57}$$

Finally, inserting Eq. (5.57) into Eq. (5.40) gives

$$\boxed{V_a(r) = \frac{e}{4\pi\varepsilon_0}\left[\frac{Z}{r} - \frac{1}{r}\int_0^r 4\pi r'^2 n_a(r')\, dr' - \int_r^\infty 4\pi r' n_a(r')\, dr'\right].} \tag{5.58}$$

Equation (5.58) has a direct electrostatic interpretation. The electron cloud can be regarded as a set of infinitesimally thin spherical shells.

All electronic charge at radii $r' < r$ acts, at the observation point r, as though it were concentrated at the origin. This produces the term

$$-\frac{1}{r}\int_0^r 4\pi r'^2 n_a(r')\,dr'. \tag{5.59}$$

A spherical shell at radius $r' > r$, on the other hand, produces a constant potential everywhere inside that shell. Summing those outer shells gives

$$-\int_r^\infty 4\pi r' n_a(r')\,dr'. \tag{5.60}$$

Equation (5.58) is therefore the continuous-charge version of the familiar electrostatic shell theorem.

Close to the nucleus,

$$\boxed{V_a(r) \sim \frac{Ze}{4\pi\varepsilon_0 r}, \qquad r \to 0,} \tag{5.61}$$

because the point-like nuclear charge dominates. The exact electrostatic potential therefore contains a Coulomb singularity at the nuclear position.

At larger distances, more of the nuclear charge is screened by the bound electrons. For a neutral, spherically symmetric atom, the total monopole charge vanishes and

$$V_a(r) \longrightarrow 0 \qquad (r \to \infty). \tag{5.62}$$

For an ion containing N_e bound electrons, however, the net charge is

$$Q = (Z - N_e)e, \tag{5.63}$$

and the long-range potential contains the unscreened Coulomb term

$$\boxed{V_a(r) \sim \frac{Q}{4\pi\varepsilon_0 r}.} \tag{5.64}$$

This distinction is particularly relevant for biological matter. Carboxylates, protonated amino groups, phosphates, metal ions, counterions, and other charged groups cannot in general be represented exactly by the potential of a neutral isolated atom.

The screened atomic potential therefore contains two characteristic spatial regimes. Close to the nucleus, the bare nuclear Coulomb field dominates. At larger distances, the bound-electron cloud progressively screens the nuclear charge, and the detailed behavior becomes increasingly sensitive to the

spatial distribution of the valence electrons and to the ionic state of the atom.

We shall return to this distinction once the atomic potential has been expressed in reciprocal space through the electron scattering factor. There it will acquire a particularly transparent interpretation in terms of low- and high-angle scattering.

The word *screening* here refers to microscopic compensation of the nuclear field by bound electronic charge. It should not be confused with introducing a macroscopic dielectric constant for a continuous medium. At the atomic scales relevant for high-resolution cryo-EM, $V(\mathbf{r})$ is generated by the actual nuclear and electronic charge distributions of the macromolecule, solvent, ions, and supporting material.

5.3.3 Electron scattering factors as a Fourier representation of the atomic potential

The electrostatic potential $V_a(\mathbf{r})$ is the natural description of an atom from the viewpoint of the electron Hamiltonian. Electron microscopy, however, very often describes the same atomic interaction in reciprocal space through the *electron atomic scattering factor*.

We shall first establish the reciprocal-space representation of the atomic potential itself, and only afterwards introduce the scattering factor.

Throughout the book, spatial frequency is denoted by $\mathbf{R}$ and is measured in cycles per metre. In the present three-dimensional setting,

$$\mathbf{R} = (R_x, R_y, R_z), \qquad R = \|\mathbf{R}\|. \tag{5.65}$$

The corresponding wavevector is

$$\mathbf{k} = 2\pi\mathbf{R}, \tag{5.66}$$

with $\mathbf{k}$ measured in rad·m^{-1}.

Using the Fourier-transform convention introduced previously in the book, the three-dimensional Fourier transform of the atomic potential is

$$\boxed{\widehat{V}_a(\mathbf{R}) = \int_{\mathbb{R}^3} V_a(\mathbf{r}) \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}\right)\, d^3r.} \tag{5.67}$$

Since V_a has units of volts, $\widehat{V}_a$ has units V·m^3.

What does spatial frequency mean for an atomic potential?

Equation (5.67) decomposes the atomic potential into spatial variations occurring on different length scales.

A spatial frequency of magnitude R corresponds roughly to variations on a characteristic length scale

$$d \sim \frac{1}{R}. \tag{5.68}$$

Consequently,

$$\boxed{\begin{array}{lcl} R \text{ large} & \Longleftrightarrow & \text{rapid spatial variations and short-distance structure,} \\ R \text{ small} & \Longleftrightarrow & \text{slow spatial variations and long-distance structure.} \end{array}} \tag{5.69}$$

This gives a reciprocal-space interpretation of the screening picture developed in Sec. 5.3.2.

Close to the nucleus, the potential changes extremely rapidly and approaches the Coulomb singularity of Eq. (5.61). These short-range variations generate substantial high-spatial-frequency content.

At larger distances, the potential is determined by the compensation between the positive nuclear charge and the more extended bound electron cloud. Screening, valence-electron redistribution, bonding, and ionic charge therefore affect primarily the longer-range structure of the potential and consequently have a particularly strong influence at low spatial frequency.

We may therefore anticipate the qualitative correspondence

$$\boxed{\begin{array}{lcl} \text{large spatial frequency} & \Longleftrightarrow & \text{short-range structure, increasingly sensitive to the nucleus,} \\ \text{small spatial frequency} & \Longleftrightarrow & \text{screening, valence charge, bonding, and ionic state.} \end{array}} \tag{5.70}$$

At this stage, this statement should be understood qualitatively. A more precise result will follow from the Mott–Bethe relation in Sec. 5.3.4, where the nuclear and electronic contributions can be separated explicitly in reciprocal space.

Fourier transform of a spherically symmetric atomic potential

For an isolated atom in a spherically symmetric state,

$$V_a(\mathbf{r}) = V_a(r), \tag{5.71}$$

and its Fourier transform can depend only on

$$R = \|\mathbf{R}\|. \tag{5.72}$$

Fourier transform of a spherically symmetric atomic potential.

Starting from Eq. (5.67),

$$\widehat{V}_a(\mathbf{R}) = \int_{\mathbb{R}^3} V_a(r) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}}\, d^3r, \tag{5.73}$$

choose the polar axis of the spherical coordinates of $\mathbf{r}$ along $\mathbf{R}$. Then

$$\mathbf{R}\cdot\mathbf{r} = Rr\cos\theta, \tag{5.74}$$

and

$$d^3r = r^2 \sin\theta\, dr\, d\theta\, d\varphi. \tag{5.75}$$

Therefore,

$$\widehat{V}_a(R) = \int_0^\infty V_a(r) r^2\, dr \int_0^{2\pi} d\varphi \int_0^\pi e^{-j2\pi Rr\cos\theta} \sin\theta\, d\theta. \tag{5.76}$$

The azimuthal integral gives 2π. For the polar integral, let

$$u = \cos\theta, \qquad du = -\sin\theta\, d\theta. \tag{5.77}$$

Then

$$\int_0^\pi e^{-j2\pi Rr\cos\theta} \sin\theta\, d\theta = \int_{-1}^1 e^{-j2\pi Rru}\, du \tag{5.78}$$

$$= \frac{e^{j2\pi Rr} - e^{-j2\pi Rr}}{j2\pi Rr} \tag{5.79}$$

$$= 2\frac{\sin(2\pi Rr)}{2\pi Rr}. \tag{5.80}$$

Hence,

$$\boxed{\widehat{V}_a(R) = 4\pi \int_0^\infty V_a(r) \frac{\sin(2\pi Rr)}{2\pi Rr} r^2 \, dr.} \tag{5.81}$$

Thus the three-dimensional Fourier transform of a radial atomic potential reduces to a one-dimensional radial integral.

Definition of the electron atomic scattering factor

Electron-microscopy tables traditionally do not tabulate $\widehat{V}_a(\mathbf{R})$ directly. Instead, they use a rescaled quantity called the *electron atomic scattering factor*, which we define in this book as

$$\boxed{f_e(\mathbf{R}) \equiv \frac{m_0 e}{2\pi\hbar^2} \widehat{V}_a(\mathbf{R}).} \tag{5.82}$$

Equivalently,

$$\boxed{\widehat{V}_a(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} f_e(\mathbf{R}).} \tag{5.83}$$

The factor multiplying $\widehat{V}_a$ has been chosen so that f_e has units of length. Indeed,

$$[f_e] = \mathrm{m}. \tag{5.84}$$

This particular rescaling is not arbitrary: when elastic scattering is derived later from the wave equation in the first Born approximation, the same combination of constants appears naturally in the scattering amplitude. We postpone that derivation until the scattering problem itself has been introduced.

For the present purpose, the important point is simply that $f_e(\mathbf{R})$ and $\widehat{V}_a(\mathbf{R})$ contain exactly the same information. They differ only by a constant multiplicative factor:

$$\boxed{V_a(\mathbf{r}) \quad \xleftrightarrow{\mathcal{F}_3} \quad \widehat{V}_a(\mathbf{R}) \quad \xleftrightarrow{f_e = \frac{m_0 e}{2\pi\hbar^2}\widehat{V}_a} \quad f_e(\mathbf{R}).} \tag{5.85}$$

For a spherically symmetric atom, substitution of Eq. (5.81) into Eq. (5.82) gives

$$\boxed{f_e(R) = \frac{m_0 e}{2\pi\hbar^2} 4\pi \int_0^\infty V_a(r) \frac{\sin(2\pi Rr)}{2\pi Rr} r^2 \, dr.} \tag{5.86}$$

Thus an electron atomic scattering factor is, apart from a fixed physical prefactor, simply the radial Fourier transform of the atomic electrostatic potential.

What the scattering factor does and does not represent

It is useful to keep three quantities conceptually distinct:

$$\boxed{\begin{array}{rcl} V_a(\mathbf{r}) & : & \text{electrostatic potential in real space,} \\ \widehat{V}_a(\mathbf{R}) & : & \text{Fourier representation of that potential,} \\ f_e(\mathbf{R}) & : & \text{conventionally rescaled Fourier representation used in EM.} \end{array}} \tag{5.87}$$

The electron scattering factor is therefore not a new physical potential, nor is it by itself a scattering probability. It is a reciprocal-space representation of the atomic interaction.

Later, when the scattering of a fast electron by $V_a(\mathbf{r})$ is derived explicitly, we shall see how $f_e(\mathbf{R})$ becomes connected with a measurable scattering amplitude. At that point it will also become necessary to distinguish the construction of the atomic potential from the approximations used to propagate the electron through a complete specimen.

5.3.4 The Mott–Bethe relation: why electrons are especially sensitive to charge

The electron atomic scattering factor introduced in Sec. 5.3.3 is determined by the Fourier transform of the *electrostatic potential* of the atom. It therefore contains contributions from both the positive nucleus and the negative electron cloud.

X-ray scattering provides a useful comparison. In ordinary non-resonant elastic X-ray scattering, the dominant scattering centers are the bound electrons themselves. The nuclei contribute negligibly to the X-ray scattering amplitude because their much larger masses make their Thomson scattering amplitudes correspondingly very small.

Before deriving the relation between X-ray and electron scattering, we therefore first establish why the conventional X-ray atomic scattering factor is the Fourier transform of the electronic number density.

Why the X-ray atomic scattering factor is the Fourier transform of the electron density

Consider an incident X-ray plane wave with wavevector $\mathbf{k}_i$ that is elastically scattered into an outgoing wavevector $\mathbf{k}_f$. Since elastic scattering does not change the photon wavelength,

$$\|\mathbf{k}_i\| = \|\mathbf{k}_f\|. \tag{5.88}$$

It is convenient to describe the momentum transfer using the spatial frequency

$$\boxed{\mathbf{R} \equiv \frac{\mathbf{k}_f - \mathbf{k}_i}{2\pi},} \tag{5.89}$$

so that $\mathbf{R}$ has the same units, cycles·m^{-1}, as the spatial-frequency variable used throughout this book.

The complete derivation of the scattering amplitude of a single electron from the electromagnetic interaction is not required here. The only standard result we need is that, sufficiently far from absorption edges, each bound electron contributes essentially the same single-electron elastic scattering amplitude. The position of the electron enters through the relative phase of the incident and outgoing waves.

For an electron located at $\mathbf{r}_j$, this phase factor is

$$\exp\left[-j(\mathbf{k}_f - \mathbf{k}_i)\cdot\mathbf{r}_j\right] = \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}_j\right). \tag{5.90}$$

Thus the elastic X-ray scattering amplitude of an atom contains a coherent sum of these phase factors over all its bound electrons.

The result may be formulated directly using the quantum-mechanical language developed in Appendix A.

From individual bound electrons to the X-ray atomic scattering factor.

Let an atom contain N_e bound electrons, with position operators $\hat{\mathbf{r}}_j$. Apart from the common single-electron Thomson scattering amplitude, the operator describing their coherent contribution at spatial frequency $\mathbf{R}$ is

$$\boxed{\hat{F}_X(\mathbf{R}) = \sum_{j=1}^{N_e} \exp\left(-j2\pi\mathbf{R}\cdot\hat{\mathbf{r}}_j\right).} \tag{5.91}$$

For elastic scattering, the atom remains in the same internal state $|\Xi_0\rangle$. According to the matrix-element construction developed in Sec. A.12, the corresponding atomic factor is therefore the diagonal matrix element

$$f_X(\mathbf{R}) = \langle\Xi_0|\hat{F}_X(\mathbf{R})|\Xi_0\rangle. \tag{5.92}$$

The electronic number-density operator is

$$\hat{n}_a(\mathbf{r}) = \sum_{j=1}^{N_e} \delta\left(\mathbf{r} - \hat{\mathbf{r}}_j\right). \tag{5.93}$$

Its expectation value in the atomic state is the ordinary electronic number density,

$$\boxed{n_a(\mathbf{r}) = \langle \Xi_0 | \hat{n}_a(\mathbf{r}) | \Xi_0 \rangle.} \tag{5.94}$$

Now Fourier transform this density:

$$\int n_a(\mathbf{r}) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}}\, d^3r = \left\langle \Xi_0 \left| \sum_j \int \delta\left(\mathbf{r} - \hat{\mathbf{r}}_j\right) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}}\, d^3r \right| \Xi_0 \right\rangle \tag{5.95}$$

$$= \left\langle \Xi_0 \left| \sum_j e^{-j2\pi \mathbf{R}\cdot\hat{\mathbf{r}}_j} \right| \Xi_0 \right\rangle. \tag{5.96}$$

Comparing with Eq. (5.92), we obtain

$$\boxed{f_X(\mathbf{R}) = \int n_a(\mathbf{r}) \exp\left(-j2\pi \mathbf{R}\cdot\mathbf{r}\right)\, d^3r.} \tag{5.97}$$

Thus the X-ray atomic scattering factor is, apart from the common single-electron scattering amplitude that has been factored out, the Fourier transform of the electronic number density.

This result has a simple limiting interpretation. At $\mathbf{R} = \mathbf{0}$, all electronic contributions have the same phase, and therefore

$$\boxed{f_X(0) = \int n_a(\mathbf{r})\, d^3r = N_e.} \tag{5.98}$$

For a neutral atom, $N_e = Z$, so

$$\boxed{f_X(0) = Z.} \tag{5.99}$$

At nonzero spatial frequency, electrons at different positions contribute different phases. Their partial cancellation causes $f_X(R)$ to decrease as progressively finer spatial variations are probed.

The single-electron scattering result used above is the standard Thomson, or non-resonant, X-ray scattering result. Resonant corrections close to atomic absorption edges introduce additional energy-dependent terms and are outside the scope of the present discussion.

From X-ray to electron scattering: the Mott–Bethe relation

We can now compare the two experiments.

X-rays respond primarily to the bound-electron density and are therefore described by $f_X(\mathbf{R})$. Fast electrons instead interact with the total electrostatic potential generated jointly by the positive nucleus and the negative electron cloud.

The connection between the two descriptions follows directly from Poisson's equation.

From Poisson's equation to the Mott–Bethe relation.

For an atom whose nucleus is placed at the origin, the charge density was given in Eq. (5.38) as

$$\rho_a(\mathbf{r}) = Ze\,\delta(\mathbf{r}) - e\,n_a(\mathbf{r}). \tag{5.100}$$

Poisson's equation, Eq. (5.32), is

$$\nabla^2 V_a(\mathbf{r}) = -\frac{\rho_a(\mathbf{r})}{\varepsilon_0}. \tag{5.101}$$

Using the Fourier-transform convention of Eq. (5.67), spatial differentiation gives

$$\mathcal{F}\left\{\nabla^2 V_a\right\} = -4\pi^2 R^2 \widehat{V}_a(\mathbf{R}). \tag{5.102}$$

Therefore,

$$-4\pi^2 R^2 \widehat{V}_a(\mathbf{R}) = -\frac{\widehat{\rho}_a(\mathbf{R})}{\varepsilon_0}, \tag{5.103}$$

and, for $R \neq 0$,

$$\boxed{\widehat{V}_a(\mathbf{R}) = \frac{\widehat{\rho}_a(\mathbf{R})}{4\pi^2 \varepsilon_0 R^2}.} \tag{5.104}$$

The Fourier transform of the nuclear point charge is simply

$$\mathcal{F}73\left\{Ze\,\delta(\mathbf{r})\right\} = Ze. \tag{5.105}$$

From Eq. (5.97),

$$\widehat{n}_a(\mathbf{R}) = f_X(\mathbf{R}). \tag{5.106}$$

Consequently,

$$\widehat{\rho}_a(\mathbf{R}) = Ze - ef_X(\mathbf{R}) \tag{5.107}$$
$$= e\left[Z - f_X(\mathbf{R})\right]. \tag{5.108}$$

Substitution into Eq. (5.104) gives

$$\boxed{\widehat{V}_a(\mathbf{R}) = \frac{e}{4\pi^2\varepsilon_0 R^2}\left[Z - f_X(\mathbf{R})\right].} \tag{5.109}$$

The electron atomic scattering factor was defined in Eq. (5.82) by

$$f_e(\mathbf{R}) = \frac{m_0 e}{2\pi\hbar^2}\widehat{V}_a(\mathbf{R}). \tag{5.110}$$

Hence,

$$f_e(\mathbf{R}) = \frac{m_0 e^2}{8\pi^3\varepsilon_0\hbar^2}\frac{Z - f_X(\mathbf{R})}{R^2}. \tag{5.111}$$

Introducing the Bohr radius,

$$\boxed{a_0 = \frac{4\pi\varepsilon_0\hbar^2}{m_0 e^2},} \tag{5.112}$$

we finally obtain

$$\boxed{f_e(\mathbf{R}) = \frac{1}{2\pi^2 a_0}\frac{Z - f_X(\mathbf{R})}{R^2}.} \tag{5.113}$$

For a spherically symmetric atom, both scattering factors depend only on $R = \|\mathbf{R}\|$, and the result becomes

$$\boxed{f_e(R) = \frac{1}{2\pi^2 a_0}\frac{Z - f_X(R)}{R^2}.} \tag{5.114}$$

The Mott–Bethe relation makes mathematically explicit the screening picture developed in Sec. 5.3.2. Electron scattering depends on the *difference* between the nuclear and electronic contributions,

$$Z - f_X(R), \tag{5.115}$$

whereas ordinary X-ray scattering depends directly on the electronic density through $f_X(R)$.

What happens at high spatial frequency?

The electron cloud is spatially extended. Consequently, its Fourier transform decreases as the spatial frequency becomes large:

$$f_X(R) \longrightarrow 0, \qquad R \longrightarrow \infty. \tag{5.116}$$

The Mott–Bethe relation then gives

$$\boxed{f_e(R) \sim \frac{Z}{2\pi^2 a_0 R^2}, \qquad R \longrightarrow \infty.} \tag{5.117}$$

Thus high-spatial-frequency electron scattering becomes increasingly dominated by the concentrated nuclear charge.

This is the reciprocal-space counterpart of the real-space result Eq. (5.61): very close to the nucleus,

$$V_a(r) \sim \frac{Ze}{4\pi\varepsilon_0 r}. \tag{5.118}$$

What happens at low spatial frequency?

For a neutral atom,

$$f_X(0) = Z. \tag{5.119}$$

Hence both the numerator and denominator of the Mott–Bethe expression vanish as $R \to 0$. To determine the limiting electron scattering factor, we must examine how $f_X(R)$ approaches Z.

For a spherically symmetric electronic density, Eq. (5.97) can be written

$$\boxed{f_X(R) = 4\pi \int_0^\infty n_a(r) \frac{\sin(2\pi R r)}{2\pi R r} r^2 \, dr.} \tag{5.120}$$

Low-frequency expansion of the X-ray atomic scattering factor.

For a spherically symmetric electronic density, the X-ray atomic scattering factor is given in Eq. (5.120).

We are interested in the behavior near the origin,

$$R \to 0. \tag{5.121}$$

For small arguments,

$$\boxed{\frac{\sin x}{x} = 1 - \frac{x^2}{6} + O(x^4).} \tag{5.122}$$

Before inserting this expansion, it is useful to identify the radial moments of the electronic density that will appear.

The electronic number density is normalized to the total number of bound electrons,

$$\boxed{\int n_a(\mathbf{r})\, d^3r = N_e.} \tag{5.123}$$

Since the density is spherically symmetric,

$$n_a(\mathbf{r}) = n_a(r), \tag{5.124}$$

and therefore

$$d^3r = r^2 \sin\theta\, dr\, d\theta\, d\varphi. \tag{5.125}$$

Integrating over the angular coordinates gives

$$\boxed{4\pi \int_0^\infty n_a(r) r^2\, dr = N_e.} \tag{5.126}$$

Because $n_a(\mathbf{r})$ integrates to N_e rather than to 1, the normalized spatial distribution associated with the one-electron density is

$$\frac{n_a(\mathbf{r})}{N_e}. \tag{5.127}$$

We may therefore define the mean squared distance of the electronic cloud from the nucleus as

$$\boxed{\langle r^2 \rangle \equiv \frac{1}{N_e} \int r^2 n_a(\mathbf{r})\, d^3r.} \tag{5.128}$$

Using spherical symmetry once again,

$$\int r^2 n_a(\mathbf{r})\, d^3r = 4\pi \int_0^\infty n_a(r) r^4\, dr. \tag{5.129}$$

Hence,

$$\boxed{4\pi \int_0^\infty n_a(r) r^4 \, dr = N_e \langle r^2 \rangle.} \tag{5.130}$$

We can now expand the scattering factor itself. Substituting

$$x = 2\pi R r \tag{5.131}$$

into Eq. (5.122) gives

$$\frac{\sin(2\pi R r)}{2\pi R r} = 1 - \frac{(2\pi R r)^2}{6} + O(R^4). \tag{5.132}$$

Therefore,

$$f_X(R) = 4\pi \int_0^\infty n_a(r) \left[1 - \frac{(2\pi R r)^2}{6} + O(R^4) \right] r^2 \, dr \tag{5.133}$$

$$= 4\pi \int_0^\infty n_a(r) r^2 \, dr - \frac{(2\pi R)^2}{6} 4\pi \int_0^\infty n_a(r) r^4 \, dr + O(R^4). \tag{5.134}$$

Using Eqs. (5.126) and (5.130),

$$f_X(R) = N_e - \frac{(2\pi R)^2}{6} N_e \langle r^2 \rangle + O(R^4) \tag{5.135}$$

$$= N_e \left[1 - \frac{2\pi^2}{3} \langle r^2 \rangle R^2 + O(R^4) \right]. \tag{5.136}$$

Thus,

$$\boxed{f_X(R) = N_e \left[1 - \frac{2\pi^2}{3} \langle r^2 \rangle R^2 + O(R^4) \right].} \tag{5.137}$$

For a neutral atom,

$$N_e = Z, \tag{5.138}$$

so

$$\boxed{f_X(R) = Z \left[1 - \frac{2\pi^2}{3} \langle r^2 \rangle R^2 + O(R^4) \right].} \tag{5.139}$$

This shows that the value at the origin is fixed by the number of bound electrons,

$$f_X(0) = N_e, \tag{5.140}$$

whereas the initial curvature of $f_X(R)$ is determined by the spatial extent of the electron cloud through $\langle r^2 \rangle$.

Substitution into the Mott–Bethe relation gives

$$f_e(R) \simeq \frac{1}{2\pi^2 a_0} \frac{Z \frac{2\pi^2}{3} \langle r^2 \rangle R^2}{R^2}, \tag{5.141}$$

and therefore

$$\boxed{f_e(0) = \frac{Z \langle r^2 \rangle}{3a_0}.} \tag{5.142}$$

The result is important. Even at zero spatial frequency, the electron scattering factor of a neutral atom is not determined by Z alone. It depends on the spatial extent of the electronic cloud.

The contrast between the two limiting regimes can now be stated on a firm mathematical basis:

$$\boxed{\begin{array}{lcl} \text{large } R & \Longleftrightarrow & \text{short-range structure and increasing nuclear dominance,} \\ \text{small } R & \Longleftrightarrow & \text{screening and sensitivity to the extended electronic cloud.} \end{array}} \tag{5.143}$$

Since valence electrons are the part of the electronic density most strongly modified by bonding, protonation, and chemical environment, these effects appear particularly strongly in the low-frequency part of the electron scattering factor.

Ions and long-range electrostatic fields

For an ion containing N_e bound electrons, Eq. (5.98) gives

$$f_X(0) = N_e. \tag{5.144}$$

If $N_e \neq Z$, the numerator of the Mott–Bethe relation does not vanish at the origin:

$$Z - f_X(0) = Z - N_e. \tag{5.145}$$

Consequently,

$$\boxed{f_e(R) \sim \frac{Z - N_e}{2\pi^2 a_0 R^2}, \qquad R \to 0.} \tag{5.146}$$

The $1/R^2$ divergence is not pathological. It is simply the Fourier counterpart of the long-range Coulomb potential of a net charge,

$$V_a(r) \sim \frac{(Z - N_e)e}{4\pi\varepsilon_0 r}, \tag{5.147}$$

derived in Eq. (5.64).

The Mott–Bethe relation therefore explains why electron scattering is especially sensitive to ionization, protonation state, and redistribution of valence charge (Yonekura and Maki-Yonekura, 2016; Yonekura et al., 2018).

It also explains why chemical bonding may modify electron-scattering signals at spatial frequencies much lower than those required to resolve an individual bond geometrically.

Care is nevertheless required when an isolated ionic scattering factor is assigned to an atom inside a biological macromolecule. Charged residues and ions are embedded in the electrostatic field of the rest of the protein, solvent, counterions, and supporting material. At sufficiently low spatial frequency, it is the charge distribution of the *complete system*, rather than an isolated atomic ion, that determines the electrostatic potential.

5.3.5 Atomic scattering factors in Fourier and real space

The electron atomic scattering factor $f_e(\mathbf{R})$ introduced in Sec. 5.3.3 is a reciprocal-space representation of the atomic electrostatic potential. For a spherically symmetric atom it depends only on

$$R = \|\mathbf{R}\|, \tag{5.148}$$

and we write it simply as $f_e(R)$.

In practical electron microscopy, atomic scattering factors are often represented by simple analytical functions rather than stored as dense numerical tables. A particularly influential parameterization was developed by Peng, Ren, Dudarev and Whelan Peng et al. (1996). They showed that elastic electron atomic scattering factors for neutral atoms can be represented accurately over the spatial-frequency range relevant to electron diffraction and

microscopy by a sum of five Gaussians,

$$\boxed{f_e(R) = \sum_{i=1}^{N_G} a_i \exp(-b_i R^2),} \tag{5.149}$$

with $N_G = 5$ in their parameterization.

The coefficients a_i and b_i depend on the chemical element. Peng et al. determined them by fitting numerical elastic scattering factors using a combination of simulated annealing and least-squares optimization Peng et al. (1996). Their parameterization covered the neutral atoms and spatial frequencies up to approximately $6\,\text{Å}^{-1}$.

Because the exponential must be dimensionless,

$$[b_i R^2] = 1. \tag{5.150}$$

Thus, if R is expressed in m^{-1},

$$[b_i] = \text{m}^2. \tag{5.151}$$

Since f_e has dimensions of length,

$$[a_i] = \text{m}. \tag{5.152}$$

If instead the tabulation uses R in Å^{-1}, as is common in electron microscopy, then a_i and b_i must be interpreted in the corresponding Å and Å^2 units.

Why is the scattering factor largest at low spatial frequency?

The typical shape of an electron atomic scattering factor is worth understanding before transforming it back to real space. Its magnitude is generally largest at low spatial frequency and decreases as R increases.

This behavior is already visible mathematically in Eq. (5.149). At the origin,

$$\boxed{f_e(0) = \sum_{i=1}^{N_G} a_i.} \tag{5.153}$$

As R increases, each Gaussian factor

$$\exp(-b_i R^2) \tag{5.154}$$

decreases. Consequently, the contributions represented by these terms progressively disappear as finer and finer spatial variations are probed.

There is also a direct physical interpretation.

Recall from Sec. 5.3.4 that the electron scattering factor is related to the nuclear charge and the electronic cloud through the Mott–Bethe relation,

$$f_e(R) = \frac{1}{2\pi^2 a_0} \frac{Z - f_X(R)}{R^2}. \tag{5.155}$$

At low spatial frequency, the scattered waves produced by charge distributed over relatively large distances vary slowly in phase and therefore contribute coherently. The scattering factor consequently contains the integrated effect of the extended screened atomic potential.

As R increases, the Fourier phase

$$\exp(-j2\pi \mathbf{R} \cdot \mathbf{r}) \tag{5.156}$$

oscillates increasingly rapidly across the atom. Contributions from different regions of the extended electron cloud then cancel one another more strongly.

At very high spatial frequency, only very short-range variations of the potential contribute efficiently. As derived in Eq. (5.117),

$$\boxed{f_e(R) \sim \frac{Z}{2\pi^2 a_0 R^2}, \qquad R \to \infty.} \tag{5.157}$$

Thus even though the high-frequency signal becomes increasingly dominated by the concentrated nuclear charge, its magnitude still decreases approximately as $1/R^2$.

The two observations are not contradictory:

$$\boxed{\begin{array}{c} \text{low } R: \text{ strong total scattering from the extended screened potential} \\ \Downarrow \\ \text{increasing } R: \text{ stronger cancellation of spatially extended contributions} \\ \Downarrow \\ \text{high } R: \text{ weaker scattering, increasingly dominated by the nucleus.} \end{array}} \tag{5.158}$$

This explains the characteristic strongly decreasing radial curves seen in tabulations of electron atomic scattering factors.

From the scattering factor back to the atomic potential

The electron atomic scattering factor does not contain different physical information from the atomic electrostatic potential. As established in Eq. (5.83),

their reciprocal-space representations are related by

$$\boxed{\widehat{V}_a(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} f_e(\mathbf{R}).} \tag{5.159}$$

Therefore, once the scattering factor is known, the real-space atomic potential is recovered by an inverse three-dimensional Fourier transform:

$$\boxed{V_a(\mathbf{r}) = \frac{2\pi\hbar^2}{m_0 e} \int_{\mathbb{R}^3} f_e(\mathbf{R}) \exp\left(j2\pi\mathbf{R}\cdot\mathbf{r}\right)\, d^3R.} \tag{5.160}$$

Equivalently,

$$\boxed{V_a(\mathbf{r}) = \frac{2\pi\hbar^2}{m_0 e} \mathcal{F}^{-1}\left\{f_e(\mathbf{R})\right\}.} \tag{5.161}$$

For the isolated atoms considered here, the potential is assumed to be spherically symmetric,

$$V_a(\mathbf{r}) = V_a(r), \qquad r = \|\mathbf{r}\|, \tag{5.162}$$

and the scattering factor consequently depends only on

$$R = \|\mathbf{R}\|. \tag{5.163}$$

The Gaussian parameterization introduced in Eq. (5.149),

$$f_e(R) = \sum_{i=1}^{N_G} a_i \exp(-b_i R^2), \tag{5.164}$$

is particularly convenient because every term can be transformed back to real space analytically.

Atomic potential corresponding to a Gaussian scattering factor.

Consider first a single Gaussian term of the scattering factor,

$$f_i(\mathbf{R}) = a_i \exp(-b_i R^2), \qquad R^2 = R_x^2 + R_y^2 + R_z^2. \tag{5.165}$$

Its inverse Fourier transform, using the convention adopted throughout this book, is

$$\mathcal{F}^{-1}\left\{f_i(\mathbf{R})\right\} = a_i \int_{\mathbb{R}^3} \exp(-b_i R^2) \exp\left(j2\pi\mathbf{R}\cdot\mathbf{r}\right)\, d^3R. \tag{5.166}$$

Writing

$$\mathbf{r} = (x, y, z), \qquad \mathbf{R} = (R_x, R_y, R_z), \tag{5.167}$$

we have

$$\mathbf{R} \cdot \mathbf{r} = R_x x + R_y y + R_z z. \tag{5.168}$$

Moreover,

$$\exp(-b_i R^2) = \exp(-b_i R_x^2) \exp(-b_i R_y^2) \exp(-b_i R_z^2). \tag{5.169}$$

The three-dimensional integral therefore separates into three one-dimensional integrals:

$$\begin{aligned} \mathcal{F}^{-1}\{f_i(\mathbf{R})\} = a_i & \left[\int_{-\infty}^{\infty} \exp(-b_i R_x^2) \exp(j2\pi R_x x)\, dR_x\right] \\ & \times \left[\int_{-\infty}^{\infty} \exp(-b_i R_y^2) \exp(j2\pi R_y y)\, dR_y\right] \\ & \times \left[\int_{-\infty}^{\infty} \exp(-b_i R_z^2) \exp(j2\pi R_z z)\, dR_z\right]. \end{aligned} \tag{5.170}$$

We now use the one-dimensional Gaussian Fourier-transform identity

$$\boxed{\int_{-\infty}^{\infty} \exp(-bR^2) \exp(j2\pi Rx)\, dR = \sqrt{\frac{\pi}{b}} \exp\left(-\frac{\pi^2 x^2}{b}\right).} \tag{5.171}$$

Applying Eq. (5.171) independently to the three Cartesian directions gives

$$\mathcal{F}^{-1}\{f_i(\mathbf{R})\} = a_i \left(\frac{\pi}{b_i}\right)^{3/2} \exp\left[-\frac{\pi^2}{b_i}\left(x^2 + y^2 + z^2\right)\right] \tag{5.172}$$

$$= a_i \left(\frac{\pi}{b_i}\right)^{3/2} \exp\left(-\frac{\pi^2 r^2}{b_i}\right). \tag{5.173}$$

Since the Fourier transform is linear, the complete scattering factor of Eq. (5.149) gives

$$\boxed{\mathcal{F}^{-1}\{f_e(\mathbf{R})\} = \sum_{i=1}^{N_G} a_i \left(\frac{\pi}{b_i}\right)^{3/2} \exp\left(-\frac{\pi^2 r^2}{b_i}\right).} \tag{5.174}$$

Finally, substituting this result into Eq. (5.161) yields the corresponding atomic electrostatic potential:

$$\boxed{V_a(r) = \frac{2\pi\hbar^2}{m_0 e} \sum_{i=1}^{N_G} a_i \left(\frac{\pi}{b_i}\right)^{3/2} \exp\left(-\frac{\pi^2 r^2}{b_i}\right).} \tag{5.175}$$

Equation (5.175) gives a direct route from tabulated electron atomic scattering factors to the real-space atomic potential required by the electron propagation model.

No additional physical object has been introduced in this process. The three quantities

$$V_a(\mathbf{r}), \qquad \widehat{V}_a(\mathbf{R}), \qquad f_e(\mathbf{R}) \tag{5.176}$$

describe the same atomic interaction in different representations or normalizations. Their relationship is

$$\boxed{V_a(\mathbf{r}) \quad \xleftrightarrow{\mathcal{F}_3} \quad \widehat{V}_a(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} f_e(\mathbf{R}).} \tag{5.177}$$

There is also an instructive relation between the widths of the Gaussians in reciprocal and real space. A term with a large value of b_i,

$$\exp(-b_i R^2), \tag{5.178}$$

is narrow in reciprocal space, but its corresponding real-space factor

$$\exp\left(-\frac{\pi^2 r^2}{b_i}\right) \tag{5.179}$$

is broad. Conversely, a component extending to large spatial frequencies corresponds to a narrow real-space feature.

This is another manifestation of the Fourier-space relation discussed earlier:

$$\boxed{\text{narrow in reciprocal space} \quad \Longleftrightarrow \quad \text{broad in real space,}} \tag{5.180}$$

and

$$\boxed{\text{broad in reciprocal space} \quad \Longleftrightarrow \quad \text{localized in real space.}} \tag{5.181}$$

For an atom, the slowly varying, spatially extended part of the screened potential is therefore represented mainly at low spatial frequencies, whereas the sharp variation close to the nucleus requires components extending to much larger values of R.

What the Gaussian parameterization means physically

The five-Gaussian representation of Eq. (5.149) should not be interpreted as saying that an atom consists physically of five Gaussian charge clouds. The Gaussians are an analytical approximation to the reciprocal-space electron atomic scattering factor.

Their main practical advantage is that every Gaussian in reciprocal space has an analytical Gaussian counterpart in real space, as derived in Eq. (5.175). This makes it possible to move efficiently between the two representations

$$f_e(R) \quad \longleftrightarrow \quad V_a(r), \tag{5.182}$$

and therefore to construct specimen potentials directly from atomic coordinates.

The shapes of these functions are themselves physically informative. Figure 5.2 shows the elastic electron atomic scattering factors of several elements commonly found in biological specimens, calculated from the five-Gaussian parameterization of Peng et al. (1996). The corresponding real-space electrostatic potentials are shown in Fig. 5.3.

Decay in reciprocal space. The scattering factor is largest at low spatial frequency and decreases rapidly as R increases. This behavior follows from the Fourier interpretation developed in Sec. 5.3.3: slowly varying, spatially extended components of the screened atomic potential contribute mainly at low R, whereas increasingly fine spatial structure is required to produce substantial amplitude at large R.

To make the rate of decay quantitative, define the normalized scattering factor

$$\boxed{\eta_f(R) \equiv \frac{f_e(R)}{f_e(0)}.} \tag{5.183}$$

For the H, C, N, O, P, S and Fe parameterizations shown in Fig. 5.2,

$$\eta_f(R) < 10^{-2} \tag{5.184}$$

once

$$\boxed{R \simeq 4.4\text{–}6.2\ \text{Å}^{-1},} \tag{5.185}$$

depending on the element.

Thus, beyond approximately

$$R \simeq 6\ \text{Å}^{-1}, \tag{5.186}$$

the elastic scattering factor of these biologically relevant elements has already fallen to around one percent, or less, of its value at the origin.

The Peng parameterization used here was fitted up to

$$s = 6\ \text{Å}^{-1}, \tag{5.187}$$

where $s = \sin\theta/\lambda$. Since the spatial-frequency convention used throughout this book is

$$R = 2s, \tag{5.188}$$

the fitted interval extends to

$$\boxed{R_{\max} = 12\ \text{Å}^{-1}.} \tag{5.189}$$

At this upper limit, all the elements displayed in Fig. 5.2 satisfy approximately

$$\boxed{\eta_f(12\ \text{Å}^{-1}) < 2 \times 10^{-3}.} \tag{5.190}$$

In other words, throughout the range represented by the Peng parameterization, the reciprocal-space contribution has fallen by more than a factor of roughly 500 relative to its zero-frequency value.

This gives a useful practical interpretation:

$$\boxed{\begin{array}{lcl} R \lesssim 1\ \text{Å}^{-1} & : & \text{strong atomic scattering contribution,} \\ R \simeq 5\text{–}6\ \text{Å}^{-1} & : & \text{typically of order 1\% of } f_e(0), \\ R \simeq 12\ \text{Å}^{-1} & : & \text{typically below 0.2\% of } f_e(0). \end{array}} \tag{5.191}$$

The scattering factor therefore has a long but rapidly decreasing high-frequency tail. As discussed in Sec. 5.3.4, this tail becomes increasingly dominated by the concentrated nuclear charge even though its absolute magnitude continues to decrease.

Decay in real space and comparison with conventional atomic radii. The corresponding real-space behavior is shown in Fig. 5.3. To quantify the radial extent of the interaction, define

$$\boxed{\eta_V(r) \equiv \frac{V_a(r)}{V_a(0)}.} \tag{5.192}$$

For the elements shown in Fig. 5.3, the radii at which the potential has fallen to 1%, 0.1%, and 0.01% of its central value are approximately those listed in Table 5.1.

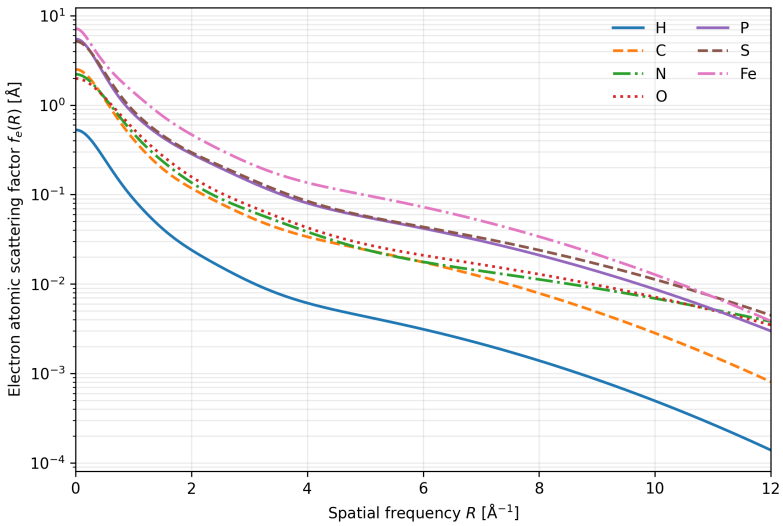


Figure 5.2: Elastic electron atomic scattering factors for several elements commonly encountered in biological specimens, calculated from the five-Gaussian parameterization of Peng et al. Peng et al. (1996). Peng et al. use the crystallographic scattering variable $s = \sin\theta/\lambda$, whereas the spatial-frequency convention of this book is $R = 2s$. Consequently, their fitted interval $0 \leq s \leq 6$ Å^{-1} corresponds to $0 \leq R \leq 12$ Å^{-1}. The scattering factor is expressed in Å. For the elements shown, $f_e(R)$ falls below approximately 1% of $f_e(0)$ at $R \simeq 4.4$–6.2 Å^{-1}, and at $R = 12$ Å^{-1} it is below approximately 0.2% of its zero-frequency value.

It is useful to compare these distances with the conventional "atomic radii" encountered in chemistry. There is no unique atomic radius. Two particularly common definitions are the *covalent radius*, inferred approximately from bond lengths, and the *van der Waals radius*, inferred from the closest approach of non-bonded atoms. Representative values are taken here from Cordero et al. (2008) for covalent radii and Alvarez (2013) for van der Waals radii.

Table 5.1: Comparison between conventional atomic radii and the radial decay of the elastic atomic electrostatic potential obtained from the Peng five-Gaussian parameterization. The quantities $r_{1\%}$, $r_{0.1\%}$, and $r_{0.01\%}$ are defined by $V_a(r)/V_a(0) = 10^{-2}, 10^{-3}$, and 10^{-4}, respectively. All distances are in Å.

Element	r_{cov}	r_{vdW}	$r_{1\%}$	$r_{0.1\%}$	$r_{0.01\%}$
H	0.31	1.20	0.57	1.12	1.68
C	0.76	1.77	0.55	1.08	1.65
N	0.71	1.66	0.43	0.88	1.38
O	0.66	1.50	0.42	0.82	1.26
P	1.07	1.90	0.47	1.09	1.75
S	1.05	1.89	0.45	1.01	1.59
Fe	1.32	2.44	0.42	0.96	1.79

Several points are immediately apparent.

First, the radii defined from the decay of the electrostatic potential are not expected to coincide with either the covalent or van der Waals radii. These quantities describe different physical properties.

The covalent radius characterizes the typical position of a neighboring atom participating in a chemical bond. It is therefore related to the geometry of the bonded electron density rather than to a boundary of the electrostatic field.

The van der Waals radius is larger and characterizes the distance at which non-bonded atoms begin to experience strong short-range repulsion. Again, it should not be interpreted as the point at which the atomic charge distribution or electrostatic potential terminates.

By contrast, $r_{1\%}$, $r_{0.1\%}$, and $r_{0.01\%}$ are purely operational radii: they specify how far from the nucleus one must move before the *elastic electrostatic potential* has fallen below a chosen fraction of its value at the origin.

For most elements in Table 5.1, the 1% radius is smaller than the covalent

radius. For example, carbon has a covalent radius of about

$$0.76 \text{ Å}, \tag{5.193}$$

whereas its atomic potential has already fallen to 1% of its central value at approximately

$$0.55 \text{ Å}. \tag{5.194}$$

At the covalent radius itself, the Peng model gives only about

$$3.7 \times 10^{-3} \tag{5.195}$$

of the central potential.

The same tendency is even stronger for the heavier elements. For iron, for example,

$$r_{\text{cov}} \simeq 1.32 \text{ Å}, \tag{5.196}$$

whereas

$$r_{1\%} \simeq 0.42 \text{ Å}. \tag{5.197}$$

At the covalent radius, the potential is only about

$$3.5 \times 10^{-4} \tag{5.198}$$

of $V_a(0)$.

This apparent discrepancy does not mean that the chemical radius of the atom has been overestimated. It reflects the fact that $V_a(0)$ is dominated by the very strong potential close to the nucleus. Normalizing the entire radial profile by this large central value makes the relative potential appear small already at distances comparable with ordinary bond lengths.

The comparison with the van der Waals radius is still more striking. For C, N, O, P and S, the potential at the van der Waals radius is already of order

$$10^{-5}\text{–}10^{-4} \tag{5.199}$$

of its central value. For Fe it is about

$$2 \times 10^{-5}. \tag{5.200}$$

Thus the conventional van der Waals surface lies well into the weak tail of the single-atom electrostatic potential.

This should not be interpreted as saying that the tail is physically irrelevant. In a macromolecule, the potentials of many atoms overlap and add. A contribution that is tiny compared with the central potential of one isolated

atom can still contribute appreciably when thousands of atomic tails are superimposed.

There is also an important conceptual distinction:

$$\boxed{\begin{array}{lcl} r_{\mathrm{cov}} & : & \text{characteristic bonded geometry,} \\ r_{\mathrm{vdW}} & : & \text{characteristic non-bonded contact geometry,} \\ r_{\varepsilon} & : & \text{chosen truncation radius for the electrostatic potential.} \end{array}} \tag{5.201}$$

Only the last of these is directly relevant to deciding how far a single-atom kernel must be evaluated in an image-formation calculation.

For the Peng parameterization used here, the potential falls below 10^{-3} of its central value at approximately

$$r \simeq 0.8\text{–}1.1\ \text{Å}, \tag{5.202}$$

and below 10^{-4} at approximately

$$r \simeq 1.3\text{–}1.8\ \text{Å}. \tag{5.203}$$

By

$$r = 2\ \text{Å}, \tag{5.204}$$

all seven potentials satisfy

$$\boxed{\eta_V(2\ \text{Å}) < 6.1 \times 10^{-5}.} \tag{5.205}$$

Thus a numerical cutoff of order

$$2\text{–}3\ \text{Å} \tag{5.206}$$

is already very conservative for the isolated-atom kernels represented by this finite Gaussian model.

Two qualifications are important. First, the threshold radii depend on the chosen tolerance; there is no physically unique point at which an atomic electrostatic potential "ends." Second, the normalization by $V_a(0)$ is especially stringent because the central value is strongly weighted by the nuclear region. The quantities in Table 5.1 should therefore be interpreted as numerical decay lengths of the atomic potential, not as alternative definitions of chemical atomic size.

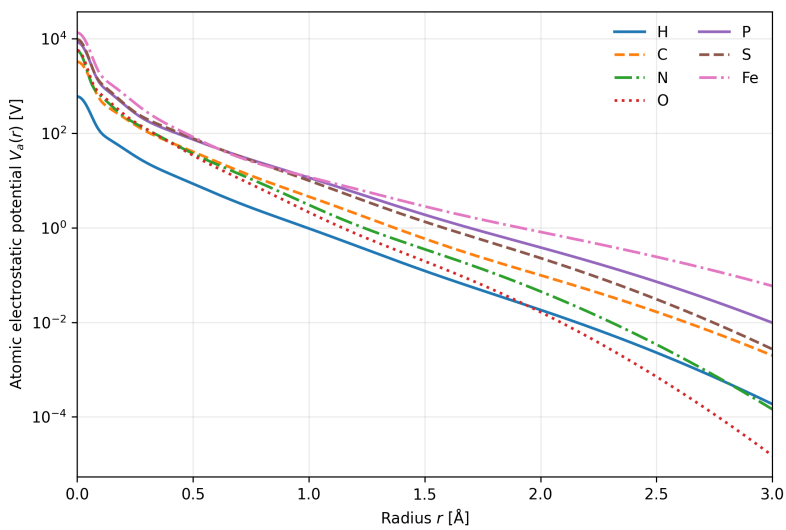


Figure 5.3: Real-space atomic electrostatic potentials corresponding to the scattering factors of Fig. 5.2. The potentials were obtained from Eq. (5.161), including the physical factor $2\pi\hbar^2/(m_0 e)$, and are expressed in volts. For the elements shown, the potential falls below approximately 0.1% of its central value at radii of only 0.8–1.1 Å and below 0.01% at approximately 1.3–1.8 Å. At $r = 2$ Å the relative contribution is already below approximately 6.1×10^{-5} for all seven elements. Because the finite five-Gaussian parameterization does not contain the exact arbitrarily high-frequency tail, it smooths the Coulomb singularity at the nuclear position.

Choice and limitations of analytical scattering-factor parameterizations

The Gaussian representation of Eq. (5.149) is only one of several analytical forms that have been proposed for electron atomic scattering factors. It is computationally convenient and can reproduce tabulated scattering factors very accurately over a prescribed finite range of spatial frequencies, but the analytical form itself need not obey the exact asymptotic behavior of an atom.

This distinction becomes important because the previous sections have already established two physical constraints on the scattering factor.

At low spatial frequency,

$$\boxed{f_e(R) \longrightarrow f_e(0), \qquad R \to 0,} \tag{5.207}$$

where, for a neutral spherical atom, Eq. (5.139) together with the Mott–Bethe relation gives

$$\boxed{f_e(0) = \frac{Z\langle r^2\rangle}{3a_0}.} \tag{5.208}$$

At the opposite extreme, the electronic X-ray scattering factor tends to zero as $R \to \infty$, and Eq. (5.117) therefore gives

$$\boxed{f_e(R) \sim \frac{Z}{2\pi^2 a_0 R^2}, \qquad R \to \infty.} \tag{5.209}$$

Equation (5.209) is especially important. Through Eq. (5.83), a $1/R^2$ high-frequency tail is precisely what is required to recover the Coulomb behavior

$$\boxed{V_a(r) \sim \frac{Ze}{4\pi\varepsilon_0 r}, \qquad r \to 0,} \tag{5.210}$$

derived previously in Eq. (5.61).

Thus an analytical parameterization should ideally reproduce not only the numerical values of $f_e(R)$ over some fitted interval, but also its correct behavior at both ends of the spatial-frequency axis.

Pure Gaussian parameterizations. The oldest and computationally simplest family writes

$$\boxed{f_e(R) = \sum_i a_i \exp(-b_i R^2),} \tag{5.211}$$

with convention-dependent numerical values of a_i and b_i.

Forms of this type have been used by Smith and Burge, Doyle and Turner, and later Peng and coworkers (Doyle and Turner, 1968; Peng et al., 1996). The five-Gaussian Peng parameterization used in the preceding sections belongs to this family.

Its advantages are clear. The function is finite at $R = 0$, smooth, easy to evaluate, and has an analytical Fourier transform. Moreover, over the finite interval on which the coefficients are fitted it can approximate numerical scattering-factor data extremely accurately.

Its limitation appears outside that interval. Every Gaussian satisfies

$$\exp(-b_i R^2) \longrightarrow 0 \tag{5.212}$$

exponentially as $R \to \infty$. Therefore,

$$f_e(R) = o(R^{-n}) \tag{5.213}$$

for every finite power n, and in particular it cannot reproduce the physical

$$f_e(R) \propto \frac{1}{R^2} \tag{5.214}$$

tail of Eq. (5.209).

The consequence in real space is equally direct. As derived in Sec. 5.3.5, a Gaussian scattering factor transforms into a Gaussian real-space potential. Hence

$$V_a(0) < \infty \tag{5.215}$$

for any finite Gaussian sum, whereas the exact electrostatic potential must diverge as $1/r$ according to Eq. (5.210).

This does not make the Peng parameterization inaccurate over its fitted range. It means only that the analytical continuation of the fitted function to arbitrarily large R is not physically correct.

Weickenmeier–Kohl parameterization. One way to incorporate the correct high-frequency behavior is the form introduced by Weickenmeier and Kohl (1991). Apart from convention-dependent rescalings of its coefficients, it has the structure

$$\boxed{f_e(R) = \sum_i a_i \frac{1 - \exp(-b_i R^2)}{R^2}.} \tag{5.216}$$

Although the denominator appears singular, the numerator behaves for small R as

$$1 - \exp(-b_i R^2) = b_i R^2 + O(R^4), \tag{5.217}$$

so each term has a finite limit,

$$\frac{1 - \exp(-b_i R^2)}{R^2} \longrightarrow b_i. \tag{5.218}$$

At high frequency,

$$\exp(-b_i R^2) \longrightarrow 0, \tag{5.219}$$

and therefore

$$f_e(R) \propto \frac{1}{R^2}. \tag{5.220}$$

Thus this family can reproduce both the finite value at the origin and the correct Coulombic high-frequency asymptote. Its corresponding real-space potential contains terms proportional to

$$\frac{\operatorname{erfc}(\alpha_i r)}{r}, \tag{5.221}$$

and consequently retains the $1/r$ singularity at the nuclear position.

Kirkland parameterization. Kirkland proposed another analytical compromise between accuracy and correct asymptotic behavior (Kirkland, 2010)[Appendix C]. In the spatial frequency notation of this book, its functional structure is

$$\boxed{f_e(R) = \sum_{i=1}^{N_L} \frac{a_i}{R^2 + b_i} + \sum_{i=1}^{N_G} c_i \exp(-d_i R^2).} \tag{5.222}$$

Kirkland uses three rational terms and three Gaussian terms. The rational terms are sometimes called Lorentzian terms.

At $R = 0$,

$$f_e(0) = \sum_i \frac{a_i}{b_i} + \sum_i c_i, \tag{5.223}$$

which is finite.

At large R, the Gaussian contribution disappears exponentially, whereas

$$\frac{a_i}{R^2 + b_i} \sim \frac{a_i}{R^2}. \tag{5.224}$$

The rational part therefore supplies the required $1/R^2$ tail.

The physical roles of the two families of terms are consequently different:

$$\boxed{\begin{array}{lcl} \text{Gaussian terms} & : & \text{flexibility for fitting low and intermediate } R, \\ \text{rational terms} & : & \text{correct high-frequency } 1/R^2 \text{ behavior.} \end{array}} \tag{5.225}$$

Their real-space transforms contain exponentially screened $1/r$ terms together with Gaussian terms. The resulting potential therefore retains the nuclear singularity while remaining analytically tractable.

Lobato–Van Dyck parameterization. Lobato and Van Dyck developed a parameterization specifically designed to reproduce simultaneously the scattering factor, electron density, and electrostatic potential while satisfying their physical constraints Lobato and Van Dyck (2014). Its functional form can be written, up to the scaling convention used for spatial frequency, as

$$\boxed{f_e(R) = \sum_i a_i \frac{2 + b_i R^2}{(1 + b_i R^2)^2}.} \tag{5.226}$$

At the origin,

$$f_e(0) = 2 \sum_i a_i, \tag{5.227}$$

so the scattering factor remains finite. At large spatial frequency,

$$\frac{2 + b_i R^2}{(1 + b_i R^2)^2} \sim \frac{1}{b_i R^2}, \tag{5.228}$$

and the required $1/R^2$ behavior is recovered.

The corresponding real-space potential contains exponentially screened terms multiplied by a contribution proportional to $1/r$. It therefore has the correct Coulomb singularity close to the nucleus.

The differences between the Peng, Weickenmeier–Kohl, Kirkland, and Lobato–Van Dyck families are summarized schematically in Table 5.2.

Fitting accuracy and physical asymptotes are different requirements. It is important not to confuse these two ideas. A parameterization may fit tabulated scattering factors extremely well over a finite interval without having the correct asymptotic behavior outside that interval. Conversely, imposing the correct limits at $R = 0$ and $R \to \infty$ does not by itself guarantee the smallest fitting error at intermediate spatial frequencies.

Table 5.2: Main analytical families used to parameterize elastic electron atomic scattering factors. The formulas indicate their functional structure; numerical coefficients depend on the spatial-frequency convention used in the original tabulation.

Model	Typical form	$R \to \infty$	$V_a(r)$ near $r = 0$
Doyle–Turner / Peng	$\sum_i a_i e^{-b_i R^2}$	exponential	finite
Weickenmeier–Kohl	$\sum_i a_i(1 - e^{-b_i R^2})/R^2$	$1/R^2$	$1/r$
Kirkland	$\sum_i a_i/(R^2 + b_i) + \sum_i c_i e^{-d_i R^2}$	$1/R^2$	$1/r$
Lobato–Van Dyck	$\sum_i a_i(2 + b_i R^2)/(1 + b_i R^2)^2$	$1/R^2$	$1/r$

The Peng parameterization is a good example of the first situation. It was optimized to give a highly accurate representation over a specified range of scattering variables (Peng et al., 1996). Within that range its Gaussian representation can be excellent. Its failure to reproduce the asymptotic $1/R^2$ tail becomes relevant only when the parameterization is extrapolated toward arbitrarily high spatial frequencies.

The physically constrained forms of Weickenmeier–Kohl, Kirkland, and Lobato–Van Dyck sacrifice some of the simplicity of a pure Gaussian sum in order to retain the correct behavior outside the central fitted range.

Why parameterizing $f_e(R)$ directly is preferable. There is a second numerical subtlety. One could first parameterize the X-ray scattering factor and then obtain the electron scattering factor from the Mott–Bethe relation,

$$f_e(R) = \frac{1}{2\pi^2 a_0} \frac{Z - f_X(R)}{R^2}. \tag{5.229}$$

Near $R = 0$,

$$f_X(R) = Z - O(R^2), \tag{5.230}$$

so the numerator is obtained by subtracting two quantities that become equal as $R \to 0$. Any small fitting error in $f_X(R)$ can therefore become large after division by R^2.

For numerical work it is consequently preferable either to parameterize $f_e(R)$ directly, as is done in the electron-scattering parameterizations discussed above, or to constrain the X-ray parameterization so that its low-frequency expansion satisfies exactly the conditions derived in Sec. 5.3.4.

Analytical parameterization and numerical truncation are separate approximations. Even when the analytical parameterization has

the correct asymptotic behavior, a numerical simulation must eventually represent it on a finite grid. It is useful to distinguish explicitly between

$$\boxed{\begin{array}{ll} 1. & \text{the physical isolated-atom model,} \\ 2. & \text{the analytical parameterization of } f_e(R), \\ 3. & \text{the finite numerical representation used in the simulation.} \end{array}} \tag{5.231}$$

A physically correct analytical potential containing a $1/r$ singularity cannot be represented exactly on a finite Cartesian grid. The highest spatial frequency available is finite, and the value assigned to the voxel containing the nucleus is necessarily a sampled, averaged, or otherwise regularized representation of the singular potential.

Likewise, real-space calculations normally truncate the atomic potential beyond some radius R_c. Lobato and Van Dyck Lobato and Van Dyck (2015) define this radius through an absolute potential threshold,

$$\boxed{V_a(R_c) = V_{\min}.} \tag{5.232}$$

Their comparison of several parameterizations shows that, for a fixed choice

$$V_{\min} = 0.01 \text{ V}, \tag{5.233}$$

the resulting cutoff radii are rather similar for light atoms but can differ appreciably for intermediate and heavy elements. This is an important reminder that a real-space cutoff is not a universal "atomic radius": it depends on both the chosen numerical tolerance and the scattering-factor parameterization.

An abrupt truncation,

$$V_a(r) = 0, \qquad r > R_c, \tag{5.234}$$

also introduces a discontinuity that may itself generate Fourier artifacts. Lobato and Van Dyck therefore use a smooth transition that forces both the potential and its first derivative continuously toward zero at the cutoff radius.

For an isolated Au atom they found that

$$V_{\min} = 0.01 \text{ V} \tag{5.235}$$

corresponds to approximately

$$R_c = 3.19 \text{ Å} \tag{5.236}$$

for the Lobato–Van Dyck potential, and that this cutoff was sufficient for convergence of their low-angle STEM calculation. High-angle signals were substantially less sensitive to the precise cutoff.

This observation complements the relative thresholds discussed in Sec. 5.3.5. A relative criterion such as

$$V_a(r)/V_a(0) < 10^{-4} \tag{5.237}$$

is useful for understanding the radial decay of a particular atomic profile, whereas an absolute criterion such as

$$V_a(r) < V_{\min} \tag{5.238}$$

is often more appropriate when deciding how far an atomic kernel must be evaluated in an actual numerical simulation.

Practical conclusion. There is therefore no unique "best" analytical form independently of the intended calculation.

For moderate spatial frequencies, a Gaussian representation such as Peng's is extremely convenient and can be very accurate within its fitted interval. For calculations that probe very high spatial frequencies or require an accurate description extremely close to the nucleus, parameterizations with the correct $1/R^2$ tail—such as the Weickenmeier–Kohl, Kirkland, or Lobato–Van Dyck forms—are preferable.

In single-particle cryo-EM, the experimentally accessible spatial frequencies are normally far below those required to resolve the nuclear singularity itself. Nevertheless, preserving the distinction between a fitted analytical model, the underlying physical potential, and its sampled numerical representation is essential if the model is to remain quantitatively meaningful.

5.3.6 The projected atomic potential

Many single-particle cryo-EM image-formation models introduce a projected specimen potential. We therefore establish explicitly how the three-dimensional atomic potential and its scattering factor are related to their projected counterparts.

Let

$$\boldsymbol{\rho} = (x, y) \tag{5.239}$$

denote transverse position, and let the incident electron propagate approximately along the z direction.

The projected atomic potential is obtained by integrating the three-dimensional electrostatic potential along the beam direction,

$$\boxed{V_a^{\text{proj}}(\boldsymbol{\rho}) = \int_{-\infty}^{\infty} V_a(\boldsymbol{\rho}, z)\, dz,} \tag{5.240}$$

where

$$\boldsymbol{\rho} = (x, y) \tag{5.241}$$

denotes the transverse position in the plane perpendicular to the electron beam.

Since V_a has units of volts and the integration is performed over a length, the projected potential has units

$$\boxed{\left[V_a^{\text{proj}}\right] = \mathrm{V\,m}.} \tag{5.242}$$

For a spherically symmetric atom,

$$V_a(\mathbf{r}) = V_a(r), \qquad r = \|\mathbf{r}\|, \tag{5.243}$$

and the projected potential can also depend only on the transverse distance from the nucleus,

$$\rho = \|\boldsymbol{\rho}\| = \sqrt{x^2 + y^2}. \tag{5.244}$$

The relation between the three-dimensional radius r, the transverse radius ρ, and the beam coordinate z is

$$\boxed{r^2 = \rho^2 + z^2.} \tag{5.245}$$

Equation (5.240) can therefore be reduced to a one-dimensional radial integral.

Projection of a spherically symmetric atomic potential.

For fixed transverse distance ρ, spherical symmetry gives

$$V_a(\boldsymbol{\rho}, z) = V_a\left(\sqrt{\rho^2 + z^2}\right). \tag{5.246}$$

Hence,

$$V_a^{\text{proj}}(\rho) = \int_{-\infty}^{\infty} V_a\left(\sqrt{\rho^2 + z^2}\right)\, dz. \tag{5.247}$$

The integrand is an even function of z, because it depends on z only

through z^2. Therefore,

$$V_a^{\text{proj}}(\rho) = 2\int_0^\infty V_a\left(\sqrt{\rho^2+z^2}\right)\,dz. \tag{5.248}$$

We now change the integration variable from z to the radial distance r. From Eq. (5.245),

$$r = \sqrt{\rho^2+z^2}, \tag{5.249}$$

so

$$z = \sqrt{r^2-\rho^2}. \tag{5.250}$$

Differentiating,

$$dz = \frac{r}{\sqrt{r^2-\rho^2}}\,dr. \tag{5.251}$$

When

$$z = 0, \tag{5.252}$$

the radial distance is

$$r = \rho, \tag{5.253}$$

whereas

$$z \to \infty \tag{5.254}$$

implies

$$r \to \infty. \tag{5.255}$$

Substituting Eq. (5.251) into Eq. (5.248) therefore gives

$$\boxed{V_a^{\text{proj}}(\rho) = 2\int_\rho^\infty V_a(r)\frac{r}{\sqrt{r^2-\rho^2}}\,dr.} \tag{5.256}$$

Equation (5.256) has the standard form of the Abel transform of a radially symmetric function. The Abel transform is precisely the mathematical operation that relates a spherically symmetric three-dimensional function to its line-of-sight projection.

Projected potential and the Fourier projection theorem.

Define the two-dimensional Fourier transform of the projected po-

tential,

$$\widehat{V}_a^{\text{proj}}(\mathbf{R}) = \int_{\mathbb{R}^2} V_a^{\text{proj}}(\boldsymbol{\rho}) \exp\left(-j2\pi\mathbf{R}\cdot\boldsymbol{\rho}\right)\, d^2\rho, \tag{5.257}$$

where

$$\mathbf{R} = (R_x, R_y) \tag{5.258}$$

is the transverse spatial frequency.

Substituting Eq. (5.240),

$$\widehat{V}_a^{\text{proj}}(\mathbf{R}) = \int_{\mathbb{R}^2}\int_{-\infty}^{\infty} V_a(x,y,z) \exp\left[-j2\pi(R_x x + R_y y)\right]\, dz\, dx\, dy \tag{5.259}$$

$$= \widehat{V}_a(R_x, R_y, R_z = 0). \tag{5.260}$$

Therefore,

$$\boxed{\widehat{V}_a^{\text{proj}}(\mathbf{R}) = \widehat{V}_a(\mathbf{R}, 0).} \tag{5.261}$$

This is the Fourier projection theorem: projection along z selects the plane $R_z = 0$ of the three-dimensional Fourier transform.

For a spherically symmetric atom,

$$f_e(R_x, R_y, 0) = f_e(R), \qquad R = \sqrt{R_x^2 + R_y^2}. \tag{5.262}$$

Using Eq. (5.83),

$$\boxed{\widehat{V}_a^{\text{proj}}(R) = \frac{2\pi\hbar^2}{m_0 e} f_e(R).} \tag{5.263}$$

Hence,

$$\boxed{V_a^{\text{proj}}(\boldsymbol{\rho}) = \frac{2\pi\hbar^2}{m_0 e}\mathcal{F}_2^{-1}\left\{f_e(\mathbf{R})\right\}.} \tag{5.264}$$

Thus a tabulated electron atomic scattering factor specifies directly the Fourier transform of the projected atomic potential.

For a spherically symmetric atom, the restriction of the three-dimensional scattering factor to the plane $R_z = 0$ is azimuthally symmetric. In the projected plane we write

$$\mathbf{R} = (R_x, R_y), \qquad R = \|\mathbf{R}\|, \tag{5.265}$$

and similarly

$$\boldsymbol{\rho} = (x, y), \qquad \rho = \|\boldsymbol{\rho}\|. \tag{5.266}$$

Equation (5.264) then reads

$$V_a^{\text{proj}}(\boldsymbol{\rho}) = \frac{2\pi\hbar^2}{m_0 e} \int_{\mathbb{R}^2} f_e(\mathbf{R}) \exp\left(j2\pi\mathbf{R}\cdot\boldsymbol{\rho}\right)\, d^2R. \tag{5.267}$$

Because $f_e(\mathbf{R}) = f_e(R)$ depends only on the magnitude of $\mathbf{R}$, this two-dimensional inverse Fourier transform can be reduced to a one-dimensional radial integral.

Reduction of the projected potential to a Hankel transform.

Introduce polar coordinates in the reciprocal plane,

$$\mathbf{R} = R(\cos\varphi, \sin\varphi), \tag{5.268}$$

for which

$$d^2R = R\, dR\, d\varphi. \tag{5.269}$$

Because the scattering factor is rotationally symmetric, we may choose the transverse coordinate system so that $\boldsymbol{\rho}$ lies along the x direction. Then

$$\mathbf{R}\cdot\boldsymbol{\rho} = R\rho\cos\varphi. \tag{5.270}$$

Substitution into Eq. (5.267) gives

$$V_a^{\text{proj}}(\rho) = \frac{2\pi\hbar^2}{m_0 e} \int_0^\infty \int_0^{2\pi} f_e(R) \exp\left(j2\pi R\rho\cos\varphi\right) R\, d\varphi\, dR \tag{5.271}$$

$$= \frac{2\pi\hbar^2}{m_0 e} \int_0^\infty f_e(R) \left[\int_0^{2\pi} \exp\left(j2\pi R\rho\cos\varphi\right) d\varphi\right] R\, dR. \tag{5.272}$$

The zeroth-order Bessel function of the first kind may be defined by

$$\boxed{J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} \exp\left(jx\cos\varphi\right)\, d\varphi.} \tag{5.273}$$

Therefore,

$$\int_0^{2\pi} \exp\left(j2\pi R\rho\cos\varphi\right)\, d\varphi = 2\pi J_0(2\pi R\rho). \tag{5.274}$$

Substitution gives

$$\boxed{V_a^{\text{proj}}(\rho) = \frac{2\pi\hbar^2}{m_0 e}\, 2\pi \int_0^\infty f_e(R) J_0(2\pi R\rho) R\, dR.} \tag{5.275}$$

Equation (5.275) is called a zeroth-order Hankel transform because a two-dimensional Fourier transform of a rotationally symmetric function reduces to an integral with the kernel $J_0(2\pi R\rho)$.

Gaussian scattering factors. We can now evaluate Eq. (5.275) analytically for the Gaussian scattering-factor model introduced in Eq. (5.149),

$$f_e(R) = \sum_{i=1}^{N_G} a_i \exp(-b_i R^2). \tag{5.276}$$

The calculation is the two-dimensional counterpart of Derivation 5.3.5. There, the inverse three-dimensional Fourier transform of each Gaussian term was obtained by applying the one-dimensional identity Eq. (5.171) independently along the three Cartesian directions. For the projected potential only the two transverse directions remain.

Projected potential for a Gaussian scattering factor.

Consider one Gaussian term,

$$f_i(\mathbf{R}) = a_i \exp(-b_i R^2), \tag{5.277}$$

where now

$$\mathbf{R} = (R_x, R_y), \qquad R^2 = R_x^2 + R_y^2. \tag{5.278}$$

Its two-dimensional inverse Fourier transform is

$$\mathcal{F}_2^{-1}\left\{f_i(\mathbf{R})\right\} = a_i \int_{\mathbb{R}^2} \exp(-b_i R^2) \exp\left(j2\pi \mathbf{R}\cdot\boldsymbol{\rho}\right)\, d^2R. \tag{5.279}$$

As in Derivation 5.3.5, the integral separates into one factor for each

Cartesian direction. Applying Eq. (5.171) to R_x and R_y therefore gives

$$\mathcal{F}_2^{-1}\left\{f_i(\mathbf{R})\right\} = a_i\left(\sqrt{\frac{\pi}{b_i}}\right)^2 \exp\left[-\frac{\pi^2}{b_i}\left(x^2+y^2\right)\right] \tag{5.280}$$

$$= a_i\frac{\pi}{b_i}\exp\left(-\frac{\pi^2\rho^2}{b_i}\right), \tag{5.281}$$

where

$$\rho^2 = x^2 + y^2. \tag{5.282}$$

By linearity,

$$\boxed{\mathcal{F}_2^{-1}\left\{f_e(\mathbf{R})\right\} = \sum_{i=1}^{N_G} a_i\frac{\pi}{b_i}\exp\left(-\frac{\pi^2\rho^2}{b_i}\right).} \tag{5.283}$$

Substituting this result into Eq. (5.264) yields

$$\boxed{V_a^{\mathrm{proj}}(\rho) = \frac{2\pi\hbar^2}{m_0 e}\sum_{i=1}^{N_G} a_i\frac{\pi}{b_i}\exp\left(-\frac{\pi^2\rho^2}{b_i}\right).} \tag{5.284}$$

The difference between the three-dimensional result of Eq. (5.175) and the projected result of Eq. (5.284) is therefore only the number of Cartesian Fourier integrations. In three dimensions, each Gaussian acquires the factor

$$\left(\frac{\pi}{b_i}\right)^{3/2}, \tag{5.285}$$

whereas in two dimensions the corresponding factor is

$$\frac{\pi}{b_i}. \tag{5.286}$$

This is consistent with the projection theorem derived above: integrating the three-dimensional potential along z is equivalent to evaluating its Fourier transform in the plane

$$R_z = 0. \tag{5.287}$$

The units provide a final consistency check. Since

$$[f_e] = \mathrm{m} \tag{5.288}$$

and

$$[d^2R] = \mathrm{m}^{-2}, \tag{5.289}$$

the inverse two-dimensional Fourier transform has units

$$\mathrm{m}^{-1}. \tag{5.290}$$

Multiplication by

$$\left[\frac{2\pi\hbar^2}{m_0 e}\right] = \mathrm{V\,m}^2 \tag{5.291}$$

then gives

$$\boxed{[V_a^{\mathrm{proj}}] = \mathrm{V\,m},} \tag{5.292}$$

as required by Eq. (5.240).

5.3.7 From atoms to a biological macromolecule: the independent-atom approximation

The previous sections developed several equivalent descriptions of an isolated atom:

$$\rho_a(\mathbf{r}) \quad \longleftrightarrow \quad V_a(\mathbf{r}) \quad \longleftrightarrow \quad f_{e,a}(\mathbf{R}), \tag{5.293}$$

together with the corresponding projected potential $V_a^{\mathrm{proj}}(\boldsymbol{\rho})$.

A biological macromolecule, however, contains many atoms. In principle, its electrostatic potential should be obtained from the complete electronic state of the molecule: the molecular charge density would be evaluated from Eq. (5.23), and the corresponding potential would then follow from Poisson's equation, Eq. (5.32).

Such a calculation retains the redistribution of electronic charge caused by chemical bonding, ionization, protonation, polarization, and the molecular environment. It is also considerably more demanding than the atomic description developed above.

The traditional approximation used in electron microscopy is therefore the *independent-atom model* (IAM). Each atom is represented by the electrostatic potential of an isolated atom of the same chemical element, usually taken to be neutral and spherically symmetric, and the molecular potential is constructed by linear superposition.

Let

$$\mathbf{r}_a = (\boldsymbol{\rho}_a, z_a) \tag{5.294}$$

be the position of atom a. The IAM potential is

$$\boxed{V_{\mathrm{IAM}}(\mathbf{r}) = \sum_{a=1}^{N_{\mathrm{at}}} V_a^{(0)}\left(\mathbf{r} - \mathbf{r}_a\right),} \tag{5.295}$$

where $V_a^{(0)}$ is the isolated-atom reference potential appropriate to the chemical element of atom a. The superscript (0) emphasizes that the reference state is normally the neutral isolated atom.

Equation (5.295) is the defining approximation of the classical IAM. It assumes that the molecular potential can be obtained simply by translating and adding isolated atomic potentials; any redistribution of electronic charge when the atoms form the molecule is neglected.

It is important, however, that the computational structure of the IAM does not require all atoms of the same element to have the same potential. A useful generalization is to divide the atoms into *chemical types* and assign a different atom-centered potential to each type. Let

$$t(a) \tag{5.296}$$

denote the type assigned to atom a. We may then write

$$\boxed{V_{\mathrm{gIAM}}(\mathbf{r}) = \sum_{a=1}^{N_{\mathrm{at}}} V_{t(a)}\left(\mathbf{r} - \mathbf{r}_a\right),} \tag{5.297}$$

where $V_{t(a)}$ is the potential associated with the chemical environment of atom a.

The ordinary IAM is recovered as the special case in which the atom type is determined only by the chemical element:

$$t(a) = \mathrm{element}(a). \tag{5.298}$$

Thus all carbon atoms, for example, receive the same atomic potential regardless of whether they belong to a methyl group, an aromatic ring, a carbonyl group, or some other chemical environment.

The generalized IAM allows this restriction to be relaxed while retaining the computationally convenient sum of atom-centered contributions. In reciprocal space, Eq. (5.297) becomes

$$\boxed{\widehat{V}_{\mathrm{gIAM}}(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} \sum_{c=1}^{N_{\mathrm{type}}} f_{e,c}(R) \sum_{a\in\mathcal{A}_c} \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}_a\right),} \tag{5.299}$$

where $\mathcal{A}_c$ is the set of atoms assigned to type c and $f_{e,c}(R)$ is the scattering factor associated with that type.

This form makes clear what has changed with respect to the classical IAM. The geometrical part,

$$\sum_{a\in\mathcal{A}_c} \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}_a\right), \tag{5.300}$$

is still determined entirely by the atomic coordinates, while the radial scattering factor $f_{e,c}(R)$ is allowed to depend on the chemical environment.

A recent example of this approach was developed by Shtyrov et al. Shtyrov et al. (2026). They refer to any model of the form of Eq. (5.297), in which atoms are grouped into transferable types with type-dependent scattering factors, as a *generalized independent-atom model*. In their particular atom typing scheme, atoms were required first to belong to the same chemical element and, in addition, to be covalently bonded to the same set of elements. Separate classes were also introduced for chemically distinct oxygen environments, such as the primary amide oxygens of Asn and Gln and the carboxylate oxygens of Asp and Glu.

The motivation is illustrated schematically by

classical IAM		generalized IAM
element	$\longrightarrow$	element + local chemical environment
$f_{e,\mathrm{C}}(R)$	$\longrightarrow$	$f_{e,\mathrm{C,type\,1}}(R), f_{e,\mathrm{C,type\,2}}(R), \ldots$

(5.301)

The generalized model is therefore able to absorb, at least empirically, part of the redistribution of valence charge produced by covalent bonding without requiring a full quantum-mechanical calculation of the complete macromolecule.

Shtyrov et al. (2026) took this idea one step further: rather than calculating the type-dependent scattering factors from an electronic-structure calculation, they inferred them directly from high-resolution cryo-EM reconstructions using a Bayesian statistical model. The resulting scattering factors were transferable between molecules and improved agreement with independent cryo-EM maps, particularly at low and intermediate spatial frequencies, where bonding and charge redistribution have their strongest effect.

This example is useful because it shows that the IAM should not be viewed as an all-or-nothing approximation. There is a natural hierarchy of

increasingly expressive models:

$$\boxed{\begin{array}{c}\text{neutral spherical IAM}\\ \Downarrow\\ \text{ionic or partially charged atomic factors}\\ \Downarrow\\ \text{environment-dependent generalized IAM}\\ \Downarrow\\ \text{aspherical / multipolar atomic models}\\ \Downarrow\\ \text{full molecular electronic-structure calculation.}\end{array}} \tag{5.302}$$

All these models ultimately seek to approximate the same physical quantity, the molecular electrostatic potential $V(\mathbf{r})$; they differ in how much of the molecular electronic environment is retained in the atom-centered representation.

Reciprocal-space representation. The Fourier-space form follows immediately from the translation property of the Fourier transform. For one translated atomic potential,

$$\mathcal{F}\left\{V_a^{(0)}(\mathbf{r}-\mathbf{r}_a)\right\} = \widehat{V}_a^{(0)}(\mathbf{R})\exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}_a\right). \tag{5.303}$$

Using the relation between atomic potential and electron scattering factor derived in Eq. (5.83),

$$\widehat{V}_a^{(0)}(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} f_{e,a}(R), \tag{5.304}$$

where

$$\mathbf{R} = (R_x, R_y, R_z), \qquad R = \|\mathbf{R}\|, \tag{5.305}$$

the Fourier transform of the IAM potential becomes

$$\boxed{\widehat{V}_{\mathrm{IAM}}(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e}\sum_{a=1}^{N_{\mathrm{at}}} f_{e,a}(R)\exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}_a\right).} \tag{5.306}$$

It is convenient to define

$$\boxed{F_e(\mathbf{R}) \equiv \sum_{a=1}^{N_{\mathrm{at}}} f_{e,a}(R)\exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}_a\right),} \tag{5.307}$$

which is the *electron structure factor* of the IAM model. Equation (5.306) can then be written simply as

$$\boxed{\widehat{V}_{\mathrm{IAM}}(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} F_e(\mathbf{R}).} \tag{5.308}$$

The structure factor separates two pieces of information:

$$\boxed{\underbrace{f_{e,a}(R)}_{\text{scattering strength and shape of atom } a} \qquad \underbrace{e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a}}_{\text{position of atom } a}.} \tag{5.309}$$

Thus atoms of different chemical elements contribute different radial scattering factors, while their positions enter through phase factors.

Equations (5.295) and (5.306) are therefore the real- and reciprocal-space representations of exactly the same IAM potential:

$$\boxed{V_{\mathrm{IAM}}(\mathbf{r}) = \sum_a V_a^{(0)}(\mathbf{r} - \mathbf{r}_a) \quad \xleftrightarrow{\mathcal{F}} \quad \widehat{V}_{\mathrm{IAM}}(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} F_e(\mathbf{R}).} \tag{5.310}$$

Projected IAM potential. For a projection along the z direction, the molecular projected potential is defined, as in Eq. (5.240), by

$$V_{\mathrm{IAM}}^{\mathrm{proj}}(\boldsymbol{\rho}) = \int_{-\infty}^{\infty} V_{\mathrm{IAM}}(\boldsymbol{\rho}, z)\, dz. \tag{5.311}$$

Substituting Eq. (5.295) gives

$$V_{\mathrm{IAM}}^{\mathrm{proj}}(\boldsymbol{\rho}) = \sum_a \int_{-\infty}^{\infty} V_a^{(0)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_a, z - z_a\right)\, dz. \tag{5.312}$$

For each atom, introduce the translated coordinate

$$z' = z - z_a, \qquad dz' = dz. \tag{5.313}$$

Since the integration limits remain $-\infty$ and $+\infty$,

$$\int_{-\infty}^{\infty} V_a^{(0)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_a, z - z_a\right)\, dz = \int_{-\infty}^{\infty} V_a^{(0)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_a, z'\right)\, dz' \tag{5.314}$$

$$= V_a^{\mathrm{proj}}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_a\right). \tag{5.315}$$

Therefore,

$$\boxed{V_{\mathrm{IAM}}^{\mathrm{proj}}(\boldsymbol{\rho}) = \sum_{a=1}^{N_{\mathrm{at}}} V_a^{\mathrm{proj}}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_a\right).} \tag{5.316}$$

This derivation makes explicit why the depth coordinate z_a no longer appears: translating an atomic potential along the direction over which it is integrated does not change its complete line integral.

Projected potential in Fourier space. The same result follows from the projection theorem derived in Eq. (5.261). Projection along z selects the plane

$$R_z = 0 \tag{5.317}$$

of the three-dimensional Fourier transform.

Evaluating Eq. (5.306) at $R_z = 0$ gives

$$\widehat{V}_{\mathrm{IAM}}^{\mathrm{proj}}(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} \sum_a f_{e,a}(R) \exp\left[-j2\pi\left(\mathbf{R}\cdot\boldsymbol{\rho}_a + R_z z_a\right)\right]_{R_z=0}, \tag{5.318}$$

and therefore

$$\boxed{\widehat{V}_{\mathrm{IAM}}^{\mathrm{proj}}(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} \sum_a f_{e,a}(R) \exp\left(-j2\pi\mathbf{R}\cdot\boldsymbol{\rho}_a\right).} \tag{5.319}$$

Here

$$\mathbf{R} = (R_x, R_y) \tag{5.320}$$

is the two-dimensional spatial-frequency vector in the projected plane.

The disappearance of z_a can therefore be understood in two equivalent ways:

$$\boxed{\begin{array}{lcl} \text{real space} & : & \displaystyle\int_{-\infty}^{\infty} V_a(z - z_a)\,dz \text{ is invariant under translation in } z, \\ \text{Fourier space} & : & \text{projection selects } R_z = 0, \text{ so } e^{-j2\pi R_z z_a} = 1. \end{array}} \tag{5.321}$$

This loss of depth information is a consequence of the projection model, not a fundamental property of electron–specimen scattering. Once propagation through the finite thickness of the specimen is retained, atoms located at different values of z_a need no longer be equivalent. The Ewald-sphere and multislice descriptions introduced later will restore this depth dependence.

5.3.8 What does the independent-atom approximation mean for a protein?

For a biological macromolecule, the IAM is an extraordinarily useful approximation. Most atoms are H, C, N, and O, with smaller numbers of P and S. Depending on the specimen, one may additionally encounter Na, K, Mg, Ca, Cl, Fe, Zn, Cu, Mn, and other ions or metal centers. Vitreous ice contributes a very large additional population of H and O atoms.

The attraction of the IAM is obvious. Once a small table of atomic potentials or electron scattering factors is available, an arbitrary protein can be converted into a three-dimensional specimen potential at negligible cost compared with a complete electronic-structure calculation. This is why independent-atom potentials have been so successful in electron-microscopy simulation Kirkland (2010)[Chap. 5].

Equation (5.295), however, also states precisely what the approximation discards.

In a real molecule, valence electrons do not remain attached to isolated spherical atoms. Their distribution is modified by covalent bonds, hydrogen bonds, aromatic delocalization, formal and partial charges, protonation states, coordination with metal ions, and the electrostatic environment generated by the remainder of the macromolecule and solvent.

Consequently, the nuclear coordinates may remain almost unchanged while the electronic charge density—and therefore $V(\mathbf{r})$—changes measurably.

The importance of this effect depends strongly on spatial frequency. As shown by the Mott–Bethe relation, Eq. (5.113), high-frequency electron scattering becomes increasingly dominated by the nuclear charge, whereas low-frequency scattering depends sensitively on the cancellation between nuclear and electronic contributions.

Kirkland estimated that representing bonded low-Z atoms as isolated, spherically symmetric atoms can produce errors of order 5–10% in low-angle electron scattering, while high-angle scattering is less affected Kirkland (2010)[Chap. 5]. This is particularly relevant for cryo-EM because low and intermediate spatial frequencies contain substantial information about molecular shape, interfaces, and overall image contrast.

Several more detailed models have therefore been proposed.

Yonekura and Maki-Yonekura showed that charged-atom electron scattering factors can improve refinement against electron-scattering data (Yonekura and Maki-Yonekura, 2016). Yonekura et al. subsequently calculated ionic scattering factors for biologically important elements, including C, N, O,

P, S, and H, using relativistic multi-configuration calculations, and found substantial differences between neutral and charged atoms over spatial-frequency ranges relevant to structural biology (Yonekura et al., 2018).

A more complete alternative is the *transferable aspherical atom model* (TAAM). Rather than assigning one neutral spherical density to every atom of a given element, TAAM uses environment-dependent multipolar pseudoatoms. Kulik, Chodkiewicz, and Dominiak showed that such models reproduce molecular electrostatic potentials of proteins more accurately than IAM, particularly around charged residues and chemically anisotropic environments (Kulik et al., 2022).

More recent calculations have emphasized a result that may at first seem counterintuitive: improving the description of valence charge does not necessarily become less important as the nominal resolution becomes poorer. Kulik and Dominiak found that electrostatic-potential Fourier maps remain sensitive to charge redistribution at resolutions of several angstroms and that the relative differences between TAAM and IAM can become particularly large at low spatial frequencies around charged groups (Kulik and Dominiak, 2025). This behavior follows naturally from the nuclear–electronic cancellation in Eq. (5.113).

Most directly relevant to single-particle cryo-EM, Shtyrov et al. estimated environment-dependent atomic scattering factors from experimental SPA electrostatic-potential maps (Shtyrov et al., 2026). Their results indicate that scattering factors inferred for atoms in their molecular environment provide a measurably better description of experimental cryo-EM maps than conventional environment-independent neutral-atom factors. Thus chemical bonding and molecular environment can leave observable signatures in sufficiently accurate SPA reconstructions.

A separate issue should not be confused with the scattering-factor model itself. The atomic coordinates may also depend on the physical state in which the specimen was measured. Ravikumar et al. reported systematically more dispersed side-chain packing in cryo-EM-derived protein structures than in X-ray structures at comparable resolution (Ravikumar et al., 2022).

Consequently, a cryo-EM forward model constructed from an X-ray-derived atomic structure can differ from the actual cryogenic specimen for

two conceptually distinct reasons:

$$\boxed{\begin{array}{c}\text{the atomic coordinates may differ,}\\ \text{and}\\ \text{the electrostatic potential assigned to those coordinates may differ.}\end{array}} \tag{5.322}$$

5.3.9 The specimen potential as the interface to image formation

The preceding sections have described several levels at which the elastic electrostatic potential of a specimen may be modeled. In the simplest independent-atom approximation of Sec. 5.3.7, isolated atomic potentials are superposed. The generalized IAM introduced there allows the atomic contribution to depend on the local chemical environment. More detailed descriptions may incorporate ionic states, charge redistribution, multipolar densities, or ultimately obtain the electrostatic potential from a molecular electronic-structure calculation.

These alternatives differ in how accurately they describe the specimen, but they all serve the same purpose in the image-formation problem: they provide an electrostatic potential

$$\boxed{V(\mathbf{r}).} \tag{5.323}$$

It is useful to regard this quantity as the interface between two levels of description. On one side lies the microscopic quantum state of the specimen; on the other lies the propagation of the imaging electron through the resulting electrostatic field.

The connection developed in this chapter can be summarized as

$$\boxed{\begin{array}{c}|\Xi_0\rangle_{\mathrm{s}}\\ \text{specimen quantum state}\\ \Downarrow\\ \rho(\mathbf{r}) = {}_{\mathrm{s}}\langle\Xi_0|\hat{\rho}_{\mathrm{s}}(\mathbf{r})|\Xi_0\rangle_{\mathrm{s}}\\ \text{elastic charge density}\\ \Downarrow\\ V(\mathbf{r})\\ \text{elastic electrostatic potential.}\end{array}} \tag{5.324}$$

The first step, Eq. (5.324), compresses the many-body quantum state into the expectation value of the specimen charge-density operator, as derived in Sec. 5.2. The second step is governed by Poisson's equation, Eq. (5.32),

$$\nabla^2 V(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\varepsilon_0}. \tag{5.325}$$

The many-body quantum mechanics has therefore not disappeared. Rather, for elastic scattering it has been condensed into the ordinary function $V(\mathbf{r})$ that acts on the imaging electron.

For an isolated spherical atom, the same information can be represented in reciprocal space by the electron atomic scattering factor. From Eq. (5.83),

$$\boxed{\widehat{V}_a(\mathbf{R}) = \frac{2\pi\hbar^2}{m_0 e} f_e(R),} \tag{5.326}$$

where

$$R = \|\mathbf{R}\|. \tag{5.327}$$

Thus $f_e(R)$ is not another physical interaction added to the electrostatic potential. It is a convenient reciprocal-space representation of the same isolated-atom potential.

For a macromolecule, the potential may then be assembled approximately from atomic contributions, as in the IAM or generalized IAM of Sec. 5.3.7, or obtained directly from a more complete electronic-structure model. Schematically,

$$\boxed{\begin{array}{c}\text{atomic or electronic-structure model} \quad \Longrightarrow \quad V(\mathbf{r}) \\ \text{IAM, generalized IAM,} \\ \text{or molecular electronic structure}\end{array}} \tag{5.328}$$

Once $V(\mathbf{r})$ has been specified, the details of how it was constructed no longer alter the form of the elastic propagation equations. They alter the numerical value and spatial structure of the potential, but not the mathematical machinery used to propagate the electron wave through it.

There remains, however, an important choice in how much of the three-dimensional potential is retained. Under the projection approximation,

$$\boxed{V^{\text{proj}}(\boldsymbol{\rho}) = \int_{-\infty}^{\infty} V(\boldsymbol{\rho}, z)\, dz,} \tag{5.329}$$

so the complete depth dependence is replaced by a two-dimensional projected potential. As shown in Sec. 5.3.7, this eliminates the individual atomic coordinates z_a from the projected IAM model.

More complete propagation theories retain

$$V(\boldsymbol{\rho}, z) \tag{5.330}$$

throughout the specimen and therefore preserve the ordering of atoms along the beam direction. This distinction will become essential when we move from a simple projection model to three-dimensional scattering and multislice propagation.

The problem can therefore be separated conceptually into two stages:

$$\boxed{\underbrace{\text{specimen physics} \longrightarrow V(\mathbf{r})}_{\text{developed in the present sections}} \qquad \longrightarrow \qquad \underbrace{\text{electron scattering and propagation}}_{\text{developed next}}.} \tag{5.331}$$

We now turn to the second stage. Given an incident electron and a known potential $V(\mathbf{r})$, we ask how the electron is scattered by that potential. This will introduce the momentum and spatial-frequency transfer associated with an elastic event, the scattering amplitude, the first Born approximation and its range of validity, and the Ewald sphere. These concepts will then lead naturally to propagation models that retain the full depth dependence of the specimen, culminating in the multislice description.

5.4 Elastic scattering by a static potential

The previous sections answered the first part of the specimen image-formation problem: given a microscopic description of the specimen, how do we obtain the electrostatic potential

$$V(\mathbf{r}) \tag{5.332}$$

experienced by the imaging electron? We now ask the complementary question: given this potential, how does an incident electron scatter from it?

The purpose of the present section is not to develop scattering theory in full generality. Rather, we need three results that will be used throughout the remainder of the book:

1. the geometrical relation between the incident and scattered electron wavevectors;

2. the connection between the scattering amplitude and the differential scattering cross section;

3. the first-Born relation between the scattering amplitude and the Fourier transform of the electrostatic potential.

The last result will provide the physical justification for the electron scattering factor introduced previously in Eq. (5.82).

5.4.1 Incident and scattered electron waves

Consider first an electron incident on a localized, static electrostatic potential $V(\mathbf{r})$. Far from the specimen, before the interaction takes place, the incident electron may be represented by the plane wave

$$\boxed{\psi_{\mathrm{i}}(\mathbf{r}) = \exp\left(j\mathbf{k}_{\mathrm{i}} \cdot \mathbf{r}\right),} \tag{5.333}$$

where $\mathbf{k}_{\mathrm{i}}$ is the incident electron wavevector.

As established previously,

$$k_{\mathrm{i}} = \|\mathbf{k}_{\mathrm{i}}\| = \frac{2\pi}{\lambda}, \tag{5.334}$$

and $\mathbf{k}$ is measured in rad·m^{-1}.

A plane-wave component emerging in a particular scattered direction has wavevector $\mathbf{k}_{\mathrm{f}}$,

$$\psi_{\mathrm{f}}(\mathbf{r}) \propto \exp\left(j\mathbf{k}_{\mathrm{f}} \cdot \mathbf{r}\right). \tag{5.335}$$

Because the potential is static and the scattering event is elastic, the electron does not exchange energy with the specimen. Consequently,

$$E_{\mathrm{f}} = E_{\mathrm{i}}, \tag{5.336}$$

and therefore the wavelength is unchanged:

$$\boxed{k_{\mathrm{f}} = k_{\mathrm{i}} = k = \frac{2\pi}{\lambda}.} \tag{5.337}$$

Elastic scattering therefore changes the *direction* of the wavevector but not its magnitude.

Wavevector and momentum transfer. The incident and scattered waves are characterized by the wavevectors $\mathbf{k}_{\mathrm{i}}$ and $\mathbf{k}_{\mathrm{f}}$. Their difference,

$$\boxed{\mathbf{q} \equiv \mathbf{k}_{\mathrm{f}} - \mathbf{k}_{\mathrm{i}},} \tag{5.338}$$

is the *wavevector transfer* associated with the scattering event.

The relation between angular wavevector and spatial frequency was established previously in Eq. (3.6),

$$\mathbf{k} = 2\pi\mathbf{R}, \tag{5.339}$$

while the connection between wavevector and mechanical momentum in a field-free region was derived in the electron-optics chapter, Eq. (4.671),

$$\mathbf{p} = \hbar\mathbf{k}. \tag{5.340}$$

We therefore introduce the spatial-frequency transfer

$$\boxed{\mathbf{R} \equiv \frac{\mathbf{q}}{2\pi} = \frac{\mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}}{2\pi}.} \tag{5.341}$$

The same quantity may equivalently be expressed through the momentum transfer:

$$\begin{aligned} \Delta\mathbf{p} &= \mathbf{p}_\mathrm{f} - \mathbf{p}_\mathrm{i} && (5.342)\\ &= \hbar\left(\mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}\right) && (5.343)\\ &= \hbar\mathbf{q} = h\mathbf{R}. && (5.344) \end{aligned}$$

Hence,

$$\boxed{\Delta\mathbf{p} = \hbar\mathbf{q} = h\mathbf{R}.} \tag{5.345}$$

Thus the same scattering event can be described in three equivalent ways:

$\Delta\mathbf{p}$	$\longleftrightarrow$	$\mathbf{q}$	$\longleftrightarrow$	$\mathbf{R}$
momentum transfer		wavevector transfer		spatial-frequency transfer
$\mathrm{kg\,m\,s^{-1}}$		$\mathrm{rad\,m^{-1}}$		$\mathrm{cycles\,m^{-1}}$.

(5.346)

Relation to the scattering angle. Let Θ denote the angle between the incident and scattered wavevectors,

$$\cos\Theta = \frac{\mathbf{k}_\mathrm{i}\cdot\mathbf{k}_\mathrm{f}}{k^2}. \tag{5.347}$$

Using Eq. (5.337),

$$\begin{aligned} q^2 &= \left|\mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}\right|^2 && (5.348)\\ &= k^2 + k^2 - 2k^2\cos\Theta && (5.349)\\ &= 2k^2(1-\cos\Theta) && (5.350)\\ &= 4k^2\sin^2\left(\frac{\Theta}{2}\right). && (5.351) \end{aligned}$$

Hence,

$$\boxed{q = 2k \sin\left(\frac{\Theta}{2}\right),} \tag{5.352}$$

and, because $k = 2\pi/\lambda$,

$$\boxed{R = \frac{2}{\lambda} \sin\left(\frac{\Theta}{2}\right).} \tag{5.353}$$

For the small scattering angles characteristic of much of cryo-EM,

$$\sin\left(\frac{\Theta}{2}\right) \simeq \frac{\Theta}{2}, \tag{5.354}$$

so

$$\boxed{R \simeq \frac{\Theta}{\lambda}.} \tag{5.355}$$

Thus progressively larger scattering angles correspond to progressively higher spatial frequencies of the specimen.

5.4.2 The scattering state and the scattering amplitude

The wave equation that governs the electron inside an electrostatic field has already been developed in the electron-optics chapter. In Sec. 4.8.3, we showed that, in a region where the magnetic vector potential vanishes, a stationary electron satisfies a local Helmholtz equation. In particular, Eqs. (4.916) and (4.918) may be written as

$$\boxed{\nabla^2 \psi(\mathbf{r}) + k^2(\mathbf{r})\psi(\mathbf{r}) = 0, \qquad k(\mathbf{r}) = \frac{p(\mathbf{r})}{\hbar}.} \tag{5.356}$$

Thus the specimen does not require a new wave equation. Its effect is to modify the local electron momentum, and hence the local wavenumber, through the electrostatic potential $V(\mathbf{r})$ constructed in the preceding sections.

Far from the specimen the electrostatic potential produced by the specimen vanishes. The momentum magnitude then approaches the constant incident value p, and

$$k(\mathbf{r}) \longrightarrow k = \frac{p}{\hbar}. \tag{5.357}$$

Equation (5.356) consequently reduces to the field-free electron Helmholtz equation Eq. (3.159),

$$\left(\nabla^2 + k^2\right)\psi = 0. \tag{5.358}$$

Inside the specimen, the local wavenumber $k(\mathbf{r})$ differs from the asymptotic field-free value k. Starting from the local Helmholtz equation derived in the electron-optics chapter, Eq. (5.356),

$$\nabla^2\psi(\mathbf{r}) + k^2(\mathbf{r})\psi(\mathbf{r}) = 0, \tag{5.359}$$

we add and subtract the constant quantity $k^2\psi(\mathbf{r})$:

$$\nabla^2\psi(\mathbf{r}) + k^2\psi(\mathbf{r}) + \left[k^2(\mathbf{r}) - k^2\right]\psi(\mathbf{r}) = 0. \tag{5.360}$$

Therefore,

$$\boxed{\left(\nabla^2 + k^2\right)\psi(\mathbf{r}) = -\left[k^2(\mathbf{r}) - k^2\right]\psi(\mathbf{r}).} \tag{5.361}$$

No approximation has been introduced in this rewriting. We have merely separated the field-free Helmholtz operator,

$$\nabla^2 + k^2, \tag{5.362}$$

from the change in local wavenumber produced by the specimen.

The quantity

$$-\left[k^2(\mathbf{r}) - k^2\right]\psi(\mathbf{r}) \tag{5.363}$$

plays the role of an *effective scattering source* for the field-free Helmholtz operator. It is not an independently prescribed source, because it contains the unknown total wavefunction $\psi(\mathbf{r})$. This is why the resulting scattering equation is still exact and implicit.

The right-hand side vanishes wherever the local electron wavenumber equals its asymptotic field-free value. Thus it is the departure of $k(\mathbf{r})$ from k, produced by the specimen potential, that generates the scattered component.

Relation to the electrostatic specimen potential. The relation between $k(\mathbf{r})$ and the electrostatic potential also follows directly from the relativistic electron-optical derivation of Eq. (4.910).

Let the specimen potential vanish sufficiently far from the specimen, and let $\mathcal{E}_0$ denote the conserved relativistic total energy of the incident electron in that region. Inside the specimen, Eq. (4.910) gives

$$p^2(\mathbf{r}) = \frac{\left[\mathcal{E}_0 + eV(\mathbf{r})\right]^2 - m_0^2c^4}{c^2}, \tag{5.364}$$

whereas outside the specimen

$$p^2 = \frac{\mathcal{E}_0^2 - m_0^2 c^4}{c^2}. \tag{5.365}$$

Subtracting,

$$p^2(\mathbf{r}) - p^2 = \frac{2\mathcal{E}_0 eV(\mathbf{r}) + e^2 V^2(\mathbf{r})}{c^2}. \tag{5.366}$$

For a fast imaging electron, the electrostatic energy associated with the specimen is very small compared with the total electron energy,

$$|eV| \ll \mathcal{E}_0. \tag{5.367}$$

The term quadratic in V may therefore be neglected. Introducing the relativistic mass

$$m = \frac{\mathcal{E}_0}{c^2} = \gamma m_0, \tag{5.368}$$

we obtain

$$p^2(\mathbf{r}) - p^2 \simeq 2meV(\mathbf{r}). \tag{5.369}$$

Since $k = p/\hbar$ according to Eq. (4.671),

$$\boxed{k^2(\mathbf{r}) - k^2 \simeq \frac{2me}{\hbar^2} V(\mathbf{r}).} \tag{5.370}$$

Substitution into Eq. (5.361) gives the standard fast-electron scattering equation

$$\boxed{\left(\nabla^2 + k^2\right)\psi(\mathbf{r}) = -\frac{2me}{\hbar^2} V(\mathbf{r})\psi(\mathbf{r}).} \tag{5.371}$$

This is the same electrostatic potential $V(\mathbf{r})$ constructed in the preceding sections. Equivalently, using the potential energy $U_{\text{stat}} = -eV$ derived in Eq. (5.26), Eq. (5.371) may be written as

$$\left(\nabla^2 + k^2\right)\psi(\mathbf{r}) = \frac{2m}{\hbar^2} U_{\text{stat}}(\mathbf{r})\psi(\mathbf{r}). \tag{5.372}$$

No Born approximation has been made here. The unknown total wavefunction ψ still appears on the right-hand side.

Incident and scattered components. The scattering experiment is specified by prescribing an incoming free-electron state. For the plane-wave illumination introduced in Eq. (5.333),

$$\psi_{\mathrm{i}}(\mathbf{r}) = \exp\left(j\mathbf{k}_{\mathrm{i}} \cdot \mathbf{r}\right), \tag{5.373}$$

and, because $|\mathbf{k}_{\mathrm{i}}| = k$, this wave satisfies the field-free Helmholtz equation

$$\boxed{\left(\nabla^2 + k^2\right)\psi_{\mathrm{i}}(\mathbf{r}) = 0.} \tag{5.374}$$

We now *define* the scattered component by

$$\boxed{\psi_{\mathrm{s}}(\mathbf{r}) \equiv \psi(\mathbf{r}) - \psi_{\mathrm{i}}(\mathbf{r}),} \tag{5.375}$$

or equivalently

$$\psi = \psi_{\mathrm{i}} + \psi_{\mathrm{s}}. \tag{5.376}$$

This is a decomposition of a single quantum state, not a statement that one part of the electron passes through the specimen unaffected while another part is created. The prescribed incident wave specifies the incoming boundary condition; ψ_{s} is the complete outgoing component generated by the interaction with the specimen.

Subtracting Eq. (5.374) from Eq. (5.371) gives

$$\boxed{\left(\nabla^2 + k^2\right)\psi_{\mathrm{s}}(\mathbf{r}) = -\frac{2me}{\hbar^2}V(\mathbf{r})\psi(\mathbf{r}).} \tag{5.377}$$

Equation (5.377) gives a precise meaning to the statement that the specimen generates the scattered wave. The source is not simply $V(\mathbf{r})$: it is the product

$$V(\mathbf{r})\psi(\mathbf{r}), \tag{5.378}$$

because the wave generated at each position depends both on the local potential and on the electron wave that actually reaches that position.

Outgoing solution. We already encountered exactly this mathematical problem in the wave-diffraction chapter. The outgoing Green function of the Helmholtz operator was introduced in Eq. (3.86),

$$G(\mathbf{r}, \mathbf{r}') = \frac{\exp\left(jk\|\mathbf{r} - \mathbf{r}'\|\right)}{\|\mathbf{r} - \mathbf{r}'\|}, \tag{5.379}$$

and satisfies Eq. (3.87),

$$\left(\nabla^2 + k^2\right) G(\mathbf{r}, \mathbf{r}') = -4\pi\delta(\mathbf{r} - \mathbf{r}'). \tag{5.380}$$

The Sommerfeld radiation condition of Eq. (3.202) selects the outgoing solution, as appropriate for a scattering experiment in which no additional waves arrive from infinity.

Applying this Green function to Eq. (5.377) gives

$$\boxed{\psi_{\mathrm{s}}(\mathbf{r}) = \frac{me}{2\pi\hbar^2} \int \frac{\exp\left(jk\|\mathbf{r} - \mathbf{r}'\|\right)}{\|\mathbf{r} - \mathbf{r}'\|} V(\mathbf{r}')\psi(\mathbf{r}')\, d^3r'.} \tag{5.381}$$

Within the scalar static-potential model, Eq. (5.381) is an exact integral form of the scattering problem. It is also implicit: the unknown total wave-function ψ occurs inside the integral. This point will become essential when the first Born approximation is introduced below.

Far-field scattering amplitude. Let

$$r = \|\mathbf{r}\| \tag{5.382}$$

be the distance from the scattering region to the observation point, and define the corresponding unit vector

$$\boxed{\widehat{\mathbf{r}} \equiv \frac{\mathbf{r}}{r}.} \tag{5.383}$$

Thus $\widehat{\mathbf{r}}$ specifies only the observation direction, whereas r specifies the distance to the observation point.

For an observation point far from the spatial extent of the specimen,

$$r \gg \|\mathbf{r}'\|, \tag{5.384}$$

the same far-field expansion used previously in diffraction theory gives

$$\|\mathbf{r} - \mathbf{r}'\| \simeq r - \widehat{\mathbf{r}} \cdot \mathbf{r}'. \tag{5.385}$$

Therefore,

$$\frac{\exp\left(jk\|\mathbf{r} - \mathbf{r}'\|\right)}{\|\mathbf{r} - \mathbf{r}'\|} \simeq \frac{e^{jkr}}{r} \exp\left(-jk\widehat{\mathbf{r}} \cdot \mathbf{r}'\right). \tag{5.386}$$

Since the scattered wave propagating toward that observation direction has wavenumber magnitude k, its outgoing wavevector is

$$\boxed{\mathbf{k}_\mathrm{f} = k\hat{\mathbf{r}}.} \tag{5.387}$$

Substitution of Eq. (5.386) into Eq. (5.381) gives

$$\psi_\mathrm{s}(\mathbf{r}) \underset{r\to\infty}{\sim} \frac{e^{jkr}}{r}\frac{me}{2\pi\hbar^2}\int V(\mathbf{r}')\psi(\mathbf{r}')e^{-j\mathbf{k}_\mathrm{f}\cdot\mathbf{r}'}\,d^3r'. \tag{5.388}$$

This motivates the definition of the *elastic scattering amplitude*,

$$\boxed{f(\mathbf{k}_\mathrm{f},\mathbf{k}_\mathrm{i}) \equiv \frac{me}{2\pi\hbar^2}\int V(\mathbf{r})\psi(\mathbf{r})\exp\left(-j\mathbf{k}_\mathrm{f}\cdot\mathbf{r}\right)\,d^3r.} \tag{5.389}$$

Its dependence on the incident wavevector $\mathbf{k}_\mathrm{i}$ is contained implicitly in the scattering state $\psi(\mathbf{r})$.

The complete scattered component therefore has the asymptotic form

$$\boxed{\psi_\mathrm{s}(\mathbf{r}) \underset{r\to\infty}{\sim} f(\mathbf{k}_\mathrm{f},\mathbf{k}_\mathrm{i})\frac{e^{jkr}}{r}.} \tag{5.390}$$

Combining this result with the prescribed incident plane wave gives the familiar asymptotic scattering state,

$$\boxed{\psi(\mathbf{r}) \underset{r\to\infty}{\sim} \exp\left(j\mathbf{k}_\mathrm{i}\cdot\mathbf{r}\right) + f(\mathbf{k}_\mathrm{f},\mathbf{k}_\mathrm{i})\frac{e^{jkr}}{r}.} \tag{5.391}$$

What the scattering amplitude represents. Equation (5.391) should not be interpreted as describing an unmodified incident electron plus a second electron produced by the specimen. It is the asymptotic decomposition of one stationary scattering state into its prescribed incoming component and its outgoing scattered component.

The scattered wave is also not a single outgoing plane wave. The factor

$$\frac{e^{jkr}}{r} \tag{5.392}$$

describes radial propagation away from the specimen, whereas $f(\mathbf{k}_\mathrm{f},\mathbf{k}_\mathrm{i})$ depends on the observation direction through $\mathbf{k}_\mathrm{f} = k\hat{\mathbf{r}}$. Different directions therefore carry different scattered amplitudes.

With the unit-amplitude normalization adopted for the incident plane wave, ψ is dimensionless. Since e^{jkr}/r has units of inverse length, Eq. (5.390) requires

$$\boxed{[f] = \mathrm{m}.} \tag{5.393}$$

The next step is to relate this amplitude to a measurable scattering probability. We shall show that its squared modulus determines the differential scattering cross section. Only after that will we introduce the first Born approximation, which consists of replacing the unknown total wavefunction ψ inside Eq. (5.389) by the incident plane wave.

5.4.3 From scattering amplitude to differential cross section

The preceding subsection introduced the scattering amplitude $f(\Omega)$ through the far-field form of the elastic scattering state. The amplitude is a complex quantity and therefore contains both magnitude and phase information.

A scattering experiment, however, measures how many electrons emerge within a given range of directions. We must therefore connect the scattering amplitude with the outgoing electron flux.

The relation between an electron wave and its probability flux was already developed in Sec. 3.11. There, the relativistic probability-current density was introduced in Eq. (3.709) and shown to satisfy the continuity equation Eq. (3.714). For a narrow, approximately monochromatic electron beam, the longitudinal flux reduces to

$$J \propto v|\psi|^2, \tag{5.394}$$

as discussed immediately before Eq. (3.724).

Thus, apart from a common normalization factor, the probability flux carried by an electron wave is its probability density multiplied by the electron velocity. This is all that is required to relate the far-field scattering amplitude to a measurable cross section.

The meaning of a differential cross section. Consider an incident beam carrying probability flux J_i. The *differential scattering cross section* $d\sigma/d\Omega$ is defined so that the probability per unit time scattered into a solid-angle element $d\Omega$ is

$$\boxed{d\Gamma_\mathrm{s} = J_\mathrm{i}\frac{d\sigma}{d\Omega}d\Omega.} \tag{5.395}$$

The differential cross section therefore has dimensions of area per unit solid angle and measures the effective area of the incident beam that contributes to scattering into the chosen direction.

Differential cross section from the elastic scattering amplitude.

For the unit-amplitude incident plane wave

$$\psi_{\mathrm{i}}(\mathbf{r}) = e^{j\mathbf{k}_{\mathrm{i}}\cdot\mathbf{r}}, \tag{5.396}$$

the incident probability flux is proportional to

$$J_{\mathrm{i}} \propto v_{\mathrm{i}}. \tag{5.397}$$

The scattered component of the far-field elastic wave was derived in Eq. (5.390). Since the outgoing wavevector $\mathbf{k}_{\mathrm{f}}$ is fixed by the observation direction through Eq. (5.387), we may write the same result as

$$\psi_{\mathrm{s}}(\mathbf{r}) \underset{r\to\infty}{\sim} f(\Omega)\frac{e^{jkr}}{r}. \tag{5.398}$$

Its probability density is therefore

$$|\psi_{\mathrm{s}}(\mathbf{r})|^2 \simeq \frac{|f(\Omega)|^2}{r^2}. \tag{5.399}$$

Using the flux relation discussed in Sec. 3.11, the outward radial probability flux is therefore

$$J_{\mathrm{s},r} \propto v_{\mathrm{f}}\frac{|f(\Omega)|^2}{r^2}. \tag{5.400}$$

A surface element on a sphere of radius r subtending the solid angle $d\Omega$ has area

$$dA = r^2 d\Omega. \tag{5.401}$$

Hence the probability per unit time crossing that surface element is

$$d\Gamma_{\mathrm{s}} = J_{\mathrm{s},r}\, dA \tag{5.402}$$

$$\propto v_{\mathrm{f}} \frac{|f(\Omega)|^2}{r^2} r^2 d\Omega \tag{5.403}$$

$$= v_{\mathrm{f}} |f(\Omega)|^2 d\Omega. \tag{5.404}$$

Dividing by the incident flux, Eq. (5.397), gives

$$\frac{d\sigma}{d\Omega} = \frac{v_{\mathrm{f}}}{v_{\mathrm{i}}} |f(\Omega)|^2. \tag{5.405}$$

For elastic scattering, Eq. (5.337) gives

$$k_{\mathrm{f}} = k_{\mathrm{i}}. \tag{5.406}$$

The initial and final electron energies, momentum magnitudes, and velocities are therefore equal:

$$v_{\mathrm{f}} = v_{\mathrm{i}}. \tag{5.407}$$

Consequently,

$$\boxed{\frac{d\sigma}{d\Omega} = |f(\Omega)|^2.} \tag{5.408}$$

Equation (5.408) also fixes the dimensions of the scattering amplitude. Since a solid angle is dimensionless in SI and a cross section has dimensions of area,

$$\boxed{[f] = \mathrm{m}.} \tag{5.409}$$

This is consistent with the asymptotic scattered wave

$$f(\Omega) \frac{e^{jkr}}{r}, \tag{5.410}$$

because f/r must be dimensionless.

Integrating over all outgoing directions gives the total elastic scattering cross section,

$$\boxed{\sigma_{\mathrm{el}} = \int_{4\pi} |f(\Omega)|^2 d\Omega.} \tag{5.411}$$

The distinction between amplitude and cross section is important. The complex scattering amplitude $f(\Omega)$ retains phase information, whereas the

differential cross section contains only its squared modulus. In a direct scattering measurement, $d\sigma/d\Omega$ determines how many electrons are scattered into a given angular interval. In coherent imaging, however, the phase of $f(\Omega)$ remains essential because scattered waves can interfere with one another and with the unscattered reference wave.

5.4.4 The first Born approximation

The exact scattered wave and scattering amplitude were derived above, in particular in Eqs. (5.381) and (5.389). The measurable consequence of the scattering amplitude was then established in Sec. 5.4.3, where Eq. (5.408) gave

$$\frac{d\sigma}{d\Omega} = |f(\Omega)|^2. \tag{5.412}$$

We now turn to the remaining difficulty: calculating the scattering amplitude itself. The exact expression depends on the unknown total wave inside the specimen. The first Born approximation replaces this self-consistent problem by a first-order one.

In particular, Eq. (5.381) showed that the scattered component is

$$\psi_{\mathrm{s}}(\mathbf{r}) = \frac{me}{2\pi\hbar^2} \int \frac{\exp\left(jk\|\mathbf{r}-\mathbf{r}'\|\right)}{\|\mathbf{r}-\mathbf{r}'\|} V(\mathbf{r}')\psi(\mathbf{r}')\, d^3r'. \tag{5.413}$$

Here $V(\mathbf{r})$ is the electrostatic specimen potential constructed in Sec. 5.3. Its relation to the potential energy of the imaging electron,

$$U(\mathbf{r}) = -eV(\mathbf{r}), \tag{5.414}$$

was established previously in Eq. (A.480).

Using the decomposition of the scattering state, Eq. (5.375),

$$\psi = \psi_{\mathrm{i}} + \psi_{\mathrm{s}}, \tag{5.415}$$

Eq. (5.413) may be written as

$$\boxed{\psi(\mathbf{r}) = \psi_{\mathrm{i}}(\mathbf{r}) + C \int G_0(\mathbf{r},\mathbf{r}')V(\mathbf{r}')\psi(\mathbf{r}')\, d^3r',} \tag{5.416}$$

where, for compactness, we have introduced

$$C \equiv \frac{me}{2\pi\hbar^2} \tag{5.417}$$

and the outgoing free-space Green function

$$G_0(\mathbf{r}, \mathbf{r}') \equiv \frac{\exp\left(jk\|\mathbf{r} - \mathbf{r}'\|\right)}{\|\mathbf{r} - \mathbf{r}'\|}. \tag{5.418}$$

This is the same outgoing Helmholtz Green function introduced previously in Eq. (3.86).

Equation (5.416) is still exact within the scalar model, and it is called the Lippmann–Schwinger equation. Its difficulty is that the wave $\psi(\mathbf{r}')$ illuminating each point of the specimen is itself the unknown wave already modified by the specimen.

Successive substitution and the Born approximation. The origin of the Born approximation is most easily understood as an iterative solution of the exact scattering equation.

Equation (5.416) has the form

$$\boxed{\psi = \psi_{\mathrm{i}} + \mathcal{K}_V \psi,} \tag{5.419}$$

where the operator $\mathcal{K}_V$ is defined by

$$\boxed{\left(\mathcal{K}_V \phi\right)(\mathbf{r}) = C \int G_0(\mathbf{r}, \mathbf{r}') V(\mathbf{r}') \phi(\mathbf{r}')\, d^3 r',} \tag{5.420}$$

with

$$C = \frac{me}{2\pi\hbar^2} \tag{5.421}$$

from Eq. (5.417), and $G_0(\mathbf{r}, \mathbf{r}')$ the outgoing free-space Green function defined in Eq. (5.418).

Equation (5.419) is implicit because the unknown wavefunction appears on both sides. A natural way of solving it is by successive substitution.

We start from the wave that would exist in the absence of the specimen,

$$\boxed{\psi^{(0)} = \psi_{\mathrm{i}}.} \tag{5.422}$$

Using this as a first estimate of the unknown wavefunction on the right-hand side of Eq. (5.419) gives

$$\boxed{\psi^{(1)} = \psi_{\mathrm{i}} + \mathcal{K}_V \psi_{\mathrm{i}}.} \tag{5.423}$$

Substituting this improved estimate once more gives

$$\psi^{(2)} = \psi_{\mathrm{i}} + \mathcal{K}_V \psi^{(1)} \tag{5.424}$$

$$= \psi_{\mathrm{i}} + \mathcal{K}_V \psi_{\mathrm{i}} + \mathcal{K}_V^2 \psi_{\mathrm{i}}. \tag{5.425}$$

A further substitution gives

$$\psi^{(3)} = \psi_{\rm i} + \mathcal{K}_V \psi_{\rm i} + \mathcal{K}_V^2 \psi_{\rm i} + \mathcal{K}_V^3 \psi_{\rm i}, \tag{5.426}$$

and continuing indefinitely produces the formal series

$$\boxed{\psi = \psi_{\rm i} + \mathcal{K}_V \psi_{\rm i} + \mathcal{K}_V^2 \psi_{\rm i} + \mathcal{K}_V^3 \psi_{\rm i} + \cdots .} \tag{5.427}$$

This is the *Born series*. Mathematically, it is the Neumann series associated with Eq. (5.419),

$$\psi = (I - \mathcal{K}_V)^{-1} \psi_{\rm i}, \tag{5.428}$$

with the formal expansion

$$(I - \mathcal{K}_V)^{-1} = I + \mathcal{K}_V + \mathcal{K}_V^2 + \cdots . \tag{5.429}$$

Why may the series be truncated? Successive substitution is useful only if each additional application of $\mathcal{K}_V$ produces a smaller modification of the wave. Let Ω_s denote the volume occupied by the specimen, and measure the magnitude of a wave inside this region by

$$\|\phi\|_{\Omega_s} \equiv \left[\int_{\Omega_s} |\phi(\mathbf{r})|^2 \, d^3 r \right]^{1/2} . \tag{5.430}$$

Suppose that repeated scattering is sufficiently weak that there exists a number $0 < \eta < 1$ such that

$$\boxed{\|\mathcal{K}_V \phi\|_{\Omega_s} \leq \eta \|\phi\|_{\Omega_s}} \tag{5.431}$$

for the waves encountered during the iteration. Then

$$\|\mathcal{K}_V^n \psi_{\rm i}\|_{\Omega_s} \leq \eta^n \, \|\psi_{\rm i}\|_{\Omega_s} , \tag{5.432}$$

so the successive terms of the Born series decrease at least geometrically.

The condition $\eta < 1$ is sufficient for convergence of the iterative series. The stronger condition

$$\boxed{\eta \ll 1} \tag{5.433}$$

is what makes it reasonable to neglect the second and higher terms and retain only the first Born contribution.

The parameter η should not be regarded as a universal or directly measurable property of the specimen. Rather, it summarizes the strength of one

additional interaction-and-propagation step for the particular specimen and electron energy under consideration. Thus Eq. (5.431) is best understood as a sufficient mathematical condition for the successive corrections to decrease. Physically, the Born approximation is favoured when the specimen modifies the incident wave only weakly: weak potentials, high incident electron energy, short interaction distances, and cancellation between the rapidly varying phases of the Green function all tend to make η smaller, whereas strong potentials and thick specimens make higher-order terms more important.

The simplest approximation is therefore to stop after the first iteration,

$$\boxed{\psi \simeq \psi^{(1)} = \psi_{\mathrm{i}} + \mathcal{K}_V \psi_{\mathrm{i}}.} \tag{5.434}$$

This is the *first Born approximation.*

The first Born scattering amplitude. The first Born approximation consists of evaluating the scattering generated by the specimen as if the wave illuminating every possible interaction point were still the unperturbed incident wave. Thus, inside the scattering integral of Eq. (5.389), we make the replacement

$$\boxed{\psi(\mathbf{r}') \simeq \psi_{\mathrm{i}}(\mathbf{r}') = \exp\left(j\mathbf{k}_{\mathrm{i}} \cdot \mathbf{r}'\right),} \tag{5.435}$$

where the incident plane wave was introduced in Eq. (5.333).

The qualification "inside the scattering integral" is important: Eq. (5.435) does not state that the total wave equals the incident wave. Rather, it neglects the modification of the illuminating wave by previous interactions when calculating the first scattered contribution. Nor is a particular interaction point selected: the interaction may occur anywhere in the specimen, and the corresponding complex amplitudes are summed coherently.

The exact far-field scattering amplitude derived in Eq. (5.389) is

$$f(\mathbf{k}_{\mathrm{f}}, \mathbf{k}_{\mathrm{i}}) = C \int V(\mathbf{r}')\psi(\mathbf{r}') \exp\left(-j\mathbf{k}_{\mathrm{f}} \cdot \mathbf{r}'\right)\, d^3r', \tag{5.436}$$

with C defined in Eq. (5.417). Substituting Eq. (5.435) gives

$$f_{\mathrm{B}}(\mathbf{k}_{\mathrm{f}}, \mathbf{k}_{\mathrm{i}}) = C \int V(\mathbf{r}') \exp\left(j\mathbf{k}_{\mathrm{i}} \cdot \mathbf{r}'\right) \exp\left(-j\mathbf{k}_{\mathrm{f}} \cdot \mathbf{r}'\right)\, d^3r' \tag{5.437}$$

$$= C \int V(\mathbf{r}') \exp\left[-j\left(\mathbf{k}_{\mathrm{f}} - \mathbf{k}_{\mathrm{i}}\right) \cdot \mathbf{r}'\right]\, d^3r'. \tag{5.438}$$

Using the wavevector transfer introduced in Eq. (5.338),

$$\mathbf{q} = \mathbf{k}_{\mathrm{f}} - \mathbf{k}_{\mathrm{i}}, \tag{5.439}$$

and the corresponding spatial-frequency transfer defined in Eq. (5.341),

$$\mathbf{R} = \frac{\mathbf{q}}{2\pi}, \tag{5.440}$$

which is consistent with Eq. (3.6),

$$\mathbf{k} = 2\pi\mathbf{R}, \tag{5.441}$$

we obtain

$$f_{\mathrm{B}}(\mathbf{R}) = C \int V(\mathbf{r}') \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}'\right)\, d^3r'. \tag{5.442}$$

The integral is exactly the three-dimensional Fourier transform of the specimen potential according to the convention established in Eq. (3.2),

$$\widehat{V}(\mathbf{R}) = \int V(\mathbf{r}') \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}'\right)\, d^3r'. \tag{5.443}$$

Since

$$C = \frac{me}{2\pi\hbar^2}, \tag{5.444}$$

from Eq. (5.417), the first Born scattering amplitude is therefore

$$\boxed{f_{\mathrm{B}}(\mathbf{R}) = \frac{me}{2\pi\hbar^2}\widehat{V}(\mathbf{R}).} \tag{5.445}$$

Thus, in the first Born approximation, scattering with transfer $\mathbf{R}$ is produced by the coherent contribution of the complete specimen, and the relative phase of the contribution from position $\mathbf{r}'$ is

$$\exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}'\right). \tag{5.446}$$

The coherent sum over all possible interaction positions is therefore precisely the Fourier component $\widehat{V}(\mathbf{R})$.

It is sometimes useful to make this coherent summation local. An infinitesimal volume element around $\mathbf{r}'$ contributes

$$\boxed{df_{\mathrm{B}}(\mathbf{r}') = \frac{me}{2\pi\hbar^2} V(\mathbf{r}') \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}'\right) d^3r'.} \tag{5.447}$$

The total first-Born amplitude is then simply the coherent sum of these differential contributions over the complete specimen,

$$f_{\mathrm{B}}(\mathbf{R}) = \int df_{\mathrm{B}}(\mathbf{r}'). \tag{5.448}$$

Thus $df_{\mathrm{B}}(\mathbf{r}')$ should be interpreted as a *complex amplitude contribution*, not as a probability of scattering at $\mathbf{r}'$. Only after all such contributions have been added coherently is the squared modulus taken to obtain a scattering probability or differential cross section.

Connection with the atomic scattering factor. Equation (5.445) also explains the electron atomic scattering factor introduced previously. For an isolated atom, Eq. (5.82) defined

$$\boxed{f_e(\mathbf{R}) = \frac{m_0 e}{2\pi\hbar^2}\widehat{V}_a(\mathbf{R}).} \tag{5.449}$$

The rest mass m_0 is used in this conventional tabulation so that the atomic factor itself is independent of the incident electron energy. With the relativistically corrected scalar mass $m = \gamma m_0$ used in the present scattering equation, the first-Born amplitude at a particular beam energy is therefore

$$\boxed{f_{\mathrm{B}}(\mathbf{R}; E) = \frac{m}{m_0} f_e(\mathbf{R}) = \gamma f_e(\mathbf{R}).} \tag{5.450}$$

This is the conventional relativistic mass correction discussed, for example, by Kirkland Kirkland (2010)[Chap. 5].

Thus the tabulated atomic scattering factor and the actual scattering amplitude at a specified beam energy should not be confused.

Physical interpretation of the Born series and its limitations. The formal Born series of Eq. (5.427) admits a useful physical interpretation in terms of successive interactions with the specimen potential. The zeroth-order term ψ_{i} is the incident wave. The first-order term $\mathcal{K}_V\psi_{\mathrm{i}}$ contains one interaction with $V(\mathbf{r}')$, followed by propagation from the interaction point through the free-space Green function introduced in Eq. (3.86). The second-order term $\mathcal{K}_V^2\psi_{\mathrm{i}}$ contains two factors of the potential: the wave interacts at one position, propagates to another position, and interacts again. Higher powers of $\mathcal{K}_V$ similarly contain successively larger numbers of interactions.

This interpretation should not be read as a classical trajectory of an individual electron bouncing through the specimen. The Born series is a stationary-wave expansion of the solution of Eq. (5.416). Each term is a coherent sum over all possible positions at which the corresponding number of interactions may occur.

At first sight this raises an important question. If one interaction occurs at $\mathbf{r}''$ and the wave subsequently propagates to $\mathbf{r}'$, why does the next application of $\mathcal{K}_V$ again integrate over the complete specimen rather than only over points farther along the optical axis?

The reason is that the exact Lippmann–Schwinger equation is a three-dimensional stationary scattering equation, not a forward-marching propa-

gation equation. Its Green function,

$$G_0(\mathbf{r}, \mathbf{r}') = \frac{\exp\left(jk\|\mathbf{r} - \mathbf{r}'\|\right)}{\|\mathbf{r} - \mathbf{r}'\|}, \tag{5.451}$$

connects any two points of the specimen and does not impose an ordering of their longitudinal coordinates. Thus successive interaction coordinates are integrated independently over the complete specimen volume. In particular, the exact formulation also contains scattering paths for which the longitudinal component of the propagation reverses direction.

This generality is appropriate for the exact three-dimensional scattering problem, but it is more than is normally required for the high-energy electrons used in cryo-electron microscopy. As derived in the electron-source chapter, Eq. (2.100), a 300 keV electron travels at a substantial fraction of the speed of light, and elastic electron scattering is strongly concentrated around the forward direction. The dominant scattering paths therefore remain close to the optical axis.

This observation motivates the forward, or paraxial, approximation developed later. Instead of using the full spherical Green function to couple arbitrary points in three dimensions, the electron wave is propagated predominantly along the optical axis. The specimen potential can then be organized into successive longitudinal slices, and the wave entering each slice is the wave already modified by all preceding slices:

$$\boxed{\psi_0 \longrightarrow V_1 \longrightarrow \text{propagation} \longrightarrow V_2 \longrightarrow \text{propagation} \longrightarrow V_3 \longrightarrow \cdots.} \tag{5.452}$$

This is the conceptual bridge to the multislice method. Unlike the first Born approximation, multislice does not reset the illuminating wave to ψ_{i} after each interaction. Each slice acts on the wave produced by all previous propagation and scattering, so multiple elastic scattering is retained. What changes with respect to the full Lippmann–Schwinger formulation is primarily the organization of propagation: the exact three-dimensional problem is replaced by a predominantly forward, longitudinal evolution.

Consequences and limitations of the first Born approximation. Equation (5.445) immediately shows that the first-Born scattering amplitude is linear in the specimen potential:

$$V = V_1 + V_2 \qquad \Longrightarrow \qquad f_{\mathrm{B}}[V] = f_{\mathrm{B}}[V_1] + f_{\mathrm{B}}[V_2]. \tag{5.453}$$

A second consequence concerns the phase of the scattering amplitude. Consider first a real atomic potential centered at the origin and having

inversion symmetry,

$$V_a(\mathbf{r}) = V_a(-\mathbf{r}). \tag{5.454}$$

Hence, its Fourier transform must be real valued

$$\boxed{\widehat{V}_a(\mathbf{R}) \in \mathbb{R},} \tag{5.455}$$

and Eq. (5.445) implies

$$\boxed{f_{\mathrm{B}}(\mathbf{R}) \in \mathbb{R}} \tag{5.456}$$

for such a centered atom.

This absence of phase is a property of the *first Born approximation*; it is not a general property of elastic electron scattering. The origin of the additional phase can be seen directly from the exact scattering equation.

How higher-order elastic scattering generates phase.

The exact integral equation, Eq. (5.416), may be written as

$$\psi(\mathbf{r}) = \psi_{\mathrm{i}}(\mathbf{r}) + C \int G_0(\mathbf{r}, \mathbf{r}') V(\mathbf{r}') \psi(\mathbf{r}')\, d^3 r', \tag{5.457}$$

where C was defined in Eq. (5.417) and the outgoing free-space Green function is

$$G_0(\mathbf{r}, \mathbf{r}') = \frac{\exp\left(jk\|\mathbf{r} - \mathbf{r}'\|\right)}{\|\mathbf{r} - \mathbf{r}'\|}. \tag{5.458}$$

The first Born approximation replaces the unknown wave inside this integral by the incident wave, as stated in Eq. (5.435). To examine the first correction beyond this approximation, retain one additional iteration:

$$\psi(\mathbf{r}) \simeq \psi_{\mathrm{i}}(\mathbf{r}) + C \int G_0(\mathbf{r}, \mathbf{r}') V(\mathbf{r}') \psi_{\mathrm{i}}(\mathbf{r}')\, d^3 r'. \tag{5.459}$$

Substituting Eq. (5.459) into the exact scattering amplitude of Eq. (5.389) gives

$$f \simeq f^{(1)} + f^{(2)}, \tag{5.460}$$

where

$$f^{(1)} = f_{\mathrm{B}}, \tag{5.461}$$

and

$$f^{(2)}(\mathbf{k}_\mathrm{f}, \mathbf{k}_\mathrm{i}) = C^2 \int d^3r \int d^3r' \, V(\mathbf{r}) e^{-j\mathbf{k}_\mathrm{f}\cdot\mathbf{r}} \times G_0(\mathbf{r}, \mathbf{r}') V(\mathbf{r}') e^{j\mathbf{k}_\mathrm{i}\cdot\mathbf{r}'}. \tag{5.462}$$

The crucial difference with the first-Born term is the propagator between the two interaction positions. From Eq. (5.458),

$$G_0(\mathbf{r}, \mathbf{r}') = \frac{\cos\left(k\|\mathbf{r}-\mathbf{r}'\|\right)}{\|\mathbf{r}-\mathbf{r}'\|} + j\frac{\sin\left(k\|\mathbf{r}-\mathbf{r}'\|\right)}{\|\mathbf{r}-\mathbf{r}'\|}. \tag{5.463}$$

Thus, even when $V(\mathbf{r})$ is everywhere real, the propagation between successive interactions introduces a complex phase. Nothing requires the imaginary contribution in Eq. (5.462) to cancel in general. Higher iterations contain further such propagators and generate additional phase.

The complex phase of the complete elastic scattering amplitude therefore does not require a complex electrostatic potential. It arises naturally when the electron wave is allowed to propagate between successive interactions with a real potential. The sign of this propagation phase is fixed by the outgoing-wave boundary condition already used in Eq. (3.86).

A different source of phase should be distinguished from this higher-order *dynamical* phase. A real potential that is not centrosymmetric need not have a real Fourier transform. In particular, if a centered atomic potential $V_a^{(0)}(\mathbf{r})$ is translated to $\mathbf{r}_a$,

$$V_a(\mathbf{r}) = V_a^{(0)}\left(\mathbf{r} - \mathbf{r}_a\right), \tag{5.464}$$

then its first-Born amplitude contains the translation phase

$$\widehat{V}_a(\mathbf{R}) = e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a}\widehat{V}_a^{(0)}(\mathbf{R}). \tag{5.465}$$

This is a geometrical phase arising from the position of the atom, not from repeated scattering. The same position-dependent phases already appeared in the independent-atom structure factor derived previously in Sec. 5.3.7.

The distinction is therefore

geometrical phase	:	already possible in first Born through spatial arrangement,
dynamical phase	:	generated by propagation and repeated scattering.

(5.466)

Going beyond first Born therefore does not mean simply allowing f_B to be complex. It means abandoning the replacement of Eq. (5.435) and allowing the wave that interacts with the specimen to have already been modified

by previous interactions. Later, for the predominantly forward propagation relevant to electron microscopy, this physics will be incorporated naturally by propagation methods such as multislice rather than by assigning an arbitrary complex value to the first-Born scattering factor.

5.5 The fast-electron equation inside a three-dimensional specimen

The preceding section described elastic scattering in terms of an incident wave, a specimen potential, and an outgoing scattering amplitude. For image formation, however, we ultimately need a different description: given the three-dimensional electrostatic potential $V(x, y, z)$ of the specimen, how does the electron wave evolve as it travels through it?

This is particularly important once we go beyond the first Born approximation. In the first Born approximation, Eq. (5.435), the field illuminating every possible interaction point is replaced by the original incident wave. A more complete propagation model must instead allow the wave at depth z to have already been modified by the specimen encountered before that depth.

For the high-energy, predominantly forward-moving electrons of transmission electron microscopy, this evolution can be described by a particularly simple first-order equation.

5.5.1 Starting from the scalar electron wave equation

In the electron-optics chapter we derived the scalar relativistic wave equation in an electromagnetic field in Sec. 4.8. For a stationary electron in an electrostatic region, it reduces to the local Helmholtz equation Eq. (4.916),

$$\nabla^2\Psi(\mathbf{r}) + k^2(\mathbf{r})\Psi(\mathbf{r}) = 0, \tag{5.467}$$

with

$$k(\mathbf{r}) = \frac{p(\mathbf{r})}{\hbar}, \tag{5.468}$$

as established in Eq. (4.918).

In the elastic-scattering section, this equation was specialized to an electron incident on the specimen with asymptotic wavenumber k_0. After retaining only the term linear in the specimen electrostatic potential $V(\mathbf{r})$, we obtained

$$\boxed{\left(\nabla^2 + k_0^2\right)\Psi(\mathbf{r}) = -\frac{2me}{\hbar^2}V(\mathbf{r})\Psi(\mathbf{r}),} \tag{5.469}$$

which is Eq. (5.371). Here

$$m = \gamma m_0 \tag{5.470}$$

is the relativistic mass parameter evaluated for the incident beam. The relativistic kinematics leading to this relation were developed in Sec. 2.3.

Far from the specimen,

$$k_0 = \frac{p}{\hbar} = \frac{2\pi}{\lambda}, \tag{5.471}$$

using the wavevector–momentum relation Eq. (4.671) and the de Broglie relation Eq. (2.101).

Equation (5.469) still describes a three-dimensional stationary wave. We now exploit the fact that, in an electron microscope, most of its rapid phase variation corresponds simply to propagation along the optical axis.

5.5.2 Separating the rapidly varying carrier

Let the incident beam propagate along the positive z direction. We factor the stationary wave into a rapidly varying plane-wave carrier and a more slowly varying envelope:

$$\boxed{\Psi(x, y, z) = \psi(x, y, z) \exp(jk_0 z)\,.} \tag{5.472}$$

Here Ψ denotes the complete stationary electron wave, whereas ψ describes departures from the uniform incident plane wave. In the absence of the specimen and for propagation exactly along the optical axis,

$$\psi = 1. \tag{5.473}$$

The factorization in Eq. (5.472) is only a change of variables. No paraxial approximation has yet been made.

From the three-dimensional electron wave equation to the fast-electron equation.

Introduce the transverse Laplacian

$$\boxed{\nabla_\perp^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.} \tag{5.474}$$

Then

$$\nabla^2 = \nabla_\perp^2 + \frac{\partial^2}{\partial z^2}. \tag{5.475}$$

From Eq. (5.472),

$$\frac{\partial \Psi}{\partial z} = e^{jk_0 z}\left(\frac{\partial \psi}{\partial z} + jk_0\psi\right), \tag{5.476}$$

and therefore

$$\frac{\partial^2 \Psi}{\partial z^2} = e^{jk_0 z}\left(\frac{\partial^2 \psi}{\partial z^2} + 2jk_0\frac{\partial \psi}{\partial z} - k_0^2\psi\right). \tag{5.477}$$

The transverse derivatives act only on the envelope,

$$\nabla_\perp^2 \Psi = e^{jk_0 z}\nabla_\perp^2 \psi. \tag{5.478}$$

Substitution into Eq. (5.469) gives

$$\begin{aligned}\nabla_\perp^2 \psi + \frac{\partial^2 \psi}{\partial z^2} + 2jk_0\frac{\partial \psi}{\partial z} - k_0^2\psi + k_0^2\psi \\ = -\frac{2me}{\hbar^2}V(x,y,z)\psi.\end{aligned} \tag{5.479}$$

The two carrier terms cancel exactly, leaving

$$\boxed{\frac{\partial^2 \psi}{\partial z^2} + 2jk_0\frac{\partial \psi}{\partial z} + \nabla_\perp^2 \psi + \frac{2me}{\hbar^2}V(x,y,z)\psi = 0.} \tag{5.480}$$

Equation (5.480) is the exact envelope equation corresponding to Eq. (5.469). The paraxial approximation enters only now.

As discussed in Sec. 4.10, the relevant electron waves propagate at small angles with respect to the optical axis. In the angular-spectrum description this condition was written in Eq. (3.369) as

$$\boxed{\lambda R_\perp \ll 1.} \tag{5.481}$$

After removal of the rapidly varying carrier $e^{jk_0 z}$, the remaining envelope therefore varies much more slowly along z. Quantitatively, we require

$$\boxed{\left|\frac{\partial^2\psi}{\partial z^2}\right| \ll 2k_0\left|\frac{\partial\psi}{\partial z}\right|.} \tag{5.482}$$

Neglecting the second longitudinal derivative in Eq. (5.480) gives

$$2jk_0\frac{\partial\psi}{\partial z} + \nabla_\perp^2\psi + \frac{2me}{\hbar^2}V(x,y,z)\psi = 0. \tag{5.483}$$

Solving for the longitudinal derivative,

$$\frac{\partial\psi}{\partial z} = \frac{j}{2k_0}\nabla_\perp^2\psi + j\frac{me}{k_0\hbar^2}V(x,y,z)\psi. \tag{5.484}$$

Finally, using Eq. (5.471),

$$k_0 = \frac{2\pi}{\lambda}, \tag{5.485}$$

we obtain

$$\boxed{\frac{\partial\psi(x,y,z)}{\partial z} = \frac{j\lambda}{4\pi}\nabla_\perp^2\psi(x,y,z) + j\sigma V(x,y,z)\psi(x,y,z),} \tag{5.486}$$

where

$$\boxed{\sigma \equiv \frac{me\lambda}{2\pi\hbar^2}.} \tag{5.487}$$

The only approximation made in Derivation 5.5.2 was the neglect of the second longitudinal derivative,

$$\left|\frac{\partial^2\psi}{\partial z^2}\right| \ll 2k_0\left|\frac{\partial\psi}{\partial z}\right|, \tag{5.488}$$

in Eq. (5.482). The following derivation shows why this condition is naturally satisfied for a fast, predominantly forward-propagating electron.

Why the electron envelope is slowly varying.

Consider locally a plane-wave component with transverse spatial fre-

quency

$$\mathbf{R}_\perp = (R_x, R_y), \qquad R_\perp = \|\mathbf{R}_\perp\|. \tag{5.489}$$

Using the relation between spatial frequency and wavevector established in Eq. (3.6), its transverse wavevector magnitude is

$$k_\perp = 2\pi R_\perp. \tag{5.490}$$

Inside the specimen, the local wavenumber differs slightly from the field-free value k_0. This change was obtained previously in Eq. (5.370),

$$k^2(\mathbf{r}) - k_0^2 \simeq \frac{2me}{\hbar^2} V(\mathbf{r}). \tag{5.491}$$

For a locally plane-wave-like component,

$$k_z^2 + k_\perp^2 = k^2(\mathbf{r}), \tag{5.492}$$

and therefore

$$k_z = \sqrt{k_0^2 + \frac{2me}{\hbar^2} V(\mathbf{r}) - 4\pi^2 R_\perp^2}. \tag{5.493}$$

Consider a locally plane-wave-like component whose complete longitudinal dependence is

$$\Psi(z) \propto e^{jk_z z}. \tag{5.494}$$

In Eq. (5.472) the same wave was written as

$$\Psi(z) = \psi(z) e^{jk_0 z}, \tag{5.495}$$

where $e^{jk_0 z}$ represents the rapid phase accumulation of the reference electron propagating along the optical axis.

Comparing the two expressions,

$$\psi(z) e^{jk_0 z} \propto e^{jk_z z}, \tag{5.496}$$

and therefore

$$\psi(z) \propto e^{j(k_z - k_0) z}. \tag{5.497}$$

Thus the envelope does not oscillate with the full longitudinal wavenumber k_z. It contains only the difference between the actual longitudinal

wavenumber of that component and the reference carrier wavenumber k_0. We therefore define

$$\boxed{\Delta k_z \equiv k_z - k_0.} \tag{5.498}$$

With this notation,

$$\psi(z) \propto e^{j\Delta k_z z}. \tag{5.499}$$

If the electron remains close to the reference propagation direction and the specimen changes its local momentum only weakly, then

$$|\Delta k_z| \ll k_0, \tag{5.500}$$

so the envelope varies along z much more slowly than the carrier $e^{jk_0 z}$.

Factoring k_0^2 from Eq. (5.493) gives

$$k_z = k_0\sqrt{1 + \frac{2meV}{\hbar^2 k_0^2} - \frac{4\pi^2 R_\perp^2}{k_0^2}}. \tag{5.501}$$

For a fast electron, both the modification produced by the specimen potential and the transverse wavevector are small compared with the large longitudinal carrier. Using

$$\sqrt{1+\epsilon} \simeq 1 + \frac{\epsilon}{2}, \qquad |\epsilon| \ll 1, \tag{5.502}$$

we obtain

$$\Delta k_z \simeq \frac{me}{\hbar^2 k_0} V(\mathbf{r}) - \frac{2\pi^2}{k_0} R_\perp^2. \tag{5.503}$$

Using Eq. (5.471),

$$k_0 = \frac{2\pi}{\lambda}, \tag{5.504}$$

and the interaction constant defined in Eq. (5.487),

$$\sigma = \frac{me}{\hbar^2 k_0}, \tag{5.505}$$

this becomes

$$\boxed{\Delta k_z \simeq \sigma V(\mathbf{r}) - \pi\lambda R_\perp^2.} \tag{5.506}$$

This result identifies the two mechanisms responsible for the longitudinal variation of the envelope:

$$\boxed{\underbrace{-\pi\lambda R_\perp^2}_{\text{diffraction}} \quad + \quad \underbrace{\sigma V}_{\text{specimen interaction}} \quad .} \tag{5.507}$$

The first term is the same paraxial longitudinal phase correction obtained previously from the angular-spectrum propagation law in Sec. 3.6; see Eq. (3.371).

For an envelope component varying locally as

$$\psi \propto e^{j\Delta k_z z}, \tag{5.508}$$

we have

$$\frac{\partial \psi}{\partial z} = j\Delta k_z \psi, \tag{5.509}$$

and

$$\frac{\partial^2 \psi}{\partial z^2} = -(\Delta k_z)^2 \psi. \tag{5.510}$$

Hence

$$\frac{|\partial^2\psi/\partial z^2|}{2k_0\,|\partial\psi/\partial z|} = \frac{|\Delta k_z|}{2k_0}. \tag{5.511}$$

Therefore, whenever

$$\boxed{|\Delta k_z| \ll k_0,} \tag{5.512}$$

the second longitudinal derivative is negligible:

$$\boxed{\left|\frac{\partial^2\psi}{\partial z^2}\right| \ll 2k_0 \left|\frac{\partial\psi}{\partial z}\right|.} \tag{5.513}$$

This is precisely the slowly varying envelope condition used in Eq. (5.482).

Thus removing the carrier $e^{jk_0 z}$ separates the large longitudinal wavenumber k_0 from the much smaller residual variation produced by transverse diffraction and by the specimen potential. It is this scale separation that permits the second longitudinal derivative to be neglected.

5.5.3 The interaction constant

The quantity σ in Eq. (5.487) determines how strongly the electron phase responds to the electrostatic potential. It is useful to connect it with quan-

tities derived previously.

The constant introduced in the elastic-scattering section, Eq. (5.417), was

$$C = \frac{me}{2\pi\hbar^2}. \tag{5.514}$$

Therefore,

$$\boxed{\sigma = \lambda C.} \tag{5.515}$$

The same relativistically corrected coupling therefore appears in both the Born scattering amplitude and the longitudinal propagation equation.

Using $h = 2\pi\hbar$, Eq. (5.487) may also be written as

$$\sigma = \frac{2\pi me\lambda}{h^2}. \tag{5.516}$$

Furthermore, Sec. 2.3 established

$$p = \gamma m_0 v = mv, \tag{5.517}$$

while the de Broglie relation Eq. (2.101) gives

$$p = \frac{h}{\lambda}. \tag{5.518}$$

Hence

$$m\lambda = \frac{h}{v}, \tag{5.519}$$

and

$$\boxed{\sigma = \frac{e}{\hbar v}.} \tag{5.520}$$

This form makes the energy dependence particularly transparent: higher-energy electrons travel faster and therefore accumulate less phase per unit electrostatic potential and unit propagation distance. The electron velocity as a function of accelerating voltage was derived previously in Eq. (2.100).

The units of σ are

$$\boxed{[\sigma] = \mathrm{V}^{-1}\,\mathrm{m}^{-1},} \tag{5.521}$$

so that

$$\sigma V\, dz \tag{5.522}$$

is dimensionless, as required for a phase.

5.5.4 Physical meaning of the fast-electron equation

Equation (5.486) can be written schematically as

$$\boxed{\frac{\partial \psi}{\partial z} = (\mathcal{D} + \mathcal{I}(z))\,\psi,} \tag{5.523}$$

where

$$\boxed{\mathcal{D} \equiv \frac{j\lambda}{4\pi}\nabla_\perp^2} \tag{5.524}$$

describes transverse diffraction during propagation, while

$$\boxed{\mathcal{I}(z) \equiv j\sigma V(x, y, z)} \tag{5.525}$$

describes the local interaction with the specimen.

These two terms have very different physical roles.

If $V = 0$, Eq. (5.486) reduces to

$$\frac{\partial \psi}{\partial z} = \frac{j\lambda}{4\pi}\nabla_\perp^2 \psi. \tag{5.526}$$

Taking the two-dimensional Fourier transform in x and y, and using the Fourier-derivative properties derived previously in Eqs. (3.308) and (3.309), gives

$$\frac{\partial \widehat{\psi}(\mathbf{R}, z)}{\partial z} = -j\pi\lambda R_\perp^2 \widehat{\psi}(\mathbf{R}, z), \tag{5.527}$$

where

$$R_\perp^2 = R_x^2 + R_y^2. \tag{5.528}$$

For every fixed transverse spatial frequency $\mathbf{R}$, this is an ordinary differential equation in z.

Free propagation of one transverse spatial-frequency component.

For fixed $\mathbf{R}$, define

$$\alpha(\mathbf{R}) \equiv -j\pi\lambda R_\perp^2. \tag{5.529}$$

Then Eq. (5.527) becomes

$$\frac{\partial \widehat{\psi}}{\partial z} = \alpha(\mathbf{R})\widehat{\psi}. \tag{5.530}$$

Separating variables,

$$\frac{d\widehat{\psi}}{\widehat{\psi}} = \alpha(\mathbf{R})\, dz. \tag{5.531}$$

Integrating from z to $z + \Delta z$ gives

$$\int_{\widehat{\psi}(\mathbf{R},z)}^{\widehat{\psi}(\mathbf{R},z+\Delta z)} \frac{d\widehat{\psi}}{\widehat{\psi}} = \alpha(\mathbf{R}) \int_z^{z+\Delta z} dz', \tag{5.532}$$

and therefore

$$\ln\left[\frac{\widehat{\psi}(\mathbf{R}, z+\Delta z)}{\widehat{\psi}(\mathbf{R}, z)}\right] = \alpha(\mathbf{R})\Delta z. \tag{5.533}$$

Exponentiating both sides,

$$\widehat{\psi}(\mathbf{R}, z+\Delta z) = \exp\left[\alpha(\mathbf{R})\Delta z\right] \widehat{\psi}(\mathbf{R}, z). \tag{5.534}$$

Substituting the definition of $\alpha(\mathbf{R})$ finally gives

$$\boxed{\widehat{\psi}\left(\mathbf{R}, z+\Delta z\right) = \exp\left(-j\pi\lambda\Delta z R_\perp^2\right) \widehat{\psi}\left(\mathbf{R}, z\right).} \tag{5.535}$$

Thus each transverse spatial-frequency component evolves independently and accumulates the quadratic propagation phase $-\pi\lambda\Delta z R_\perp^2$. Equation (5.535) is precisely the Fresnel transfer function derived previously in Eq. (3.373), after removal of the common carrier phase $e^{jk_0\Delta z}$.

Thus the first term of Eq. (5.486) is the differential form of the Fresnel propagation already developed in the wave-diffraction chapter.

The second term,

$$j\sigma V(x, y, z)\psi, \tag{5.536}$$

changes the wave locally according to the electrostatic potential. Unlike the first Born approximation, Eq. (5.435), the wave multiplying $V(x, y, z)$ is not reset to the incident plane wave. It is the current wave $\psi(x, y, z)$, already modified by everything that happened at smaller z.

This is the essential reason why Eq. (5.486) can describe repeated elastic scattering within the paraxial, forward-propagating approximation.

5.5.5 What has been approximated?

It is useful to state explicitly what assumptions have led to Eq. (5.486).

First, for the elastic channel the microscopic many-body specimen has already been replaced by the prescribed scalar electrostatic potential $V(\mathbf{r})$, as developed in Sec. 5.2 and summarized in Sec. 5.3.9. From this point onward, the elastic propagation problem no longer requires us to follow the internal quantum state of the specimen explicitly.

Second, the spin-independent scalar relativistic description developed in Sec. 4.8 is being used. Spin-dependent terms are therefore outside the present model.

Third, the specimen potential is treated as a small perturbation of the large kinetic energy of the incident electron, so that the local relativistic momentum may be linearized in V. This is the same approximation already used to obtain Eq. (5.469).

Fourth, we have made the paraxial, or slowly varying envelope, approximation Eq. (5.482). The electron is therefore treated as propagating predominantly in the positive z direction.

No projection approximation has been made:

$$V = V(x, y, z) \tag{5.537}$$

still retains the complete depth dependence of the specimen.

No weak-phase approximation has been made either. The interaction term in Eq. (5.486) is proportional to the current wave itself and may accumulate a finite phase.

Finally, Eq. (5.486) is not a first-Born propagation equation. First Born replaces the wave inside the interaction by the fixed incident wave, Eq. (5.435). Here, by contrast, $\psi(x, y, z)$ evolves continuously with depth, and the specimen acts on the wave produced by all preceding evolution.

We have therefore reached the central propagation equation for fast elastic electrons in a three-dimensional specimen:

$$\boxed{\frac{\partial \psi}{\partial z} = \frac{j\lambda}{4\pi}\nabla_\perp^2 \psi + j\sigma V(x, y, z)\psi.} \tag{5.538}$$

Its structure already suggests the next steps. If transverse propagation is neglected while the electron crosses the specimen, the interaction term leads to the phase-object approximation (Sec. 5.6). If instead the diffraction and interaction terms are applied successively over thin longitudinal intervals, the result is the multislice method (Sec. 5.8).

5.6 The phase-object and projection approximations as limiting cases

The fast-electron equation derived in Eq. (5.486),

$$\boxed{\frac{\partial \psi}{\partial z} = \frac{j\lambda}{4\pi}\nabla_\perp^2 \psi + j\sigma V(x,y,z)\psi,} \tag{5.539}$$

contains two simultaneous processes. The first term describes transverse diffraction as the electron propagates through the specimen; the second describes its local interaction with the electrostatic potential.

This equation provides a natural starting point for several approximations commonly used in cryo-EM. In particular, the phase-object and projection approximations arise when the specimen is sufficiently thin that transverse propagation inside it may be neglected. The weak-phase-object approximation is a further approximation made only afterwards.

5.6.1 Neglecting transverse propagation inside the specimen

Suppose that the specimen extends from $z = z_{\text{in}}$ to $z = z_{\text{out}}$. If the transverse diffraction accumulated over this distance is negligible, we omit the first term of Eq. (5.486),

$$\frac{j\lambda}{4\pi}\nabla_\perp^2 \psi \simeq 0. \tag{5.540}$$

The fast-electron equation then reduces to

$$\boxed{\frac{\partial \psi(x,y,z)}{\partial z} = j\sigma V(x,y,z)\psi(x,y,z).} \tag{5.541}$$

An important consequence of Eq. (5.541) is that different transverse positions no longer communicate with one another while the electron crosses the specimen. For every fixed (x, y), the equation is simply an ordinary differential equation along z.

The phase-object transmission function.

For fixed (x, y), Eq. (5.541) may be written as

$$\frac{1}{\psi}\frac{\partial \psi}{\partial z} = j\sigma V(x,y,z). \tag{5.542}$$

Integrating from the entrance to the exit of the specimen gives

$$\int_{\psi(x,y,z_{\text{in}})}^{\psi(x,y,z_{\text{out}})} \frac{d\psi}{\psi} = j\sigma \int_{z_{\text{in}}}^{z_{\text{out}}} V(x,y,z)\,dz. \tag{5.543}$$

Therefore,

$$\ln\left[\frac{\psi(x,y,z_{\text{out}})}{\psi(x,y,z_{\text{in}})}\right] = j\sigma \int_{z_{\text{in}}}^{z_{\text{out}}} V(x,y,z)\,dz. \tag{5.544}$$

Exponentiating,

$$\psi(x,y,z_{\text{out}}) = \exp\left[j\sigma \int_{z_{\text{in}}}^{z_{\text{out}}} V(x,y,z)\,dz\right] \psi(x,y,z_{\text{in}}). \tag{5.545}$$

Define the projected specimen potential

$$\boxed{V^{\text{proj}}(x,y) \equiv \int_{z_{\text{in}}}^{z_{\text{out}}} V(x,y,z)\,dz.} \tag{5.546}$$

This generalizes to the complete specimen the projected atomic potential introduced previously in Eq. (5.240).

The exit wave is consequently

$$\boxed{\psi_{\text{exit}}(x,y) = t(x,y)\psi_{\text{in}}(x,y),} \tag{5.547}$$

where

$$\boxed{t(x,y) = \exp\left[j\sigma V^{\text{proj}}(x,y)\right]} \tag{5.548}$$

is the *specimen transmission function*.

Equation (5.548) is therefore not introduced as an independent model. It is a limiting solution of the three-dimensional fast-electron equation, Eq. (5.486), obtained by neglecting transverse diffraction while the electron traverses the specimen.

For the real electrostatic potential considered here,

$$|t(x,y)| = 1. \tag{5.549}$$

Thus the specimen changes the phase of the electron wave but not its amplitude in this approximation. This is the origin of the name *phase object*.

5.6.2 Why this is also a projection approximation

The same derivation has another important consequence. The complete three-dimensional function $V(x, y, z)$ enters Eq. (5.548) only through the line integral

$$V^{\text{proj}}(x, y) = \int V(x, y, z)\, dz. \tag{5.550}$$

All information about where along z a particular contribution to the potential occurred has disappeared.

Consequently, any two three-dimensional potentials satisfying

$$\int V_1(x, y, z)\, dz = \int V_2(x, y, z)\, dz \tag{5.551}$$

produce exactly the same transmission function within this approximation:

$$\boxed{V_1^{\text{proj}}(x, y) = V_2^{\text{proj}}(x, y) \quad \Longrightarrow \quad t_1(x, y) = t_2(x, y).} \tag{5.552}$$

This is the *projection approximation*. The specimen is replaced, for the purpose of electron propagation, by its two-dimensional projected electrostatic potential.

The terminology deserves some care. In the present paraxial model, the phase-object and projection descriptions arise from the same limiting operation—neglecting transverse propagation while the electron is inside the specimen—but they emphasize different consequences:

$$\boxed{\begin{aligned} \text{phase-object viewpoint} \quad &: \quad \psi_{\text{exit}} = t\, \psi_{\text{in}}, \\ \text{projection viewpoint} \quad &: \quad V(x, y, z) \longrightarrow V^{\text{proj}}(x, y). \end{aligned}} \tag{5.553}$$

The projection theorem derived previously for atomic potentials, Eq. (5.261), is a mathematical identity once a projected potential has been defined. The *projection approximation* made here is different: it is the physical assumption that this projected potential contains all the specimen information required to determine the exit wave.

5.6.3 When may propagation through the thickness be neglected?

The word "thin" should not be interpreted only as a geometrical statement. Whether propagation through the specimen matters depends also on the electron wavelength and on the transverse spatial frequencies that must be represented.

The free-space propagation law derived from the fast-electron equation in Eq. (5.535) is

$$\widehat{\psi}\left(\mathbf{R}, z+\Delta z\right) = \exp\left(-j\pi\lambda\Delta z R^2\right)\widehat{\psi}\left(\mathbf{R}, z\right), \tag{5.554}$$

where here

$$\mathbf{R} = (R_x, R_y), \qquad R = \|\mathbf{R}\| \tag{5.555}$$

is the transverse spatial frequency.

If a specimen of thickness

$$T = z_{\text{out}} - z_{\text{in}} \tag{5.556}$$

is to be treated as though no appreciable transverse propagation occurs while the electron crosses it, the Fresnel phase accumulated over this distance should be small for the spatial frequencies of interest:

$$\boxed{\pi\lambda T R^2 \ll 1.} \tag{5.557}$$

Equivalently, for a real-space period

$$d = \frac{1}{R}, \tag{5.558}$$

the condition is

$$\boxed{T \ll \frac{d^2}{\pi\lambda}.} \tag{5.559}$$

The criterion in Eq. (5.557) becomes quite restrictive at high resolution. For example, at an accelerating voltage of 300 keV the electron wavelength is approximately

$$\lambda \simeq 1.97\,\text{pm}. \tag{5.560}$$

The propagation phase accumulated across a specimen of thickness T at real-space period $d = 1/R$ is therefore

$$\phi_{\text{prop}} = \frac{\pi\lambda T}{d^2}. \tag{5.561}$$

For a specimen 100 nm thick, this phase is approximately

d	ϕ_{prop}
10 Å	0.62 rad
5 Å	2.47 rad
3 Å	6.87 rad
2 Å	15.5 rad.

(5.562)

Thus the same 100 nm particle may be reasonably close to a projection at nanometre resolution, while internal propagation cannot be ignored at near-atomic resolution.

The thickness dependence is equally important. At 3 Å resolution and 300 keV,

T	ϕ_{prop}
10 nm	0.69 rad
20 nm	1.37 rad
50 nm	3.44 rad
100 nm	6.87 rad.

(5.563)

These numbers illustrate an important point: "thin" is not an absolute property of the specimen. The validity of the projection approximation depends simultaneously on specimen thickness, electron wavelength, and the spatial resolution being considered. A particle may satisfy the projection approximation for low spatial frequencies while violating it strongly at the high spatial frequencies required for near-atomic reconstruction.

The condition $\pi\lambda T/d^2 \ll 1$ should be interpreted as an order-of-magnitude criterion rather than as a sharp boundary. Its value measures the magnitude of the depth-dependent Fresnel phase that is being neglected.

This criterion anticipates a result that will be developed later from the geometry of the Ewald sphere (Sec. 5.9): finite specimen depth becomes observable through a depth-dependent propagation phase.

5.6.4 The weak-phase-object approximation

The result above is the starting point of the standard *weak-phase-object approximation* (WPOA) widely used in transmission electron microscopy.

Within the phase-object approximation, the specimen transmission function is

$$t(x, y) = \exp\left[j\varphi(x, y)\right], \tag{5.564}$$

where, from Eq. (5.546),

$$\boxed{\varphi(x, y) = \sigma V^{\text{proj}}(x, y) = \sigma \int V(x, y, z)\, dz.} \tag{5.565}$$

The phase-object approximation by itself does not require $\varphi(x, y)$ to be small. The additional assumption defining the WPOA is

$$\boxed{|\varphi(x, y)| \ll 1.} \tag{5.566}$$

Expanding the exponential to first order then gives

$$\boxed{t(x,y) \simeq 1 + j\varphi(x,y) = 1 + j\sigma V^{\text{proj}}(x,y).} \tag{5.567}$$

For uniform plane-wave illumination,

$$\psi_{\text{in}}(x,y) = 1, \tag{5.568}$$

the exit wave is therefore

$$\boxed{\psi_{\text{exit}}(x,y) \simeq 1 + j\sigma V^{\text{proj}}(x,y).} \tag{5.569}$$

This is the standard weak-phase-object expression used in conventional TEM and cryo-EM image-formation theory.

5.6.5 The microscopic origin of the specimen phase

Equation (5.569) closes a connection that was deliberately left open in the electron-optics chapter. In Sec. 4.10.10, we introduced the wave entering the objective lens in the form

$$\psi_i(\mathbf{r}) \simeq 1 + j\varphi(\mathbf{r}), \qquad |\varphi| \ll 1, \tag{5.570}$$

in Eq. (4.1205), and explicitly postponed the physical origin of φ to the treatment of the electron–specimen interaction.

We can now identify that phase. Within the projection and weak-phase approximations,

$$\boxed{\varphi(x,y) = \sigma V^{\text{proj}}(x,y) = \sigma \int V(x,y,z)\, dz.} \tag{5.571}$$

Thus the phase modulation assumed in the first CTF derivation is the electrostatic phase accumulated by the fast electron while traversing the projected specimen potential.

This also explains why a pure phase object does not produce an intensity variation immediately at the specimen exit plane. From Eq. (5.548),

$$|t(x,y)|^2 = 1, \tag{5.572}$$

so, for uniform illumination,

$$|\psi_{\text{exit}}(x,y)|^2 = 1. \tag{5.573}$$

The spatial information resides in the *phase* of the exit wave. As shown subsequently in Sec. 4.10.10, the electron-optical system converts this phase modulation into measurable intensity contrast by changing the relative phases of the spatial-frequency components before they interfere at the image plane.

5.6.6 Three approximations that should not be confused

The derivations above also clarify three approximations that are often used together but have different physical meanings.

The *first Born approximation*, introduced in Sec. 5.4.4, truncates the Born series after the first scattered contribution, Eq. (5.434). Operationally, when evaluating the scattering source, it replaces the total wave reaching each interaction point by the incident wave,

$$\psi(\mathbf{r}') \simeq \psi_{\mathrm{i}}(\mathbf{r}'), \tag{5.574}$$

as stated explicitly in Eq. (5.435). It therefore neglects the modification of the illuminating wave by previous interactions. Importantly, first-Born scattering is still intrinsically three-dimensional: the scattering amplitude contains the three-dimensional Fourier transform $\widehat{V}(\mathbf{R})$ in Eq. (5.445). It does not, by itself, project the specimen along z.

The *projection approximation*, introduced in Sec. 5.6.2, is a different assumption. It follows from neglecting transverse propagation while the electron traverses the specimen,

$$\frac{j\lambda}{4\pi}\nabla_{\perp}^{2}\psi \simeq 0, \tag{5.575}$$

as stated in Eq. (5.540). The three-dimensional potential then enters the exit wave only through the projected potential

$$\boxed{V^{\mathrm{proj}}(x,y) = \int V(x,y,z)\,dz,} \tag{5.576}$$

defined in Eq. (5.546). The depth at which each contribution occurs is therefore lost. This approximation does not require the accumulated phase σV^{proj} to be small. A practical condition for its validity was obtained subsequently in Sec. 5.6.3, Eq. (5.557).

Finally, the *weak-phase-object approximation* (WPOA), introduced in Sec. 5.6.4, makes an additional approximation to the phase-object transmission function. It assumes

$$\boxed{|\varphi(x,y)| \ll 1,} \tag{5.577}$$

as stated in Eq. (5.566), where $\varphi = \sigma V^{\mathrm{proj}}$ from Eq. (5.565). The exponential transmission function can then be linearized,

$$\boxed{e^{j\sigma V^{\mathrm{proj}}} \simeq 1 + j\sigma V^{\mathrm{proj}},} \tag{5.578}$$

as derived in Eq. (5.567).

The distinction may be summarized as

$$\boxed{\begin{aligned} \text{first Born} \quad &: \quad \psi \text{ inside the scattering source} \to \psi_{\mathrm{i}}, \\ \text{projection} \quad &: \quad \nabla_{\perp}^{2}\psi \to 0 \quad \Longrightarrow \quad V(x,y,z) \to \int V(x,y,z)\,dz, \\ \text{weak phase} \quad &: \quad e^{j\varphi} \to 1 + j\varphi. \end{aligned}} \tag{5.579}$$

Therefore,

$$\boxed{\text{first Born} \neq \text{projection} \neq \text{weak phase.}} \tag{5.580}$$

Their predictions may coincide in a regime where all three approximations are simultaneously valid, but their assumptions are logically distinct. In particular, the appearance of a term linear in V^{proj} in Eq. (5.569) should not by itself be interpreted as a first-Born assumption.

The next section will relax the projection approximation while preserving the forward paraxial description. Instead of collapsing the complete specimen into a single projected potential, we shall divide $V(x,y,z)$ into thin longitudinal slices and alternate local interaction with Fresnel propagation. This leads naturally to the multislice method.

5.7 From the specimen exit wave to the standard Single Particle Analysis image model

We have now obtained the wave emerging from a thin biological specimen. Under the weak-phase-object approximation, Sec. 5.6.4, Eq. (5.569) gives

$$\boxed{\psi_{\mathrm{exit}}(x,y) \simeq 1 + j\sigma V^{\mathrm{proj}}(x,y).} \tag{5.581}$$

This result is expressed in terms of an electron wavefunction. The experimental image, however, is not a measurement of the complex wavefunction. A cryo-EM detector records individual electrons arriving at different positions after the exit wave has propagated through the objective-lens imaging system.

We must therefore complete the chain

$$\boxed{V(\mathbf{r}) \longrightarrow \psi_{\mathrm{exit}} \longrightarrow \psi_{\mathrm{obs}} \longrightarrow I_{\mathrm{obs}} \longrightarrow \text{electron detections.}} \tag{5.582}$$

This will lead directly to the linear image-formation model normally used in single-particle analysis.

5.7.1 Propagation of the exit wave through the microscope

The action of the objective-lens imaging system on a coherent electron wave was derived previously in Sec. 4.10.9. For an approximately shift-invariant coherent system, Eq. (4.1197) gives

$$\boxed{\widehat{\psi}_{\text{obs}}(\mathbf{R}) = H(\mathbf{R})\widehat{\psi}_{\text{exit}}(\mathbf{R}),} \tag{5.583}$$

where $H(\mathbf{R})$ is the coherent *wave-transfer function*.

This distinction in terminology is important. The function $H(\mathbf{R})$ acts on complex electron amplitudes. The contrast transfer function, or CTF, will appear only after those amplitudes have interfered and the detector forms an intensity.

From the weak-phase exit wave to the CTF-modulated intensity.

Taking the two-dimensional Fourier transform of Eq. (5.581) gives

$$\widehat{\psi}_{\text{exit}}(\mathbf{R}) = \delta(\mathbf{R}) + j\sigma\widehat{V}^{\text{proj}}(\mathbf{R}). \tag{5.584}$$

The delta term represents the strong unscattered reference beam.

Substituting Eq. (5.584) into Eq. (5.583),

$$\widehat{\psi}_{\text{obs}}(\mathbf{R}) = H(\mathbf{R})\delta(\mathbf{R}) + j\sigma H(\mathbf{R})\widehat{V}^{\text{proj}}(\mathbf{R}). \tag{5.585}$$

Assuming that the direct beam is transmitted and normalized so that

$$H(\mathbf{0}) = 1, \tag{5.586}$$

we obtain

$$\boxed{\widehat{\psi}_{\text{obs}}(\mathbf{R}) = \delta(\mathbf{R}) + j\sigma H(\mathbf{R})\widehat{V}^{\text{proj}}(\mathbf{R}).} \tag{5.587}$$

Define

$$g(\mathbf{r}) = \mathcal{F}^{-1}\left\{H(\mathbf{R})\widehat{V}^{\text{proj}}(\mathbf{R})\right\}. \tag{5.588}$$

Then

$$\psi_{\text{obs}}(\mathbf{r}) = 1 + j\sigma g(\mathbf{r}). \tag{5.589}$$

The detector is sensitive to electron flux rather than directly to the complex amplitude. As derived in Sec. 3.11, Eq. (3.724), for the narrow monochromatic beam considered here,

$$I_{\text{obs}}(\mathbf{r}) \propto |\psi_{\text{obs}}(\mathbf{r})|^2 . \tag{5.590}$$

Hence

$$\begin{aligned} I_{\text{obs}} &= |1 + j\sigma g|^2 \\ &= 1 + j\sigma g - j\sigma g^* + \sigma^2 |g|^2. \end{aligned} \tag{5.591}$$

Because the WPOA retains only terms first order in the specimen phase, the final term is neglected:

$$\boxed{\delta I(\mathbf{r}) \equiv I_{\text{obs}}(\mathbf{r}) - 1 \simeq j\sigma \left[g(\mathbf{r}) - g^*(\mathbf{r})\right] .} \tag{5.592}$$

Since the projected electrostatic potential is real,

$$\widehat{V}^{\text{proj}}(-\mathbf{R}) = \widehat{V}^{\text{proj}\,*}(\mathbf{R}). \tag{5.593}$$

Taking the Fourier transform of Eq. (5.592) therefore gives

$$\boxed{\widehat{\delta I}(\mathbf{R}) = j\sigma \left[H(\mathbf{R}) - H^*(-\mathbf{R})\right] \widehat{V}^{\text{proj}}(\mathbf{R}).} \tag{5.594}$$

This is precisely the general weak-phase transfer relation already derived in Eq. (4.1219), now with the microscopic identification

$$\varphi = \sigma V^{\text{proj}} \tag{5.595}$$

from Eq. (5.565).

Define the coherent phase-contrast CTF by

$$\boxed{\text{CTF}_{\text{coh}}(\mathbf{R}) \equiv \frac{j}{2} \left[H(\mathbf{R}) - H^*(-\mathbf{R})\right] .} \tag{5.596}$$

Then

$$\boxed{\widehat{\delta I}(\mathbf{R}) = 2\sigma\, \text{CTF}_{\text{coh}}(\mathbf{R}) \widehat{V}^{\text{proj}}(\mathbf{R}).} \tag{5.597}$$

Equation (5.597) is the microscopic origin of the standard linear cryo-EM image-formation equation.

If

$$h_{\mathrm{CTF}}(\mathbf{r}) = \mathcal{F}^{-1}\left\{\mathrm{CTF}(\mathbf{R})\right\}, \tag{5.598}$$

then, apart from normalization,

$$\boxed{\delta I(\mathbf{r}) \propto h_{\mathrm{CTF}} * V^{\mathrm{proj}}(\mathbf{r}).} \tag{5.599}$$

Thus the often-used statement that "the cryo-EM image is the projected potential convolved with the CTF" is not a fundamental postulate. It is the consequence of

$$\boxed{\begin{array}{c}\text{projection approximation}\\ + \\ \text{weak-phase-object approximation}\\ + \\ \text{coherent electron-wave transfer}\\ + \\ \text{linearization of the detected intensity.}\end{array}} \tag{5.600}$$

The more complete CTF derived in Sec. 4.15, Eq. (4.2117), additionally incorporates the objective aperture, deterministic aberrations, amplitude contrast, and the attenuation envelopes associated with partial coherence and motion. Those effects modify the transfer function but preserve the same linear structure as long as the weak-object approximation remains valid.

5.7.2 From intensity to detected electrons

Equation (5.597) gives an expected spatial electron intensity. A real micrograph, however, consists of discrete electron detections.

The relation between electron flux and expected counts was established in Eq. (3.727). Let D denote the mean incident electron dose, measured in electrons per unit area, and normalize the unscattered image intensity to unity. Then the expected number of electrons detected in pixel p, whose area is A_p, is

$$\boxed{\mu_p = D \int_{A_p} \left[1 + \delta I(\mathbf{r})\right] d^2 r.} \tag{5.601}$$

For pixels sufficiently small that the intensity varies little across one pixel,

$$\mu_p \simeq D A_p \left[1 + \delta I(\mathbf{r}_p)\right]. \tag{5.602}$$

Using Eq. (5.599),

$$\boxed{\mu_p \simeq DA_p \left[1 + a \left(h_{\mathrm{CTF}} * V^{\mathrm{proj}}\right)(\mathbf{r}_p)\right],} \tag{5.603}$$

where the factor a absorbs the normalization convention, including the factor 2σ appearing in Eq. (5.597).

In the ideal independent-arrival limit, the actual number of electrons recorded in the pixel is a random variable,

$$\boxed{N_p \sim \mathrm{Poisson}(\mu_p).} \tag{5.604}$$

Thus the CTF determines the spatial modulation of the *expected* electron counts, whereas the recorded micrograph is one random realization of those detection events.

A real direct-electron detector additionally introduces finite efficiency and spatial spreading, summarized by its MTF and DQE as discussed in Sec. 3.11. These detector effects may be included as an additional transfer factor and as a modification of the noise statistics, but they do not change the specimen-side derivation above.

5.7.3 The standard single-particle forward model

We can now write the form normally used in single-particle analysis.

Let $V_0(\mathbf{r})$ denote the three-dimensional electrostatic potential of the particle in a reference orientation, and let $\mathcal{R}_i \in SO(3)$ describe the orientation of particle i with respect to the electron beam. Define its projected potential as

$$\boxed{(\mathcal{P}_{\mathcal{R}_i} V_0)(x, y) = \int V_0 \left[\mathcal{R}_i^{-1} \begin{pmatrix} x \\ y \\ z \end{pmatrix}\right] dz.} \tag{5.605}$$

If the particle also has an in-plane translation $\mathbf{t}_i$, the background-subtracted and normalized image may be written

$$\boxed{y_i(\mathbf{r}) = a_i \left[h_i * \mathcal{P}_{\mathcal{R}_i} V_0\right](\mathbf{r} - \mathbf{t}_i) + n_i(\mathbf{r}),} \tag{5.606}$$

where

$$\begin{aligned} h_i &= \mathcal{F}^{-1}\{\mathrm{CTF}_i\}, \\ a_i &= \text{overall image scale}, \\ n_i &= \text{shot noise and other experimental noise.} \end{aligned} \tag{5.607}$$

Equivalently, in Fourier space,

$$\boxed{\widehat{y}_i(\mathbf{R}) = a_i \,\mathrm{CTF}_i(\mathbf{R}) e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i} \widehat{\mathcal{P}_{\mathcal{R}_i} V_0}(\mathbf{R}) + \widehat{n}_i(\mathbf{R}).} \tag{5.608}$$

This is the standard linear image-formation model underlying much of single-particle cryo-EM.

Using the Fourier projection theorem derived in Eq. (5.261), the Fourier transform of each projected potential corresponds, within the projection approximation, to a central plane through the three-dimensional Fourier transform of the particle potential. We shall later examine how this statement is modified when the curvature of the Ewald sphere (Sec. 5.9) can no longer be neglected.

5.7.4 What changes for a phase object that is not weak?

The linear SPA model of Eq. (5.606) does *not* follow from the phase-object approximation alone.

Without the weak-phase linearization, the specimen exit wave is, from Sec. 5.6.1 and Eq. (5.548),

$$\boxed{\psi_{\mathrm{exit}}(\mathbf{r}) = \exp\left[j\sigma V^{\mathrm{proj}}(\mathbf{r})\right].} \tag{5.609}$$

The microscope still acts linearly on the *wave*. If $h(\mathbf{r})$ is its coherent amplitude impulse response,

$$\psi_{\mathrm{obs}} = h * \exp\left[j\sigma V^{\mathrm{proj}}\right]. \tag{5.610}$$

The detected intensity is therefore

$$\boxed{I_{\mathrm{PO}}(\mathbf{r}) = \left|h * \exp\left[j\sigma V^{\mathrm{proj}}\right]\right|^2.} \tag{5.611}$$

This expression is nonlinear in the projected potential. The coherent wave-transfer function $H(\mathbf{R})$ remains perfectly valid, but a single scalar CTF multiplying $\widehat{V}^{\mathrm{proj}}(\mathbf{R})$ is no longer a complete description of the image.

The origin of the nonlinearity is already visible in

$$e^{j\varphi} = 1 + j\varphi - \frac{\varphi^2}{2} - j\frac{\varphi^3}{6} + \cdots. \tag{5.612}$$

Terms such as φ^2 and φ^3 mix spatial frequencies before the objective-lens transfer and before the final squared-modulus detection operation.

Thus

$$\boxed{\text{phase object} \quad \not\Longrightarrow \quad \text{linear CTF image model.}} \tag{5.613}$$

The WPOA is precisely the additional approximation that removes these nonlinear terms and restores a linear mapping from projected potential to detected image.

5.7.5 What changes under the first Born approximation?

The first Born approximation requires a different distinction. As introduced in Sec. 5.4.4 and expressed explicitly in Eq. (5.435), it linearizes the *scattering process* by replacing the wave illuminating each interaction point by the incident wave.

Let

$$\psi_{\mathrm{B},s}^{\mathrm{exit}} = \mathcal{B}[V] \tag{5.614}$$

denote the first-Born scattered wave evaluated at the specimen exit surface. The operator $\mathcal{B}$ is linear in the complete three-dimensional potential,

$$\boxed{\mathcal{B}[V_1 + V_2] = \mathcal{B}[V_1] + \mathcal{B}[V_2],} \tag{5.615}$$

as follows directly from Eq. (5.445).

With the unscattered reference retained,

$$\psi_{\mathrm{exit}} \simeq 1 + \mathcal{B}[V]. \tag{5.616}$$

After coherent imaging,

$$\psi_{\mathrm{obs}} \simeq 1 + h * \mathcal{B}[V], \tag{5.617}$$

and therefore

$$\begin{aligned} I_{\mathrm{B}} &= |1 + h * \mathcal{B}[V]|^2 \\ &= 1 + 2\,\mathrm{Re}\left\{h * \mathcal{B}[V]\right\} + |h * \mathcal{B}[V]|^2 . \end{aligned} \tag{5.618}$$

Since

$$\mathcal{B}[V] = O(V), \tag{5.619}$$

retaining the detected image only to first order in the potential gives

$$\boxed{\delta I_{\mathrm{B}} \simeq 2\,\mathrm{Re}\left\{h * \mathcal{B}[V]\right\}.} \tag{5.620}$$

Thus first-Born imaging is also linear in the specimen potential to first order. However, first Born alone does *not* imply the standard SPA projection model. The operator $\mathcal{B}$ acts on the complete three-dimensional potential and, in general, retains information about where along z the scattering occurred.

Only when the projection approximation of Sec. 5.6.2 is imposed as an additional approximation does the first-order exit-wave perturbation reduce to

$$\mathcal{B}[V] \longrightarrow j\sigma V^{\text{proj}}, \tag{5.621}$$

and Eq. (5.620) reduces to the same linear CTF model obtained from the WPOA.

5.7.6 How the three specimen models relate to the SPA equation

The relation among the different models can now be stated precisely:

specimen model	exit wave	image model
phase object	$e^{j\sigma V^{\text{proj}}}$	$\left\lvert h * e^{j\sigma V^{\text{proj}}} \right\rvert^2$
weak phase object	$1 + j\sigma V^{\text{proj}}$	$V^{\text{proj}} * h_{\text{CTF}}$
first Born	$1 + \mathcal{B}[V]$	$2\,\text{Re}\,\{h * \mathcal{B}[V]\}$

(5.622)

The standard SPA model,

$$\boxed{\text{image} = \text{projected potential} * \text{CTF} + \text{noise},} \tag{5.623}$$

therefore occupies a particular intersection of approximations. It is obtained naturally from the WPOA. It is also recovered from first-Born scattering when the additional projection approximation is made. It is *not*, however, the exact image model of a general phase object.

This distinction will become important in the next developments. The multislice method will abandon the projection approximation and allow the electron wave to evolve through successive depths of the three-dimensional specimen, while the subsequent objective-lens and detector physics remain governed by the same principles developed above.

5.8 The multislice method

The projection approximation of Sec. 5.6.2 replaced the complete three-dimensional specimen by a single projected potential, because transverse propagation while the electron crossed the specimen was neglected. This led to the phase-object transmission function Eq. (5.548).

For sufficiently thick specimens, or at sufficiently high spatial resolution, this approximation eventually fails, as quantified by Eq. (5.557). Contributions arising at different depths can no longer be regarded as though they acted in the same transverse plane.

The fast-electron equation, Eq. (5.486), already contains the required physics:

$$\boxed{\frac{\partial \psi}{\partial z} = \frac{j\lambda}{4\pi}\nabla_\perp^2 \psi + j\sigma V(x,y,z)\psi.} \tag{5.624}$$

Its two terms describe two different processes:

$$\boxed{\underbrace{\frac{j\lambda}{4\pi}\nabla_\perp^2}_{\text{transverse propagation}} \quad + \quad \underbrace{j\sigma V(x,y,z)}_{\text{interaction with the specimen}} \quad .} \tag{5.625}$$

The multislice method follows by dividing the specimen into sufficiently thin slabs and approximating the simultaneous action of these two processes by their successive action within each slab:

$$\boxed{\text{interaction} \longrightarrow \text{propagation} \longrightarrow \text{interaction} \longrightarrow \text{propagation} \longrightarrow \cdots .} \tag{5.626}$$

This is an *operator-splitting approximation* to Eq. (5.624). It is this viewpoint that makes clear both why multislice works and where its errors arise (Kirkland, 2010, Secs. 6.4–6.11).

5.8.1 Propagation and interaction as operators

Using the operators already introduced in Eqs. (5.524) and (5.525), define

$$\boxed{\mathcal{D} \equiv \frac{j\lambda}{4\pi}\nabla_\perp^2} \tag{5.627}$$

and

$$\boxed{\mathcal{I}(z) \equiv j\sigma V(x,y,z).} \tag{5.628}$$

Equation (5.624) becomes

$$\boxed{\frac{\partial \psi}{\partial z} = [\mathcal{D} + \mathcal{I}(z)]\,\psi.} \tag{5.629}$$

If only $\mathcal{D}$ were present, the electron would propagate freely. If only $\mathcal{I}$ were present, the wave would acquire a local phase from the electrostatic potential. Multislice exploits the fact that each of these two subproblems can be solved very simply.

There is, however, an important mathematical point:

$$\boxed{\mathcal{D}\mathcal{I} \neq \mathcal{I}\mathcal{D}} \tag{5.630}$$

in general. Diffraction and interaction therefore cannot simply be separated exactly over an arbitrary finite specimen thickness. Their separation becomes an approximation whose accuracy improves as the longitudinal step is reduced.

5.8.2 Dividing the specimen into slices

Let the specimen be divided by planes

$$z_0 < z_1 < \cdots < z_N, \tag{5.631}$$

with slice n extending from z_n to z_{n+1} and having thickness

$$\boxed{\Delta z_n = z_{n+1} - z_n.} \tag{5.632}$$

The potential associated with this slice is projected only over its own thickness:

$$\boxed{V_n^{\text{proj}}(x, y) \equiv \int_{z_n}^{z_{n+1}} V(x, y, z)\, dz.} \tag{5.633}$$

Notice the important difference with Eq. (5.546). There the complete specimen was projected into one plane. Here each projection covers only a thin slab:

$$V^{\text{proj}}(x, y) = \sum_{n=0}^{N-1} V_n^{\text{proj}}(x, y), \tag{5.634}$$

but the separate V_n^{proj} are retained so that propagation can occur between them.

Multislice as operator splitting.

Consider the propagation of the envelope from z_n to z_{n+1}. Formally, Eq. (5.629) defines an evolution operator

$$\psi(z_{n+1}) = \mathcal{U}_n \psi(z_n). \tag{5.635}$$

Let

$$\mathcal{A}(z) \equiv \mathcal{D} + \mathcal{I}(z), \tag{5.636}$$

so that Eq. (5.629) becomes

$$\frac{\partial \psi(z)}{\partial z} = \mathcal{A}(z)\psi(z). \tag{5.637}$$

We suppress the transverse coordinates here for clarity.

If $\mathcal{A}$ were independent of z, the solution over one slice would simply be

$$\psi(z_{n+1}) = \exp\left[\Delta z_n \mathcal{A}\right] \psi(z_n). \tag{5.638}$$

When $\mathcal{A}$ depends on z, however, this expression cannot in general be used because the operators acting at different depths need not commute.

To see the correct form of the solution, integrate Eq. (5.637) from z_n to an arbitrary depth z:

$$\psi(z) = \psi(z_n) + \int_{z_n}^{z} \mathcal{A}(\zeta)\psi(\zeta)\, d\zeta. \tag{5.639}$$

This equation is still implicit because the unknown wave appears inside the integral.

To expose its structure, rename the outer integration variable ζ_2 and substitute Eq. (5.639) for $\psi(\zeta_2)$:

$$\begin{aligned}\psi(z) = \psi(z_n) &+ \int_{z_n}^{z} \mathcal{A}(\zeta_2)\, d\zeta_2\, \psi(z_n) \\ &+ \int_{z_n}^{z} d\zeta_2 \int_{z_n}^{\zeta_2} d\zeta_1\, \mathcal{A}(\zeta_2)\mathcal{A}(\zeta_1)\psi(\zeta_1).\end{aligned} \tag{5.640}$$

Repeating the substitution generates

$$\psi(z_{n+1}) = \Bigg[\mathbb{I} + \int_{z_n}^{z_{n+1}} \mathcal{A}(\zeta_1)\, d\zeta_1 + \int_{z_n}^{z_{n+1}} d\zeta_2 \int_{z_n}^{\zeta_2} d\zeta_1\, \mathcal{A}(\zeta_2)\mathcal{A}(\zeta_1) + \int_{z_n}^{z_{n+1}} d\zeta_3 \int_{z_n}^{\zeta_3} d\zeta_2 \int_{z_n}^{\zeta_2} d\zeta_1\, \mathcal{A}(\zeta_3)\mathcal{A}(\zeta_2)\mathcal{A}(\zeta_1) + \cdots\Bigg]\psi(z_n). \tag{5.641}$$

The nested integration limits now have an especially transparent interpretation:

$$\boxed{z_n \le \zeta_1 \le \zeta_2 \le \zeta_3 \le \cdots \le z_{n+1}.} \tag{5.642}$$

Thus increasing subscripts correspond to increasing depth along the direction of propagation.

The interaction at ζ_1 occurs first, followed by that at ζ_2, then that at ζ_3, and so on. Since operators act on a state from right to left, the corresponding product is

$$\boxed{\mathcal{A}(\zeta_3)\mathcal{A}(\zeta_2)\mathcal{A}(\zeta_1)\psi(z_n).} \tag{5.643}$$

Indeed,

$$\psi \xrightarrow{\mathcal{A}(\zeta_1)} \mathcal{A}(\zeta_1)\psi \xrightarrow{\mathcal{A}(\zeta_2)} \mathcal{A}(\zeta_2)\mathcal{A}(\zeta_1)\psi \xrightarrow{\mathcal{A}(\zeta_3)} \mathcal{A}(\zeta_3)\mathcal{A}(\zeta_2)\mathcal{A}(\zeta_1)\psi. \tag{5.644}$$

This longitudinal ordering matters because, in general,

$$\mathcal{A}(\zeta_2)\mathcal{A}(\zeta_1) \neq \mathcal{A}(\zeta_1)\mathcal{A}(\zeta_2). \tag{5.645}$$

The importance of longitudinal ordering can be expressed using the *commutator*. For two operators $\mathcal{A}$ and $\mathcal{B}$, the commutator is defined as

$$\boxed{[\mathcal{A}, \mathcal{B}] \equiv \mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}.} \tag{5.646}$$

If

$$[\mathcal{A}, \mathcal{B}] = 0, \tag{5.647}$$

then the order of the two operations does not matter,

$$\mathcal{AB} = \mathcal{BA}, \tag{5.648}$$

and the operators are said to commute.

For the present problem,

$$\mathcal{A}(z) = \mathcal{D} + \mathcal{I}(z). \tag{5.649}$$

At two different depths ζ_1 and ζ_2,

$$\begin{aligned}[\mathcal{A}(\zeta_1), \mathcal{A}(\zeta_2)] &= [\mathcal{D} + \mathcal{I}(\zeta_1), \mathcal{D} + \mathcal{I}(\zeta_2)] \\ &= [\mathcal{D}, \mathcal{D}] + [\mathcal{D}, \mathcal{I}(\zeta_2)] + [\mathcal{I}(\zeta_1), \mathcal{D}] + [\mathcal{I}(\zeta_1), \mathcal{I}(\zeta_2)].\end{aligned} \tag{5.650}$$

The first term vanishes trivially,

$$[\mathcal{D}, \mathcal{D}] = 0. \tag{5.651}$$

Moreover, from the definition Eq. (5.646),

$$[\mathcal{I}(\zeta_1), \mathcal{D}] = -[\mathcal{D}, \mathcal{I}(\zeta_1)]. \tag{5.652}$$

The interaction operators at two different depths also commute. Indeed, since

$$\mathcal{I}(z) = j\sigma V(x, y, z) \tag{5.653}$$

acts simply by multiplying the wave by a scalar function, we have

$$\begin{aligned}\mathcal{I}(\zeta_1)\mathcal{I}(\zeta_2)\psi &= (j\sigma)^2 V(x, y, \zeta_1) V(x, y, \zeta_2)\psi \\ &= (j\sigma)^2 V(x, y, \zeta_2) V(x, y, \zeta_1)\psi \\ &= \mathcal{I}(\zeta_2)\mathcal{I}(\zeta_1)\psi.\end{aligned} \tag{5.654}$$

Hence

$$\boxed{[\mathcal{I}(\zeta_1), \mathcal{I}(\zeta_2)] = 0.} \tag{5.655}$$

We therefore obtain

$$\boxed{[\mathcal{A}(\zeta_1), \mathcal{A}(\zeta_2)] = [\mathcal{D}, \mathcal{I}(\zeta_2)] - [\mathcal{D}, \mathcal{I}(\zeta_1)].} \tag{5.656}$$

In contrast, the diffraction operator $\mathcal{D}$ does not generally commute with the interaction operator $\mathcal{I}$. The reason is that $\mathcal{D}$ contains transverse

derivatives, whereas $\mathcal{I}$ multiplies by the spatially varying potential. Differentiating the product $V\psi$ is not the same operation as differentiating ψ first and multiplying the result by V afterwards. Consequently,

$$[\mathcal{D}, \mathcal{I}] \neq 0 \tag{5.657}$$

in general.

The series in Eq. (5.641) is conventionally written in the compact form

$$\boxed{\mathcal{U}_n = \mathcal{T}_z \exp\left[\int_{z_n}^{z_{n+1}} \mathcal{A}(z)\, dz\right],} \tag{5.658}$$

or, restoring the two operators,

$$\boxed{\mathcal{U}_n = \mathcal{T}_z \exp\left[\int_{z_n}^{z_{n+1}} (\mathcal{D} + \mathcal{I}(z))\, dz\right].} \tag{5.659}$$

Here $\mathcal{T}_z$ is not an additional physical operation. It is notation specifying that, when the exponential is expanded, operators at smaller z must act first and therefore appear farther to the right, exactly as in Eq. (5.641).

If all operators at different depths commuted,

$$[\mathcal{A}(\zeta_1), \mathcal{A}(\zeta_2)] = 0 \qquad \text{for all } \zeta_1, \zeta_2, \tag{5.660}$$

the ordering symbol would be unnecessary, and the solution would reduce to the ordinary exponential

$$\mathcal{U}_n = \exp\left[\int_{z_n}^{z_{n+1}} \mathcal{A}(z)\, dz\right]. \tag{5.661}$$

We now separate the evolution into two parts. The interaction operator is

$$\mathcal{I}(z) = j\sigma V(x, y, z). \tag{5.662}$$

Since interaction operators at different depths commute, their interaction-only evolution over one slice can be written exactly as

$$\begin{aligned} \mathcal{T}_n &= \exp\left[\int_{z_n}^{z_{n+1}} \mathcal{I}(z)\, dz\right] \\ &= \exp\left[j\sigma \int_{z_n}^{z_{n+1}} V(x, y, z)\, dz\right]. \end{aligned} \tag{5.663}$$

Using Eq. (5.633),

$$\boxed{\mathcal{T}_n = \exp\left[j\sigma V_n^{\text{proj}}(x,y)\right].} \tag{5.664}$$

The propagation-only evolution over the same distance is

$$\boxed{\mathcal{P}_{\Delta z_n} = \exp\left[\Delta z_n \mathcal{D}\right].} \tag{5.665}$$

The exact evolution contains an interleaved action of $\mathcal{D}$ and $\mathcal{I}$. For a sufficiently small slice, the first-order Lie splitting approximates this evolution by

$$\boxed{\mathcal{U}_n \simeq \mathcal{P}_{\Delta z_n}\mathcal{T}_n.} \tag{5.666}$$

Thus

$$\boxed{\psi_{n+1} = \mathcal{P}_{\Delta z_n}\left\{\mathcal{T}_n \psi_n\right\},} \tag{5.667}$$

where

$$\psi_n(x,y) \equiv \psi(x,y,z_n). \tag{5.668}$$

5.8.3 The transmission function of one slice

Since $\mathcal{T}_n$ acts by multiplication in real space, Eq. (5.664) defines the slice transmission function

$$\boxed{t_n(x,y) = \exp\left[j\sigma V_n^{\text{proj}}(x,y)\right] = \exp\left[j\sigma \int_{z_n}^{z_{n+1}} V(x,y,z)\,dz\right].} \tag{5.669}$$

The wave immediately after interaction with slice n is therefore

$$\boxed{\psi_n^+(x,y) = t_n(x,y)\psi_n(x,y).} \tag{5.670}$$

Equation (5.669) has the same form as the phase-object transmission function Eq. (5.548), but it applies only to one thin slice.

A potentially confusing terminology should be clarified here. Each slice is treated as a *thin phase grating*, but the weak-phase-object approximation of Sec. 5.6.4 has not been made. In particular, multislice retains the complete exponential

$$e^{j\sigma V_n^{\text{proj}}}, \tag{5.671}$$

rather than replacing it by

$$1 + j\sigma V_n^{\text{proj}}. \tag{5.672}$$

Thus

$$\boxed{\text{thin slice} \quad \neq \quad \text{weak-phase slice.}} \tag{5.673}$$

5.8.4 Propagation between consecutive slices

After interaction with slice n, the wave is propagated over Δz_n. Free propagation was already derived from the fast-electron equation in Eq. (5.535). In transverse Fourier space,

$$\boxed{P_{\Delta z_n}(\mathbf{R}) = \exp\left(-j\pi\lambda\Delta z_n R^2\right),} \tag{5.674}$$

where

$$\mathbf{R} = (R_x, R_y), \qquad R^2 = R_x^2 + R_y^2. \tag{5.675}$$

Consequently,

$$\boxed{\widehat{\psi}_{n+1}(\mathbf{R}) = P_{\Delta z_n}(\mathbf{R})\,\mathcal{F}_{xy}\left\{t_n(x,y)\psi_n(x,y)\right\}.} \tag{5.676}$$

Equivalently, in real space,

$$\boxed{\psi_{n+1}(x,y) = \mathcal{F}_{xy}^{-1}\left\{P_{\Delta z_n}(\mathbf{R})\,\mathcal{F}_{xy}\left[t_n(x,y)\psi_n(x,y)\right]\right\}.} \tag{5.677}$$

Equations (5.669) and (5.677) are the standard multislice recursion.

The computational representation is particularly efficient because the two operators are simple in complementary domains:

$$\boxed{\begin{array}{rcl} \mathcal{T}_n & : & \text{multiplication in real space,} \\ \mathcal{P}_{\Delta z_n} & : & \text{multiplication in Fourier space.} \end{array}} \tag{5.678}$$

The FFT provides the rapid transformation between these two representations (Kirkland, 2010, Sec. 6.6).

5.8.5 Why multislice contains multiple elastic scattering

The recursion in Eq. (5.667) resolves an important limitation of the first Born approximation.

In first Born, Sec. 5.4.4, the wave illuminating every interaction point is replaced by the original incident wave, Eq. (5.435). In multislice, by contrast, the wave that reaches slice $n+1$ is

$$\psi_{n+1} = \mathcal{P}_{\Delta z_n}\mathcal{T}_n\psi_n, \tag{5.679}$$

and ψ_n already contains the effects of all preceding slices.

Thus the interaction sequence is

$$\boxed{\psi_0 \longrightarrow t_0\psi_0 \longrightarrow \mathcal{P}_0\{t_0\psi_0\} \longrightarrow t_1\mathcal{P}_0\{t_0\psi_0\} \longrightarrow \cdots.} \tag{5.680}$$

This is not merely a qualitative statement. The multiple-scattering terms can be exposed explicitly.

How multiple scattering appears in multislice.

For clarity, consider two slices and denote their propagation operators by $\mathcal{P}_0$ and $\mathcal{P}_1$. The exit wave is

$$\psi_2 = \mathcal{P}_1 \mathcal{T}_1 \mathcal{P}_0 \mathcal{T}_0 \psi_0. \tag{5.681}$$

To identify the scattering orders, expand the slice transmission functions formally in powers of the potential:

$$\mathcal{T}_n = 1 + j\sigma V_n^{\text{proj}} + O(V^2). \tag{5.682}$$

Substitution into Eq. (5.681) gives

$$\begin{aligned} \psi_2 = \mathcal{P}_1 \mathcal{P}_0 \psi_0 \\ &+ j\sigma \mathcal{P}_1 V_1^{\text{proj}} \mathcal{P}_0 \psi_0 \\ &+ j\sigma \mathcal{P}_1 \mathcal{P}_0 V_0^{\text{proj}} \psi_0 \\ &+ (j\sigma)^2 \mathcal{P}_1 V_1^{\text{proj}} \mathcal{P}_0 V_0^{\text{proj}} \psi_0 + \cdots . \end{aligned} \tag{5.683}$$

The first line is the unscattered propagation. The next two terms contain one interaction with the specimen and are first order in the potential. The fourth term contains two interactions separated by propagation:

$$\boxed{V_0^{\text{proj}} \quad \longrightarrow \quad \mathcal{P}_0 \quad \longrightarrow \quad V_1^{\text{proj}},} \tag{5.684}$$

and is therefore a double-scattering contribution. Continuing the recursion generates still higher orders automatically.

Multislice therefore does not impose the first Born approximation. Multiple elastic scattering is generated simply because every slice acts on the wave already produced by the preceding slices.

There is nevertheless an important difference from the complete Born series of Eq. (5.427). The latter arises from the full three-dimensional Green-function scattering equation and does not intrinsically impose a longitudinal ordering of the interaction positions. Multislice is derived instead from the paraxial fast-electron equation and retains predominantly forward propagation. Its multiple-scattering paths are therefore naturally ordered along the optical axis:

$$\boxed{z_0 < z_1 < z_2 < \cdots .} \tag{5.685}$$

5.8.6 Where the multislice approximation enters

It is useful to identify precisely what has been approximated.

The fast-electron equation Eq. (5.486) already contains the forward/paraxial approximation derived in Sec. 5.5. Multislice introduces an additional numerical approximation: over each finite slice, the simultaneous action of diffraction and interaction is replaced by their successive action.

The origin of the corresponding error is the noncommutativity in Eq. (5.630). For two operators $\mathcal{D}$ and $\mathcal{I}$, the Baker–Campbell–Hausdorff expansion gives, schematically,

$$e^{\Delta z \mathcal{D}} e^{\Delta z \mathcal{I}} = \exp\left[\Delta z(\mathcal{D}+\mathcal{I}) + \frac{\Delta z^2}{2}[\mathcal{D},\mathcal{I}] + O(\Delta z^3)\right]. \tag{5.686}$$

Thus the leading local splitting error is controlled by

$$[\mathcal{D},\mathcal{I}]. \tag{5.687}$$

For the present operators this commutator can be evaluated explicitly.

The commutator between diffraction and interaction.

Using Eqs. (5.627) and (5.628),

$$\begin{aligned}[\mathcal{D},\mathcal{I}]\psi &= \mathcal{D}\left(j\sigma V\psi\right) - j\sigma V\mathcal{D}\psi \\ &= -\frac{\lambda\sigma}{4\pi}\left[\nabla_\perp^2(V\psi) - V\nabla_\perp^2\psi\right].\end{aligned} \tag{5.688}$$

Using the product rule,

$$\nabla_\perp^2(V\psi) = V\nabla_\perp^2\psi + 2\nabla_\perp V\cdot\nabla_\perp\psi + (\nabla_\perp^2 V)\psi, \tag{5.689}$$

so

$$\boxed{[\mathcal{D},\mathcal{I}]\psi = -\frac{\lambda\sigma}{4\pi}\left[2\nabla_\perp V\cdot\nabla_\perp\psi + (\nabla_\perp^2 V)\psi\right].} \tag{5.690}$$

Equation (5.690) gives a useful physical interpretation of slice-thickness errors. Interaction and propagation fail to commute most strongly where the potential changes rapidly in the transverse direction and where the electron wave itself contains strong transverse variations. Atomic potentials, which are sharply peaked around nuclei, therefore place demanding requirements on the slicing and spatial sampling.

This also explains why the accuracy of multislice depends not merely on the nominal value of Δz, but on how the three-dimensional potential is distributed within each slice (Kirkland, 2010, Secs. 6.7 and 6.11).

5.8.7 Choosing the slice thickness

There is no universal slice thickness that is optimal for every specimen and every desired resolution.

A useful indication comes from the propagation phase derived in Eq. (5.535). Across one slice, a transverse spatial frequency R accumulates the phase

$$\Delta\phi_{\text{prop}} = \pi\lambda\Delta z\, R^2. \tag{5.691}$$

If this phase changes substantially over one slice for frequencies that must be represented accurately, interaction and propagation are being separated over too large a longitudinal interval.

Thus

$$\boxed{\pi\lambda\Delta z\, R_{\max}^2 \ll 1} \tag{5.692}$$

is a useful conservative indicator, although it is not by itself a complete convergence criterion because the spatial variation of $V(x, y, z)$ also matters.

In practice, the appropriate Δz should therefore be established by convergence: the calculation is repeated with progressively smaller slices until the exit wave, or the final observable of interest, changes negligibly.

For biological specimens there is usually no natural crystallographic layering along the beam direction. The slicing should consequently be regarded as a numerical discretization of the continuous three-dimensional electrostatic potential rather than as a statement that the specimen itself consists of physical layers.

5.8.8 A symmetric splitting

The first-order splitting in Eq. (5.666) assigns the entire interaction to one side of the propagation interval. A more symmetric approximation is obtained by placing the interaction at the center of the propagation step:

$$\boxed{\mathcal{U}_n \simeq \mathcal{P}_{\Delta z_n/2}\, \mathcal{T}_n\, \mathcal{P}_{\Delta z_n/2}.} \tag{5.693}$$

This is the symmetric, or Strang, splitting. For sufficiently smooth operators it cancels the leading asymmetric commutator error: its local error is of order

$$O(\Delta z_n^3), \tag{5.694}$$

rather than

$$O(\Delta z_n^2) \tag{5.695}$$

for the first-order splitting.

Written explicitly,

$$\boxed{\psi_{n+1} \simeq \mathcal{P}_{\Delta z_n/2}\left\{t_n\,\mathcal{P}_{\Delta z_n/2}\left[\psi_n\right]\right\}.} \tag{5.696}$$

For a sequence of equal slices, adjacent half-step propagations combine naturally. Symmetric and other higher-order slice formulations are discussed by Kirkland (2010, Sec. 6.11).

The improvement in formal order does not eliminate the need for adequate representation of the sharply varying atomic potential. Spatial sampling and potential subslicing can remain the dominant sources of error.

5.8.9 Spatial sampling, bandwidth, and aliasing

Multislice is normally implemented on a discrete transverse grid. Suppose the sampling intervals are Δx and Δy. The corresponding Nyquist spatial frequencies are

$$R_{x,\mathrm{N}} = \frac{1}{2\Delta x}, \qquad R_{y,\mathrm{N}} = \frac{1}{2\Delta y}. \tag{5.697}$$

The interaction step multiplies two functions in real space,

$$\psi_n^+ = t_n \psi_n. \tag{5.698}$$

According to the Fourier product–convolution relation, this multiplication becomes a convolution in spatial-frequency space. Consequently, even if both factors are individually band limited, their product can contain spatial frequencies beyond the Nyquist boundary.

On a discrete FFT grid these frequencies do not simply disappear. They wrap into the represented frequency interval and become *aliased* frequencies, producing potentially severe numerical artifacts.

A standard multislice implementation therefore applies an explicit bandwidth limit. Kirkland describes the commonly used symmetric 2/3 rule: the wave and transmission function are restricted so that their convolution does not contaminate the retained frequency band (Kirkland, 2010, Sec. 6.8).

This numerical bandwidth should not be confused with a physical objective aperture. It is an anti-aliasing device required by the discrete representation.

5.8.10 Periodic boundaries and the simulation box

The FFT implementation introduces another important numerical assumption. A discretely Fourier-transformed function is implicitly periodic in the transverse directions. Thus a potential represented in a finite box of dimensions $L_x \times L_y$ is mathematically repeated:

$$V(x + L_x, y, z) = V(x, y, z), \tag{5.699}$$

and similarly in y.

For a crystal this periodicity can be physically appropriate. For an isolated biological particle it is not. A box containing one particle actually represents an infinite periodic array of copies of that particle unless the copies are sufficiently separated that they do not affect the region of interest.

The practical solution is to embed the particle and surrounding solvent in a transverse simulation box sufficiently large that propagation from one periodic copy does not contaminate another:

$$\boxed{\text{isolated particle} \quad \longrightarrow \quad \text{sufficiently padded periodic supercell.}} \tag{5.700}$$

Both the real-space support of the potential and the lateral spreading of the electron wave during Fresnel propagation must be considered when choosing this padding. Periodic-continuation artifacts and the use of larger supercells are discussed in Kirkland (2010, Sec. 6.9).

5.8.11 From the incident wave to the exit wave

Starting from the incident envelope

$$\psi_0(x, y) = \psi_{\text{in}}(x, y), \tag{5.701}$$

the complete multislice calculation is the ordered product

$$\boxed{\psi_{\text{exit}} = \mathcal{P}_{N-1}\mathcal{T}_{N-1} \cdots \mathcal{P}_1\mathcal{T}_1\mathcal{P}_0\mathcal{T}_0\psi_{\text{in}}.} \tag{5.702}$$

For plane-wave illumination,

$$\psi_{\text{in}}(x, y) = 1. \tag{5.703}$$

The result is a generally complex exit wave,

$$\psi_{\text{exit}}(x, y), \tag{5.704}$$

whose amplitude and phase both contain information about the three-dimensional specimen and about multiple elastic scattering.

At this point the specimen calculation is complete. The subsequent electron-optical imaging and detection do not depend on whether the exit wave was obtained from the WPOA, a phase-object model, first Born, or multislice. As discussed in Sec. 5.7, the objective lens acts on the exit wave through its coherent wave-transfer function, after which the detector measures electron intensity and ultimately discrete electron arrivals.

5.8.12 What multislice has changed

The progression from the projection approximation to multislice can now be stated very simply.

Under the projection approximation,

$$V(x, y, z) \longrightarrow V^{\text{proj}}(x, y), \tag{5.705}$$

and the complete specimen acts at once:

$$\psi_{\text{exit}} = \exp\left[j\sigma V^{\text{proj}}\right] \psi_{\text{in}}. \tag{5.706}$$

In multislice, depth is retained:

$$\boxed{\begin{array}{l} \psi_{\text{in}} \xrightarrow{t_0} \text{interaction} \xrightarrow{\mathcal{P}_0} \text{propagation} \xrightarrow{t_1} \cdots \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \xrightarrow{\mathcal{P}_{N-1}} \psi_{\text{exit}}. \end{array}} \tag{5.707}$$

The essential advance is not merely that the specimen has been divided into slices. It is that the electron wave is allowed to evolve *between* successive interactions. Each slice therefore sees the wave produced by all preceding slices.

Multislice consequently retains multiple elastic scattering while remaining within the forward/paraxial approximation used to derive the fast-electron equation. It abandons the projection approximation, but it does not return to the unrestricted three-dimensional Lippmann–Schwinger problem.

This distinction will be particularly useful when we examine the three-dimensional geometry of elastic scattering and the curvature of the Ewald sphere in next section.

5.9 The Ewald sphere and genuinely three-dimensional elastic scattering

The projection approximation of Sec. 5.6.2 removes the depth coordinate from the specimen:

$$V(x, y, z) \longrightarrow V^{\text{proj}}(x, y) = \int V(x, y, z)\, dz. \tag{5.708}$$

In Fourier space, the projection theorem, Eq. (5.261), states that this corresponds to retaining the plane

$$R_z = 0 \tag{5.709}$$

of the three-dimensional Fourier transform of the potential.

Multislice, in contrast, retains the depth ordering of the specimen and allows the electron wave to propagate between successive depths. We now examine the same distinction from reciprocal space.

The central geometrical fact is remarkably simple: elastic scattering does not allow an arbitrary change in the electron wavevector. Because the electron energy is unchanged, Eq. (5.337) gives

$$\boxed{\|\mathbf{k}_{\text{f}}\| = \|\mathbf{k}_{\text{i}}\| = k = \frac{2\pi}{\lambda}.} \tag{5.710}$$

Consequently, all possible final wavevectors lie on a sphere of radius $2\pi/\lambda$ in wavevector space. This is the *Ewald sphere*.

For the present discussion it is particularly useful to express the same geometry in the spatial-frequency-transfer convention already introduced in Eq. (5.341).

5.9.1 The Ewald sphere in the spatial-frequency space

Let the incident beam propagate along the positive z direction,

$$\mathbf{k}_{\text{i}} = k\hat{\mathbf{z}}. \tag{5.711}$$

The wavevector transfer was defined previously in Eq. (5.338) as

$$\mathbf{q} = \mathbf{k}_{\text{f}} - \mathbf{k}_{\text{i}}, \tag{5.712}$$

and the corresponding spatial-frequency transfer in Eq. (5.341) as

$$\boxed{\mathbf{R} = \frac{\mathbf{q}}{2\pi}.} \tag{5.713}$$

Thus

$$\boxed{\mathbf{k}_\mathrm{f} = \mathbf{k}_\mathrm{i} + 2\pi\mathbf{R}.} \tag{5.714}$$

In this section,

$$\mathbf{R} = (\mathbf{R}_\perp, R_z), \qquad \mathbf{R}_\perp = (R_x, R_y), \qquad R_\perp = \|\mathbf{R}_\perp\|. \tag{5.715}$$

Equation of the Ewald sphere in transfer-frequency coordinates.

Elasticity requires

$$\|\mathbf{k}_\mathrm{i} + 2\pi\mathbf{R}\|^2 = k^2. \tag{5.716}$$

Using Eq. (5.711),

$$\mathbf{k}_\mathrm{i} + 2\pi\mathbf{R} = \begin{pmatrix} 2\pi R_x \\ 2\pi R_y \\ k + 2\pi R_z \end{pmatrix}. \tag{5.717}$$

Therefore,

$$(2\pi)^2 R_\perp^2 + (k + 2\pi R_z)^2 = k^2. \tag{5.718}$$

Since

$$k = \frac{2\pi}{\lambda}, \tag{5.719}$$

division by $(2\pi)^2$ gives

$$\boxed{R_\perp^2 + \left(R_z + \frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.} \tag{5.720}$$

Thus, in spatial-frequency-transfer coordinates, the Ewald sphere has radius

$$\frac{1}{\lambda} \tag{5.721}$$

and is centered at

$$\left(0, 0, -\frac{1}{\lambda}\right). \tag{5.722}$$

Importantly, it passes through the origin $\mathbf{R} = \mathbf{0}$, which corresponds to the unscattered beam.

Solving Eq. (5.720) for R_z gives two branches:

$$\boxed{R_z = -\frac{1}{\lambda} \pm \sqrt{\frac{1}{\lambda^2} - R_\perp^2}.} \tag{5.723}$$

For $R_\perp = 0$, the positive sign gives

$$R_z = 0, \tag{5.724}$$

corresponding to forward propagation, whereas the negative sign gives

$$R_z = -\frac{2}{\lambda}, \tag{5.725}$$

corresponding to exact backscattering. The forward/paraxial theory developed in Sec. 5.5 retains the branch near $R_z = 0$:

$$\boxed{R_z^{\mathrm{E}}(R_\perp) = -\frac{1}{\lambda} + \sqrt{\frac{1}{\lambda^2} - R_\perp^2}.} \tag{5.726}$$

The Ewald sphere is tangent at the forward-scattering point to the plane

$$\boxed{R_z = 0.} \tag{5.727}$$

Thus the projection approximation may be regarded geometrically as the replacement of the curved Ewald sphere by its tangent plane at the origin. The departure from this plane increases quadratically with the transverse spatial frequency $R_\perp$.

Equation (5.720) deserves a notational clarification. Two different spatial-frequency vectors are involved in scattering, and although they have the same units, they have different physical meanings.

For any plane wave with wavevector $\mathbf{k}$, define its *absolute spatial-frequency vector* as

$$\mathbf{R}^{\mathrm{abs}} \equiv \frac{\mathbf{k}}{2\pi}. \tag{5.728}$$

Thus the incident and scattered waves have absolute spatial frequencies

$$\mathbf{R}_{\mathrm{i}}^{\mathrm{abs}} = \frac{\mathbf{k}_{\mathrm{i}}}{2\pi}, \qquad \mathbf{R}_{\mathrm{f}}^{\mathrm{abs}} = \frac{\mathbf{k}_{\mathrm{f}}}{2\pi}. \tag{5.729}$$

For elastic scattering,

$$\left\|\mathbf{R}_{\mathrm{i}}^{\mathrm{abs}}\right\| = \left\|\mathbf{R}_{\mathrm{f}}^{\mathrm{abs}}\right\| = \frac{1}{\lambda}. \tag{5.730}$$

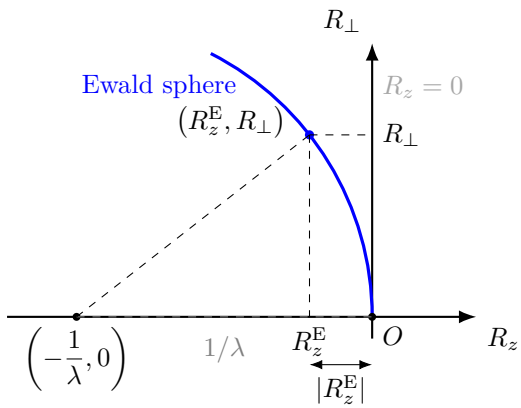


Figure 5.4: Meridional section of the Ewald sphere in transfer-frequency coordinates $(R_z, R_\perp)$. The sphere has radius $1/\lambda$ and center $(-1/\lambda, 0)$. Its forward branch passes through the origin and is tangent there to the plane $R_z = 0$, which in this two-dimensional section coincides with the vertical axis. A scattered component with transverse spatial frequency $R_\perp$ lies at $\left(R_z^{\mathrm{E}}, R_\perp\right)$ on the curved Ewald sphere.

The quantity introduced previously in Eq. (5.341) is different. It is the *spatial-frequency transfer*,

$$\boxed{\mathbf{R} \equiv \frac{\mathbf{k}_{\mathrm{f}} - \mathbf{k}_{\mathrm{i}}}{2\pi} = \mathbf{R}_{\mathrm{f}}^{\mathrm{abs}} - \mathbf{R}_{\mathrm{i}}^{\mathrm{abs}}.} \tag{5.731}$$

The absolute spatial frequency specifies the propagation direction and wavelength of a particular plane wave. The transfer spatial frequency instead specifies the change produced by scattering: it is zero for the unscattered beam and is directly related to the momentum transfer through Eq. (5.341).

For an incident beam along the positive z direction,

$$\mathbf{R}_{\mathrm{i}}^{\mathrm{abs}} = \begin{pmatrix} 0 \\ 0 \\ 1/\lambda \end{pmatrix}. \tag{5.732}$$

Using Eq. (5.731), the absolute spatial frequency of the scattered wave is therefore

$$\boxed{\mathbf{R}_{\mathrm{f}}^{\mathrm{abs}} = \begin{pmatrix} R_x \\ R_y \\ 1/\lambda + R_z \end{pmatrix}.} \tag{5.733}$$

Elasticity requires

$$\left\|\mathbf{R}_{\mathrm{f}}^{\mathrm{abs}}\right\| = \frac{1}{\lambda}, \tag{5.734}$$

and hence

$$R_x^2 + R_y^2 + \left(R_z + \frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}, \tag{5.735}$$

which is exactly Eq. (5.720).

5.9.2 Connection with the scattering angle

The same geometry can be expressed directly through the scattering angle Θ introduced before Eq. (5.352). If the outgoing wavevector forms the angle Θ with the positive z axis, its transverse and longitudinal components are

$$k_{\mathrm{f},\perp} = k\sin\Theta, \qquad k_{\mathrm{f},z} = k\cos\Theta. \tag{5.736}$$

Spatial-frequency transfer as a function of the scattering angle.

Let the incident beam propagate along the positive z axis. Its wavevector is therefore

$$\boxed{\mathbf{k}_{\mathrm{i}} = \begin{pmatrix} 0 \\ 0 \\ k \end{pmatrix}, \qquad k = \frac{2\pi}{\lambda}.} \tag{5.737}$$

Let the scattered wave make an angle Θ with the incident beam, and let ϕ denote its azimuthal angle around the z axis. Since elastic scattering preserves the magnitude of the wavevector, Eq. (5.710), the final wavevector has the same magnitude k:

$$\boxed{\mathbf{k}_{\mathrm{f}} = k\begin{pmatrix} \sin\Theta\cos\phi \\ \sin\Theta\sin\phi \\ \cos\Theta \end{pmatrix}.} \tag{5.738}$$

The wavevector transfer was defined in Eq. (5.338) as

$$\mathbf{q} = \mathbf{k}_{\mathrm{f}} - \mathbf{k}_{\mathrm{i}}. \tag{5.739}$$

Substituting Eqs. (5.737) and (5.738),

$$\mathbf{q} = k \begin{pmatrix} \sin\Theta\cos\phi \\ \sin\Theta\sin\phi \\ \cos\Theta \end{pmatrix} - k \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{5.740}$$

$$= k \begin{pmatrix} \sin\Theta\cos\phi \\ \sin\Theta\sin\phi \\ \cos\Theta - 1 \end{pmatrix}. \tag{5.741}$$

The corresponding spatial-frequency transfer is, from Eq. (5.341),

$$\mathbf{R} = \frac{\mathbf{q}}{2\pi}. \tag{5.742}$$

Using $k = 2\pi/\lambda$,

$$\boxed{\mathbf{R} = \frac{1}{\lambda} \begin{pmatrix} \sin\Theta\cos\phi \\ \sin\Theta\sin\phi \\ \cos\Theta - 1 \end{pmatrix}.} \tag{5.743}$$

We can now identify separately its transverse and longitudinal components. The transverse part is

$$\mathbf{R}_\perp = \frac{\sin\Theta}{\lambda} \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix}. \tag{5.744}$$

Its magnitude is therefore

$$R_\perp = \sqrt{R_x^2 + R_y^2} \tag{5.745}$$

$$= \frac{\sin\Theta}{\lambda}\sqrt{\cos^2\phi + \sin^2\phi} \tag{5.746}$$

$$= \frac{\sin\Theta}{\lambda}. \tag{5.747}$$

Hence

$$\boxed{R_\perp = \frac{\sin\Theta}{\lambda}.} \tag{5.748}$$

The longitudinal component follows directly from the third component of Eq. (5.743):

$$\boxed{R_z = \frac{\cos\Theta - 1}{\lambda}.} \tag{5.749}$$

Using the trigonometric identity

$$1 - \cos\Theta = 2\sin^2\left(\frac{\Theta}{2}\right), \tag{5.750}$$

we obtain

$$\boxed{R_z = -\frac{2}{\lambda}\sin^2\left(\frac{\Theta}{2}\right).} \tag{5.751}$$

Notice that $R_z \leq 0$ for the forward branch. The scattered wave still propagates predominantly along $+z$, but its longitudinal component is slightly smaller than that of the incident wave because part of its wavevector now lies in the transverse direction.

Finally, the magnitude of the complete three-dimensional spatial-frequency transfer is

$$R = \|\mathbf{R}\| = \sqrt{R_\perp^2 + R_z^2}. \tag{5.752}$$

Substituting Eqs. (5.748) and (5.751),

$$R^2 = \frac{\sin^2\Theta}{\lambda^2} + \frac{4}{\lambda^2}\sin^4\left(\frac{\Theta}{2}\right). \tag{5.753}$$

To simplify the first term, use

$$\sin\Theta = 2\sin\left(\frac{\Theta}{2}\right)\cos\left(\frac{\Theta}{2}\right). \tag{5.754}$$

Therefore,

$$\begin{aligned} R^2 &= \frac{4}{\lambda^2}\sin^2\left(\frac{\Theta}{2}\right)\cos^2\left(\frac{\Theta}{2}\right) \\ &\quad + \frac{4}{\lambda^2}\sin^4\left(\frac{\Theta}{2}\right) && (5.755) \\ &= \frac{4}{\lambda^2}\sin^2\left(\frac{\Theta}{2}\right)\left[\cos^2\left(\frac{\Theta}{2}\right) + \sin^2\left(\frac{\Theta}{2}\right)\right] && (5.756) \\ &= \frac{4}{\lambda^2}\sin^2\left(\frac{\Theta}{2}\right). && (5.757) \end{aligned}$$

Taking the positive square root,

$$\boxed{R = \frac{2}{\lambda} \sin\left(\frac{\Theta}{2}\right).} \tag{5.758}$$

This reproduces exactly the scattering-vector result obtained previously in Eq. (5.353).

5.9.3 The paraxial curvature of the Ewald sphere

For the small scattering angles relevant to cryo-EM,

$$\lambda R_\perp \ll 1. \tag{5.759}$$

The forward Ewald branch, Eq. (5.726), may then be expanded explicitly.

Paraxial curvature of the Ewald sphere.

Rewrite Eq. (5.726) as

$$R_z^{\mathrm{E}} = \frac{1}{\lambda}\left[-1 + \sqrt{1 - \lambda^2 R_\perp^2}\right]. \tag{5.760}$$

Using

$$\sqrt{1-x} = 1 - \frac{x}{2} - \frac{x^2}{8} + O(x^3), \tag{5.761}$$

with

$$x = \lambda^2 R_\perp^2, \tag{5.762}$$

gives

$$R_z^{\mathrm{E}} = -\frac{\lambda R_\perp^2}{2} - \frac{\lambda^3 R_\perp^4}{8} + O(\lambda^5 R_\perp^6). \tag{5.763}$$

To paraxial order,

$$\boxed{R_z^{\mathrm{E}} \simeq -\frac{\lambda R_\perp^2}{2}.} \tag{5.764}$$

5.9.4 What the projection approximation removes

The significance of this curvature becomes apparent when we return to the first-Born relation derived in Eq. (5.445),

$$f_{\mathrm{B}}(\mathbf{R}) = \frac{me}{2\pi\hbar^2}\widehat{V}(\mathbf{R}). \tag{5.765}$$

This equation states that a scattering transfer $\mathbf{R}$ probes the corresponding three-dimensional Fourier component of the specimen potential. Elasticity, however, restricts the allowed values of $\mathbf{R}$ to the Ewald sphere. Hence, for a given transverse spatial frequency $\mathbf{R}_\perp$,

$$\boxed{f_{\mathrm{B}}^{\mathrm{E}}(\mathbf{R}_\perp) = \frac{me}{2\pi\hbar^2}\widehat{V}\left(\mathbf{R}_\perp, R_z^{\mathrm{E}}(R_\perp)\right).} \tag{5.766}$$

By contrast, projection along z gives, by the Fourier projection theorem of Eq. (5.261),

$$\widehat{V}^{\mathrm{proj}}(\mathbf{R}_\perp) = \widehat{V}(\mathbf{R}_\perp, 0). \tag{5.767}$$

The projected first-Born description therefore uses

$$\boxed{f_{\mathrm{B}}^{\mathrm{proj}}(\mathbf{R}_\perp) = \frac{me}{2\pi\hbar^2}\widehat{V}(\mathbf{R}_\perp, 0).} \tag{5.768}$$

Comparing Eqs. (5.766) and (5.768) gives a major conceptual result:

$$\boxed{\begin{array}{c}\text{within the single-scattering description,}\\ \text{the projection approximation is the}\\ \text{flat-Ewald-sphere approximation.}\end{array}} \tag{5.769}$$

More precisely, the curved sampling surface

$$R_z = R_z^{\mathrm{E}}(R_\perp) \tag{5.770}$$

is replaced by its tangent plane

$$R_z = 0. \tag{5.771}$$

The qualification "within the single-scattering description" is important. The Ewald sphere itself follows from elastic energy conservation and does not require the Born approximation. What the first Born approximation adds is the particularly simple identification

$$f_{\mathrm{B}}(\mathbf{R}) \propto \widehat{V}(\mathbf{R}), \tag{5.772}$$

so that elastic scattering can be described as sampling the three-dimensional Fourier transform of the specimen potential on the Ewald sphere.

Multislice does not require a separate Ewald-sphere sampling operation. Instead, it evolves the electron wave through the depth-dependent potential. The Fresnel propagator between slices introduces the longitudinal phase associated with Ewald-sphere curvature automatically. Indeed, in the single-scattering limit, the multislice depth phase

$$\exp\left(j\pi\lambda z R_\perp^2\right) \tag{5.773}$$

is, using Eq. (5.764),

$$\exp\left(-j2\pi R_z^{\mathrm{E}} z\right), \tag{5.774}$$

and the coherent sum over depth therefore reproduces sampling of $\widehat{V}$ on the paraxial Ewald sphere.

Beyond single scattering, however, the simple picture of sampling the original specimen Fourier transform once on a single Ewald surface no longer applies. Each slice acts on the wave already modified by preceding slices, so spatial-frequency components are successively generated, propagated, and mixed. Multislice therefore incorporates the same depth-dependent phase physics that gives rise to Ewald curvature, while also retaining multiple elastic scattering within the forward/paraxial approximation.

Standard multislice normally uses the Fresnel propagator and therefore contains the *paraxial* form of the Ewald curvature. If the exact forward angular-spectrum propagator were used instead,

$$\exp\left[j\Delta z\left(\sqrt{k^2-(2\pi R_\perp)^2}-k\right)\right], \tag{5.775}$$

the exact forward Ewald geometry would be retained rather than its paraxial approximation.

5.9.5 Depth appears as a phase

The loss of the z coordinate under projection can now be seen directly.

Define the two-dimensional Fourier transform of the potential at each depth,

$$\boxed{\widehat{V}_\perp(\mathbf{R}_\perp, z) = \int_{\mathbb{R}^2} V(\boldsymbol{\rho}, z)\exp\left(-j2\pi\mathbf{R}_\perp\cdot\boldsymbol{\rho}\right)d^2\rho,} \tag{5.776}$$

where

$$\boldsymbol{\rho} = (x, y). \tag{5.777}$$

The three-dimensional Fourier transform can then be written as

$$\boxed{\widehat{V}(\mathbf{R}_\perp, R_z) = \int \widehat{V}_\perp(\mathbf{R}_\perp, z) \exp(-j2\pi R_z z)\, dz.} \tag{5.778}$$

On the Ewald sphere this becomes

$$\widehat{V}\left(\mathbf{R}_\perp, R_z^{\mathrm{E}}\right) = \int \widehat{V}_\perp(\mathbf{R}_\perp, z) \exp\left[-j2\pi R_z^{\mathrm{E}}(R_\perp) z\right] dz. \tag{5.779}$$

Using the paraxial curvature Eq. (5.764),

$$\boxed{\exp\left[-j2\pi R_z^{\mathrm{E}} z\right] \simeq \exp\left[j\pi\lambda z R_\perp^2\right].} \tag{5.780}$$

Thus contributions from different depths are not equivalent. They enter the scattering amplitude with different phases.

Under the projection approximation,

$$R_z = 0, \tag{5.781}$$

and therefore

$$\exp(-j2\pi R_z z) = 1. \tag{5.782}$$

All depths then contribute with the same longitudinal phase and are collapsed into

$$\int \widehat{V}_\perp(\mathbf{R}_\perp, z)\, dz. \tag{5.783}$$

We can therefore state the physical content of Ewald curvature as

$$\boxed{\text{Ewald curvature} \quad \Longleftrightarrow \quad \text{depth-dependent phase.}} \tag{5.784}$$

A single image does not thereby become a direct measurement of the three-dimensional potential. Rather, the important point is that the image is no longer generated by a pure geometrical projection: the depth distribution of the specimen affects the complex scattering amplitudes.

5.9.6 Why a very flat Ewald sphere can still matter

For high-energy electrons the Ewald sphere is extremely large. At 300 keV,

$$\lambda \simeq 1.97\,\mathrm{pm} = 0.0197\,\text{Å}, \tag{5.785}$$

so its radius in spatial-frequency coordinates is approximately

$$\frac{1}{\lambda} \simeq 50.8\,\text{Å}^{-1}. \tag{5.786}$$

At first sight such a large radius suggests that the sphere should be essentially flat over the spatial-frequency range relevant to biological structure.

Geometrically this is true, but phase is sensitive to the product of the small curvature and the specimen thickness.

Consider, for example, a real-space period

$$d = 3\,\text{Å}, \qquad R_\perp = \frac{1}{d} \simeq 0.333\,\text{Å}^{-1}. \tag{5.787}$$

Equation (5.764) gives

$$R_z^{\mathrm{E}} \simeq -\frac{0.0197}{2(3)^2}\,\text{Å}^{-1} \simeq -1.09 \times 10^{-3}\,\text{Å}^{-1}. \tag{5.788}$$

This longitudinal frequency is tiny compared with $R_\perp$. Nevertheless, across a particle of thickness

$$T = 100\,\text{nm} = 1000\,\text{Å}, \tag{5.789}$$

the phase difference between the front and back of the particle is

$$\Delta\phi_{\mathrm{E}} = 2\pi \left|R_z^{\mathrm{E}}\right| T \tag{5.790}$$

$$\simeq \pi\lambda T R_\perp^2 \tag{5.791}$$

$$\simeq 6.9\,\text{rad}. \tag{5.792}$$

Thus a geometrically minute Ewald curvature can generate many radians of relative phase across a large biological particle at high resolution.

More generally,

$$\boxed{\Delta\phi_{\mathrm{E}} \simeq \pi\lambda T R_\perp^2 = \frac{\pi\lambda T}{d^2}.} \tag{5.793}$$

This is exactly the quantity introduced earlier when deriving the projection criterion in Eq. (5.557). The numerical values in Eq. (5.562) can therefore now be interpreted geometrically as the phase error produced by flattening the Ewald sphere.

The two apparently different arguments are the same physics:

real space		reciprocal space
neglect propagation through depth	$\Longleftrightarrow$	flatten the Ewald sphere,
$\pi\lambda T R_\perp^2 \ll 1$	$\Longleftrightarrow$	$2\pi\lvert R_z^{\mathrm{E}}\rvert T \ll 1.$

(5.794)

The magnitude of the curvature can also be compared with the sampling of a discrete three-dimensional Fourier grid. If the real-space map is sampled

with pixel size a and contains N voxels along one dimension, its field of view is

$$L = Na, \tag{5.795}$$

and the corresponding Fourier-grid spacing is

$$\boxed{\Delta R = \frac{1}{L} = \frac{1}{Na}.} \tag{5.796}$$

Thus the displacement of the Ewald sphere from the projection plane $R_z = 0$, expressed in Fourier voxels, is

$$\boxed{n_{\mathrm{E}} = \frac{|R_z^{\mathrm{E}}|}{\Delta R} \simeq \frac{\lambda N a}{2d^2}.} \tag{5.797}$$

For a real-space box of width

$$L = 50\,\mathrm{nm} = 500\,\text{Å} \tag{5.798}$$

sampled at

$$a = 0.75\,\text{Å/pixel}, \tag{5.799}$$

the number of real-space pixels is approximately

$$N = \frac{L}{a} = \frac{500\,\text{Å}}{0.75\,\text{Å/pixel}} \simeq 667. \tag{5.800}$$

The corresponding Fourier-grid spacing is

$$\boxed{\Delta R = \frac{1}{L} = \frac{1}{500\,\text{Å}} = 2.0 \times 10^{-3}\,\text{Å}^{-1}.} \tag{5.801}$$

At 3 Å resolution, Eq. (5.764) gives

$$|R_z^{\mathrm{E}}| \simeq 1.09 \times 10^{-3}\,\text{Å}^{-1}. \tag{5.802}$$

The displacement from the central projection plane, expressed in Fourier voxels, is therefore

$$\boxed{n_{\mathrm{E}}(3\,\text{Å}) = \frac{|R_z^{\mathrm{E}}|}{\Delta R} \simeq \frac{1.09 \times 10^{-3}}{2.0 \times 10^{-3}} \simeq 0.55 \quad \text{Fourier voxels.}} \tag{5.803}$$

At 2 Å resolution,

$$|R_z^{\mathrm{E}}| \simeq 2.46 \times 10^{-3}\,\text{Å}^{-1}, \tag{5.804}$$

and hence

$$\boxed{n_{\mathrm{E}}(2\,\text{Å}) \simeq \frac{2.46 \times 10^{-3}}{2.0 \times 10^{-3}} \simeq 1.23 \quad \text{Fourier voxels.}} \tag{5.805}$$

Thus, for a 50 nm reconstruction box, the Ewald sphere is displaced by approximately half a Fourier voxel from the central plane at 3 Å resolution, but by more than one Fourier voxel at 2 Å resolution. The curvature is therefore already comparable with the discrete Fourier sampling at near-atomic resolution.

The number of Fourier voxels is not itself a physical measure of Ewald curvature. For fixed physical curvature, $|R_z^{\mathrm{E}}|$, it depends on the chosen box size through $\Delta R = 1/L$. The real-space pixel size determines how many pixels are required to represent that box, but once L is fixed it does not change the Fourier-voxel spacing.

5.9.7 Ewald curvature and Fresnel propagation are the same paraxial phase

The connection with multislice can be made even more explicit. Multislice retained the longitudinal coordinate precisely by allowing the wave produced at one depth to propagate before interacting with subsequent slices. The Fresnel propagator derived in Eq. (5.535) is

$$P_{\Delta z}(\mathbf{R}_\perp) = \exp\left(-j\pi\lambda\Delta z R_\perp^2\right). \tag{5.806}$$

In the single-scattering limit of multislice, this propagation phase reproduces exactly the paraxial curvature of the Ewald sphere.

From depth-dependent Fresnel propagation to the Ewald sphere.

Let the specimen occupy

$$0 \le z \le T. \tag{5.807}$$

Consider a weak scattering contribution generated in an infinitesimal layer at depth z. To first order in the potential, the scattered Fourier amplitude generated by that layer is proportional to

$$j\sigma \widehat{V}_\perp(\mathbf{R}_\perp, z)\, dz. \tag{5.808}$$

Before reaching the exit surface at T, this contribution must propagate through the remaining distance $T-z$. From Eq. (5.535), it therefore acquires the phase

$$\exp\left[-j\pi\lambda(T-z)R_\perp^2\right]. \tag{5.809}$$

Adding coherently the contributions generated at all depths gives

$$\delta\widehat{\psi}_{\text{exit}}(\mathbf{R}_\perp) \propto j\sigma \int_0^T \widehat{V}_\perp(\mathbf{R}_\perp, z) \times \exp\left[-j\pi\lambda(T-z)R_\perp^2\right] dz. \tag{5.810}$$

Separating the part independent of z,

$$\delta\widehat{\psi}_{\text{exit}}(\mathbf{R}_\perp) \propto j\sigma e^{-j\pi\lambda T R_\perp^2} \times \int_0^T \widehat{V}_\perp(\mathbf{R}_\perp, z) e^{j\pi\lambda z R_\perp^2}\, dz. \tag{5.811}$$

We now use the paraxial Ewald-sphere relation derived in Eq. (5.764),

$$R_z^{\text{E}} \simeq -\frac{\lambda R_\perp^2}{2}. \tag{5.812}$$

Multiplying by $-j2\pi z$ gives

$$-j2\pi R_z^{\text{E}} z \simeq -j2\pi\left(-\frac{\lambda R_\perp^2}{2}\right) z \tag{5.813}$$

$$= j\pi\lambda z R_\perp^2. \tag{5.814}$$

Therefore,

$$\boxed{e^{j\pi\lambda z R_\perp^2} = e^{-j2\pi R_z^{\text{E}} z}} \tag{5.815}$$

to paraxial order.

Substituting this result into Eq. (5.811),

$$\delta\widehat{\psi}_{\text{exit}}(\mathbf{R}_\perp) \propto j\sigma e^{-j\pi\lambda T R_\perp^2} \times \int_0^T \widehat{V}_\perp(\mathbf{R}_\perp, z) e^{-j2\pi R_z^{\text{E}} z}\, dz. \tag{5.816}$$

The integral in Eq. (5.816) should now be compared with the definition of the three-dimensional Fourier transform. From Eq. (5.776),

$$\widehat{V}_\perp(\mathbf{R}_\perp, z) = \int_{\mathbb{R}^2} V(\boldsymbol{\rho}, z) e^{-j2\pi \mathbf{R}_\perp \cdot \boldsymbol{\rho}}\, d^2\rho, \tag{5.817}$$

and therefore Eq. (5.778) gives

$$\widehat{V}(\mathbf{R}_\perp, R_z) = \int_{-\infty}^{+\infty} \widehat{V}_\perp(\mathbf{R}_\perp, z) e^{-j2\pi R_z z}\, dz. \tag{5.818}$$

If the specimen occupies $0 \leq z \leq T$, then

$$V(\boldsymbol{\rho}, z) = 0 \qquad \text{outside } 0 \leq z \leq T, \tag{5.819}$$

so the integral over the whole z axis reduces to

$$\widehat{V}(\mathbf{R}_\perp, R_z) = \int_0^T \widehat{V}_\perp(\mathbf{R}_\perp, z) e^{-j2\pi R_z z}\, dz. \tag{5.820}$$

Evaluating this Fourier transform at $R_z = R_z^{\mathrm{E}}$ therefore gives

$$\boxed{\int_0^T \widehat{V}_\perp(\mathbf{R}_\perp, z) e^{-j2\pi R_z^{\mathrm{E}} z}\, dz = \widehat{V}\left(\mathbf{R}_\perp, R_z^{\mathrm{E}}\right).} \tag{5.821}$$

Substitution into Eq. (5.816) finally gives

$$\boxed{\delta\widehat{\psi}_{\mathrm{exit}}(\mathbf{R}_\perp) \propto j\sigma e^{-j\pi\lambda T R_\perp^2} \widehat{V}\left(\mathbf{R}_\perp, R_z^{\mathrm{E}}\right).} \tag{5.822}$$

The factor

$$e^{-j\pi\lambda T R_\perp^2} \tag{5.823}$$

in Eq. (5.822) is common to all contributions when the exit surface is chosen as the reference plane. The physically important depth discrimination is contained in

$$e^{j\pi\lambda z R_\perp^2}, \tag{5.824}$$

which is precisely the paraxial Ewald-curvature phase of Eq. (5.780).

Thus we arrive at another important equivalence:

$$\boxed{\begin{array}{c}\text{paraxial Ewald curvature in reciprocal space}\\ \Longleftrightarrow \\ \text{depth-dependent Fresnel propagation in real space.}\end{array}} \tag{5.825}$$

This explains why multislice restores information that the projection approximation discarded. The separate slices are not important merely because they preserve a record of the atomic z coordinates. They are important because scattering generated at different depths undergoes different amounts of propagation before reaching the exit surface.

The equivalence above applies to the single-scattering limit of multislice. The complete multislice recursion developed in Sec. 5.8 additionally allows the propagated wave to interact again with subsequent slices and therefore includes multiple elastic scattering within the paraxial approximation.

5.9.8 From the Ewald sphere to a real-valued cryo-EM image

So far we have followed one elastically scattered plane wave. For a given transverse spatial frequency $\mathbf{R}_\perp$, its three-dimensional spatial-frequency transfer lies on the forward branch of the Ewald sphere,

$$\mathbf{R} = \left(\mathbf{R}_\perp, R_z^{\mathrm{E}}\right), \tag{5.826}$$

where Eq. (5.726) gives

$$R_z^{\mathrm{E}} < 0. \tag{5.827}$$

This description is sufficient for discussing the complex scattered wave, but an electron microscope does not directly record that complex wave. The detector ultimately records an intensity. We must therefore ask how the scattered waves on the Ewald sphere combine to form the real-valued image.

A first important point is that there are two scattering directions to consider. They correspond to opposite transverse wavevectors,

$$+\,\mathbf{R}_\perp \qquad \text{and} \qquad -\,\mathbf{R}_\perp. \tag{5.828}$$

Both are *forward*-scattered waves: their longitudinal wavevector component remains positive. They should not be confused with forward and backward scattering along the optical axis.

For convenience define the positive magnitude of the longitudinal departure from the central plane as

$$\boxed{\Delta R_z^{\mathrm{E}} \equiv \left|R_z^{\mathrm{E}}\right| = -R_z^{\mathrm{E}}.} \tag{5.829}$$

The two physical scattering points on the forward Ewald sphere are therefore

$$\boxed{\begin{aligned} \mathbf{R}_+ &= \left(+\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right), \\ \mathbf{R}_- &= \left(-\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right). \end{aligned}} \tag{5.830}$$

These two points are mirror images with respect to the optical axis, but they are *not* a Hermitian pair in the three-dimensional Fourier transform. A Hermitian pair would require reversal of all three frequency coordinates.

Why two Ewald-sphere samples contribute to one image Fourier coefficient.

Within the weak-scattering description, write the electron wave in the image plane as

$$\psi_{\mathrm{obs}}(\boldsymbol{\rho}) = 1 + s(\boldsymbol{\rho}), \tag{5.831}$$

where the unscattered beam has been normalized to unit amplitude and s contains the weak scattered contribution.

The detector records the intensity introduced previously in Eq. (3.724),

$$I(\boldsymbol{\rho}) = |1 + s(\boldsymbol{\rho})|^2 \tag{5.832}$$

$$= 1 + s(\boldsymbol{\rho}) + s^*(\boldsymbol{\rho}) + |s(\boldsymbol{\rho})|^2. \tag{5.833}$$

Neglecting the second-order term $|s|^2$ gives

$$\boxed{\delta I(\boldsymbol{\rho}) \simeq s(\boldsymbol{\rho}) + s^*(\boldsymbol{\rho}).} \tag{5.834}$$

Taking the two-dimensional Fourier transform,

$$\boxed{\widehat{\delta I}(\mathbf{R}_\perp) = \widehat{s}(\mathbf{R}_\perp) + \widehat{s}^*(-\mathbf{R}_\perp).} \tag{5.835}$$

This equation explains why the second scattered direction cannot be ignored. The Fourier coefficient of the *real image* at $+\mathbf{R}_\perp$ contains both

$$\widehat{s}(+\mathbf{R}_\perp) \tag{5.836}$$

and the complex conjugate of the scattered wave at

$$-\mathbf{R}_\perp. \tag{5.837}$$

Let $H(\mathbf{R}_\perp)$ denote the coherent wave-transfer function of the objective lens, introduced in Eq. (4.1197). In the first-order Ewald-sphere description,

$$\widehat{s}(+\mathbf{R}_\perp) = j\sigma H(+\mathbf{R}_\perp)\widehat{V}\left(+\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right). \tag{5.838}$$

Similarly,

$$\widehat{s}(-\mathbf{R}_\perp) = j\sigma H(-\mathbf{R}_\perp)\widehat{V}\left(-\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right). \tag{5.839}$$

Because the electrostatic potential $V(\mathbf{r})$ is real, its three-dimensional Fourier transform satisfies Hermitian symmetry,

$$\boxed{\widehat{V}(-\mathbf{R}) = \widehat{V}^*(\mathbf{R}).} \tag{5.840}$$

Therefore,

$$\begin{aligned}\widehat{V}^*&\left(-\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right)\\ &= \widehat{V}\left(+\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right).\end{aligned} \tag{5.841}$$

Taking the complex conjugate of Eq. (5.839) and substituting into Eq. (5.835) gives (See Fig. 5.5)

$$\begin{aligned}\widehat{\delta I}(\mathbf{R}_\perp) = j\sigma H(\mathbf{R}_\perp)&\widehat{V}\left(\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right)\\ &- j\sigma H^*(-\mathbf{R}_\perp)\widehat{V}\left(\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right).\end{aligned} \tag{5.842}$$

Thus a single two-dimensional Fourier coefficient of the measured image contains information from *two distinct points* of the three-dimensional specimen transform:

$$\boxed{\left(\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right) \qquad \text{and} \qquad \left(\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right).} \tag{5.843}$$

Their separation along the longitudinal spatial-frequency direction is

$$\boxed{2\Delta R_z^{\mathrm{E}}.} \tag{5.844}$$

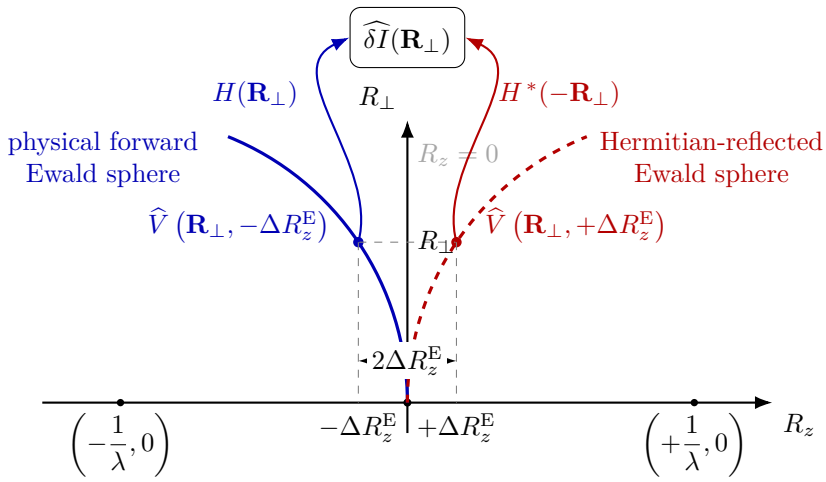


Figure 5.5: Why two three-dimensional specimen Fourier components contribute to one Fourier coefficient of the real-valued image. The solid curve is the physical forward Ewald sphere. For the scattered wave at $+\mathbf{R}_\perp$, the specimen is sampled at $\left(\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right)$. The contribution from the opposite transverse scattering direction enters the image as $\widehat{s}^*(-\mathbf{R}_\perp)$. Because the specimen potential is real, Hermitian symmetry maps this contribution to $\widehat{V}(\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}})$, represented here on the dashed, Hermitian-reflected copy of the Ewald sphere. Thus the single measured image coefficient $\widehat{\delta I}(\mathbf{R}_\perp)$ contains information from two three-dimensional Fourier components separated by $2\Delta R_z^{\mathrm{E}}$ along the longitudinal spatial-frequency direction. The dashed sphere is a graphical representation of the Hermitian mapping and is not a second physical scattering sphere.

In the flat-Ewald-sphere approximation,

$$\Delta R_z^{\mathrm{E}} \longrightarrow 0. \tag{5.845}$$

The two distinct three-dimensional samples of Eq. (5.843) then collapse onto the same point,

$$\widehat{V}\left(\mathbf{R}_\perp, \pm\Delta R_z^{\mathrm{E}}\right) \longrightarrow \widehat{V}\left(\mathbf{R}_\perp, 0\right). \tag{5.846}$$

Equation (5.842) consequently reduces to

$$\widehat{\delta I}(\mathbf{R}_\perp) = j\sigma\left[H(\mathbf{R}_\perp) - H^*(-\mathbf{R}_\perp)\right]\widehat{V}(\mathbf{R}_\perp, 0), \tag{5.847}$$

which is precisely the conventional linear CTF result derived in Eq. (4.1219).

Thus the conventional CTF model implicitly assumes not only weak scattering, but also that the two Ewald-sideband samples can be identified with the same central-section Fourier coefficient.

The two complex sidebands of the CTF. Equation (5.842) already contains the two complex sidebands of the image transfer. For a pure weak-phase object, it reads

$$\begin{aligned}\widehat{\delta I}(\mathbf{R}_\perp) = {} & j\sigma H(\mathbf{R}_\perp)\widehat{V}\left(\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right) \\ & - j\sigma H^*(-\mathbf{R}_\perp)\widehat{V}\left(\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right).\end{aligned} \tag{5.848}$$

It is therefore natural to define the two coherent single-sideband transfer factors as

$$\boxed{H_P(\mathbf{R}_\perp) \equiv jH(\mathbf{R}_\perp),} \tag{5.849}$$

and

$$\boxed{H_Q(\mathbf{R}_\perp) \equiv -jH^*(-\mathbf{R}_\perp).} \tag{5.850}$$

With these definitions, Eq. (5.848) becomes

$$\begin{aligned}\widehat{\delta I}(\mathbf{R}_\perp) = {} & \sigma H_P(\mathbf{R}_\perp)\widehat{V}\left(\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right) \\ & + \sigma H_Q(\mathbf{R}_\perp)\widehat{V}\left(\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right).\end{aligned} \tag{5.851}$$

The factors $+j$ and $-j$ in Eqs. (5.849) and (5.850) are not additional lens phases. They arise from the weak-phase interaction with the specimen.

Origin of the $\pi/2$ in the two sidebands.

Within the weak-phase approximation, the specimen transmission function was derived in Eq. (5.567) as

$$t(\boldsymbol{\rho}) \simeq 1 + j\varphi(\boldsymbol{\rho}). \tag{5.852}$$

Thus the scattered part of the exit wave is

$$s_{\mathrm{exit}}(\boldsymbol{\rho}) = j\varphi(\boldsymbol{\rho}). \tag{5.853}$$

The factor j means that, before the objective-lens phase is applied, the weak phase-scattered wave is in quadrature with the unscattered reference wave:

$$\boxed{j = e^{j\pi/2}.} \tag{5.854}$$

After coherent propagation through the objective, Eq. (4.1197) multiplies this component by $H(\mathbf{R}_\perp)$. Hence the positive-frequency sideband carries the factor

$$jH(\mathbf{R}_\perp). \tag{5.855}$$

When the detector forms the real intensity, the conjugate contribution enters through $s^*(\boldsymbol{\rho})$. Its Fourier coefficient at $+\mathbf{R}_\perp$ is obtained from the scattered wave at $-\mathbf{R}_\perp$ and contains

$$-jH^*(-\mathbf{R}_\perp), \tag{5.856}$$

because

$$\boxed{-j = e^{-j\pi/2}.} \tag{5.857}$$

Therefore the $\pm\pi/2$ appearing when the two sidebands are written as complex exponentials is simply another way of writing the factors $\pm j$ already derived from weak-phase image formation. It is not an additional aberration of the objective lens.

For example, if the coherent transfer contains only a symmetric aberration phase,

$$H(\mathbf{R}_\perp) = P(\mathbf{R}_\perp)e^{-j\chi_{\mathrm{sym}}(\mathbf{R}_\perp)}, \tag{5.858}$$

then

$$H_P = jPe^{-j\chi_{\mathrm{sym}}} \tag{5.859}$$

$$= Pe^{-j(\chi_{\mathrm{sym}}-\pi/2)}, \tag{5.860}$$

whereas

$$H_Q = -jPe^{+j\chi_{\mathrm{sym}}} \tag{5.861}$$

$$= Pe^{+j(\chi_{\mathrm{sym}}-\pi/2)}. \tag{5.862}$$

The precise appearance of $\chi \pm \pi/2$ depends on the sign convention used for the aberration phase and the coherent transfer function. The physical statement is convention independent:

$$\boxed{H_P = jH, \qquad H_Q = -jH^*(-\mathbf{R}_\perp).} \tag{5.863}$$

Including symmetric and asymmetric aberrations and a phase plate. We next make the coherent transfer function H explicit. As derived in Sec. 4.15, the deterministic aberration phase can be separated into symmetric and antisymmetric parts,

$$\chi(\mathbf{R}_\perp) = \chi_{\text{sym}}(\mathbf{R}_\perp) + \chi_{\text{asym}}(\mathbf{R}_\perp), \tag{5.864}$$

with

$$\chi_{\text{sym}}(-\mathbf{R}_\perp) = \chi_{\text{sym}}(\mathbf{R}_\perp), \tag{5.865}$$

and

$$\chi_{\text{asym}}(-\mathbf{R}_\perp) = -\chi_{\text{asym}}(\mathbf{R}_\perp) \tag{5.866}$$

[Eq. (4.2075)].

A phase plate introduces an additional *physical* phase difference between the scattered and unscattered waves. Let

$$\phi_{\text{PP}}(\mathbf{R}_\perp) \tag{5.867}$$

denote this relative phase shift after the phase of the unscattered reference beam has been chosen as zero. For the usual rotationally symmetric phase-plate model,

$$\phi_{\text{PP}}(-\mathbf{R}_\perp) = \phi_{\text{PP}}(\mathbf{R}_\perp). \tag{5.868}$$

The coherent wave-transfer function may then be written

$$\boxed{\begin{aligned} H(\mathbf{R}_\perp) = P(\mathbf{R}_\perp) e^{-j\chi_{\text{asym}}(\mathbf{R}_\perp)} \\ \times e^{-j[\chi_{\text{sym}}(\mathbf{R}_\perp)+\phi_{\text{PP}}(\mathbf{R}_\perp)]}. \end{aligned}} \tag{5.869}$$

Here ϕ_{PP} is conceptually different from the $\pi/2$ derived in Sec. 5.9.8. The former is deliberately introduced by a physical device in the microscope; the latter arises from the factor j in the weak-phase specimen wave.

For compactness, define the total symmetric coherent phase

$$\boxed{\Phi(\mathbf{R}_\perp) \equiv \chi_{\text{sym}}(\mathbf{R}_\perp) + \phi_{\text{PP}}(\mathbf{R}_\perp).} \tag{5.870}$$

Using Eq. (5.869),

$$\boxed{H_P = jPe^{-j\chi_{\text{asym}}}e^{-j\Phi},} \tag{5.871}$$

whereas

$$\boxed{H_Q = -jPe^{-j\chi_{\text{asym}}}e^{+j\Phi}.} \tag{5.872}$$

Notice that the antisymmetric phase has the same sign in the two sidebands, whereas the symmetric phase changes sign. This is precisely the origin of the factorization derived previously in Eq. (4.2081).

Including amplitude contrast. The pure-phase result is still not the complete weak-object model. As derived in Sec. 4.15, a weak specimen may modify both the amplitude and the phase of the electron wave,

$$\psi_i \simeq 1 + a + j\varphi. \tag{5.873}$$

Under the conventional one-parameter amplitude-contrast approximation, the two object modulations are taken to have the same spatial dependence. Using the normalized amplitude-contrast coefficient A defined in Eq. (4.1795), the weak object perturbation can therefore be written, apart from an irrelevant overall normalization, with the two coefficients

$$\boxed{A + j\sqrt{1 - A^2}} \tag{5.874}$$

and its complex conjugate

$$\boxed{A - j\sqrt{1 - A^2}.} \tag{5.875}$$

Normalized amplitude contrast in the two complex sidebands.

The weak amplitude–phase object was written in Eq. (4.1767) as

$$\psi_i(\mathbf{r}) \simeq 1 + a(\mathbf{r}) + j\varphi(\mathbf{r}), \tag{5.876}$$

where $a(\mathbf{r})$ is the weak amplitude modulation and $\varphi(\mathbf{r})$ is the weak phase modulation. In Fourier space, Eq. (4.1773) gives

$$\widehat{\psi}_i(\mathbf{R}) = \delta(\mathbf{R}) + \widehat{a}(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R}). \tag{5.877}$$

The conventional one-parameter amplitude-contrast model makes the

additional approximation derived in Eq. (4.1793),

$$\widehat{a}(\mathbf{R}) \simeq Q\,\widehat{\varphi}(\mathbf{R}), \tag{5.878}$$

where, in the usual model, Q is taken to be independent of spatial frequency over the range of interest. The weak scattered part of the positive-frequency component can therefore be written

$$\widehat{a}(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R}) \simeq Q\widehat{\varphi}(\mathbf{R}) + j\widehat{\varphi}(\mathbf{R}) \tag{5.879}$$

$$= (Q + j)\widehat{\varphi}(\mathbf{R}). \tag{5.880}$$

The magnitude of $Q + j$ is

$$|Q + j| = \sqrt{Q^2 + 1}. \tag{5.881}$$

It is therefore convenient to factor this common magnitude out:

$$Q + j = \sqrt{1 + Q^2}\left[\frac{Q}{\sqrt{1 + Q^2}} + j\frac{1}{\sqrt{1 + Q^2}}\right]. \tag{5.882}$$

Using the normalized amplitude-contrast coefficient introduced in Eq. (4.1795),

$$A = \frac{Q}{\sqrt{1 + Q^2}}, \qquad \sqrt{1 - A^2} = \frac{1}{\sqrt{1 + Q^2}}, \tag{5.883}$$

we obtain

$$\boxed{Q + j = \sqrt{1 + Q^2}\left[A + j\sqrt{1 - A^2}\right].} \tag{5.884}$$

The factor $\sqrt{1 + Q^2}$ is a common scale factor in the conventional one-parameter model and may be absorbed into the overall normalization of the weak object function. The normalized complex coefficient of the positive-frequency perturbation is therefore

$$\boxed{A + j\sqrt{1 - A^2}.} \tag{5.885}$$

The second term entering the real image arises from the complex conjugate weak perturbation. As shown in the derivation of Eq. (4.1783), its object factor is

$$\widehat{a}(\mathbf{R}) - j\widehat{\varphi}(\mathbf{R}). \tag{5.886}$$

Using the same proportionality,

$$\widehat{a}(\mathbf{R}) - j\widehat{\varphi}(\mathbf{R}) \simeq (Q - j)\widehat{\varphi}(\mathbf{R}) \tag{5.887}$$

$$= \sqrt{1+Q^2}\left[A - j\sqrt{1-A^2}\right]\widehat{\varphi}(\mathbf{R}). \tag{5.888}$$

Hence its normalized coefficient is the complex conjugate

$$\boxed{A - j\sqrt{1-A^2}.} \tag{5.889}$$

Thus the two coefficients are not independent assumptions. They are the normalized forms of the two weak-object factors

$$Q + j \qquad \text{and} \qquad Q - j, \tag{5.890}$$

which appear because the real-valued image intensity contains the scattered wave and its complex conjugate.

Since

$$A^2 + \left(\sqrt{1-A^2}\right)^2 = 1, \tag{5.891}$$

both normalized coefficients have unit modulus:

$$\left|A \pm j\sqrt{1-A^2}\right| = 1. \tag{5.892}$$

Introducing the amplitude-contrast phase ϕ_A from Eq. (4.1798),

$$\sin\phi_A = A, \qquad \cos\phi_A = \sqrt{1-A^2}, \tag{5.893}$$

they may equivalently be written as

$$\boxed{A \pm j\sqrt{1-A^2} = e^{\pm j(\pi/2-\phi_A)}.} \tag{5.894}$$

Amplitude contrast therefore rotates the ideal weak-phase factors $\pm j$ away from exact quadrature rather than introducing an independent attenuation of either sideband.

Consequently, the two coherent sideband transfer factors generalize to

$$\boxed{H_P^{(A)}(\mathbf{R}_\perp) = \left[A + j\sqrt{1-A^2}\right] H(\mathbf{R}_\perp),} \tag{5.895}$$

and

$$\boxed{H_Q^{(A)}(\mathbf{R}_\perp) = \left[A - j\sqrt{1-A^2}\right] H^*(-\mathbf{R}_\perp).} \tag{5.896}$$

For $A = 0$, these expressions reduce exactly to Eqs. (5.849) and (5.850). Substituting Eq. (5.869),

$$\boxed{H_P^{(A)} = Pe^{-j\chi_{\text{asym}}}\left[A + j\sqrt{1-A^2}\right]e^{-j\Phi},} \tag{5.897}$$

whereas

$$\boxed{H_Q^{(A)} = Pe^{-j\chi_{\text{asym}}}\left[A - j\sqrt{1-A^2}\right]e^{+j\Phi}.} \tag{5.898}$$

Including the experimental envelope. The transfer function H describes one coherent realization of the microscope. The attenuation envelope has a different origin. As derived in Sec. 4.13, temporal coherence, spatial coherence, and specimen motion result from averaging different coherent realizations at the detector. Under the factorized approximation used in the complete CTF model, their combined effect is

$$E_{\text{tot}}(\mathbf{R}_\perp) \tag{5.899}$$

[Eq. (4.2115)].

Thus the effective single-sideband factors at the image-transfer level are

$$\boxed{\text{CTF}_P(\mathbf{R}_\perp) \equiv E_{\text{tot}}(\mathbf{R}_\perp)H_P^{(A)}(\mathbf{R}_\perp),} \tag{5.900}$$

and

$$\boxed{\text{CTF}_Q(\mathbf{R}_\perp) \equiv E_{\text{tot}}(\mathbf{R}_\perp)H_Q^{(A)}(\mathbf{R}_\perp).} \tag{5.901}$$

The Ewald-sphere image equation therefore becomes

$$\begin{aligned}\widehat{\delta I}(\mathbf{R}_\perp) \propto\ & \text{CTF}_P(\mathbf{R}_\perp)\widehat{V}\left(\mathbf{R}_\perp, -\Delta R_z^{\text{E}}\right) \\ & + \text{CTF}_Q(\mathbf{R}_\perp)\widehat{V}\left(\mathbf{R}_\perp, +\Delta R_z^{\text{E}}\right).\end{aligned} \tag{5.902}$$

Recovery of the complete CTF in the flat-Ewald limit.

When the Ewald sphere is approximated by its tangent plane,

$$\Delta R_z^{\text{E}} \longrightarrow 0, \tag{5.903}$$

the two specimen Fourier samples in Eq. (5.902) become identical:

$$\widehat{V}\left(\mathbf{R}_\perp, -\Delta R_z^{\text{E}}\right) = \widehat{V}\left(\mathbf{R}_\perp, +\Delta R_z^{\text{E}}\right) = \widehat{V}(\mathbf{R}_\perp, 0). \tag{5.904}$$

The two sideband transfer factors can therefore be added.

Using Eqs. (5.897) and (5.898),

$$\begin{aligned}\mathrm{CTF}_P + \mathrm{CTF}_Q = E_{\mathrm{tot}} P e^{-j\chi_{\mathrm{asym}}} \\ \times \left\{ \left[A + j\sqrt{1-A^2}\right] e^{-j\Phi} \right. \\ \left. + \left[A - j\sqrt{1-A^2}\right] e^{+j\Phi} \right\}.\end{aligned} \tag{5.905}$$

Collect first the amplitude-contrast terms:

$$A\left(e^{-j\Phi} + e^{+j\Phi}\right) = 2A\cos\Phi. \tag{5.906}$$

The phase-contrast terms give

$$j\sqrt{1-A^2}\left(e^{-j\Phi} - e^{+j\Phi}\right) = 2\sqrt{1-A^2}\sin\Phi. \tag{5.907}$$

Therefore,

$$\begin{aligned}\mathrm{CTF}_P + \mathrm{CTF}_Q = 2E_{\mathrm{tot}} P e^{-j\chi_{\mathrm{asym}}} \\ \times \left[\sqrt{1-A^2}\sin\Phi + A\cos\Phi\right].\end{aligned} \tag{5.908}$$

Absorbing the same irrelevant global factor of two used in the definition of the conventional CTF, we obtain

$$\boxed{\begin{aligned}\mathrm{CTF}(\mathbf{R}_\perp) = E_{\mathrm{tot}}(\mathbf{R}_\perp) P(\mathbf{R}_\perp) e^{-j\chi_{\mathrm{asym}}(\mathbf{R}_\perp)} \\ \times \left[\sqrt{1-A^2}\,\sin\left(\chi_{\mathrm{sym}} + \phi_{\mathrm{PP}}\right)\right. \\ \left. + A\,\cos\left(\chi_{\mathrm{sym}} + \phi_{\mathrm{PP}}\right)\right].\end{aligned}} \tag{5.909}$$

For $\phi_{\mathrm{PP}} = 0$, this reduces exactly to the complete CTF derived previously in Eq. (4.2117).

Using $\phi_A = \arcsin A$ from Eq. (4.1798), the same result can be written

$$\boxed{\begin{aligned}\mathrm{CTF}(\mathbf{R}_\perp) = E_{\mathrm{tot}}(\mathbf{R}_\perp) P(\mathbf{R}_\perp) e^{-j\chi_{\mathrm{asym}}(\mathbf{R}_\perp)} \\ \times \sin\left[\chi_{\mathrm{sym}}(\mathbf{R}_\perp) + \phi_{\mathrm{PP}}(\mathbf{R}_\perp) + \phi_A\right].\end{aligned}} \tag{5.910}$$

Why defocus separates the two sidebands in real space. The two Ewald contributions in Eq. (5.902) need not remain superimposed spatially

in a defocused image. Their displacement is determined by the spatial-frequency dependence of the *phase* of the two complex sideband transfer factors CTF_P and CTF_Q.

From Eqs. (5.900) and (5.901), the complete sideband factors are

$$\begin{aligned}\mathrm{CTF}_P(\mathbf{R}_\perp) = E_{\mathrm{tot}}(\mathbf{R}_\perp)P(\mathbf{R}_\perp)e^{-j\chi_{\mathrm{asym}}(\mathbf{R}_\perp)} \\ \times \left[A + j\sqrt{1-A^2}\right] e^{-j\Phi(\mathbf{R}_\perp)},\end{aligned} \tag{5.911}$$

and

$$\begin{aligned}\mathrm{CTF}_Q(\mathbf{R}_\perp) = E_{\mathrm{tot}}(\mathbf{R}_\perp)P(\mathbf{R}_\perp)e^{-j\chi_{\mathrm{asym}}(\mathbf{R}_\perp)} \\ \times \left[A - j\sqrt{1-A^2}\right] e^{+j\Phi(\mathbf{R}_\perp)},\end{aligned} \tag{5.912}$$

where

$$\Phi(\mathbf{R}_\perp) = \chi_{\mathrm{sym}}(\mathbf{R}_\perp) + \phi_{\mathrm{PP}}(\mathbf{R}_\perp) \tag{5.913}$$

was defined in Eq. (5.870).

Using Eq. (5.894),

$$A \pm j\sqrt{1-A^2} = e^{\pm j(\pi/2-\phi_A)}, \tag{5.914}$$

the phases of the two sideband factors are

$$\boxed{\begin{aligned}\theta_P(\mathbf{R}_\perp) &= -\chi_{\mathrm{asym}}(\mathbf{R}_\perp) - \Phi(\mathbf{R}_\perp) + \frac{\pi}{2} - \phi_A, \\ \theta_Q(\mathbf{R}_\perp) &= -\chi_{\mathrm{asym}}(\mathbf{R}_\perp) + \Phi(\mathbf{R}_\perp) - \frac{\pi}{2} + \phi_A.\end{aligned}} \tag{5.915}$$

The factors $E_{\mathrm{tot}}(\mathbf{R}_\perp)$ and $P(\mathbf{R}_\perp)$ are real in the present model and therefore modify the amplitudes of the sidebands but not their phase gradients. Likewise, if the amplitude-contrast coefficient A is taken to be independent of spatial frequency, ϕ_A is constant and produces no real-space displacement.

Real-space displacement produced by a sideband phase gradient.

Consider a narrow Fourier wave packet centered at $\mathbf{R}_{\perp,0}$ whose transfer factor contains the phase

$$e^{j\theta(\mathbf{R}_\perp)}. \tag{5.916}$$

Over the narrow support of the packet, expand the phase to first order:

$$\theta(\mathbf{R}_\perp) \simeq \theta(\mathbf{R}_{\perp,0}) + \nabla_{\mathbf{R}_\perp}\theta(\mathbf{R}_{\perp,0}) \cdot (\mathbf{R}_\perp - \mathbf{R}_{\perp,0}) \,. \tag{5.917}$$

Apart from an irrelevant constant phase, the transfer factor therefore contains

$$\exp\left[j\nabla_{\mathbf{R}_\perp}\theta \cdot \mathbf{R}_\perp\right]. \tag{5.918}$$

With the Fourier convention of Multiplication by

$$e^{j2\pi\mathbf{R}_\perp\cdot\mathbf{r}_0} \tag{5.919}$$

translates the corresponding real-space pattern by $-\mathbf{r}_0$. Hence the displacement associated with a phase $\theta(\mathbf{R}_\perp)$ is

$$\boxed{\Delta\boldsymbol{\rho} = -\frac{1}{2\pi}\nabla_{\mathbf{R}_\perp}\theta.} \tag{5.920}$$

Applying Eq. (5.920) to Eq. (5.915) gives

$$\boxed{\begin{aligned}\Delta\boldsymbol{\rho}_P &= \frac{1}{2\pi}\nabla_{\mathbf{R}_\perp}\left(\chi_{\text{asym}} + \Phi\right), \\ \Delta\boldsymbol{\rho}_Q &= \frac{1}{2\pi}\nabla_{\mathbf{R}_\perp}\left(\chi_{\text{asym}} - \Phi\right).\end{aligned}} \tag{5.921}$$

Thus the asymmetric aberration phase produces the *same* displacement of both sidebands, whereas the symmetric phase produces opposite displacements.

For an ideal rotationally symmetric phase plate, ϕ_{PP} is approximately constant over the scattered frequencies of interest and therefore

$$\nabla_{\mathbf{R}_\perp}\phi_{\text{PP}} \simeq 0. \tag{5.922}$$

Such a constant phase shift changes contrast transfer but does not delocalize a Fourier component.

Similarly, in the absence of asymmetric aberrations,

$$\chi_{\text{asym}} = 0, \tag{5.923}$$

Eq. (5.921) reduces to

$$\boxed{\Delta\boldsymbol{\rho}_P = +\frac{1}{2\pi}\nabla_{\mathbf{R}_\perp}\Phi, \qquad \Delta\boldsymbol{\rho}_Q = -\frac{1}{2\pi}\nabla_{\mathbf{R}_\perp}\Phi.} \tag{5.924}$$

The two sidebands are therefore displaced by equal amounts in opposite directions.

Keeping only defocus and spherical aberration in χ_{sym}, the symmetric phase is

$$\chi_{\text{sym}}(R_\perp) = -\pi\lambda\Delta f R_\perp^2 + \frac{\pi}{2} C_s \lambda^3 R_\perp^4, \tag{5.925}$$

with the present sign convention for defocus.

Its gradient is radial:

$$\nabla_{\mathbf{R}_\perp} \chi_{\text{sym}} = \left(-2\pi\lambda\Delta f R_\perp + 2\pi C_s \lambda^3 R_\perp^3\right) \widehat{\mathbf{R}}_\perp. \tag{5.926}$$

Therefore,

$$\boxed{\begin{aligned} \Delta\boldsymbol{\rho}_P &= \left(-\lambda\Delta f R_\perp + C_s\lambda^3 R_\perp^3\right) \widehat{\mathbf{R}}_\perp, \\ \Delta\boldsymbol{\rho}_Q &= -\Delta\boldsymbol{\rho}_P, \end{aligned}} \tag{5.927}$$

when the phase-plate phase is constant and asymmetric aberrations are neglected.

When defocus dominates over spherical aberration at the spatial frequency considered,

$$\left|C_s\lambda^3 R_\perp^3\right| \ll \left|\lambda\Delta f R_\perp\right|, \tag{5.928}$$

the magnitude of the displacement of either sideband becomes

$$\boxed{\Delta\rho \simeq |\Delta f|\lambda R_\perp.} \tag{5.929}$$

At resolution d, for which

$$R_\perp \simeq \frac{1}{d}, \tag{5.930}$$

this gives

$$\boxed{\Delta\rho \simeq \frac{|\Delta f|\lambda}{d}.} \tag{5.931}$$

Since the two sidebands are displaced in opposite directions, their centre-to-centre separation is

$$\boxed{\Delta\rho_{\text{sep}} \simeq \frac{2|\Delta f|\lambda}{d}.} \tag{5.932}$$

For a particle of diameter D, the two high-resolution fringe packets are therefore fully separated approximately when

$$\frac{2|\Delta f|\lambda}{d} > D, \tag{5.933}$$

or

$$\boxed{|\Delta f| > \frac{Dd}{2\lambda}.} \tag{5.934}$$

This is the origin of the spatially separated high-resolution fringe packets demonstrated experimentally by Russo and Henderson (2018b). The Ewald-sphere curvature makes the two sidebands correspond to different three-dimensional specimen Fourier components, whereas the phase gradient produced mainly by defocus makes those two contributions appear at different positions in the real-space image.

Why conventional CTF correction produces twin images. The preceding analysis also explains the displaced *twin images* that appear when a conventionally CTF-corrected image contains spatially separated sidebands.

It is useful first to distinguish the sideband *transfer factors* CTF_P and CTF_Q, introduced in Eqs. (5.900) and (5.901), from the phase corrections used to undo their deterministic delocalization. Write

$$\mathrm{CTF}_P(\mathbf{R}_\perp) = M_P(\mathbf{R}_\perp)e^{j\theta_P(\mathbf{R}_\perp)}, \tag{5.935}$$

and

$$\mathrm{CTF}_Q(\mathbf{R}_\perp) = M_Q(\mathbf{R}_\perp)e^{j\theta_Q(\mathbf{R}_\perp)}, \tag{5.936}$$

where the moduli contain, among other factors, the objective pupil and the total envelope E_{tot}, while the phases θ_P and θ_Q were given in Eq. (5.915).

Where the sidebands have nonzero amplitude, the corresponding phase-only recentering operators are

$$\boxed{\mathcal{C}_P(\mathbf{R}_\perp) = e^{-j\theta_P(\mathbf{R}_\perp)}, \qquad \mathcal{C}_Q(\mathbf{R}_\perp) = e^{-j\theta_Q(\mathbf{R}_\perp)}.} \tag{5.937}$$

They undo the phase gradients responsible for the real-space delocalization. They do not restore attenuation represented by E_{tot}.

Consider the P sideband, initially displaced by $+\Delta\boldsymbol{\rho}_{\mathrm{sym}}$. Applying its own correction $\mathcal{C}_P$ cancels the phase gradient that generated this displacement, and therefore brings the packet back to its correct origin:

$$+\,\Delta\boldsymbol{\rho}_{\mathrm{sym}} - \Delta\boldsymbol{\rho}_{\mathrm{sym}} = \mathbf{0}. \tag{5.938}$$

If instead the opposite phase branch $\mathcal{C}_Q$ acts on the same packet, its phase gradient has the opposite sign and the displacement is reinforced:

$$+\,\Delta\boldsymbol{\rho}_{\mathrm{sym}} + \Delta\boldsymbol{\rho}_{\mathrm{sym}} = +2\Delta\boldsymbol{\rho}_{\mathrm{sym}}. \tag{5.939}$$

Likewise, the Q sideband produces one correctly recentered contribution and one contribution displaced to

$$-2\Delta\boldsymbol{\rho}_{\mathrm{sym}}. \tag{5.940}$$

A conventional CTF correction does not preserve the two complex branches separately. In the flat-Ewald description they are combined into the ordinary CTF, so that both phase branches act simultaneously. Consequently, once the sidebands have been spatially resolved, the corrected image contains both the desired recentered contribution and displaced twin-image contributions:

$$\boxed{\begin{array}{lcl} P \text{ sideband} & \longrightarrow & \mathbf{0},\ +2\Delta\boldsymbol{\rho}_{\mathrm{sym}}, \\ Q \text{ sideband} & \longrightarrow & \mathbf{0},\ -2\Delta\boldsymbol{\rho}_{\mathrm{sym}}. \end{array}} \tag{5.941}$$

When defocus dominates the symmetric aberration phase, Eq. (5.931) gives

$$\left|\Delta\boldsymbol{\rho}_{\mathrm{sym}}\right| \simeq \frac{|\Delta f|\lambda}{d}. \tag{5.942}$$

The twin contributions generated by conventional CTF processing are therefore displaced from the correctly recentered particle by approximately

$$\boxed{2\left|\Delta\boldsymbol{\rho}_{\mathrm{sym}}\right| \simeq \frac{2|\Delta f|\lambda}{d}.} \tag{5.943}$$

These are the displaced high-resolution "ghosts" observed by Russo and Henderson (2018b). Their origin should not be confused with the Ewald curvature itself. Ewald curvature makes the P and Q sidebands correspond to two different three-dimensional specimen Fourier samples; the objective-lens phase gradient delocalizes those sidebands in real space; and recombining the two phase-correction branches produces the twin images.

Neither the total envelope E_{tot} nor a spatially constant amplitude-contrast phase or phase-plate shift changes these geometric positions: they modify the amplitudes or constant phases of the sidebands. Spatial-frequency-dependent phase terms determine the delocalization through their gradients.

5.9.9 Ewald-sphere curvature and absolute handedness.

Ewald-sphere curvature has another important consequence: in principle, it breaks the fundamental handedness ambiguity of the projection approximation.

Let the electrostatic potential of the specimen be $V(\mathbf{r})$. An inverted copy,

$$\boxed{V^{\text{inv}}(\mathbf{r}) = V(-\mathbf{r}),} \tag{5.944}$$

has the opposite handedness. Any improper three-dimensional transformation can be written as an inversion followed by a proper rotation, so that, because particle orientations are free parameters in single-particle analysis, it is sufficient to consider this inversion.

Why the projection approximation cannot determine handedness.

The projected potential of the original object is

$$V^{\text{proj}}(\boldsymbol{\rho}) = \int_{-\infty}^{+\infty} V(\boldsymbol{\rho}, z)\, dz. \tag{5.945}$$

For the inverted object,

$$V_{\text{inv}}^{\text{proj}}(\boldsymbol{\rho}) = \int_{-\infty}^{+\infty} V(-\boldsymbol{\rho}, -z)\, dz. \tag{5.946}$$

Introducing

$$z' = -z, \qquad dz' = -dz, \tag{5.947}$$

and reversing the integration limits gives

$$V_{\text{inv}}^{\text{proj}}(\boldsymbol{\rho}) = \int_{-\infty}^{+\infty} V(-\boldsymbol{\rho}, z')\, dz' \tag{5.948}$$

$$= V^{\text{proj}}(-\boldsymbol{\rho}). \tag{5.949}$$

Hence

$$\boxed{V_{\text{inv}}^{\text{proj}}(\boldsymbol{\rho}) = V^{\text{proj}}(-\boldsymbol{\rho}).} \tag{5.950}$$

The transformation $\boldsymbol{\rho} \to -\boldsymbol{\rho}$ is simply a 180° rotation in the image plane. Since such an in-plane rotation can be absorbed into the particle orientation, a set of ideal projections cannot distinguish a three-dimensional object from its inverted copy.

The situation changes when the finite curvature of the Ewald sphere is

retained. For a given transverse spatial frequency $\mathbf{R}_\perp$, define

$$\boxed{\begin{aligned} V_-(\mathbf{R}_\perp) &\equiv \widehat{V}\left(\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right), \\ V_+(\mathbf{R}_\perp) &\equiv \widehat{V}\left(\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right). \end{aligned}} \tag{5.951}$$

How inversion exchanges the two Ewald samples.

The Fourier transform of the inverted object is

$$\widehat{V}^{\mathrm{inv}}(\mathbf{R}) = \widehat{V}(-\mathbf{R}). \tag{5.952}$$

As in the projection case, compensate for the apparent inversion in the image plane by a 180° rotation, $\mathbf{R}_\perp \to -\mathbf{R}_\perp$. The Ewald sample of the inverted object at the lower longitudinal coordinate is then

$$\widehat{V}^{\mathrm{inv}}\left(-\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right) = \widehat{V}\left(\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right) \tag{5.953}$$

$$= V_+(\mathbf{R}_\perp). \tag{5.954}$$

Similarly,

$$\widehat{V}^{\mathrm{inv}}\left(-\mathbf{R}_\perp, +\Delta R_z^{\mathrm{E}}\right) = \widehat{V}\left(\mathbf{R}_\perp, -\Delta R_z^{\mathrm{E}}\right) \tag{5.955}$$

$$= V_-(\mathbf{R}_\perp). \tag{5.956}$$

Thus inversion of the three-dimensional structure, followed by the in-plane rotation that was sufficient to hide the inversion in the projection approximation, now exchanges the two Ewald samples:

$$\boxed{V_- \longleftrightarrow V_+.} \tag{5.957}$$

If the Ewald sphere were flat, $\Delta R_z^{\mathrm{E}} = 0$, these two coefficients would coincide and the exchange would have no observable consequence. Finite Ewald curvature therefore provides the geometrical mechanism by which the hand ambiguity can be broken.

Using Eq. (5.902), the image coefficient for the original hand is

$$\widehat{\delta I}(\mathbf{R}_\perp) = \mathrm{CTF}_P(\mathbf{R}_\perp)V_-(\mathbf{R}_\perp) + \mathrm{CTF}_Q(\mathbf{R}_\perp)V_+(\mathbf{R}_\perp). \tag{5.958}$$

For the inverted structure, after the compensating 180° in-plane rotation, Eq. (5.957) gives

$$\widehat{\delta I}^{180}_{\text{inv}}(\mathbf{R}_\perp) = \text{CTF}_P(\mathbf{R}_\perp)V_+(\mathbf{R}_\perp) + \text{CTF}_Q(\mathbf{R}_\perp)V_-(\mathbf{R}_\perp). \tag{5.959}$$

Subtracting the two expressions,

$$\boxed{\begin{aligned}&\widehat{\delta I}(\mathbf{R}_\perp) - \widehat{\delta I}^{180}_{\text{inv}}(\mathbf{R}_\perp)\\ &\qquad = \left[\text{CTF}_P(\mathbf{R}_\perp) - \text{CTF}_Q(\mathbf{R}_\perp)\right]\left[V_-(\mathbf{R}_\perp) - V_+(\mathbf{R}_\perp)\right].\end{aligned}} \tag{5.960}$$

Ewald-sphere curvature removes the exact geometrical equivalence between the two hands that exists under the projection approximation. Whether and how optimally this weak parity-sensitive information should be extracted from experimental particle images is an image-processing problem and will be considered separately.

5.10 Depth is defocus: the three-dimensional origin of depth-dependent phase

The projection approximation removes the longitudinal coordinate of the specimen by replacing

$$V(x, y, z) \tag{5.961}$$

with the projected potential

$$V^{\text{proj}}(x, y) = \int V(x, y, z)\, dz \tag{5.962}$$

defined in Eq. (5.546). Once this approximation is made, the complete particle is treated as if all of its scattering originated in a single transverse plane. It is then natural to associate the whole particle with one value of defocus.

A three-dimensional specimen is different. Scattering generated at different depths must propagate different distances before reaching the exit surface, and therefore acquires different propagation phases. In the paraxial regime, the resulting phase difference has exactly the same quadratic dependence on spatial frequency as objective-lens defocus.

This is the meaning of the shorthand statement

$$\boxed{\text{depth is defocus.}} \tag{5.963}$$

It does not mean that the objective-lens setting changes with depth. The lens has one physical setting. Rather, *defocus is defined relative to an object plane*, and different planes inside a thick specimen have different effective defocus values with respect to the same microscope setting.

5.10.1 Different specimen planes propagate different distances

Let the electron beam propagate along the positive z direction, and let the specimen occupy

$$z_{\text{in}} \leq z \leq z_{\text{exit}}. \tag{5.964}$$

Consider a weak scattered contribution generated in a thin plane at depth z. After its generation, that contribution must propagate the remaining distance

$$\boxed{L(z) = z_{\text{exit}} - z} \tag{5.965}$$

before reaching the exit surface.

The Fresnel propagator derived in Eq. (5.535) therefore gives

$$\boxed{P_{z_{\text{exit}}-z}(\mathbf{R}_\perp) = \exp\left[-j\pi\lambda\left(z_{\text{exit}} - z\right)R_\perp^2\right],} \tag{5.966}$$

where

$$R_\perp = \|\mathbf{R}_\perp\| = \sqrt{R_x^2 + R_y^2}. \tag{5.967}$$

Thus two otherwise identical scattering events occurring at depths z_1 and z_2 do not arrive at the exit surface with the same phase. Their propagation factors are

$$P_1 = \exp\left[-j\pi\lambda(z_{\text{exit}} - z_1)R_\perp^2\right] \tag{5.968}$$

and

$$P_2 = \exp\left[-j\pi\lambda(z_{\text{exit}} - z_2)R_\perp^2\right]. \tag{5.969}$$

Their ratio is

$$\frac{P_2}{P_1} = \exp\left[-j\pi\lambda\left(z_{\text{exit}} - z_2 - z_{\text{exit}} + z_1\right)R_\perp^2\right] \tag{5.970}$$

$$= \boxed{\exp\left[j\pi\lambda(z_2 - z_1)R_\perp^2\right].} \tag{5.971}$$

The relative phase accumulated between the two contributions is therefore

$$\boxed{\Delta\phi_{21}(R_\perp) = \pi\lambda(z_2 - z_1)R_\perp^2.} \tag{5.972}$$

The effect grows linearly with the separation between the two planes and quadratically with spatial frequency. Low-resolution information is therefore relatively insensitive to specimen depth, whereas high-resolution information becomes increasingly sensitive to the longitudinal position at which it was generated.

5.10.2 A defocus value refers to a reference plane

To connect Eq. (5.972) with the objective-lens defocus, recall that the cryo-EM defocus convention was defined in Eq. (4.1188), with positive Δf denoting underfocus. The corresponding defocus aberration phase is

$$\chi_{\mathrm{def}}(\mathbf{R}_\perp) = -\pi\lambda\Delta f R_\perp^2 \tag{5.973}$$

from Eq. (4.1191). Consequently, the coherent wave-transfer function contains the factor

$$\exp\left[j\pi\lambda\Delta f R_\perp^2\right]. \tag{5.974}$$

Choose now a reference plane inside the specimen at

$$z = z_{\mathrm{ref}}, \tag{5.975}$$

and let

$$\Delta f_{\mathrm{ref}} \tag{5.976}$$

be the defocus assigned to that plane. In single-particle cryo-EM this reference plane is commonly chosen near the centre of the particle, or more precisely near the effective plane to which the fitted defocus refers.

Why changing specimen depth is equivalent to changing defocus.

A contribution generated at depth z carries the exit-surface propagation factor of Eq. (5.966),

$$P(z) = e^{-j\pi\lambda(z_{\mathrm{exit}}-z)R_\perp^2}. \tag{5.977}$$

The corresponding factor for the chosen reference plane is

$$P(z_{\mathrm{ref}}) = e^{-j\pi\lambda(z_{\mathrm{exit}}-z_{\mathrm{ref}})R_\perp^2}. \tag{5.978}$$

Therefore,

$$\frac{P(z)}{P(z_{\mathrm{ref}})} = \exp\left[j\pi\lambda(z - z_{\mathrm{ref}})R_\perp^2\right]. \tag{5.979}$$

The objective-lens transfer associated with the reference plane contains the defocus factor

$$\exp\left[j\pi\lambda\Delta f_{\mathrm{ref}}R_\perp^2\right]. \tag{5.980}$$

Multiplying by Eq. (5.979) gives

$$\begin{aligned}\exp\left[j\pi\lambda\Delta f_{\mathrm{ref}}R_\perp^2\right]&\exp\left[j\pi\lambda(z - z_{\mathrm{ref}})R_\perp^2\right]\\ &= \exp\left\{j\pi\lambda\left[\Delta f_{\mathrm{ref}} + z - z_{\mathrm{ref}}\right]R_\perp^2\right\}.\end{aligned} \tag{5.981}$$

This is precisely the defocus transfer factor of a plane whose effective defocus is

$$\boxed{\Delta f_{\mathrm{eff}}(z) = \Delta f_{\mathrm{ref}} + z - z_{\mathrm{ref}}.} \tag{5.982}$$

Consequently,

$$\boxed{\Delta f_{\mathrm{eff}}(z_2) - \Delta f_{\mathrm{eff}}(z_1) = z_2 - z_1.} \tag{5.983}$$

Equation (5.982) depends on the sign conventions adopted here: z increases along electron propagation and positive Δf denotes underfocus. Reversing either convention changes the corresponding sign. The invariant physical statement is the quadratic relative phase of Eq. (5.972).

If the reference plane is chosen at the centre of a particle of thickness T, its front and back surfaces lie approximately at

$$z - z_{\mathrm{ref}} = -\frac{T}{2} \qquad \text{and} \qquad +\frac{T}{2}. \tag{5.984}$$

Their effective defoci are therefore

$$\boxed{\Delta f_{\mathrm{front}} \simeq \Delta f_{\mathrm{ref}} - \frac{T}{2}, \qquad \Delta f_{\mathrm{back}} \simeq \Delta f_{\mathrm{ref}} + \frac{T}{2}.} \tag{5.985}$$

The complete particle therefore spans a deterministic defocus range of approximately

$$\boxed{\Delta f_{\mathrm{back}} - \Delta f_{\mathrm{front}} \simeq T.} \tag{5.986}$$

It is important not to confuse this with the statistical defocus spread introduced earlier in the treatment of temporal coherence. The latter arises

because different electrons or different times during the exposure experience different microscope conditions and are averaged incoherently. Here the microscope setting is fixed. Different depths of the *same specimen* contribute deterministic phases and must be added coherently.

5.10.3 What the single-defocus image model neglects

The conventional weak-object image model assigns one CTF to the complete projected specimen. Suppressing translations and overall scale factors, its Fourier-space form is

$$\boxed{\widehat{\delta I}_{\text{proj}}(\mathbf{R}_\perp) = \text{CTF}\,(\mathbf{R}_\perp; \Delta f_{\text{ref}})\,\widehat{V}^{\text{proj}}(\mathbf{R}_\perp),} \tag{5.987}$$

where the complete CTF may contain all the terms developed previously in Eq. (4.2117).

Since

$$\widehat{V}^{\text{proj}}(\mathbf{R}_\perp) = \int \widehat{V}_\perp(\mathbf{R}_\perp, z)\,dz, \tag{5.988}$$

Eq. (5.987) amounts to assigning the same defocus to every depth before the contributions are added.

Within the weak, linear image-formation model, retaining the depth coordinate instead gives

$$\boxed{\begin{aligned}\widehat{\delta I}(\mathbf{R}_\perp) \simeq \int &\text{CTF}\,[\mathbf{R}_\perp; \Delta f_{\text{eff}}(z)] \\ &\times \widehat{V}_\perp(\mathbf{R}_\perp, z)\,dz,\end{aligned}} \tag{5.989}$$

with $\Delta f_{\text{eff}}(z)$ given by Eq. (5.982). All other microscope parameters can remain unchanged; it is the defocus term that shifts systematically with depth.

The difference is particularly transparent for a specimen consisting of only two weak planes. If their transverse potentials are $\widehat{V}_1$ and $\widehat{V}_2$, the depth-aware model gives

$$\boxed{\begin{aligned}\widehat{\delta I} \simeq\ &\text{CTF}(\mathbf{R}_\perp; \Delta f_1)\widehat{V}_1(\mathbf{R}_\perp) \\ &+ \text{CTF}(\mathbf{R}_\perp; \Delta f_2)\widehat{V}_2(\mathbf{R}_\perp),\end{aligned}} \tag{5.990}$$

whereas the projection model replaces this by

$$\boxed{\widehat{\delta I}_{\text{proj}} \simeq \text{CTF}(\mathbf{R}_\perp; \Delta f_{\text{ref}})\left[\widehat{V}_1(\mathbf{R}_\perp) + \widehat{V}_2(\mathbf{R}_\perp)\right].} \tag{5.991}$$

These expressions become equivalent only when the phase change produced by the depth-dependent defocus is negligible over the spatial-frequency range of interest.

5.10.4 Multislice contains the depth dependence automatically

The multislice method of Sec. 5.8 does not assign an independent CTF to each slice. Instead, depth enters more fundamentally through the propagation operators between successive slices.

Recall the multislice recurrence

$$\psi_{n+1} = \mathcal{P}_{\Delta z_n} \left\{ \mathcal{T}_n \psi_n \right\} \tag{5.992}$$

from Eq. (5.667), with Fourier-space propagator

$$P_{\Delta z_n}(\mathbf{R}_\perp) = \exp\left(-j\pi\lambda\Delta z_n R_\perp^2\right) \tag{5.993}$$

from Eq. (5.674).

A wave generated or modified in an early slice is therefore propagated through every subsequent interval before reaching the exit surface, whereas a contribution generated close to the exit surface undergoes much less propagation. No additional "depth correction" has to be added to multislice: the different propagation distances are already contained in its ordered sequence of transmission and propagation operators (Kirkland, 2010).

This becomes especially transparent if multislice is linearized to first order in the specimen potential. Using

$$\mathcal{T}_n \simeq 1 + j\sigma V_n^{\mathrm{proj}}, \tag{5.994}$$

and retaining only terms containing one interaction with the specimen, the scattered exit wave has the form

$$\boxed{\delta\widehat{\psi}_{\mathrm{exit}}(\mathbf{R}_\perp) \simeq j\sigma \sum_n \widehat{V}_n^{\mathrm{proj}}(\mathbf{R}_\perp) \exp\left[-j\pi\lambda\left(z_{\mathrm{exit}} - z_n\right) R_\perp^2\right],} \tag{5.995}$$

where z_n denotes a representative depth of slice n.

Single-scattering limit of multislice as a coherent sum over depth.

Consider the multislice recurrence derived in Eq. (5.667),

$$\psi_{n+1} = \mathcal{P}_{\Delta z_n} \left\{ \mathcal{T}_n \psi_n \right\}, \tag{5.996}$$

where

$$\mathcal{T}_n = \exp\left(j\sigma V_n^{\text{proj}} \right) \tag{5.997}$$

is the transmission operator of slice n, and $\mathcal{P}_{\Delta z_n}$ propagates the wave through the distance Δz_n.

For a weak interaction in each slice,

$$\mathcal{T}_n = \exp\left(j\sigma V_n^{\text{proj}} \right) \simeq 1 + j\sigma V_n^{\text{proj}}. \tag{5.998}$$

We now retain only terms that contain one factor of the specimen potential. Terms containing products such as

$$V_m^{\text{proj}} V_n^{\text{proj}} \tag{5.999}$$

represent two or more scattering events and are discarded in this single-scattering limit.

To see the structure explicitly, consider first three slices. The exit wave is

$$\psi_{\text{exit}} = \mathcal{P}_2 \mathcal{T}_2 \mathcal{P}_1 \mathcal{T}_1 \mathcal{P}_0 \mathcal{T}_0 \psi_{\text{in}}, \tag{5.1000}$$

where, for compactness, $\mathcal{P}_n \equiv \mathcal{P}_{\Delta z_n}$. Substituting Eq. (5.998),

$$\begin{aligned} \psi_{\text{exit}} \simeq\ & \mathcal{P}_2 \left(1 + j\sigma V_2^{\text{proj}} \right) \mathcal{P}_1 \left(1 + j\sigma V_1^{\text{proj}} \right) \\ & \times \mathcal{P}_0 \left(1 + j\sigma V_0^{\text{proj}} \right) \psi_{\text{in}}. \end{aligned} \tag{5.1001}$$

Keeping only terms of zeroth and first order in the potential gives

$$\begin{aligned} \psi_{\text{exit}} \simeq\ & \mathcal{P}_2 \mathcal{P}_1 \mathcal{P}_0 \psi_{\text{in}} \\ & + j\sigma \mathcal{P}_2 \mathcal{P}_1 \mathcal{P}_0 \left\{ V_0^{\text{proj}} \psi_{\text{in}} \right\} \\ & + j\sigma \mathcal{P}_2 \mathcal{P}_1 \left\{ V_1^{\text{proj}} \mathcal{P}_0 \psi_{\text{in}} \right\} \\ & + j\sigma \mathcal{P}_2 \left\{ V_2^{\text{proj}} \mathcal{P}_1 \mathcal{P}_0 \psi_{\text{in}} \right\}. \end{aligned} \tag{5.1002}$$

For the normally incident plane wave used here,

$$\psi_{\text{in}} = 1. \tag{5.1003}$$

Free-space propagation leaves this constant component unchanged because its transverse spatial frequency is zero:

$$\mathcal{P}_{\Delta z}\{1\} = 1. \tag{5.1004}$$

Therefore, Eq. (5.1002) reduces to

$$\begin{aligned}\psi_{\text{exit}} \simeq 1 &+ j\sigma\mathcal{P}_2\mathcal{P}_1\mathcal{P}_0\left\{V_0^{\text{proj}}\right\} \\ &+ j\sigma\mathcal{P}_2\mathcal{P}_1\left\{V_1^{\text{proj}}\right\} + j\sigma\mathcal{P}_2\left\{V_2^{\text{proj}}\right\}.\end{aligned} \tag{5.1005}$$

The physical meaning is now explicit: a contribution generated in an early slice propagates through all the slices that remain before the exit surface, whereas a contribution generated near the exit surface propagates only a short distance.

More generally, let $\bar{z}_n$ denote a representative depth of slice n. The contribution generated in that slice must propagate the remaining distance

$$z_{\text{exit}} - \bar{z}_n. \tag{5.1006}$$

Since free-space propagation operators compose according to

$$\mathcal{P}_{L_2}\mathcal{P}_{L_1} = \mathcal{P}_{L_1+L_2}, \tag{5.1007}$$

the first-order scattered exit wave can be written

$$\boxed{\delta\psi_{\text{exit}} \simeq j\sigma\sum_n \mathcal{P}_{z_{\text{exit}}-\bar{z}_n}\left\{V_n^{\text{proj}}\right\}.} \tag{5.1008}$$

Taking the two-dimensional Fourier transform and using the Fresnel propagator of Eq. (5.674),

$$P_L(\mathbf{R}_\perp) = \exp\left(-j\pi\lambda L R_\perp^2\right), \tag{5.1009}$$

gives

$$\boxed{\delta\widehat{\psi}_{\text{exit}}(\mathbf{R}_\perp) \simeq j\sigma\sum_n \widehat{V}_n^{\text{proj}}(\mathbf{R}_\perp)\exp\left[-j\pi\lambda\left(z_{\text{exit}} - \bar{z}_n\right)R_\perp^2\right].} \tag{5.1010}$$

To obtain the continuous-depth limit, recall from Eq. (5.633) that

$$V_n^{\text{proj}}(x,y) = \int_{z_n}^{z_{n+1}} V(x,y,z)\,dz. \tag{5.1011}$$

For a sufficiently thin slice,

$$\widehat{V}_n^{\text{proj}}(\mathbf{R}_\perp) \simeq \widehat{V}_\perp(\mathbf{R}_\perp, \bar{z}_n)\Delta z_n. \tag{5.1012}$$

Hence Eq. (5.995) becomes the Riemann sum

$$\begin{aligned}\delta\widehat{\psi}_{\text{exit}}(\mathbf{R}_\perp) \simeq j\sigma \sum_n \widehat{V}_\perp(\mathbf{R}_\perp, \bar{z}_n) \\ \times e^{-j\pi\lambda(z_{\text{exit}}-\bar{z}_n)R_\perp^2}\Delta z_n.\end{aligned} \tag{5.1013}$$

Taking the limit $\max_n \Delta z_n \to 0$ therefore gives

$$\boxed{\begin{aligned}\delta\widehat{\psi}_{\text{exit}}(\mathbf{R}_\perp) \simeq j\sigma \int_{z_{\text{in}}}^{z_{\text{exit}}} \widehat{V}_\perp(\mathbf{R}_\perp, z) \\ \times \exp\left[-j\pi\lambda(z_{\text{exit}}-z)R_\perp^2\right]\,dz.\end{aligned}} \tag{5.1014}$$

Equation (5.1014) makes clear what full multislice contains automatically. Each depth generates scattering at a different point along the beam and the resulting contribution propagates the appropriate remaining distance to the exit surface. The single-scattering limit makes these contributions additive; full multislice additionally allows the wave incident on each later slice to have already been modified by all preceding slices.

5.10.5 The same longitudinal phase appears in the Ewald sphere

We can now make the connection with the Ewald-sphere geometry derived in Sec. 5.9.

The forward branch of the Ewald sphere in transfer-frequency coordinates was found in Eq. (5.726) to be

$$R_z^{\text{E}}(R_\perp) = -\frac{1}{\lambda} + \sqrt{\frac{1}{\lambda^2} - R_\perp^2}. \tag{5.1015}$$

This quantity has a direct propagation interpretation. It is the difference between the longitudinal spatial frequency of the scattered plane wave and that of the incident carrier wave.

Depth-dependent defocus and Ewald curvature are the same longitudinal phase.

We begin with the exact angular-spectrum propagation law derived in Sec. 3.6. A transverse Fourier component $\mathbf{R}_\perp$ propagating through a field-free distance L acquires the phase

$$\exp\left[j2\pi L\sqrt{\frac{1}{\lambda^2}-R_\perp^2}\right] \tag{5.1016}$$

from Eq. (3.289). The rapidly varying incident carrier itself accumulates

$$\exp\left(j\frac{2\pi}{\lambda}L\right). \tag{5.1017}$$

After removing this carrier phase, as was done when deriving the fast-electron envelope equation, the remaining propagation factor is

$$P_L^{\text{exact}}(\mathbf{R}_\perp) = \exp\left\{j2\pi L\left[\sqrt{\frac{1}{\lambda^2}-R_\perp^2}-\frac{1}{\lambda}\right]\right\} \tag{5.1018}$$

$$= \boxed{\exp\left[j2\pi L R_z^{\text{E}}(R_\perp)\right],} \tag{5.1019}$$

where Eq. (5.1015) has been used.

Thus the longitudinal displacement of the Ewald sphere is exactly the excess longitudinal spatial frequency governing propagation of the carrier-removed scattered wave.

Now consider scattering generated at depth z. The remaining distance to the exit surface is

$$L = z_{\text{exit}} - z. \tag{5.1020}$$

Equation (5.1019) gives

$$P_{z_{\text{exit}}-z}^{\text{exact}} = e^{j2\pi R_z^{\text{E}}(z_{\text{exit}}-z)} \tag{5.1021}$$

$$= e^{j2\pi R_z^{\text{E}} z_{\text{exit}}} e^{-j2\pi R_z^{\text{E}} z}. \tag{5.1022}$$

The first factor is common to all depths of the specimen at a given $\mathbf{R}_\perp$; it depends only on the choice of exit reference plane. The second factor,

$$\boxed{e^{-j2\pi R_z^{\mathrm{E}} z},} \tag{5.1023}$$

contains the relative phase between different depths.

But the three-dimensional Fourier transform of the specimen potential was written in Eq. (5.778) as

$$\widehat{V}(\mathbf{R}_\perp, R_z) = \int \widehat{V}_\perp(\mathbf{R}_\perp, z) e^{-j2\pi R_z z}\, dz. \tag{5.1024}$$

Hence, in the single-scattering limit, summing the contributions from all depths with the propagation phase of Eq. (5.1023) produces

$$\boxed{\int \widehat{V}_\perp(\mathbf{R}_\perp, z) e^{-j2\pi R_z^{\mathrm{E}} z}\, dz = \widehat{V}\left(\mathbf{R}_\perp, R_z^{\mathrm{E}}\right).} \tag{5.1025}$$

Thus the familiar statement that first-order elastic scattering samples the three-dimensional specimen transform on the Ewald sphere is the reciprocal-space expression of the coherent propagation phases acquired by contributions generated at different depths.

We now take the paraxial limit. From Eq. (5.764),

$$R_z^{\mathrm{E}} \simeq -\frac{\lambda R_\perp^2}{2}. \tag{5.1026}$$

The depth-dependent Ewald factor therefore becomes

$$e^{-j2\pi R_z^{\mathrm{E}} z} \simeq \exp\left[-j2\pi\left(-\frac{\lambda R_\perp^2}{2}\right) z\right] \tag{5.1027}$$

$$= \boxed{e^{j\pi\lambda z R_\perp^2}.} \tag{5.1028}$$

On the other hand, the Fresnel factor from depth z to the exit surface is

$$e^{-j\pi\lambda(z_{\mathrm{exit}} - z) R_\perp^2} = e^{-j\pi\lambda z_{\mathrm{exit}} R_\perp^2} e^{j\pi\lambda z R_\perp^2}. \tag{5.1029}$$

The second factor is exactly Eq. (5.1028).

Finally, relative to a chosen specimen plane z_{ref},

$$e^{j\pi\lambda(z-z_{\mathrm{ref}})R_\perp^2} \tag{5.1030}$$

is also precisely the additional defocus factor corresponding to

$$\Delta f_{\mathrm{eff}}(z) - \Delta f_{\mathrm{ref}} = z - z_{\mathrm{ref}} \tag{5.1031}$$

from Eq. (5.983).

We have therefore identified the same phase in three different descriptions:

$$\boxed{\begin{aligned} \text{Fresnel propagation through depth} \quad &: \quad e^{j\pi\lambda(z-z_{\mathrm{ref}})R_\perp^2}, \\ \text{depth-dependent defocus} \quad &: \quad e^{j\pi\lambda(\Delta f_{\mathrm{eff}}-\Delta f_{\mathrm{ref}})R_\perp^2}, \\ \text{Ewald-sphere longitudinal phase} \quad &: \quad e^{-j2\pi R_z^{\mathrm{E}}(z-z_{\mathrm{ref}})}. \end{aligned}} \tag{5.1032}$$

In the paraxial limit these three expressions are identical because

$$R_z^{\mathrm{E}} \simeq -\frac{\lambda R_\perp^2}{2}. \tag{5.1033}$$

The result of the derivation may be summarized physically as *Ewald-sphere curvature and depth-dependent defocus are two descriptions of the same longitudinal phase.*

The connection is even stronger than the paraxial statement alone might suggest. Equation (5.1019) shows that the exact forward Ewald branch is already contained in the carrier-removed angular-spectrum propagator. The familiar quadratic defocus phase arises when both are reduced to their common paraxial form.

5.10.6 When is one defocus and the projection approximation adequate for Single Particle Analysis?

The previous derivations give a simple quantitative criterion for deciding when specimen depth can be neglected. Let D denote the effective extent of the particle along the beam direction. For an approximately spherical particle, D is close to its diameter; for an anisotropic particle, it generally depends on orientation.

Consider two scattering contributions separated by the full depth D.

From Eq. (5.972), their relative propagation phase is, in the paraxial approximation,

$$\boxed{\Delta\phi_{\text{depth}}(R_\perp) \simeq \pi\lambda D R_\perp^2.} \tag{5.1034}$$

At real-space resolution

$$d = \frac{1}{R_\perp}, \tag{5.1035}$$

this becomes

$$\boxed{\Delta\phi_{\text{depth}}(d) \simeq \frac{\pi\lambda D}{d^2}.} \tag{5.1036}$$

Equation (5.1036) measures the front-to-back phase spread across the particle. If the single defocus used in the conventional image-formation model is assigned to a reference plane near the particle centre, the largest phase difference relative to that reference plane is approximately one half of this value:

$$\boxed{\Delta\phi_{\text{centre}\to\text{edge}} \simeq \frac{\pi\lambda D}{2d^2}.} \tag{5.1037}$$

Thus a natural criterion for the single-defocus projection approximation is

$$\boxed{\pi\lambda D R_\perp^2 = \frac{\pi\lambda D}{d^2} \ll 1.} \tag{5.1038}$$

This is precisely the projection criterion derived earlier in Eq. (5.557).

The criterion should not be interpreted as a sharp resolution boundary. Rather,

$$\eta_{\text{depth}} \equiv \pi\lambda D R_\perp^2 \tag{5.1039}$$

is the dimensionless parameter controlling the approximation. Values $\eta_{\text{depth}} \ll 1$ imply an almost depth-independent phase; values of order unity indicate that depth effects are becoming significant; and values of several radians indicate that describing the complete particle by one projected potential and one defocus is no longer quantitatively accurate.

The same condition has three equivalent interpretations. If Eq. (5.1038) holds, propagation through the specimen thickness may be neglected, the different depths may be assigned essentially the same effective defocus, and the Ewald sphere may be replaced locally by its tangent plane. Thus

$$\boxed{\begin{aligned} \pi\lambda D R_\perp^2 \ll 1 \quad &\Longleftrightarrow \quad \text{negligible propagation through specimen depth,} \\ &\Longleftrightarrow \quad \text{one effective defocus is adequate,} \\ &\Longleftrightarrow \quad \text{the Ewald sphere may be replaced by its tangent plane.} \end{aligned}} \tag{5.1040}$$

For SPA this distinction is important. A fitted particle defocus should be understood as the defocus associated with a chosen effective reference plane, usually close to the particle centre. It does not mean that every atom in a three-dimensional particle is physically at the same defocus. Multislice retains the different propagation distances automatically, as shown in Sec. 5.10.4, whereas in the single-scattering reciprocal-space description the same longitudinal phase appears as Ewald-sphere curvature in Sec. 5.10.5.

Numerical values for typical SPA conditions. For commonly used acceleration voltages,

$$\lambda_{200\,\mathrm{keV}} \simeq 0.0251\,\text{Å}, \qquad \lambda_{300\,\mathrm{keV}} \simeq 0.0197\,\text{Å}. \tag{5.1041}$$

Table 5.3 gives the front-to-back phase spread predicted by Eq. (5.1036) for representative particle depths and resolutions.

Table 5.3: Front-to-back depth-dependent phase spread $\Delta\phi_{\mathrm{depth}} \simeq \pi\lambda D/d^2$ for representative SPA particle depths and resolutions. The maximum phase difference relative to a reference plane at the particle centre is approximately one half of the tabulated value.

Energy	D (Å)	4 Å	3 Å	2 Å
200 keV	100	0.49	0.88	1.97
200 keV	300	1.48	2.63	5.91
200 keV	500	2.46	4.38	9.85
300 keV	100	0.39	0.69	1.55
300 keV	300	1.16	2.06	4.64
300 keV	500	1.93	3.44	7.73

Several trends are immediately apparent. For a relatively small particle with $D = 100\,\text{Å}$ at 300 keV, the front-to-back phase spread is only

$$0.39\,\mathrm{rad} \tag{5.1042}$$

at 4 Å resolution and

$$0.69\,\mathrm{rad} \tag{5.1043}$$

at 3 Å. If the defocus reference plane lies near the particle centre, the largest phase error relative to that plane is only about 0.20 and 0.34 rad, respectively. A single projected potential with one effective defocus is therefore an excellent approximation under these conditions.

At 2 Å resolution, however, the same 100 Å particle spans approximately

$$1.55\,\mathrm{rad} \tag{5.1044}$$

from front to back. The specimen may still be regarded as physically thin, but treating all of its depths as having exactly the same phase is no longer an exact approximation.

Particle depth has an equally strong effect. For a 300 Å particle at 300 keV, the phase spreads at 4, 3, and 2 Å resolution are approximately

$$1.16, \qquad 2.06, \qquad 4.64\,\mathrm{rad}, \tag{5.1045}$$

respectively. For a 500 Å particle, even the 4 Å components span almost 2 rad from front to back, while at 2 Å the spread exceeds 7 rad.

The effect is somewhat weaker at 300 keV than at 200 keV because the electron wavelength is shorter, but the dependence on particle depth and resolution is unchanged:

$$\Delta\phi_{\mathrm{depth}} \propto \lambda D R_\perp^2 = \frac{\lambda D}{d^2}. \tag{5.1046}$$

Thus the error grows linearly with particle depth but quadratically as higher resolution is pursued.

A useful characteristic resolution. For a chosen tolerable front-to-back phase spread ϵ, Eq. (5.1036) can be inverted to give

$$\boxed{d_\epsilon = \sqrt{\frac{\pi\lambda D}{\epsilon}}.} \tag{5.1047}$$

For the illustrative choice

$$\epsilon = 1\,\mathrm{rad}, \tag{5.1048}$$

the corresponding characteristic resolutions are

D (Å)	$d_{1\,\mathrm{rad}}$ at 200 keV	$d_{1\,\mathrm{rad}}$ at 300 keV
100	2.81 Å	2.49 Å
300	4.86 Å	4.31 Å
500	6.28 Å	5.56 Å

These values are *not* resolution limits. They simply indicate where the front-to-back phase spread becomes of order one radian. A different tolerance would give a different numerical boundary, and a central reference

plane experiences only approximately half of this front-to-back spread at either edge.

The estimates nevertheless explain why the projection approximation has been so successful in conventional SPA. At moderate resolution and for relatively small particles, the longitudinal phase variation is small and the standard projected image-formation model is highly accurate. As particle size increases and resolutions approach the atomic scale, however, the same approximation becomes progressively less exact. There is no sudden transition: the controlling parameter is the continuously increasing quantity $\lambda D R_\perp^2$.

5.11 Atomic disorder, thermal motion, occupancies, and B-factors

The independent-atom model developed in Sec. 5.3.9 assigns every atom a well-defined type and position. In reciprocal space, the corresponding electron structure factor was written in Eq. (5.307) as

$$F_e(\mathbf{R}) = \sum_{a=1}^{N_{\mathrm{at}}} f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a}. \tag{5.1049}$$

A real biological specimen is not described by a single exact set of atomic coordinates. Atomic positions fluctuate, alternative conformations may be populated, some sites may be only partially occupied, and the electron beam itself progressively changes the specimen during the exposure.

These effects are often summarized, sometimes too casually, by a B-factor. The purpose of this section is to distinguish the different physical mechanisms that can produce apparently similar resolution-dependent attenuation.

A useful starting point is to replace the deterministic atomic position $\mathbf{r}_a$ by

$$\boxed{\mathbf{r}_a^{(\xi)} = \mathbf{r}_a + \mathbf{u}_a^{(\xi)},} \tag{5.1050}$$

where ξ labels one microscopic configuration and $\mathbf{u}_a^{(\xi)}$ is the displacement of atom a from its reference position. We shall first study the effect of such displacements before distinguishing their possible physical origins.

5.11.1 Atomic displacements and the Debye–Waller factor

For one configuration ξ, the contribution of atom a to the electron structure factor is

$$F_{e,a}^{(\xi)}(\mathbf{R}) = f_{e,a}(R) \exp\left[-j2\pi\mathbf{R}\cdot\left(\mathbf{r}_a + \mathbf{u}_a^{(\xi)}\right)\right]. \tag{5.1051}$$

Separating the reference position from the displacement,

$$F_{e,a}^{(\xi)}(\mathbf{R}) = f_{e,a}(R) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a} e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_a^{(\xi)}}. \tag{5.1052}$$

The first two factors are those already present in Eq. (5.307). All information about positional disorder is contained in the last factor.

The Debye–Waller factor as the characteristic function of atomic displacement.

Taking the ensemble average of Eq. (5.1052) gives

$$\begin{aligned}\left\langle F_{e,a}^{(\xi)}(\mathbf{R})\right\rangle_\xi &= f_{e,a}(R) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a} \\ &\quad\times \left\langle e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_a}\right\rangle.\end{aligned} \tag{5.1053}$$

Define

$$\boxed{D_a(\mathbf{R}) \equiv \left\langle e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_a}\right\rangle.} \tag{5.1054}$$

This quantity is the characteristic function of the probability distribution of the atomic displacement.

Thus, without assuming any particular distribution,

$$\boxed{\left\langle F_{e,a}^{(\xi)}(\mathbf{R})\right\rangle_\xi = f_{e,a}(R) D_a(\mathbf{R}) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a}.} \tag{5.1055}$$

Suppose now that the atomic displacement follows a three-dimensional Gaussian distribution,

$$\mathbf{u}_a \sim \mathcal{N}(\mathbf{0}, \mathbf{U}_a), \tag{5.1056}$$

where $\mathbf{U}_a$ is the displacement covariance matrix. The characteristic function of a Gaussian random vector is

$$\left\langle e^{-j\mathbf{q}\cdot\mathbf{u}_a}\right\rangle = \exp\left(-\frac{1}{2}\mathbf{q}^T\mathbf{U}_a\mathbf{q}\right). \tag{5.1057}$$

Setting

$$\mathbf{q} = 2\pi\mathbf{R} \tag{5.1058}$$

therefore gives

$$\boxed{D_a(\mathbf{R}) = \exp\left(-2\pi^2\mathbf{R}^T\mathbf{U}_a\mathbf{R}\right).} \tag{5.1059}$$

For isotropic disorder,

$$\mathbf{U}_a = U_a\mathbf{I}, \tag{5.1060}$$

and hence

$$D_a(R) = e^{-2\pi^2 U_a R^2}. \tag{5.1061}$$

Introducing the conventional isotropic atomic displacement parameter

$$\boxed{B_a \equiv 8\pi^2 U_a,} \tag{5.1062}$$

we finally obtain

$$\boxed{D_a(R) = \exp\left(-\frac{B_a R^2}{4}\right).} \tag{5.1063}$$

The parameter B_a has units of Å^2. Care is required when relating it to a quoted root-mean-square displacement. In the isotropic case,

$$U_a = \left\langle u_{a,x}^2 \right\rangle = \left\langle u_{a,y}^2 \right\rangle = \left\langle u_{a,z}^2 \right\rangle, \tag{5.1064}$$

whereas

$$\left\langle \|\mathbf{u}_a\|^2 \right\rangle = 3U_a. \tag{5.1065}$$

Consequently,

$$\boxed{B_a = \frac{8\pi^2}{3}\left\langle \|\mathbf{u}_a\|^2 \right\rangle.} \tag{5.1066}$$

Equation (5.1054) is more fundamental than the B-factor representation. A Gaussian B-factor is obtained only when the displacement distribution is Gaussian, or when its second moment provides an adequate approximation over the spatial frequency range of interest.

Throughout this book we use the standard crystallographic and SPA convention of Eq. (5.1063). Some alternative conventions may absorb the factor 1/4 into the definition of B.

Where does the factor $1/4$ in the Debye–Waller factor come from?.

In crystallography, the Debye–Waller factor is often written as

$$D = \exp\left[-B\left(\frac{\sin\theta}{\lambda}\right)^2\right], \tag{5.1067}$$

where θ is the Bragg angle, that is, one half of the total scattering angle. From the elastic-scattering geometry derived in Eq. (5.353),

$$R = \frac{2}{\lambda}\sin\theta. \tag{5.1068}$$

Hence

$$\frac{\sin\theta}{\lambda} = \frac{R}{2}, \tag{5.1069}$$

and therefore

$$B\left(\frac{\sin\theta}{\lambda}\right)^2 = \frac{BR^2}{4}. \tag{5.1070}$$

Thus the conventional crystallographic expression becomes

$$\boxed{D(R) = e^{-BR^2/4}.} \tag{5.1071}$$

5.11.2 Occupancy and positional disorder

Positional uncertainty is not the only source of attenuation of an atomic contribution. An atom or ligand may be present only in a fraction of the particles.

Let

$$\eta_a^{(\xi)} = \begin{cases} 1, & \text{if atom } a \text{ is present in configuration } \xi, \\ 0, & \text{otherwise,} \end{cases} \tag{5.1072}$$

and define its occupancy as

$$\boxed{q_a = \left\langle \eta_a^{(\xi)} \right\rangle, \qquad 0 \le q_a \le 1.} \tag{5.1073}$$

Assuming for the moment that occupancy and positional displacement are statistically independent, the configuration-dependent atomic contribu-

tion is

$$F_{e,a}^{(\xi)} = \eta_a^{(\xi)} f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a} e^{-j2\pi \mathbf{R}\cdot\mathbf{u}_a^{(\xi)}}. \tag{5.1074}$$

Taking its expectation,

$$\left\langle F_{e,a}^{(\xi)} \right\rangle = q_a f_{e,a}(R) D_a(\mathbf{R}) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a}. \tag{5.1075}$$

For isotropic Gaussian positional disorder,

$$\boxed{\left\langle F_e^{(\xi)}(\mathbf{R}) \right\rangle = \sum_a q_a f_{e,a}(R) e^{-B_a R^2/4} e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a}.} \tag{5.1076}$$

Equation (5.1076) is the natural extension of the IAM structure factor of Eq. (5.307) to include occupancy and isotropic atomic displacement parameters.

The roles of q_a and B_a are physically different. Occupancy multiplies the atomic contribution approximately uniformly over spatial frequency, whereas an atomic displacement parameter preferentially attenuates high spatial frequencies:

$$\boxed{\begin{array}{lcl} q_a & : & \text{fraction of the ensemble containing the site,} \\ B_a & : & \text{width of its positional distribution.} \end{array}} \tag{5.1077}$$

At finite resolution, however, the two parameters can be strongly correlated during atomic-model refinement.

Why partial occupancy does not merely remove scattered intensity.

Consider one atomic site a at position $\mathbf{r}_a$. If the site is occupied, its contribution to the electron structure factor introduced in Eq. (5.307) is

$$\boxed{F_{e,a}(\mathbf{R}) = f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a},} \tag{5.1078}$$

where $f_{e,a}(R)$ is the electron scattering factor of atom a.

Now suppose that the site is not occupied in every member of the specimen ensemble. For configuration ξ, introduce the occupancy variable already used in Eq. (5.1073),

$$\eta_a^{(\xi)} = \begin{cases} 1, & \text{if site } a \text{ is occupied,} \\ 0, & \text{if site } a \text{ is empty,} \end{cases} \tag{5.1079}$$

with

$$\Pr\left(\eta_a^{(\xi)} = 1\right) = q_a, \qquad \Pr\left(\eta_a^{(\xi)} = 0\right) = 1 - q_a. \tag{5.1080}$$

Thus

$$\left\langle \eta_a^{(\xi)} \right\rangle_\xi = q_a. \tag{5.1081}$$

The contribution of this site to the structure factor of one particular configuration is therefore

$$\boxed{F_{e,a}^{(\xi)}(\mathbf{R}) = \eta_a^{(\xi)} f_{e,a}(R) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a}.} \tag{5.1082}$$

Taking the ensemble average of the *complex scattering amplitude*,

$$\left\langle F_{e,a}^{(\xi)}(\mathbf{R}) \right\rangle_\xi = \left\langle \eta_a^{(\xi)} \right\rangle_\xi f_{e,a}(R) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a} \tag{5.1083}$$

$$= q_a f_{e,a}(R) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a}. \tag{5.1084}$$

Hence

$$\boxed{\left\langle F_{e,a}^{(\xi)}(\mathbf{R}) \right\rangle_\xi = q_a F_{e,a}(\mathbf{R}).} \tag{5.1085}$$

This is the origin of the usual occupancy factor q_a multiplying an atomic contribution in Eq. (5.1076).

The intensity associated with this *coherently averaged* amplitude is

$$\left| \left\langle F_{e,a}^{(\xi)} \right\rangle_\xi \right|^2 = q_a^2 \left| F_{e,a} \right|^2 \tag{5.1086}$$

$$= \boxed{q_a^2 |f_{e,a}(R)|^2.} \tag{5.1087}$$

The Fourier phase disappears from the modulus because

$$\left| e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a} \right| = 1. \tag{5.1088}$$

However, this is not the same as averaging the scattered intensity of the different configurations. For one configuration,

$$\left| F_{e,a}^{(\xi)}(\mathbf{R}) \right|^2 = \left(\eta_a^{(\xi)} \right)^2 |f_{e,a}(R)|^2. \tag{5.1089}$$

Because $\eta_a^{(\xi)}$ can only be 0 or 1,

$$\left(\eta_a^{(\xi)}\right)^2 = \eta_a^{(\xi)}. \tag{5.1090}$$

Therefore,

$$\left\langle \left| F_{e,a}^{(\xi)}(\mathbf{R}) \right|^2 \right\rangle_\xi = \left\langle \eta_a^{(\xi)} \right\rangle_\xi |f_{e,a}(R)|^2 \tag{5.1091}$$

$$= \boxed{q_a |f_{e,a}(R)|^2.} \tag{5.1092}$$

We can now separate this average intensity into the intensity of the coherent average plus a fluctuation term:

$$q_a |f_{e,a}|^2 = q_a^2 |f_{e,a}|^2 + \left(q_a - q_a^2\right) |f_{e,a}|^2 \tag{5.1093}$$

$$= q_a^2 |f_{e,a}|^2 + q_a(1-q_a)|f_{e,a}|^2. \tag{5.1094}$$

Thus

$$\boxed{\begin{aligned} \left\langle \left| F_{e,a}^{(\xi)} \right|^2 \right\rangle_\xi = & \left| \left\langle F_{e,a}^{(\xi)} \right\rangle_\xi \right|^2 \\ & + q_a(1-q_a)|f_{e,a}(R)|^2. \end{aligned}} \tag{5.1095}$$

The second term is the variance of the configuration-dependent scattering amplitude:

$$\boxed{\mathrm{Var}_\xi \left[F_{e,a}^{(\xi)}(\mathbf{R}) \right] = q_a(1-q_a)|f_{e,a}(R)|^2.} \tag{5.1096}$$

Therefore partial occupancy does not mean that the fraction $1-q_a$ of the scattering simply disappears. It reduces the *coherently reproducible* atomic amplitude by the factor q_a, while configuration-to-configuration fluctuations contribute the remaining intensity.

The distinction is the same one that will appear below in the frozen-configuration construction. In general,

$$\left\langle |F|^2 \right\rangle = |\langle F \rangle|^2 + \left\langle |F - \langle F \rangle|^2 \right\rangle. \tag{5.1097}$$

The first term is the part that remains phase coherent from one member of the ensemble to another; the second measures scattering associated with configurational fluctuations. For a partially occupied atomic site, these two contributions scale respectively as q_a^2 and $q_a(1-q_a)$.

5.11.3 Frozen configurations and the frozen-phonon construction

Thermal motion provides the historical origin of the Debye–Waller factor. Before introducing the *frozen-phonon* construction, it is useful to clarify what a phonon is.

What is a phonon? Atoms in a solid do not remain exactly at fixed positions. Even if the solid has a well-defined equilibrium structure, the atoms vibrate about their equilibrium coordinates. These vibrations are not, in general, independent motions of individual atoms. Because the atoms interact, the displacement of one atom affects its neighbours, and the natural motions of the solid are therefore *collective vibrational modes.*

Let

$$\mathbf{r}_a(t) = \mathbf{r}_a^{(0)} + \mathbf{u}_a(t) \tag{5.1098}$$

be the position of atom a, where $\mathbf{r}_a^{(0)}$ is its equilibrium position and $\mathbf{u}_a(t)$ its displacement. For sufficiently small displacements, the potential energy may be approximated to second order around equilibrium. The coupled equations of motion can then be diagonalized into independent normal modes. The displacement of each atom may be written schematically as

$$\boxed{\mathbf{u}_a(t) = \sum_\nu \mathbf{e}_{a\nu} Q_\nu(t),} \tag{5.1099}$$

where ν labels a vibrational mode, $\mathbf{e}_{a\nu}$ specifies how atom a participates in that mode, and $Q_\nu(t)$ is its mode amplitude.

Each normal coordinate behaves, in the harmonic approximation, as an independent oscillator,

$$\frac{d^2 Q_\nu}{dt^2} + \omega_\nu^2 Q_\nu = 0. \tag{5.1100}$$

Classically, this simply describes a collective vibration at angular frequency ω_ν.

Quantum mechanics quantizes the energy of each vibrational mode:

$$\boxed{E_{\nu,n} = \hbar\omega_\nu \left(n + \frac{1}{2}\right), \qquad n = 0, 1, 2, \ldots} \tag{5.1101}$$

A single quantum of excitation of such a vibrational mode is called a *phonon.* Thus a phonon is not another kind of particle residing inside the specimen; it is the quantum-mechanical description of one unit of excitation of a collective atomic vibration.

In a perfect crystal, translational periodicity allows these modes to be labelled by a crystal wavevector and a polarization, and the language of phonons is especially natural. In a vitrified biological specimen there is no long-range crystalline periodicity. The atoms and the surrounding amorphous ice still possess vibrational normal modes, but they need not resemble plane waves extending through a crystal. For this reason, in cryo-EM the term *frozen phonon* should be understood primarily as the historical name of a computational construction rather than as a statement that the specimen is a perfect phonon-supporting crystal.

An additional point is that cooling does not eliminate atomic displacements completely. Lower temperature reduces the thermal population of vibrational modes, but quantum zero-point motion remains even in the vibrational ground state because of the $\hbar\omega_\nu/2$ term in Eq. (5.1101).

The frozen-phonon idea. In crystalline electron microscopy, a particularly intuitive way to simulate the effect of atomic vibrations is the *frozen-phonon* construction (Loane et al., 1991; Kirkland, 2010; Muller et al., 2001; Lobato and Van Dyck, 2015).

The central idea is to exploit the very different time scales of the fast electron and the atomic motion. During the passage of one high-energy electron through the specimen, the atomic configuration may be regarded approximately as fixed. One therefore chooses one instantaneous set of displaced atomic coordinates,

$$\left\{\mathbf{r}_a^{(\xi)}\right\}_{a=1}^{N_{\mathrm{at}}}, \tag{5.1102}$$

constructs the corresponding static specimen potential

$$V^{(\xi)}(\mathbf{r}), \tag{5.1103}$$

and propagates the electron through that fixed potential.

The calculation is then repeated for many configurations ξ drawn from the statistical distribution of atomic displacements. The detector intensities, rather than the complex electron waves, are finally averaged over the configurations.

Schematically,

$$\boxed{\begin{array}{c}\text{sample displaced atomic coordinates}\\ \Downarrow\\ V^{(\xi)}(\mathbf{r})\\ \Downarrow\\ \text{elastic propagation through the frozen configuration}\\ \Downarrow\\ I^{(\xi)}(\boldsymbol{\rho})\\ \Downarrow\\ \left\langle I^{(\xi)}(\boldsymbol{\rho})\right\rangle_{\xi}.\end{array}} \tag{5.1104}$$

The adjective "frozen" therefore refers to the approximation made during each individual electron-propagation calculation: the atoms are held fixed while that electron wave traverses the specimen. Different members of the ensemble represent different snapshots of the vibrational motion.

For single-particle cryo-EM, it is useful to generalize this idea from literal thermal phonons to an ensemble of *frozen configurations*. The configurations may contain residual thermal displacements, zero-point motion, static structural disorder, or other small-scale variations of the atomic coordinates. The same computational construction can represent all of them, although their physical origins are different.

This distinction is especially important for vitrified biological specimens. A distribution of atomic positions observed across a SPA dataset should not automatically be interpreted as thermal vibration. It may instead reflect conformational heterogeneity, incomplete occupancy, different chemical states, or beam-induced evolution of the specimen. These mechanisms can produce similar attenuation of coherent high-resolution signal while corresponding to very different specimen physics.

Why frozen-configuration simulations average intensity, not wave amplitude.

Let

$$\psi_{\text{obs}}^{(\xi)}(\boldsymbol{\rho}) \tag{5.1105}$$

be the detector-plane wave calculated for frozen configuration ξ. Each such calculation may be performed using the ordinary elastic multislice equation of Sec. 5.8.

The experimentally relevant configuration-averaged intensity is

$$\boxed{\overline{I}(\boldsymbol{\rho}) = \left\langle \left| \psi_{\text{obs}}^{(\xi)}(\boldsymbol{\rho}) \right|^2 \right\rangle_\xi .} \tag{5.1106}$$

Write each wave as

$$\psi_{\text{obs}}^{(\xi)} = \overline{\psi}_{\text{obs}} + \delta\psi^{(\xi)}, \tag{5.1107}$$

where

$$\overline{\psi}_{\text{obs}} = \left\langle \psi_{\text{obs}}^{(\xi)} \right\rangle_\xi , \qquad \left\langle \delta\psi^{(\xi)} \right\rangle_\xi = 0. \tag{5.1108}$$

Then

$$\overline{I} = \left\langle \left| \overline{\psi}_{\text{obs}} + \delta\psi^{(\xi)} \right|^2 \right\rangle \tag{5.1109}$$

$$= \left| \overline{\psi}_{\text{obs}} \right|^2 + \left\langle \left| \delta\psi^{(\xi)} \right|^2 \right\rangle , \tag{5.1110}$$

because the cross terms vanish. Hence

$$\boxed{\overline{I} = \underbrace{\left| \left\langle \psi_{\text{obs}}^{(\xi)} \right\rangle \right|^2}_{\text{coherent contribution}} + \underbrace{\left\langle \left| \psi_{\text{obs}}^{(\xi)} - \left\langle \psi_{\text{obs}}^{(\xi)} \right\rangle \right|^2 \right\rangle}_{\text{configurationally diffuse contribution}} .} \tag{5.1111}$$

Therefore,

$$\left\langle |\psi|^2 \right\rangle \neq |\langle \psi \rangle|^2 \tag{5.1112}$$

unless all configurations produce exactly the same wave.

This distinction is essential. Replacing the specimen by an averaged potential and performing one multislice calculation would retain only the coherent-average part and would miss the intensity arising from configuration-to-configuration fluctuations.

Each frozen-configuration multislice calculation is an elastic calculation: the atomic potential is static and no explicit energy transfer is introduced. When used to model thermal diffuse scattering, the ensemble average nevertheless approximates the redistribution of intensity caused physically by electron–phonon interactions. It is therefore best understood as an energy-integrated approximation to phonon scattering; it does not provide the energy-loss spectrum of the phonons.

Where the Debye–Waller attenuated coherent intensity goes.

Consider one atom a at reference position $\mathbf{r}_a$. Its contribution to the electron structure factor of the reference configuration is, from Eq. (5.307),

$$F_{e,a}(\mathbf{R}) = f_{e,a}(R)e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_a}. \tag{5.1113}$$

In frozen configuration ξ, the atom is displaced according to Eq. (5.1050),

$$\mathbf{r}_a^{(\xi)} = \mathbf{r}_a + \mathbf{u}_a^{(\xi)}. \tag{5.1114}$$

Its contribution to the structure factor therefore becomes

$$F_{e,a}^{(\xi)}(\mathbf{R}) = f_{e,a}(R)e^{-j2\pi\mathbf{R}\cdot(\mathbf{r}_a+\mathbf{u}_a^{(\xi)})} \tag{5.1115}$$

$$= F_{e,a}(\mathbf{R})e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_a^{(\xi)}}. \tag{5.1116}$$

The displacement changes the phase of the atomic contribution but not its modulus, because

$$\left|e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_a^{(\xi)}}\right| = 1. \tag{5.1117}$$

Consequently, for every frozen configuration,

$$\left|F_{e,a}^{(\xi)}(\mathbf{R})\right|^2 = |F_{e,a}(\mathbf{R})|^2 \tag{5.1118}$$

$$= |f_{e,a}(R)|^2. \tag{5.1119}$$

Averaging these configuration-dependent intensities therefore gives

$$\boxed{\left\langle\left|F_{e,a}^{(\xi)}(\mathbf{R})\right|^2\right\rangle_\xi = |f_{e,a}(R)|^2.} \tag{5.1120}$$

Now consider instead the *coherently averaged complex amplitude*. Taking the ensemble average of Eq. (5.1116),

$$\left\langle F_{e,a}^{(\xi)}(\mathbf{R})\right\rangle_\xi = F_{e,a}(\mathbf{R})\left\langle e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_a^{(\xi)}}\right\rangle_\xi \tag{5.1121}$$

$$= F_{e,a}(\mathbf{R})D_a(\mathbf{R}), \tag{5.1122}$$

where the disorder factor D_a was defined in Eq. (5.1054).

For isotropic Gaussian displacements, Eq. (5.1063) gives

$$D_a(R) = \exp\left(-\frac{B_a R^2}{4}\right). \tag{5.1123}$$

Thus

$$\boxed{\left\langle F_{e,a}^{(\xi)}(\mathbf{R})\right\rangle_\xi = F_{e,a}(\mathbf{R}) e^{-B_a R^2/4}.} \tag{5.1124}$$

The intensity associated with this coherently reproducible part is therefore

$$\left|\left\langle F_{e,a}^{(\xi)}(\mathbf{R})\right\rangle_\xi\right|^2 = |F_{e,a}(\mathbf{R})|^2\, e^{-B_a R^2/2} \tag{5.1125}$$

$$= \boxed{|f_{e,a}(R)|^2\, e^{-B_a R^2/2}.} \tag{5.1126}$$

We can now use the general decomposition derived in Eq. (5.1111),

$$\left\langle |F|^2 \right\rangle = |\langle F\rangle|^2 + \left\langle |F - \langle F\rangle|^2 \right\rangle. \tag{5.1127}$$

For the displaced atomic contribution this gives

$$\left\langle \left| F_{e,a}^{(\xi)} - \left\langle F_{e,a}^{(\xi)}\right\rangle_\xi \right|^2 \right\rangle_\xi = |f_{e,a}(R)|^2 - |f_{e,a}(R)|^2\, e^{-B_a R^2/2}. \tag{5.1128}$$

Hence

$$\boxed{\begin{aligned} I_{\mathrm{conf},a}(\mathbf{R}) &\equiv \left\langle \left| F_{e,a}^{(\xi)} - \left\langle F_{e,a}^{(\xi)}\right\rangle_\xi \right|^2 \right\rangle_\xi \\ &= |f_{e,a}(R)|^2 \left[1 - e^{-B_a R^2/2}\right]. \end{aligned}} \tag{5.1129}$$

Thus the Debye–Waller factor does not describe a disappearance of scattered intensity. The ensemble average separates the scattering into a part whose complex amplitude is reproducible from one configuration to another,

$$|f_{e,a}(R)|^2\, e^{-B_a R^2/2}, \tag{5.1130}$$

and a part associated with configuration-to-configuration fluctuations,

$$|f_{e,a}(R)|^2 \left[1 - e^{-B_a R^2/2}\right]. \tag{5.1131}$$

Their sum is the configuration-averaged single-atom scattering intensity,

$$\boxed{|f_{e,a}(R)|^2 e^{-B_a R^2/2} + |f_{e,a}(R)|^2 \left[1 - e^{-B_a R^2/2}\right] = |f_{e,a}(R)|^2 .} \tag{5.1132}$$

For one isolated atom the displacement changes only the phase of its scattering amplitude, so the decomposition above is especially transparent. For a specimen containing many atoms, the individual atomic waves interfere. Changing the atomic coordinates therefore changes the interference terms as well, and the configurationally varying contribution is redistributed through reciprocal space. This is the origin of thermal diffuse scattering in the crystalline case. The frozen-configuration calculation retains this redistribution by performing the complete scattering calculation for each displaced atomic configuration before averaging the resulting intensities.

Electron scattering factor of a protein with fluctuating atomic positions.

Consider a protein containing N_{at} atoms. For the reference configuration, its electron structure factor in the independent-atom model is

$$\boxed{F_e(\mathbf{R}) = \sum_{a=1}^{N_{\text{at}}} f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a},} \tag{5.1133}$$

as introduced previously in Eq. (5.307).

In frozen configuration ξ, every atom is displaced from its reference position,

$$\mathbf{r}_a^{(\xi)} = \mathbf{r}_a + \mathbf{u}_a^{(\xi)}. \tag{5.1134}$$

The corresponding structure factor is therefore

$$F_e^{(\xi)}(\mathbf{R}) = \sum_{a=1}^{N_{\text{at}}} f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot(\mathbf{r}_a + \mathbf{u}_a^{(\xi)})} \tag{5.1135}$$

$$= \sum_{a=1}^{N_{\text{at}}} f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a} e^{-j2\pi \mathbf{R}\cdot\mathbf{u}_a^{(\xi)}}. \tag{5.1136}$$

We first calculate the ensemble-averaged *complex scattering amplitude*. By linearity of expectation,

$$\left\langle F_e^{(\xi)}(\mathbf{R}) \right\rangle_\xi = \sum_{a=1}^{N_{\text{at}}} f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a} \times \left\langle e^{-j2\pi \mathbf{R}\cdot\mathbf{u}_a^{(\xi)}} \right\rangle_\xi. \tag{5.1137}$$

Using the atomic disorder factor defined in Eq. (5.1054),

$$D_a(\mathbf{R}) = \left\langle e^{-j2\pi \mathbf{R}\cdot\mathbf{u}_a} \right\rangle, \tag{5.1138}$$

we obtain

$$\boxed{\left\langle F_e^{(\xi)}(\mathbf{R}) \right\rangle_\xi = \sum_{a=1}^{N_{\text{at}}} f_{e,a}(R) D_a(\mathbf{R}) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a}.} \tag{5.1139}$$

Notice that no assumption about independence between the atomic displacements has been required to obtain this result.

For isotropic Gaussian displacements, Eq. (5.1063) gives

$$D_a(R) = e^{-B_a R^2/4}, \tag{5.1140}$$

and hence

$$\boxed{\left\langle F_e^{(\xi)}(\mathbf{R}) \right\rangle_\xi = \sum_{a=1}^{N_{\text{at}}} f_{e,a}(R) e^{-B_a R^2/4} e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a}.} \tag{5.1141}$$

If, as a simplifying model, all atoms have the same isotropic B-factor,

$$B_a = B, \tag{5.1142}$$

the Debye–Waller factor can be taken outside the sum:

$$\left\langle F_e^{(\xi)}(\mathbf{R}) \right\rangle_\xi = e^{-BR^2/4} \sum_a f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a} \tag{5.1143}$$

$$= \boxed{F_e(\mathbf{R}) e^{-BR^2/4}.} \tag{5.1144}$$

Thus the familiar Debye–Waller factor attenuates the *coherently reproducible protein structure factor*.

Average scattered intensity of a fluctuating protein.

Starting from Eq. (5.1136), the scattered intensity of one frozen configuration is

$$\left| F_e^{(\xi)}(\mathbf{R}) \right|^2 = F_e^{(\xi)}(\mathbf{R}) F_e^{(\xi)*}(\mathbf{R}) \tag{5.1145}$$

$$\begin{aligned} &= \sum_{a=1}^{N_{\text{at}}} \sum_{b=1}^{N_{\text{at}}} f_{e,a}(R) f_{e,b}^*(R) \\ &\quad \times e^{-j2\pi \mathbf{R}\cdot(\mathbf{r}_a - \mathbf{r}_b)} \\ &\quad \times e^{-j2\pi \mathbf{R}\cdot(\mathbf{u}_a^{(\xi)} - \mathbf{u}_b^{(\xi)})}. \end{aligned} \tag{5.1146}$$

Averaging over frozen configurations gives

$$\begin{aligned} \left\langle \left| F_e^{(\xi)}(\mathbf{R}) \right|^2 \right\rangle_\xi &= \sum_{a,b} f_{e,a}(R) f_{e,b}^*(R) e^{-j2\pi \mathbf{R}\cdot(\mathbf{r}_a - \mathbf{r}_b)} \\ &\quad \times \left\langle e^{-j2\pi \mathbf{R}\cdot(\mathbf{u}_a - \mathbf{u}_b)} \right\rangle. \end{aligned} \tag{5.1147}$$

This equation is general. It shows that the average intensity depends not only on the displacement of each individual atom, but on the statistics of the *relative displacement*

$$\mathbf{u}_a - \mathbf{u}_b \tag{5.1148}$$

of every pair of atoms.

Suppose that the complete set of atomic displacements is jointly Gaussian. Define the covariance matrices

$$\mathbf{U}_{ab} \equiv \left\langle \mathbf{u}_a \mathbf{u}_b^T \right\rangle . \tag{5.1149}$$

The covariance matrix of the relative displacement is then

$$\mathbf{U}_{a-b} = \left\langle (\mathbf{u}_a - \mathbf{u}_b)(\mathbf{u}_a - \mathbf{u}_b)^T \right\rangle \tag{5.1150}$$

$$= \boxed{\mathbf{U}_{aa} + \mathbf{U}_{bb} - \mathbf{U}_{ab} - \mathbf{U}_{ba}.} \tag{5.1151}$$

Using the Gaussian characteristic function already employed in Eq. (5.1059),

$$\boxed{\left\langle e^{-j2\pi \mathbf{R}\cdot(\mathbf{u}_a - \mathbf{u}_b)} \right\rangle = \exp\left[-2\pi^2 \mathbf{R}^T \mathbf{U}_{a-b} \mathbf{R}\right].} \tag{5.1152}$$

Consequently,

$$\boxed{\begin{aligned} \left\langle |F_e^{(\xi)}(\mathbf{R})|^2 \right\rangle = & \sum_{a,b} f_{e,a}(R) f_{e,b}^*(R) e^{-j2\pi \mathbf{R}\cdot(\mathbf{r}_a - \mathbf{r}_b)} \\ & \times \exp\left[-2\pi^2 \mathbf{R}^T \mathbf{U}_{a-b} \mathbf{R}\right]. \end{aligned}} \tag{5.1153}$$

Equation (5.1153) reveals that the relevant quantity for structural coherence is the fluctuation of *relative atomic positions*. If two atoms move together,

$$\mathbf{u}_a^{(\xi)} = \mathbf{u}_b^{(\xi)}, \tag{5.1154}$$

then

$$\mathbf{U}_{a-b} = 0, \tag{5.1155}$$

and their interference term is not attenuated. A rigid translation of the complete protein therefore does not produce an internal Debye–Waller loss; it changes only the overall phase of the protein structure factor.

Independent identical atomic displacements.

Suppose now, as a simplified model, that the atomic displacements

are independent and identically distributed,

$$\left\langle \mathbf{u}_a \mathbf{u}_b^T \right\rangle = \begin{cases} U\mathbf{I}, & a = b, \\ \mathbf{0}, & a \neq b. \end{cases} \tag{5.1156}$$

For $a \neq b$, Eq. (5.1151) becomes

$$\mathbf{U}_{a-b} = 2U\mathbf{I}. \tag{5.1157}$$

Hence the pair-interference attenuation is

$$\exp\left[-2\pi^2 \mathbf{R}^T (2U\mathbf{I})\mathbf{R}\right] = e^{-4\pi^2 U R^2} \tag{5.1158}$$

$$= \boxed{e^{-BR^2/2},} \tag{5.1159}$$

where

$$B = 8\pi^2 U \tag{5.1160}$$

from Eq. (5.1062).

For $a = b$, however,

$$\mathbf{u}_a - \mathbf{u}_a = 0, \tag{5.1161}$$

so the corresponding self-scattering term is not attenuated.

Separating the diagonal and off-diagonal terms in Eq. (5.1153),

$$\begin{aligned} \left\langle |F_e^{(\xi)}|^2 \right\rangle = & \sum_a |f_{e,a}(R)|^2 \\ & + e^{-BR^2/2} \sum_{a \neq b} f_{e,a}(R) f_{e,b}^*(R) e^{-j2\pi \mathbf{R}\cdot(\mathbf{r}_a - \mathbf{r}_b)}. \end{aligned} \tag{5.1162}$$

But the squared modulus of the reference protein structure factor is

$$\begin{aligned} |F_e(\mathbf{R})|^2 = & \sum_a |f_{e,a}(R)|^2 \\ & + \sum_{a \neq b} f_{e,a}(R) f_{e,b}^*(R) e^{-j2\pi \mathbf{R}\cdot(\mathbf{r}_a - \mathbf{r}_b)}. \end{aligned} \tag{5.1163}$$

Therefore,

$$\sum_{a \neq b} \cdots = |F_e(\mathbf{R})|^2 - \sum_a |f_{e,a}(R)|^2. \tag{5.1164}$$

Substitution into Eq. (5.1162) gives

$$\left\langle |F_e^{(\xi)}|^2 \right\rangle = \sum_a |f_{e,a}|^2 + e^{-BR^2/2} \left[|F_e|^2 - \sum_a |f_{e,a}|^2 \right] \tag{5.1165}$$

$$= e^{-BR^2/2} |F_e(\mathbf{R})|^2 + \left[1 - e^{-BR^2/2} \right] \sum_a |f_{e,a}(R)|^2. \tag{5.1166}$$

Thus

$$\boxed{\left\langle |F_e^{(\xi)}(\mathbf{R})|^2 \right\rangle = \underbrace{e^{-BR^2/2} |F_e(\mathbf{R})|^2}_{\text{coherent structural contribution}} + \underbrace{\left[1 - e^{-BR^2/2} \right] \sum_a |f_{e,a}(R)|^2}_{\text{configurationally diffuse contribution}} .} \tag{5.1167}$$

Atomic disorder does not attenuate all parts of the protein scattering in the same way. The self-scattering terms $|f_{e,a}|^2$ are unaffected by positional disorder, because translating an isolated atom changes only its phase. What is attenuated are the cross terms between different atoms, which encode the coherent interference responsible for molecular structure.

In the simplified independent-displacement model, the attenuated interference is redistributed into a configurationally diffuse component. In a realistic protein the atomic motions are correlated, and Eq. (5.1153), rather than Eq. (5.1167), is the more fundamental description.

5.11.4 Static conformational heterogeneity in SPA

The atomic-displacement treatment developed above is particularly useful for understanding thermal disorder, but biological single-particle cryo-EM introduces a different type of variability. Different vitrified particles in the same dataset need not have identical structures. A flexible loop may adopt several positions, a domain may undergo a hinge motion, a ligand may be present or absent, or an entire subunit may occupy different conformational states.

The adjective *static* requires some clarification. It does not mean that the macromolecule was immobile before vitrification. Rather, vitrification traps different particles at different points of their conformational ensemble. During the acquisition of one particle image, the particle can therefore be

regarded, to first approximation, as having one particular frozen conformation. The heterogeneity appears when different particles in the dataset are compared.

Let the possible conformations be indexed by

$$c = 1, \ldots, K, \tag{5.1168}$$

with probabilities

$$p_c, \qquad \sum_{c=1}^{K} p_c = 1. \tag{5.1169}$$

The electrostatic potential of conformation c is denoted by

$$V_c(\mathbf{r}). \tag{5.1170}$$

The ensemble-averaged potential is then

$$\boxed{\overline{V}(\mathbf{r}) = \sum_{c=1}^{K} p_c V_c(\mathbf{r}).} \tag{5.1171}$$

In the weak linear image-formation regime, averaging correctly aligned observations of many particles therefore tends to reinforce the part of the structure that is reproducible throughout the ensemble and to blur features whose positions vary between conformations. In this sense, a reconstruction obtained under the assumption of a single homogeneous structure tends toward an ensemble-average structure rather than toward any particular member of the ensemble.

The ensemble average in reciprocal space. For conformation c, let

$$F_e^{(c)}(\mathbf{R}) \tag{5.1172}$$

denote the electron structure factor introduced previously in Eq. (5.307). Explicitly,

$$F_e^{(c)}(\mathbf{R}) = \sum_{a=1}^{N_{\mathrm{at}}} f_{e,a}(R) e^{-j2\pi \mathbf{R}\cdot\mathbf{r}_a^{(c)}}. \tag{5.1173}$$

The structure factor corresponding to the mean potential of Eq. (5.1171) is therefore

$$\boxed{\overline{F}_e(\mathbf{R}) = \sum_{c=1}^{K} p_c F_e^{(c)}(\mathbf{R}).} \tag{5.1174}$$

Equation (5.1174) is the reciprocal-space counterpart of Eq. (5.1171). It describes the part of the molecular scattering amplitude that is reproducible across the particle population.

Mean structure and structural variance of a conformational ensemble.

Although Eq. (5.1174) gives the mean complex structure factor, it does not describe the full variability of the population.

Define the deviation of conformation c from the ensemble mean as

$$\delta F_e^{(c)}(\mathbf{R}) = F_e^{(c)}(\mathbf{R}) - \overline{F}_e(\mathbf{R}). \tag{5.1175}$$

By definition,

$$\sum_c p_c \delta F_e^{(c)}(\mathbf{R}) = 0. \tag{5.1176}$$

The conformational variance at spatial frequency $\mathbf{R}$ is

$$S_{\text{het}}(\mathbf{R}) \equiv \sum_c p_c \left|\delta F_e^{(c)}(\mathbf{R})\right|^2. \tag{5.1177}$$

Substituting

$$F_e^{(c)} = \overline{F}_e + \delta F_e^{(c)} \tag{5.1178}$$

into the ensemble average of the squared modulus gives

$$\sum_c p_c \left|F_e^{(c)}\right|^2 = \sum_c p_c \left|\overline{F}_e + \delta F_e^{(c)}\right|^2 \tag{5.1179}$$

$$= \left|\overline{F}_e\right|^2 + \sum_c p_c \left|\delta F_e^{(c)}\right|^2, \tag{5.1180}$$

because the cross terms vanish after averaging.

Consequently,

$$\boxed{\sum_c p_c \left|F_e^{(c)}(\mathbf{R})\right|^2 = \left|\overline{F}_e(\mathbf{R})\right|^2 + S_{\text{het}}(\mathbf{R}).} \tag{5.1181}$$

Equivalently,

$$\boxed{S_{\text{het}}(\mathbf{R}) = \sum_c p_c \left|F_e^{(c)}(\mathbf{R})\right|^2 - \left|\overline{F}_e(\mathbf{R})\right|^2.} \tag{5.1182}$$

Equation (5.1181) resembles the coherent–configurational decomposition obtained for frozen atomic configurations in Eq. (5.1111), but its interpre-

tation in SPA is different. Different conformations are normally observed in different particle images rather than being simultaneously averaged within one detector exposure. Thus S_{het} is most naturally interpreted as *particle-to-particle structural variability*. If this variability is ignored during reconstruction, it behaves as a source of non-reproducible structural signal and reduces the coherent information that can be reinforced by averaging.

Correlated motion matters. The derivation of Eq. (5.1153) showed that, when all atoms move, the interference between atoms a and b depends on the distribution of their *relative displacement*

$$\mathbf{u}_a - \mathbf{u}_b, \tag{5.1183}$$

rather than on their individual displacements separately.

This observation is especially important for biological conformational heterogeneity. Atomic motions inside a protein are usually strongly correlated. A secondary-structure element may move almost rigidly, a domain may rotate about a hinge, and an entire subunit may translate relative to the rest of a complex. Describing such motions by independent atomic B-factors discards these correlations.

For two atoms that move together,

$$\mathbf{u}_a^{(c)} = \mathbf{u}_b^{(c)}, \tag{5.1184}$$

their relative displacement is unchanged:

$$\mathbf{u}_a^{(c)} - \mathbf{u}_b^{(c)} = 0. \tag{5.1185}$$

Their mutual interference is therefore preserved even if the pair moves through a large absolute distance.

This gives an important physical principle:

$$\boxed{\begin{array}{c}\text{conformational disorder attenuates structural coherence}\\ \text{according to fluctuations of relative atomic positions,}\\ \text{not simply according to absolute atomic displacements.}\end{array}} \tag{5.1186}$$

A rigidly moving domain preserves its internal structure.

Divide the protein into a fixed part A and a mobile domain B. Let

the reference structure factor be

$$F_e(\mathbf{R}) = F_A(\mathbf{R}) + F_B(\mathbf{R}). \tag{5.1187}$$

Suppose that in conformation c the whole domain B undergoes a rigid translation $\mathbf{d}_c$, while its internal atomic coordinates remain unchanged. A translation multiplies its structure factor by the Fourier phase factor

$$e^{-j2\pi\mathbf{R}\cdot\mathbf{d}_c}, \tag{5.1188}$$

so that

$$\boxed{F_e^{(c)}(\mathbf{R}) = F_A(\mathbf{R}) + F_B(\mathbf{R})e^{-j2\pi\mathbf{R}\cdot\mathbf{d}_c}.} \tag{5.1189}$$

For one conformation,

$$\begin{aligned}\left|F_e^{(c)}\right|^2 &= |F_A|^2 + |F_B|^2 \\ &\quad + F_A F_B^* e^{+j2\pi\mathbf{R}\cdot\mathbf{d}_c} + F_A^* F_B e^{-j2\pi\mathbf{R}\cdot\mathbf{d}_c}.\end{aligned} \tag{5.1190}$$

The term

$$|F_B(\mathbf{R})|^2 \tag{5.1191}$$

is independent of $\mathbf{d}_c$. Thus the interference among atoms *inside* the moving domain is completely preserved by a rigid translation.

Only the cross terms between the fixed part and the moving domain depend on $\mathbf{d}_c$. Averaging over conformations gives

$$\begin{aligned}\left\langle |F_e^{(c)}|^2 \right\rangle_c &= |F_A|^2 + |F_B|^2 \\ &\quad + F_A F_B^* \left\langle e^{+j2\pi\mathbf{R}\cdot\mathbf{d}_c} \right\rangle_c \\ &\quad + F_A^* F_B \left\langle e^{-j2\pi\mathbf{R}\cdot\mathbf{d}_c} \right\rangle_c .\end{aligned} \tag{5.1192}$$

Therefore rigid-body conformational variability preserves the internal structure of each rigid domain while progressively reducing the coherence between domains whose relative positions fluctuate.

This result is more informative for biological macromolecules than a model in which every atom fluctuates independently. A flexible protein may contain regions with excellent internal high-resolution order even though their position relative to the remainder of the molecule is poorly defined.

When can conformational heterogeneity look like a B-factor? A Gaussian B-factor can represent conformational heterogeneity only under restrictive conditions. If the position of an atom, or the rigid translation of a small domain, follows an approximately Gaussian distribution, then its ensemble-averaged Fourier contribution contains the characteristic function derived in Eq. (5.1059). For isotropic Gaussian displacements this gives

$$e^{-BR^2/4}. \tag{5.1193}$$

Mathematically, this Gaussian attenuation does not reveal whether the underlying positional variance arose from thermal vibration or from static particle-to-particle heterogeneity. Fourier attenuation alone cannot determine its physical origin.

However, large-amplitude, correlated, or multimodal conformational changes are not generally Gaussian and therefore cannot be described faithfully by one B-factor.

A two-state domain motion is not, in general, a B-factor.

Consider again a protein composed of a fixed part A and a mobile domain B. Suppose that the mobile domain occupies two equally probable positions,

$$\mathbf{d}_{\pm} = \pm\frac{\Delta\mathbf{d}}{2}. \tag{5.1194}$$

From Eq. (5.1189), the two structure factors are

$$F_e^{(\pm)}(\mathbf{R}) = F_A(\mathbf{R}) + F_B(\mathbf{R})e^{\mp j\pi\mathbf{R}\cdot\Delta\mathbf{d}}. \tag{5.1195}$$

The ensemble-averaged structure factor is

$$\overline{F}_e(\mathbf{R}) = \frac{1}{2}\left[F_e^{(+)}(\mathbf{R}) + F_e^{(-)}(\mathbf{R})\right] \tag{5.1196}$$

$$= F_A(\mathbf{R}) + \frac{F_B(\mathbf{R})}{2}\left[e^{-j\pi\mathbf{R}\cdot\Delta\mathbf{d}} + e^{+j\pi\mathbf{R}\cdot\Delta\mathbf{d}}\right] \tag{5.1197}$$

$$= \boxed{F_A(\mathbf{R}) + F_B(\mathbf{R})\cos\left(\pi\mathbf{R}\cdot\Delta\mathbf{d}\right).} \tag{5.1198}$$

The contribution of the mobile domain is therefore modulated by a cosine, not by a Gaussian envelope. The modulation is also direction-dependent through $\mathbf{R}\cdot\Delta\mathbf{d}$.

Only when

$$|\mathbf{R} \cdot \Delta\mathbf{d}| \ll 1 \tag{5.1199}$$

may we use

$$\cos x \simeq 1 - \frac{x^2}{2} \simeq e^{-x^2/2}, \tag{5.1200}$$

so that the two-state motion appears approximately Gaussian over a restricted low-frequency range.

The two-state example also shows why replacing conformational heterogeneity by independent atomic B-factors can be misleading. Every atom in the mobile domain may undergo a large displacement, yet the internal structure factor $F_B(\mathbf{R})$ remains fully coherent. What is uncertain is the *relative placement* of that domain with respect to the rest of the molecule.

Interpretation of large B-factors in SPA. A large atomic or local B-factor in an SPA model should therefore not automatically be interpreted as large-amplitude thermal vibration. It may reflect genuine atomic fluctuations, unresolved conformational heterogeneity, partial occupancy, imperfect alignment, or uncertainty in the atomic model (Afonine et al., 2018; Bharadwaj and Jakobi, 2022).

Moreover, a B-factor contains only a second-moment description of positional uncertainty. It does not retain the correlations between atomic motions that distinguish, for example, a rigidly moving domain from a collection of independently fluctuating atoms.

Classification and continuous-heterogeneity methods attempt to recover this missing information by separating, or explicitly modelling, the different conformational states before averaging. When successful, features that would otherwise appear as blurred density or large effective B-factors can become structurally resolved.

Thus, for SPA,

$$\boxed{\begin{array}{c}\text{a } B\text{-factor describes positional spread,}\\ \text{whereas conformational heterogeneity describes an ensemble}\\ \text{of correlated three-dimensional structures.}\end{array}} \tag{5.1201}$$

The former may approximate the latter in special limits, but the two concepts are not equivalent.

5.11.5 Radiation-induced motion and radiation damage

Thermal or conformational disorder should also be distinguished from changes induced by the imaging electron beam.

Radiation-induced motion is not simply another member of a pre-existing equilibrium ensemble. The specimen evolves as electron fluence accumulates. Atomic coordinates may move, entire particles may translate or rotate, the surrounding vitreous ice may deform, and radiation chemistry may alter the molecular structure itself.

Direct electron detectors make this distinction experimentally visible because an exposure is recorded as a sequence of movie frames. Early and late frames therefore do not observe exactly the same specimen. High spatial frequencies are particularly sensitive to this evolution, which motivates frame alignment and resolution-dependent exposure weighting (Grant and Grigorieff, 2015).

Some components of residual motion again produce a Gaussian attenuation and can therefore masquerade as a B-factor.

A Gaussian distribution of residual translations produces a B-factor-like envelope.

Suppose that after motion correction a particle retains a random residual in-plane displacement

$$\delta\boldsymbol{\rho} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\rho\right). \tag{5.1202}$$

A translation multiplies a Fourier coefficient by

$$e^{-j2\pi\mathbf{R}_\perp\cdot\delta\boldsymbol{\rho}}. \tag{5.1203}$$

Averaging over residual translations gives

$$\boxed{\left\langle e^{-j2\pi\mathbf{R}_\perp\cdot\delta\boldsymbol{\rho}}\right\rangle = e^{-2\pi^2\mathbf{R}_\perp^T\boldsymbol{\Sigma}_\rho\mathbf{R}_\perp}.} \tag{5.1204}$$

For isotropic residual motion with per-coordinate variance σ_ρ^2,

$$\boldsymbol{\Sigma}_\rho = \sigma_\rho^2\mathbf{I}, \tag{5.1205}$$

so

$$E_{\text{motion}}(R_\perp) = e^{-2\pi^2\sigma_\rho^2 R_\perp^2}. \tag{5.1206}$$

Defining

$$\boxed{B_{\text{motion}} = 8\pi^2\sigma_\rho^2,} \tag{5.1207}$$

we obtain

$$\boxed{E_{\text{motion}}(R_\perp) = e^{-B_{\text{motion}}R_\perp^2/4}.} \tag{5.1208}$$

Thus the same mathematical envelope that describes Gaussian atomic disorder also describes Gaussian residual image motion. Their physical origins are entirely different.

Moreover, radiation damage cannot in general be represented only as motion. Bond breakage, decarboxylation, loss of radiation-sensitive side chains, and other chemical changes alter the specimen potential itself. All these effects will be further explored in Sec. 5.13.

5.11.6 What does a B-factor mean in Single Particle Analysis?

The same symbol B is used for several quantities that should not be identified physically.

At the atomic-model level, an isotropic atomic displacement parameter B_a modifies the contribution of atom a according to Eq. (5.1076),

$$f_{e,a}(R) \longrightarrow f_{e,a}(R)e^{-B_a R^2/4}. \tag{5.1209}$$

This factor represents positional uncertainty of that particular atom within the adopted model.

It should not be confused with the intrinsic frequency dependence of $f_{e,a}(R)$ itself. The atomic scattering factors derived in Sec. 5.3.9 already account for the finite spatial extent and screening of the atomic electrostatic potential. The B_a factor represents an *additional* positional distribution of that atom.

At the level of a reconstructed map, one often introduces an overall Gaussian attenuation

$$\boxed{F_{\text{obs}}(\mathbf{R}) \simeq F_{\text{ideal}}(\mathbf{R}) \exp\left(-\frac{B_{\text{overall}}R^2}{4}\right).} \tag{5.1210}$$

This B_{overall} is an effective parameter. It can contain contributions from structural disorder, imperfect particle alignment, residual motion, radiation

damage, microscope coherence, reconstruction errors, and other resolution-dependent losses (Rosenthal and Henderson, 2003; Sorzano et al., 2017; Bharadwaj and Jakobi, 2022).

The distinction between amplitude and power should also be kept explicit. Equation (5.1210) implies

$$\boxed{|F_{\text{obs}}(\mathbf{R})|^2 \simeq |F_{\text{ideal}}(\mathbf{R})|^2 e^{-B_{\text{overall}}R^2/2}.} \tag{5.1211}$$

Therefore a Guinier plot based on logarithmic amplitudes and one based on logarithmic powers differ by a factor of two in their slopes.

The microscope envelopes derived in Sec. 4.13 provide another example. Their total effect was represented by $E_{\text{tot}}(\mathbf{R})$ in Eq. (4.2115). Over a restricted frequency range, one may sometimes fit

$$E_{\text{tot}}(\mathbf{R}) \simeq e^{-B_{\text{opt}}R^2/4}, \tag{5.1212}$$

but this does not make B_{opt} an atomic displacement parameter. The underlying physical envelope may be non-Gaussian, anisotropic, or have a different dependence on spatial frequency (Sorzano et al., 2017).

Why different Gaussian blurring mechanisms are inseparable from one fitted B-factor.

Suppose several independent mechanisms attenuate Fourier amplitudes according to

$$E_k(R) = \exp\left(-\frac{B_k R^2}{4}\right). \tag{5.1213}$$

Their combined attenuation is

$$E_{\text{tot}}(R) = \prod_k E_k(R) \tag{5.1214}$$

$$= \prod_k \exp\left(-\frac{B_k R^2}{4}\right) \tag{5.1215}$$

$$= \exp\left[-\frac{R^2}{4}\sum_k B_k\right]. \tag{5.1216}$$

Hence

$$\boxed{B_{\text{eff}} = \sum_k B_k.} \tag{5.1217}$$

> A measurement of B_{eff} alone therefore cannot determine how much arose from atomic displacement, conformational heterogeneity, motion, alignment error, microscope coherence, or another Gaussian blur.

This is one reason that the phrase "temperature factor" can be misleading in modern cryo-EM. An atomic B-factor may reflect genuine thermal motion, but it may also absorb static positional disorder or unresolved conformational variability. A global map B-factor has an even broader meaning: it is usually an effective description of the combined loss of high-resolution signal produced by several specimen, microscope, acquisition, and processing effects.

For cryogenic biological specimens, it is therefore useful to keep the physical mechanisms in Table 5.4 conceptually distinct.

The mathematical similarity of some of these Fourier-space attenuations should therefore not be mistaken for physical equivalence. In particular, observing an approximately Gaussian decay $e^{-BR^2/4}$ does not by itself identify the mechanism responsible for that decay.

5.12 Inelastic scattering: transitions of the specimen

The quantum description of the specimen was already introduced in Sec. 5.2. There we considered an initial internal specimen state $|\Xi_0\rangle_{\text{s}}$ and showed that, if the specimen is finally found in the state $|\Xi_m\rangle_{\text{s}}$, the corresponding interaction channel acting on the imaging electron is

$$\boxed{\hat{U}_{m0} = {}_{\text{s}}\langle\Xi_m|\hat{H}_{\text{int}}|\Xi_0\rangle_{\text{s}},} \tag{5.1218}$$

as given in Eq. (5.12).

We also distinguished the two possibilities

$$m = 0 \qquad \text{and} \qquad m \neq 0, \tag{5.1219}$$

corresponding respectively to elastic and inelastic channels in Eq. (5.13).

Until now, however, the rest of the chapter has concentrated almost entirely on the elastic channel. We now return to the off-diagonal channels and develop their scattering physics.

Table 5.4: Different sources of structural variability and high-resolution signal loss in single-particle cryo-EM. Although several of these mechanisms can produce an approximately Gaussian attenuation in Fourier space, their physical origins are different.

Mechanism	**Physical meaning**
Thermal displacement	Small atomic fluctuations about a local equilibrium structure. If their positional distribution is approximately Gaussian, the coherently averaged scattering amplitude acquires a Debye–Waller factor of the form $e^{-BR^2/4}$.
Static conformational heterogeneity	Different vitrified particles occupy different conformations. For sufficiently small, approximately Gaussian positional variations, the resulting attenuation may resemble a B-factor. Large, correlated, or multimodal conformational changes cannot in general be represented by a single B-factor.
Partial occupancy	An atomic site, ligand, subunit, or conformational state is present only in a fraction of the particle population. Occupancy primarily scales the coherently reproducible contribution of that structural feature, rather than producing by itself a Gaussian decay with spatial frequency.
Radiation-induced motion	The specimen moves as electron fluence accumulates during acquisition. Residual motion reduces high-resolution coherence and may produce an approximately B-factor-like envelope when the remaining displacement distribution is approximately Gaussian.
Radiation damage	The electron beam changes the molecular structure itself through radiation chemistry. This effect depends on accumulated fluence and spatial frequency and cannot, in general, be reduced to a simple positional displacement.
Global map B-factor	An effective description of the observed decay of Fourier amplitudes with spatial frequency. It may combine structural heterogeneity, particle motion, imperfect alignment, microscope envelopes, radiation damage, and reconstruction or processing effects, and therefore should not be interpreted as a direct measurement of atomic thermal motion.

5.12.1 From the static elastic potential to transition potentials

In Sec. 5.2.2, the channel decomposition was applied to the specimen charge-density operator. For the elastic channel,

$$m = 0, \tag{5.1220}$$

the charge-density matrix element reduces to the ordinary specimen charge density,

$$\rho(\mathbf{r}) = \rho_{00}(\mathbf{r}), \tag{5.1221}$$

and the interaction acting on the imaging electron becomes

$$\boxed{U_{\mathrm{el}}(\mathbf{r}) = -eV(\mathbf{r}),} \tag{5.1222}$$

as derived in Eq. (5.26).

Thus the static electrostatic potential used throughout the elastic scattering theory is the diagonal channel of the full quantum interaction.

For an inelastic channel,

$$m \neq 0, \tag{5.1223}$$

the corresponding specimen quantity is instead the off-diagonal transition charge density

$$\boxed{\rho_{m0}(\mathbf{r}) = {}_{\mathrm{s}}\langle \Xi_m | \hat{\rho}_{\mathrm{s}}(\mathbf{r}) | \Xi_0 \rangle_{\mathrm{s}},} \tag{5.1224}$$

introduced in Eq. (5.27).

Its Coulomb potential is

$$\boxed{V_{m0}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho_{m0}(\mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|}\, d^3r',} \tag{5.1225}$$

as given in Eq. (5.28).

The corresponding interaction energy of the imaging electron is

$$U_{m0}(\mathbf{r}) = -eV_{m0}(\mathbf{r}), \tag{5.1226}$$

which is the off-diagonal counterpart of the static elastic interaction.

The important distinction is therefore

$$\boxed{\begin{array}{lcl} \rho_{00} = {}_{\mathrm{s}}\langle \Xi_0 | \hat{\rho}_{\mathrm{s}} | \Xi_0 \rangle_{\mathrm{s}} & \longrightarrow & \text{static elastic potential,} \\ \rho_{m0} = {}_{\mathrm{s}}\langle \Xi_m | \hat{\rho}_{\mathrm{s}} | \Xi_0 \rangle_{\mathrm{s}} & \longrightarrow & \text{inelastic transition potential.} \end{array}} \tag{5.1227}$$

This is the same distinction summarized previously in Eq. (5.29).

The off-diagonal quantity $V_{m0}(\mathbf{r})$ should not be interpreted as an ordinary electrostatic potential generated by a static charge distribution. Rather, it describes the spatial part of the Coulomb interaction that couples two different internal quantum states of the specimen.

The quantum-mechanical meaning of such matrix elements was developed in Sec. A.12 of the appendix. We shall use that formalism here without repeating it.

5.12.2 The combined electron–specimen transition

To describe a complete scattering event, both the state of the imaging electron and the internal state of the specimen must be specified.

The tensor-product description of the combined system was introduced in Sec. A.9.2. In particular, Eq. (A.390) showed that a state of the complete system is written as the product of an electron state and a specimen state.

For an incident electron with wavevector $\mathbf{k}_\mathrm{i}$ and a specimen initially in the internal state $|\Xi_n\rangle_\mathrm{s}$, we therefore write

$$\boxed{|\mathbf{k}_\mathrm{i}, \Xi_n\rangle \equiv |\mathbf{k}_\mathrm{i}\rangle_\mathrm{e} \otimes |\Xi_n\rangle_\mathrm{s}.} \tag{5.1228}$$

After scattering, the imaging electron may emerge with wavevector $\mathbf{k}_\mathrm{f}$ while the specimen occupies the internal state $|\Xi_m\rangle_\mathrm{s}$:

$$\boxed{|\mathbf{k}_\mathrm{i}, \Xi_n\rangle \longrightarrow |\mathbf{k}_\mathrm{f}, \Xi_m\rangle.} \tag{5.1229}$$

The distinction between the two cases was discussed in detail in Sec. A.9.2:

$$\boxed{\begin{array}{lcl} m = n & : & \text{same internal specimen state: elastic channel,} \\ m \neq n & : & \text{different internal specimen state: inelastic channel.} \end{array}} \tag{5.1230}$$

Notice that elastic scattering does not imply

$$\mathbf{k}_\mathrm{f} = \mathbf{k}_\mathrm{i}. \tag{5.1231}$$

The direction of the electron may change substantially. In the negligible-recoil approximation used in the elastic-scattering section, only the magnitudes are equal:

$$\boxed{k_\mathrm{f} = k_\mathrm{i},} \tag{5.1232}$$

as derived in Eq. (5.337).

5.12.3 Energy and momentum transfer

An inelastic scattering event changes both the state of the imaging electron and the internal state of the specimen. To keep these two effects distinct, it is useful to separate the motion of the specimen as a whole from its internal degrees of freedom.

This separation was introduced in Sec. A.9.2. There, the specimen Hamiltonian was written as

$$\hat{H}_{\mathrm{s}} = \frac{\hat{\mathbf{P}}^{\,2}}{2M} + \hat{H}_{\mathrm{internal}}, \tag{5.1233}$$

in Eq. (A.410), and a specimen state was correspondingly labelled as

$$|\mathbf{P}, \Xi_n\rangle_{\mathrm{s}}, \tag{5.1234}$$

in Eq. (A.411). Here $\mathbf{P}$ describes the centre-of-mass momentum of the complete specimen, while $|\Xi_n\rangle_{\mathrm{s}}$ specifies its internal state.

The inelastic case was then considered explicitly in Sec. A.9.2. In the notation of that section, the specimen changes from one internal state to another while it may simultaneously recoil as a whole.

The appendix used $|\Xi_0\rangle_{\mathrm{s}}$ for the initial internal state. Here we use the more general label $|\Xi_n\rangle_{\mathrm{s}}$, because there is no need to assume that the initial state is the ground state. With the replacement $0 \to n$, Eq. (A.426) becomes

$$\boxed{E_{\mathrm{e},i} + \frac{P_{\mathrm{i}}^2}{2M} + E_n = E_{\mathrm{e},f} + \frac{P_{\mathrm{f}}^2}{2M} + E_m,} \tag{5.1235}$$

where

$$P_{\mathrm{i}} = \|\mathbf{P}_{\mathrm{i}}\|, \qquad P_{\mathrm{f}} = \|\mathbf{P}_{\mathrm{f}}\|, \tag{5.1236}$$

and E_n and E_m are the initial and final internal energies of the specimen.

Rearranging Eq. (5.1235) gives

$$\boxed{E_{\mathrm{e},i} - E_{\mathrm{e},f} = (E_m - E_n) + \frac{P_{\mathrm{f}}^2 - P_{\mathrm{i}}^2}{2M}.} \tag{5.1237}$$

This is the generalization of Eq. (A.427) from the appendix to an arbitrary initial internal state n.

Equation (5.1237) separates the energy lost by the imaging electron into two physically different contributions:

$$\boxed{\begin{array}{c}\text{electron energy loss}\\ = \\ \text{change of internal energy of the specimen}\\ + \\ \text{centre-of-mass recoil energy.}\end{array}} \tag{5.1238}$$

For a massive biological specimen, the centre-of-mass recoil term is normally very small compared with the energies of the electronic or vibrational excitations of interest. Neglecting it gives the same approximation derived in Eq. (A.429),

$$E_{\mathrm{e},i} - E_{\mathrm{e},f} \simeq E_m - E_n. \tag{5.1239}$$

We therefore define the internal excitation energy associated with the transition $n \to m$ as

$$\boxed{\epsilon_{mn} \equiv E_m - E_n.} \tag{5.1240}$$

Under the negligible-recoil approximation,

$$\boxed{E_{\mathrm{e},i} - E_{\mathrm{e},f} \simeq \epsilon_{mn}.} \tag{5.1241}$$

For the ordinary energy-loss events considered below,

$$\epsilon_{mn} > 0, \tag{5.1242}$$

so that

$$E_{\mathrm{e},f} < E_{\mathrm{e},i}. \tag{5.1243}$$

The imaging electron therefore emerges with a smaller momentum magnitude and a longer wavelength. In principle, $\epsilon_{mn} < 0$ is also possible if the specimen begins in an excited state and transfers energy to the imaging electron, but this is not the dominant situation considered here.

How does the energy loss change the electron wavevector? The relativistic relation between electron energy and momentum was derived in Eq. (2.68),

$$E^2 = p^2c^2 + m_0^2c^4. \tag{5.1244}$$

In a field-free region, the mechanical momentum and wavevector are related by

$$\mathbf{p} = \hbar\mathbf{k}, \tag{5.1245}$$

as established in Eq. (4.671). Hence the total relativistic energy of the imaging electron may be written as

$$\boxed{E_{\mathrm{e}}(k) = \sqrt{\hbar^2 c^2 k^2 + m_0^2 c^4}.} \tag{5.1246}$$

Although Eq. (5.1246) contains the rest energy, this causes no difficulty: the rest energy is the same before and after scattering and therefore cancels from the energy difference in Eq. (5.1241).

For an energy-loss transition,

$$E_{\mathrm{e}}(k_{\mathrm{f}}) = E_{\mathrm{e}}(k_{\mathrm{i}}) - \epsilon_{mn}. \tag{5.1247}$$

Substituting Eq. (5.1246) and solving for k_{f} gives

$$\boxed{k_{\mathrm{f}} = \frac{1}{\hbar c}\sqrt{\left[E_{\mathrm{e}}(k_{\mathrm{i}}) - \epsilon_{mn}\right]^2 - m_0^2 c^4}.} \tag{5.1248}$$

Thus, unlike elastic scattering, an inelastic event changes the magnitude of the electron wavevector. Setting

$$\epsilon_{mn} = 0 \tag{5.1249}$$

in Eq. (5.1248) immediately gives

$$k_{\mathrm{f}} = k_{\mathrm{i}}, \tag{5.1250}$$

recovering the elastic result of Eq. (5.337).

Momentum and spatial-frequency transfer. Energy transfer and momentum transfer are related by the scattering kinematics, but they are not the same quantity. The energy transfer specifies which internal excitation is produced, whereas the momentum transfer specifies how strongly different spatial variations of that excitation are probed.

No new reciprocal-space notation is required for the inelastic case. In the elastic-scattering treatment we introduced

$$\boxed{\mathbf{R} = \frac{\mathbf{k}_{\mathrm{f}} - \mathbf{k}_{\mathrm{i}}}{2\pi},} \tag{5.1251}$$

in Eq. (5.341). We retain exactly the same definition here.

The relation

$$\mathbf{k} = 2\pi\mathbf{R} \tag{5.1252}$$

was established in Eq. (3.6), while Eq. (4.671) gives

$$\mathbf{p} = \hbar \mathbf{k}. \tag{5.1253}$$

Therefore the change in momentum of the imaging electron is

$$\begin{aligned} \Delta \mathbf{p}_{\mathrm{e}} &\equiv \mathbf{p}_{\mathrm{f}} - \mathbf{p}_{\mathrm{i}} && (5.1254)\\ &= \hbar\,(\mathbf{k}_{\mathrm{f}} - \mathbf{k}_{\mathrm{i}}) && (5.1255)\\ &= h\mathbf{R}. && (5.1256) \end{aligned}$$

Hence

$$\boxed{\Delta \mathbf{p}_{\mathrm{e}} = h\mathbf{R}.} \tag{5.1257}$$

This is the same kinematic relation obtained for elastic scattering in Eq. (5.345); what changes in the inelastic case is that k_{f} need not equal k_{i}.

Momentum conservation was written explicitly for the complete electron–specimen system in Eq. (A.414). The same conservation law holds when the internal state of the specimen changes:

$$\mathbf{p}_{\mathrm{i}} + \mathbf{P}_{\mathrm{i}} = \mathbf{p}_{\mathrm{f}} + \mathbf{P}_{\mathrm{f}}. \tag{5.1258}$$

Consequently,

$$\begin{aligned} \mathbf{P}_{\mathrm{f}} - \mathbf{P}_{\mathrm{i}} &= \mathbf{p}_{\mathrm{i}} - \mathbf{p}_{\mathrm{f}} && (5.1259)\\ &= -h\mathbf{R}. && (5.1260) \end{aligned}$$

Thus the momentum transferred to the specimen is

$$\boxed{\Delta \mathbf{p}_{\mathrm{s}} = -h\mathbf{R}.} \tag{5.1261}$$

The minus sign follows directly from the convention adopted in Eq. (5.341): $\mathbf{R}$ describes the momentum change of the imaging electron, $\mathbf{p}_{\mathrm{f}} - \mathbf{p}_{\mathrm{i}}$, whereas the specimen receives the opposite momentum.

Two quantities characterize the transition. An inelastic scattering event may therefore be characterized by the pair

$$\boxed{\begin{aligned} \text{energy transferred to the specimen} &= \epsilon_{mn} = E_m - E_n,\\ \text{momentum transferred to the specimen} &= -h\mathbf{R}. \end{aligned}} \tag{5.1262}$$

The first quantity tells us *which energetic transition* has occurred. The second tells us *with which spatial-frequency transfer* the imaging electron has coupled to that transition. Both will enter the microscopic scattering probability.

5.12.4 The first-Born transition matrix element

We are therefore ready to ask the central dynamical question: what is the amplitude for the combined transition

$$|\mathbf{k}_{\mathrm{i}}, \Xi_n\rangle \longrightarrow |\mathbf{k}_{\mathrm{f}}, \Xi_m\rangle? \tag{5.1263}$$

The answer is the inelastic analogue of the first Born approximation developed for elastic scattering in Sec. 5.4.4. The essential difference is that the specimen is now allowed to change its internal quantum state.

The transition matrix element. The meaning of an expression of the form

$$\langle \psi_f | \widehat{A} | \psi_i \rangle \tag{5.1264}$$

was introduced in Sec. A.4.7 and summarized in Eq. (A.137): the operator acts on the initial state, and the resulting state is projected onto the chosen final state.

Applying this construction to the electron–specimen interaction, we define

$$\boxed{M_{mn}(\mathbf{k}_{\mathrm{f}}, \mathbf{k}_{\mathrm{i}}) \equiv \left\langle \mathbf{k}_{\mathrm{f}}, \Xi_m \middle| \widehat{H}_{\mathrm{int}} \middle| \mathbf{k}_{\mathrm{i}}, \Xi_n \right\rangle .} \tag{5.1265}$$

The quantity $M_{mn}(\mathbf{k}_{\mathrm{f}}, \mathbf{k}_{\mathrm{i}})$ is called the *transition matrix element*. It measures how strongly the interaction Hamiltonian couples the particular initial channel $(\mathbf{k}_{\mathrm{i}}, n)$ to the particular final channel $(\mathbf{k}_{\mathrm{f}}, m)$.

For the moment we suppress the centre-of-mass state of the specimen. Its recoil was accounted for kinematically in Sec. 5.12.3; here we focus on the internal transition

$$|\Xi_n\rangle \longrightarrow |\Xi_m\rangle. \tag{5.1266}$$

Equation (5.20) introduced the interaction $U_{m0}(\mathbf{r})$ associated with a transition from the reference state $|\Xi_0\rangle_{\mathrm{s}}$ to $|\Xi_m\rangle_{\mathrm{s}}$. We now remove the restriction that the initial state be $|\Xi_0\rangle_{\mathrm{s}}$ and define, for arbitrary initial and final internal states,

$$\boxed{U_{mn}(\mathbf{r}) \equiv {}_{\mathrm{s}}\langle \Xi_m | \widehat{H}_{\mathrm{int}} | \Xi_n \rangle_{\mathrm{s}}.} \tag{5.1267}$$

The subscript s emphasizes that this matrix element is taken only over the internal degrees of freedom of the specimen. The coordinate $\mathbf{r}$ of the imaging electron is left untouched, so $U_{mn}(\mathbf{r})$ remains a function acting on the imaging electron.

For $m = n$, Eq. (5.1267) is a diagonal expectation value and gives the static interaction associated with that specimen state. For $m \neq n$, it is an off-diagonal transition interaction.

A note on plane-wave normalization. The appendix represents exact momentum eigenstates by delta-normalized plane waves, as in Eq. (A.155). The elastic scattering amplitude in Sec. 5.4.4, however, was derived using the unit-amplitude plane wave

$$\psi_{\mathrm{i}}(\mathbf{r}) = e^{j\mathbf{k}_{\mathrm{i}}\cdot\mathbf{r}} \tag{5.1268}$$

of Eq. (5.333).

To compare the elastic and inelastic scattering formulas directly, we use the same unit-amplitude scattering convention here. Overall continuum-normalization factors depend on the chosen normalization of the generalized plane-wave states; they cancel appropriately when transition rates and cross sections are constructed. The spatial dependence derived below is independent of that convention.

The transition matrix element is a Fourier component of the transition interaction.

Start from the matrix element of Eq. (5.1265),

$$M_{mn} = \left\langle \mathbf{k}_{\mathrm{f}}, \Xi_m \left| \widehat{H}_{\mathrm{int}} \right| \mathbf{k}_{\mathrm{i}}, \Xi_n \right\rangle . \tag{5.1269}$$

We first take the matrix element over the internal coordinates of the specimen. By Eq. (5.1267), this converts $\widehat{H}_{\mathrm{int}}$ into the function $U_{mn}(\mathbf{r})$ acting on the imaging electron:

$$M_{mn} = \langle \mathbf{k}_{\mathrm{f}} \left| U_{mn}(\mathbf{r}) \right| \mathbf{k}_{\mathrm{i}} \rangle . \tag{5.1270}$$

We now write this matrix element in the position representation of the imaging electron. With the same unit-amplitude plane-wave convention used in the elastic-scattering derivation,

$$\langle \mathbf{r} | \mathbf{k}_{\mathrm{i}} \rangle = e^{j\mathbf{k}_{\mathrm{i}}\cdot\mathbf{r}}, \tag{5.1271}$$

and

$$\langle \mathbf{k}_{\mathrm{f}} | \mathbf{r} \rangle = e^{-j\mathbf{k}_{\mathrm{f}}\cdot\mathbf{r}}. \tag{5.1272}$$

Therefore,

$$M_{mn}(\mathbf{k}_\mathrm{f}, \mathbf{k}_\mathrm{i}) = \int e^{-j\mathbf{k}_\mathrm{f}\cdot\mathbf{r}} U_{mn}(\mathbf{r}) e^{j\mathbf{k}_\mathrm{i}\cdot\mathbf{r}}\, d^3r \tag{5.1273}$$

$$= \int U_{mn}(\mathbf{r}) \exp\left[-j\left(\mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}\right)\cdot\mathbf{r}\right]\, d^3r. \tag{5.1274}$$

The spatial-frequency transfer was defined in Eq. (5.341) as

$$\mathbf{R} = \frac{\mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}}{2\pi}. \tag{5.1275}$$

Hence

$$M_{mn}(\mathbf{R}) = \int U_{mn}(\mathbf{r}) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}}\, d^3r. \tag{5.1276}$$

Using the Fourier-transform convention of Eq. (3.2), we finally obtain

$$\boxed{M_{mn}(\mathbf{R}) = \widehat{U}_{mn}(\mathbf{R}).} \tag{5.1277}$$

Thus a transition with spatial-frequency transfer $\mathbf{R}$ probes the Fourier component $\widehat{U}_{mn}(\mathbf{R})$ of the interaction that couples the two chosen internal specimen states.

From the matrix element to the first-Born scattering amplitude. We can now make the connection with the elastic first-Born result particularly explicit.

In the elastic problem, the potential energy of the imaging electron is

$$U(\mathbf{r}) = -eV(\mathbf{r}), \tag{5.1278}$$

as derived in Eq. (5.26). Therefore the elastic first-Born amplitude of Eq. (5.445),

$$f_\mathrm{B}(\mathbf{R}) = \frac{me}{2\pi\hbar^2}\widehat{V}(\mathbf{R}), \tag{5.1279}$$

may equivalently be written in terms of the electron potential energy as

$$\boxed{f_\mathrm{B}(\mathbf{R}) = -\frac{m}{2\pi\hbar^2}\widehat{U}(\mathbf{R}).} \tag{5.1280}$$

Nothing in the free propagation from the interaction point to the far field depends on whether the specimen remained in the same internal state or changed to another one. In first order, the only change is therefore

$$U(\mathbf{r}) \longrightarrow U_{mn}(\mathbf{r}). \tag{5.1281}$$

Consequently, the scattering amplitude associated with the transition $n \to m$ is

$$\boxed{f_{mn}^{(1)}(\mathbf{R}) = -\frac{m}{2\pi\hbar^2}\widehat{U}_{mn}(\mathbf{R}).} \tag{5.1282}$$

Using Eq. (5.1277), this may also be written as

$$\boxed{f_{mn}^{(1)}(\mathbf{R}) = -\frac{m}{2\pi\hbar^2}M_{mn}(\mathbf{R}).} \tag{5.1283}$$

We have retained here the same relativistically corrected scalar mass m used in the elastic first-Born treatment. For the low-loss inelastic processes relevant to most cryo-EM imaging,

$$\epsilon_{mn} \ll E_{\mathrm{e}}, \tag{5.1284}$$

so the small change of the corresponding kinematic prefactor between the incident and outgoing energies may normally be neglected. A more exact relativistic treatment modifies this prefactor but not the central channel structure derived here.

The elastic result is the diagonal member of the same formula.

For an elastic channel the specimen remains in the same internal state, so

$$m = n. \tag{5.1285}$$

The transition interaction becomes the diagonal expectation value

$$U_{nn}(\mathbf{r}) = \langle \Xi_n | \widehat{H}_{\mathrm{int}} | \Xi_n \rangle. \tag{5.1286}$$

For the static elastic state used earlier in this chapter, Eq. (5.26) gives

$$U_{nn}(\mathbf{r}) = -eV(\mathbf{r}). \tag{5.1287}$$

Substitution into Eq. (5.1282) gives

$$f_{nn}^{(1)}(\mathbf{R}) = -\frac{m}{2\pi\hbar^2}\widehat{U}_{nn}(\mathbf{R}) \tag{5.1288}$$

$$= -\frac{m}{2\pi\hbar^2}\left[-e\widehat{V}(\mathbf{R})\right] \tag{5.1289}$$

$$= \boxed{\frac{me}{2\pi\hbar^2}\widehat{V}(\mathbf{R}),} \tag{5.1290}$$

which is precisely Eq. (5.445).

Thus the elastic first-Born scattering amplitude is not a separate theory. It is the diagonal member of the more general family of electron–specimen transition amplitudes.

One scattering machinery, different specimen channels. The relationship between elastic and inelastic first-order scattering can therefore be summarized as

$$\boxed{\begin{array}{llll} \text{elastic:} & U_{nn}(\mathbf{r}) = \langle \Xi_n | \widehat{H}_{\text{int}} | \Xi_n \rangle, & & f^{(1)}_{nn}(\mathbf{R}) = -\dfrac{m}{2\pi\hbar^2}\widehat{U}_{nn}(\mathbf{R}), \\ \text{inelastic:} & U_{mn}(\mathbf{r}) = \langle \Xi_m | \widehat{H}_{\text{int}} | \Xi_n \rangle, & m \neq n, & f^{(1)}_{mn}(\mathbf{R}) = -\dfrac{m}{2\pi\hbar^2}\widehat{U}_{mn}(\mathbf{R}). \end{array}} \tag{5.1291}$$

The propagation and Fourier geometry are therefore the same. What changes is the matrix element of the specimen interaction that acts as the source of the scattered wave.

For elastic scattering, the relevant quantity is a diagonal expectation value. It reduces to the ordinary static electrostatic potential and gives

$$f_{\text{B}}(\mathbf{R}) \propto \widehat{V}(\mathbf{R}). \tag{5.1292}$$

For inelastic scattering, the relevant quantity is an off-diagonal matrix element between two different internal states,

$$f^{(1)}_{mn}(\mathbf{R}) \propto \widehat{U}_{mn}(\mathbf{R}). \tag{5.1293}$$

The spatial frequency $\mathbf{R}$ is not arbitrary: for a specified transition $n \to m$, the allowed incident and final wavevectors must also satisfy the energy-conservation condition derived in Sec. 5.12.3. Thus the internal excitation energy determines the allowed scattering kinematics, while $\widehat{U}_{mn}(\mathbf{R})$ determines the strength of the coupling at that momentum transfer.

The first-Born approximation has the same physical meaning as in the elastic case. It retains one action of $\widehat{H}_{\text{int}}$:

$$\boxed{\begin{array}{c} |\mathbf{k}_{\text{i}}, \Xi_n\rangle \\ \Downarrow \\ \widehat{H}_{\text{int}} \\ \Downarrow \\ |\mathbf{k}_{\text{f}}, \Xi_m\rangle. \end{array}} \tag{5.1294}$$

This "one interaction" does not mean interaction with one particular atom. The operator $\widehat{H}_{\text{int}}$ acts on the specimen as a whole, and the final state $|\Xi_m\rangle$ may represent a localized electronic excitation, a molecular vibration, a collective excitation, or another many-body state.

We have therefore reduced the inelastic scattering problem to the quantity

$$U_{mn}(\mathbf{r}). \tag{5.1295}$$

The next question is what this transition interaction actually contains for a Coulomb-coupled biological specimen. This leads naturally to the transition charge density and its Fourier-space transition form factor.

5.12.5 Transition charge density and transition form factor

The previous subsection reduced the first-Born inelastic scattering problem to the transition interaction

$$U_{mn}(\mathbf{r}) = {}_{\text{s}}\langle \Xi_m | \widehat{H}_{\text{int}}(\mathbf{r}) | \Xi_n \rangle_{\text{s}}, \tag{5.1296}$$

introduced in Eq. (5.1267). We now ask what microscopic property of the specimen is contained in this quantity.

Most of the required construction has already been developed. In Sec. 5.2.1, the nuclear and electronic charges of the specimen were collected into the charge-density operator $\hat{\rho}_{\text{s}}(\mathbf{r})$ of Eq. (5.14). The microscopic Coulomb interaction was then written in the compact form

$$\widehat{H}_{\text{int}}(\mathbf{r}) = -\frac{e}{4\pi\varepsilon_0} \int \frac{\hat{\rho}_{\text{s}}(\mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|}\, d^3 r', \tag{5.1297}$$

in Eq. (5.18).

In Sec. 5.2.2, this interaction was projected onto the different internal states of the specimen. For the particular initial state $|\Xi_0\rangle_{\text{s}}$, Eq. (5.19) introduced

$$\rho_{m0}(\mathbf{r}) = {}_{\text{s}}\langle \Xi_m | \hat{\rho}_{\text{s}}(\mathbf{r}) | \Xi_0 \rangle_{\text{s}}. \tag{5.1298}$$

The diagonal channel $m = 0$ produced the ordinary elastic charge density of Eq. (5.23) and hence the static electrostatic potential of Eq. (5.25). The off-diagonal channels $m \neq 0$ were identified in Eq. (5.27) as transition charge densities.

We now remove the restriction that the initial state be $|\Xi_0\rangle_{\text{s}}$ and consider an arbitrary internal transition

$$|\Xi_n\rangle_{\text{s}} \longrightarrow |\Xi_m\rangle_{\text{s}}. \tag{5.1299}$$

Transition charge density. For an arbitrary pair of internal specimen states, define

$$\boxed{\rho_{mn}(\mathbf{r}) \equiv {}_{\mathrm{s}}\langle \Xi_m | \hat{\rho}_{\mathrm{s}}(\mathbf{r}) | \Xi_n \rangle_{\mathrm{s}}.} \tag{5.1300}$$

Equation (5.1300) is the direct generalization of Eq. (5.19) from ρ_{m0} to ρ_{mn}.

For a diagonal matrix element,

$$m = n, \tag{5.1301}$$

the quantity

$$\rho_{nn}(\mathbf{r}) = {}_{\mathrm{s}}\langle \Xi_n | \hat{\rho}_{\mathrm{s}}(\mathbf{r}) | \Xi_n \rangle_{\mathrm{s}} \tag{5.1302}$$

is an ordinary expectation value of the charge-density operator. It describes the mean charge density of the specimen when it occupies the internal state $|\Xi_n\rangle_{\mathrm{s}}$. For the reference state used previously in the elastic treatment,

$$n = 0, \tag{5.1303}$$

this reduces to the elastic charge density $\rho_{00}(\mathbf{r}) = \rho(\mathbf{r})$ introduced in Eq. (5.23).

For an off-diagonal matrix element,

$$m \neq n, \tag{5.1304}$$

the interpretation is different. The function $\rho_{mn}(\mathbf{r})$ is not an ordinary static charge distribution. Rather, it describes the spatial component of the charge-density operator that couples the initial state $|\Xi_n\rangle_{\mathrm{s}}$ to the final state $|\Xi_m\rangle_{\mathrm{s}}$.

It is therefore important not to confuse a transition charge density with the difference between the ordinary charge densities of the two states:

$$\boxed{\rho_{mn}(\mathbf{r}) \neq \rho_{mm}(\mathbf{r}) - \rho_{nn}(\mathbf{r}), \qquad m \neq n.} \tag{5.1305}$$

The difference density

$$\Delta\rho_{mn}(\mathbf{r}) \equiv \rho_{mm}(\mathbf{r}) - \rho_{nn}(\mathbf{r}) \tag{5.1306}$$

compares the *mean charge distributions* of two stationary states. The transition density ρ_{mn} is instead an off-diagonal matrix element and measures the *amplitude with which the local charge density couples the two states.*

Consequently, $\rho_{mn}(\mathbf{r})$ need not be real. It should not be interpreted as a literal instantaneous distribution of electrical charge.

From transition charge density to transition interaction.

Starting from the transition interaction of Eq. (5.1267),

$$U_{mn}(\mathbf{r}) = {}_{\mathrm{s}}\langle \Xi_m | \widehat{H}_{\mathrm{int}}(\mathbf{r}) | \Xi_n \rangle_{\mathrm{s}}, \tag{5.1307}$$

insert the charge-density representation of the Coulomb interaction from Eq. (5.18):

$$U_{mn}(\mathbf{r}) = -\frac{e}{4\pi\varepsilon_0} {}_{\mathrm{s}}\left\langle \Xi_m \left| \int \frac{\hat{\rho}_{\mathrm{s}}(\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|}\, d^3r' \right| \Xi_n \right\rangle_{\mathrm{s}}. \tag{5.1308}$$

The Coulomb kernel depends on the position $\mathbf{r}$ of the imaging electron and on the integration coordinate $\mathbf{r}'$, but not on the internal quantum coordinates over which the specimen matrix element is taken. We may therefore move the matrix element inside the integral:

$$U_{mn}(\mathbf{r}) = -\frac{e}{4\pi\varepsilon_0} \int \frac{{}_{\mathrm{s}}\langle \Xi_m | \hat{\rho}_{\mathrm{s}}(\mathbf{r}') | \Xi_n \rangle_{\mathrm{s}}}{\|\mathbf{r}-\mathbf{r}'\|}\, d^3r'. \tag{5.1309}$$

Using Eq. (5.1300),

$$\boxed{U_{mn}(\mathbf{r}) = -\frac{e}{4\pi\varepsilon_0} \int \frac{\rho_{mn}(\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|}\, d^3r'.} \tag{5.1310}$$

This is the generalization of Eq. (5.20) from the particular channel $0 \to m$ to an arbitrary channel $n \to m$.

It is convenient to define the corresponding *transition potential*

$$\boxed{V_{mn}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho_{mn}(\mathbf{r}')}{\|\mathbf{r}-\mathbf{r}'\|}\, d^3r',} \tag{5.1311}$$

so that

$$\boxed{U_{mn}(\mathbf{r}) = -eV_{mn}(\mathbf{r}).} \tag{5.1312}$$

For $m = n$, Eq. (5.1311) is the ordinary electrostatic potential associated with the charge density of state n. In particular, for $m = n = 0$ it reduces to Eq. (5.25), and Eq. (5.1312) reduces to the elastic relation Eq. (5.26).

For $m \neq n$, however, $V_{mn}(\mathbf{r})$ is not the ordinary electrostatic potential of either the initial or final state. Like ρ_{mn} itself, it is an off-diagonal transition

quantity. Its role is to determine the amplitude with which the Coulomb interaction drives the transition $n \to m$.

Electronic and nuclear contributions. The charge-density operator of Eq. (5.14) contains both nuclear and electronic position operators:

$$\hat{\rho}_{\mathrm{s}}(\mathbf{r}) = e \sum_A Z_A \delta\left(\mathbf{r} - \hat{\mathbf{R}}_A\right) - e \sum_j \delta\left(\mathbf{r} - \hat{\mathbf{r}}_j\right). \tag{5.1313}$$

Consequently, the general transition density $\rho_{mn}(\mathbf{r})$ may contain contributions from either nuclear or electronic motion, depending on the nature of the states $|\Xi_n\rangle_{\mathrm{s}}$ and $|\Xi_m\rangle_{\mathrm{s}}$.

For a purely electronic excitation treated in the fixed-nuclei approximation, the nuclear coordinates are replaced by fixed values $\mathbf{R}_A$. The nuclear contribution then acts as the identity on the electronic state. For two different orthogonal electronic states,

$$\langle \Xi_m | \Xi_n \rangle = 0, \qquad m \neq n, \tag{5.1314}$$

so its off-diagonal matrix element vanishes. The transition density is then purely electronic:

$$\boxed{\rho_{mn}(\mathbf{r}) = -e \left\langle \Xi_m \middle| \sum_j \delta\left(\mathbf{r} - \hat{\mathbf{r}}_j\right) \middle| \Xi_n \right\rangle, \qquad m \neq n.} \tag{5.1315}$$

For a vibrational transition, by contrast, the nuclear degrees of freedom are themselves part of the internal quantum state. The nuclear position operators already present in Eq. (5.14) may then have non-zero off-diagonal matrix elements. The same definition Eq. (5.1300) therefore covers both electronic and vibrational inelastic scattering; what changes is which part of the full charge-density operator connects the chosen initial and final states.

The transition form factor. The first-Born scattering amplitude obtained in Sec. 5.12.4 depends on $\widehat{U}_{mn}(\mathbf{R})$. We therefore introduce the Fourier transform of the transition charge density,

$$\boxed{F_{mn}(\mathbf{R}) \equiv \widehat{\rho}_{mn}(\mathbf{R}) = \int \rho_{mn}(\mathbf{r}) \exp\left(-j 2\pi \mathbf{R} \cdot \mathbf{r}\right) \, d^3 r.} \tag{5.1316}$$

We shall call $F_{mn}(\mathbf{R})$ the *transition charge form factor*.

With the definition of Eq. (5.1316), ρ_{mn} has units of charge per unit volume and F_{mn} has units of charge. Some scattering texts instead divide by e and use a dimensionless transition form factor. We retain the charge-valued convention because it follows directly from the charge-density notation already used in this chapter.

The transition form factor tells us how strongly the spatial structure of the transition couples to a particular spatial-frequency transfer $\mathbf{R}$.

Transition potential in reciprocal space.

The transition potential of Eq. (5.1311) obeys the same Poisson equation as the ordinary electrostatic potential, but with the static charge density replaced by the transition charge density:

$$\boxed{\nabla^2 V_{mn}(\mathbf{r}) = -\frac{\rho_{mn}(\mathbf{r})}{\varepsilon_0}.} \tag{5.1317}$$

The corresponding elastic Fourier-space relation was derived previously in Eq. (5.104). Applying the same Fourier-transform rule here, with the convention of Eq. (3.2),

$$\mathcal{F}\left\{\nabla^2 V_{mn}\right\} = -4\pi^2 R^2 \widehat{V}_{mn}(\mathbf{R}). \tag{5.1318}$$

Fourier transforming Eq. (5.1317) therefore gives

$$-4\pi^2 R^2 \widehat{V}_{mn}(\mathbf{R}) = -\frac{F_{mn}(\mathbf{R})}{\varepsilon_0}. \tag{5.1319}$$

Hence, for $R \neq 0$,

$$\boxed{\widehat{V}_{mn}(\mathbf{R}) = \frac{F_{mn}(\mathbf{R})}{4\pi^2 \varepsilon_0 R^2}.} \tag{5.1320}$$

Using Eq. (5.1312),

$$\boxed{\widehat{U}_{mn}(\mathbf{R}) = -\frac{eF_{mn}(\mathbf{R})}{4\pi^2 \varepsilon_0 R^2}.} \tag{5.1321}$$

Why the Coulomb interaction gives a $1/R^2$ factor.

The factor $1/R^2$ in Eq. (5.1321) is a direct consequence of the long-range Coulomb kernel

$$\frac{1}{\|\mathbf{r}-\mathbf{r}'\|}. \tag{5.1322}$$

To see this most directly, consider the Green function

$$G(\mathbf{r}) = \frac{1}{4\pi\|\mathbf{r}\|}. \tag{5.1323}$$

It satisfies

$$\boxed{\nabla^2 G(\mathbf{r}) = -\delta(\mathbf{r}).} \tag{5.1324}$$

Taking the Fourier transform and using the convention of Eq. (3.2),

$$\mathcal{F}\left\{\nabla^2 G\right\} = -4\pi^2 R^2 \widehat{G}(\mathbf{R}), \tag{5.1325}$$

whereas

$$\mathcal{F}\left\{\delta(\mathbf{r})\right\} = 1. \tag{5.1326}$$

Therefore,

$$-4\pi^2 R^2 \widehat{G}(\mathbf{R}) = -1, \tag{5.1327}$$

and hence

$$\boxed{\widehat{G}(\mathbf{R}) = \frac{1}{4\pi^2 R^2}.} \tag{5.1328}$$

Since

$$G(\mathbf{r}) = \frac{1}{4\pi r}, \tag{5.1329}$$

Eq. (5.1328) is equivalent to

$$\boxed{\mathcal{F}\left\{\frac{1}{r}\right\} = \frac{1}{\pi R^2}.} \tag{5.1330}$$

Thus a Coulomb potential, which is a convolution of a charge density with $1/r$ in real space, becomes multiplication by a quantity proportional to $1/R^2$ in reciprocal space.

This is precisely why Eq. (5.1320) contains

$$\widehat{V}_{mn}(\mathbf{R}) \propto \frac{F_{mn}(\mathbf{R})}{R^2}. \tag{5.1331}$$

The factor $1/R^2$ therefore does not arise from any particular property of the excitation. It is the reciprocal-space signature of the Coulomb interaction itself. The transition-dependent physics is contained in $F_{mn}(\mathbf{R})$.

Equation (5.1321) gives a direct microscopic interpretation of the transition interaction that entered the first-Born scattering amplitude. Substituting it into Eq. (5.1282) gives

$$\boxed{f^{(1)}_{mn}(\mathbf{R}) = \frac{me}{8\pi^3\varepsilon_0\hbar^2}\frac{F_{mn}(\mathbf{R})}{R^2}.} \tag{5.1332}$$

Thus the first-order inelastic scattering amplitude contains two clearly separated ingredients:

$$\boxed{f^{(1)}_{mn}(\mathbf{R}) \propto \underbrace{\frac{1}{R^2}}_{\text{Coulomb kernel}} \underbrace{F_{mn}(\mathbf{R})}_{\text{structure of the transition}}.} \tag{5.1333}$$

As shown above, the factor $1/R^2$ is the universal reciprocal-space dependence of the Coulomb kernel. The particular microscopic character of the excitation is contained instead in the transition form factor $F_{mn}(\mathbf{R})$. The transition form factor $F_{mn}(\mathbf{R})$ contains the microscopic spatial information about the particular excitation of the specimen.

Why an internal transition has no net transition charge.

At zero spatial frequency, Eq. (5.1316) gives

$$F_{mn}(\mathbf{0}) = \int \rho_{mn}(\mathbf{r})\, d^3r \tag{5.1334}$$

$$= {}_{\mathrm{s}}\left\langle \Xi_m \middle| \int \hat{\rho}_{\mathrm{s}}(\mathbf{r})\, d^3r \middle| \Xi_n \right\rangle_{\mathrm{s}}. \tag{5.1335}$$

Define the total-charge operator of the specimen as

$$\boxed{\hat{Q}_{\mathrm{s}} \equiv \int \hat{\rho}_{\mathrm{s}}(\mathbf{r})\, d^3r.} \tag{5.1336}$$

For internal transitions within a fixed total-charge sector, $\hat{Q}_{\mathrm{s}}$ acts as multiplication by the total specimen charge Q_{s}:

$$\hat{Q}_{\mathrm{s}}|\Xi_n\rangle_{\mathrm{s}} = Q_{\mathrm{s}}|\Xi_n\rangle_{\mathrm{s}}. \tag{5.1337}$$

Therefore,

$$F_{mn}(\mathbf{0}) = Q_{\mathrm{ss}}\langle\Xi_m|\Xi_n\rangle_{\mathrm{s}}. \tag{5.1338}$$

For two different orthogonal internal states,

$${}_{\mathrm{s}}\langle\Xi_m|\Xi_n\rangle_{\mathrm{s}} = 0, \qquad m \neq n, \tag{5.1339}$$

and hence

$$\boxed{F_{mn}(\mathbf{0}) = \int \rho_{mn}(\mathbf{r})\, d^3r = 0, \qquad m \neq n.} \tag{5.1340}$$

Equation (5.1340) does not say that the transition density vanishes. It says only that its spatial integral is zero. An internal excitation rearranges or oscillates charge within the specimen without changing its total charge.

Nor does Eq. (5.1340) imply that inelastic scattering must become weak at small momentum transfer. The first-Born amplitude contains the ratio

$$\frac{F_{mn}(\mathbf{R})}{R^2}, \tag{5.1341}$$

so the relevant question is how rapidly $F_{mn}(\mathbf{R})$ approaches zero as $R \to 0$. The answer depends on the multipolar character of the transition and will become important when we derive the characteristic angular distribution of low-loss inelastic scattering.

The microscopic picture. The same Coulomb scattering machinery therefore applies to diagonal and off-diagonal specimen channels:

$$\boxed{\begin{aligned} m = n \quad &: \quad \rho_{nn}(\mathbf{r}) \longrightarrow V_{nn}(\mathbf{r}) \longrightarrow f^{(1)}_{nn}(\mathbf{R}), \\ m \neq n \quad &: \quad \rho_{mn}(\mathbf{r}) \longrightarrow V_{mn}(\mathbf{r}) \longrightarrow f^{(1)}_{mn}(\mathbf{R}). \end{aligned}} \tag{5.1342}$$

For elastic scattering, the relevant source is a diagonal charge-density matrix element: the ordinary mean charge distribution of the specimen. For inelastic scattering, the relevant source is an off-diagonal transition density connecting two different internal states.

This gives a useful microscopic interpretation for biological specimens. An excitation is characterized by two complementary quantities. Its energy

$$\epsilon_{mn} = E_m - E_n \tag{5.1343}$$

identifies the energetic transition, while its transition form factor

$$F_{mn}(\mathbf{R}) \tag{5.1344}$$

describes how the spatial structure of that transition couples to the momentum transferred by the imaging electron.

5.12.6 From transition amplitudes to a measurable cross section

The previous subsections determined the amplitude for a specified inelastic transition

$$|\mathbf{k}_{\mathrm{i}}, \Xi_n\rangle \longrightarrow |\mathbf{k}_{\mathrm{f}}, \Xi_m\rangle. \tag{5.1345}$$

In particular, Eq. (5.1332) showed that

$$f^{(1)}_{mn}(\mathbf{R}) \propto \frac{F_{mn}(\mathbf{R})}{R^2}. \tag{5.1346}$$

A microscope, however, does not normally measure the internal final state $|\Xi_m\rangle_{\mathrm{s}}$ of the specimen. It measures the outgoing electron: its direction and, in an energy-loss experiment, its final energy.

We must therefore connect the transition amplitude with the probability that electrons emerge within a solid angle $d\Omega$ while losing an energy within an interval $d\epsilon$.

The quantum-mechanical connection between a matrix element and a transition rate was derived in Sec. A.12.6. Fermi's golden rule, Eq. (A.648), states that a weak interaction produces transitions at a rate proportional to the squared modulus of the matrix element,

$$\boxed{\text{transition rate} \propto |M_{mn}|^2,} \tag{5.1347}$$

with energy conservation enforced by a Dirac delta.

Thus the logical sequence is

$$\boxed{\begin{array}{c} M_{mn} \\ \Downarrow \\ |M_{mn}|^2 \\ \Downarrow \\ \text{transition rate} \\ \Downarrow \\ \text{differential scattering cross section.} \end{array}} \tag{5.1348}$$

Cross section for one particular final specimen state. Consider a particular transition

$$|\mathbf{k}_{\mathrm{i}}, \Xi_n\rangle \longrightarrow |\mathbf{k}_{\mathrm{f}}, \Xi_m\rangle. \tag{5.1349}$$

The magnitude k_{f} of the outgoing electron wavevector is fixed by the energy lost in this transition, as derived in Eq. (5.1248).

The spatial form of the outgoing wave is the same spherical-wave form derived for elastic scattering in Eq. (5.390). The difference is that the outgoing wavenumber is now k_{f} rather than the elastic value k, and the corresponding channel amplitude is $f_{mn}(\Omega)$. Thus,

$$\boxed{\psi_{\mathrm{s},m}(\mathbf{r}) \underset{r\to\infty}{\sim} f_{mn}(\Omega)\frac{e^{jk_{\mathrm{f}}r}}{r}.} \tag{5.1350}$$

The first-Born expression for this channel amplitude was obtained in Eq. (5.1282), and its transition-form-factor representation was derived in Eq. (5.1332).

The relation between wave amplitude and probability flux was developed in Sec. 3.11. In particular, Eq. (3.723) shows that the flux carried by a narrow electron beam is proportional to its probability density multiplied by its velocity.

Applying this result to Eq. (5.1350), whose probability density is

$$|\psi_{\mathrm{s},m}(\mathbf{r})|^2 \underset{r\to\infty}{\sim} \frac{|f_{mn}(\Omega)|^2}{r^2}, \tag{5.1351}$$

gives the outward radial flux

$$\boxed{J_{\mathrm{s},m}^{(r)} \propto v_{\mathrm{f}}\frac{|f_{mn}(\Omega)|^2}{r^2}.} \tag{5.1352}$$

This is the direct inelastic counterpart of the elastic radial-flux result derived in Eq. (5.400).

Similarly, the incident plane wave carries a flux

$$\boxed{J_{\mathrm{i}} \propto v_{\mathrm{i}},} \tag{5.1353}$$

which is the same incident-flux relation used in Eq. (5.397) of the elastic derivation.

A surface element subtending a solid angle $d\Omega$ at distance r has area

$$dA = r^2 d\Omega. \tag{5.1354}$$

Exactly as in Sec. 5.4.3, the factors r^2 and $1/r^2$ therefore cancel when the scattered radial flux is integrated over this element.

Dividing the outgoing rate by the incident flux gives

$$\boxed{\frac{d\sigma_{n\to m}}{d\Omega} = \frac{v_{\mathrm{f}}}{v_{\mathrm{i}}} |f_{mn}(\Omega)|^2.} \tag{5.1355}$$

This is precisely the general flux-ratio relation already obtained in Eq. (5.405); here it is applied to the particular inelastic channel $n \to m$.

For elastic scattering, Eq. (5.337) gives $k_{\mathrm{f}} = k_{\mathrm{i}}$ and hence $v_{\mathrm{f}} = v_{\mathrm{i}}$. Therefore Eq. (5.1355) reduces to

$$\frac{d\sigma}{d\Omega} = |f(\Omega)|^2, \tag{5.1356}$$

which is exactly the elastic result of Eq. (5.408).

Resolving the energy loss. The excitation energy associated with the transition $n \to m$ was defined in Eq. (5.1240) as

$$\epsilon_{mn} = E_m - E_n. \tag{5.1357}$$

Under the negligible-recoil approximation, this is also the energy lost by the imaging electron.

If the measured energy loss is denoted by ϵ, the contribution of the particular final state m is therefore restricted by

$$\delta\left(\epsilon - \epsilon_{mn}\right). \tag{5.1358}$$

Summing over all possible final specimen states gives

$$\boxed{\frac{d^2\sigma}{d\Omega\, d\epsilon} = \sum_m \frac{k_{\mathrm{f}}}{k_{\mathrm{i}}} |f_{mn}(\mathbf{R})|^2 \delta\left(\epsilon - \epsilon_{mn}\right).} \tag{5.1359}$$

Equation (5.1359) is the *double differential scattering cross section*: it gives the effective scattering area per unit solid angle and per unit energy loss.

Its dimensions are therefore

$$\left[\frac{d^2\sigma}{d\Omega\, d\epsilon}\right] = \mathrm{m}^2\,\mathrm{J}^{-1}, \tag{5.1360}$$

because the steradian is dimensionless in SI.

A dimensionless transition form factor. The transition form factor introduced in Eq. (5.1316),

$$F_{mn}(\mathbf{R}), \tag{5.1361}$$

has units of charge. It is convenient at this point to define the corresponding dimensionless quantity

$$\boxed{\mathcal{F}_{mn}(\mathbf{R}) \equiv \frac{F_{mn}(\mathbf{R})}{e}.} \tag{5.1362}$$

For a purely electronic transition in the fixed-nuclei approximation, Eq. (5.1315) gives

$$\mathcal{F}_{mn}(\mathbf{R}) = -\left\langle \Xi_m \left| \sum_j e^{-j2\pi\mathbf{R}\cdot\hat{\mathbf{r}}_j} \right| \Xi_n \right\rangle. \tag{5.1363}$$

The overall minus sign reflects the negative charge of the specimen electrons and disappears when the squared modulus is formed.

The dynamic structure factor. An experiment that detects only the imaging electron does not determine which internal final state m was produced. Different allowed final states must therefore be summed.

This motivates the definition

$$\boxed{S_n(\mathbf{R},\epsilon) \equiv \sum_m \left|\mathcal{F}_{mn}(\mathbf{R})\right|^2 \delta\left[\epsilon - (E_m - E_n)\right].} \tag{5.1364}$$

The quantity $S_n(\mathbf{R},\epsilon)$ is the *dynamic structure factor*. In the electron-energy-loss literature the closely related name *dynamic form factor* is also commonly used.

The word "dynamic" is important. A static structure factor describes how charge is distributed in space. The dynamic structure factor contains

an additional energy coordinate and therefore describes *which spatial charge fluctuations exist at which excitation energies.*

If the specimen is not in a single pure initial state but in a statistical ensemble with populations p_n, the corresponding structure factor is

$$\boxed{S(\mathbf{R},\epsilon) = \sum_n p_n \sum_m \left|\mathcal{F}_{mn}(\mathbf{R})\right|^2 \delta\left[\epsilon - (E_m - E_n)\right].} \tag{5.1365}$$

For a specimen initially in one state n, this reduces immediately to Eq. (5.1364).

The diagonal contribution $m = n$ occurs at

$$\epsilon = 0 \tag{5.1366}$$

and represents the elastic channel. For positive energy loss,

$$\epsilon > 0, \tag{5.1367}$$

the dynamic structure factor contains the inelastic transitions.

From the transition form factor to the double differential cross section.

The first-Born amplitude obtained in Eq. (5.1332) is

$$f_{mn}^{(1)}(\mathbf{R}) = \frac{me}{8\pi^3\varepsilon_0\hbar^2}\frac{F_{mn}(\mathbf{R})}{R^2}. \tag{5.1368}$$

Using Eq. (5.1362),

$$F_{mn} = e\mathcal{F}_{mn}, \tag{5.1369}$$

this becomes

$$f_{mn}^{(1)}(\mathbf{R}) = \frac{me^2}{8\pi^3\varepsilon_0\hbar^2}\frac{\mathcal{F}_{mn}(\mathbf{R})}{R^2}. \tag{5.1370}$$

The Lorentz factor was defined in Eq. (2.47), and the relativistically corrected scalar mass used in the Born treatment is

$$m = \gamma m_0. \tag{5.1371}$$

The Bohr radius was introduced in Eq. (5.112),

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{m_0 e^2}. \tag{5.1372}$$

Therefore,

$$\frac{me^2}{8\pi^3\varepsilon_0\hbar^2} = \frac{\gamma}{2\pi^2 a_0}, \tag{5.1373}$$

and hence

$$\boxed{f^{(1)}_{mn}(\mathbf{R}) = \frac{\gamma}{2\pi^2 a_0}\frac{\mathcal{F}_{mn}(\mathbf{R})}{R^2}.} \tag{5.1374}$$

Substitution into Eq. (5.1359) gives

$$\begin{aligned}\frac{d^2\sigma}{d\Omega\, d\epsilon} &= \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{\gamma^2}{4\pi^4 a_0^2 R^4} \\ &\quad\times \sum_m |\mathcal{F}_{mn}(\mathbf{R})|^2\, \delta\left[\epsilon - (E_m - E_n)\right].\end{aligned} \tag{5.1375}$$

Using Eq. (5.1364), we finally obtain

$$\boxed{\frac{d^2\sigma_\mathrm{in}}{d\Omega\, d\epsilon} = \frac{\gamma^2}{4\pi^4 a_0^2 R^4}\frac{k_\mathrm{f}}{k_\mathrm{i}} S_n(\mathbf{R},\epsilon),} \tag{5.1376}$$

for positive energy loss $\epsilon > 0$.

Connection with the conventional EELS notation. Electron-energy-loss spectroscopy is commonly written using the wavevector transfer

$$\mathbf{q} \equiv \mathbf{k}_\mathrm{i} - \mathbf{k}_\mathrm{f}. \tag{5.1377}$$

Our spatial-frequency convention, Eq. (5.341), gives

$$\boxed{\mathbf{q} = -2\pi\mathbf{R}, \qquad q = 2\pi R.} \tag{5.1378}$$

Consequently,

$$\frac{1}{R^4} = \frac{16\pi^4}{q^4}. \tag{5.1379}$$

Equation (5.1376) therefore becomes

$$\boxed{\frac{d^2\sigma_\mathrm{in}}{d\Omega\, d\epsilon} = \frac{4\gamma^2}{a_0^2 q^4}\frac{k_\mathrm{f}}{k_\mathrm{i}} S(\mathbf{q},\epsilon).} \tag{5.1380}$$

This is the standard semi-relativistic plane-wave Born expression used in electron-energy-loss spectroscopy (Egerton, 2011; Moreno et al., 2007).

What is relativistic in this expression? Equation (5.1380) is often called the *semi-relativistic plane-wave Born approximation*. The fast electron kinematics are corrected through the Lorentz factor γ, but the interaction itself is still treated as an instantaneous scalar Coulomb interaction and the incident and outgoing electron states are represented by scalar plane waves.

This level of approximation is consistent with the scalar relativistically corrected electron-scattering theory used throughout the present chapter. It captures the central dependence

$$\boxed{\frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon} \propto \frac{S(\mathbf{R}, \epsilon)}{R^4}.} \tag{5.1381}$$

A fully relativistic treatment of inelastic scattering contains additional electrodynamic effects, including retardation and coupling to transverse currents. In such a treatment the simple Coulomb denominator q^4 is modified and the transition operator contains additional terms. These corrections are important for quantitative EELS under some conditions, but they do not alter the conceptual role of the dynamic structure factor developed here (Zhang et al., 2025).

Equation (5.1376) separates the inelastic scattering problem into two conceptually different parts.

The factor

$$\frac{1}{R^4} \tag{5.1382}$$

comes from squaring the long-range Coulomb amplitude derived in Sec. 5.12.5. It strongly favours small momentum transfers.

The dynamic structure factor

$$S(\mathbf{R}, \epsilon) \tag{5.1383}$$

contains the physics of the specimen. It tells us which internal excitations exist at energy loss ϵ and how strongly each of them couples to the spatial-frequency transfer $\mathbf{R}$.

Thus,

$$\boxed{\begin{array}{c}\text{Coulomb interaction}\\ \times \\ \text{dynamic structure of the specimen}\\ \Downarrow \\ \text{measurable energy- and angle-resolved scattering.}\end{array}} \tag{5.1384}$$

The next question is therefore not how to convert an amplitude into a probability—that has now been done—but how the dynamic structure factor behaves for the low-energy excitations that dominate inelastic scattering in biological specimens.

5.12.7 Why different inelastic final states do not interfere, but an inelastic electron can still carry phase

The dynamic structure factor introduced in the previous subsection contains a sum over the possible final states of the specimen,

$$S_n(\mathbf{R}, \epsilon) = \sum_m |\mathcal{F}_{mn}(\mathbf{R})|^2 \, \delta\left[\epsilon - (E_m - E_n)\right]. \tag{5.1385}$$

An important feature of this expression is that the contributions of different final states appear as a sum of *squared amplitudes*, rather than as the squared modulus of a sum of amplitudes.

This is not an additional assumption about inelastic scattering. It is a direct consequence of the quantum state of the combined electron–specimen system.

The general quantum-mechanical origin of this result was developed in Sec. A.13. We now apply that result explicitly to the inelastic scattering channels introduced in the present section.

The outgoing electron is correlated with the specimen. Suppose that the specimen begins in the internal state $|\Xi_n\rangle_{\mathrm{s}}$. After interaction with the imaging electron, several final specimen states may be possible. The outgoing state of the complete electron–specimen system can then be written schematically as

$$\boxed{|\Psi_{\mathrm{out}}\rangle = \sum_m |\psi_m\rangle_{\mathrm{e}} \otimes |\Xi_m\rangle_{\mathrm{s}}.} \tag{5.1386}$$

This is the scattering-specific counterpart of the general combined state introduced in Eq. (A.663).

The electron state $|\psi_m\rangle_{\mathrm{e}}$ denotes the complete outgoing electron wave associated with the specimen being left in $|\Xi_m\rangle_{\mathrm{s}}$. Its normalization may contain the probability amplitude for producing that channel.

Equation (5.1386) is generally not a product of one electron state and one specimen state. Instead, the outgoing electron and specimen are correlated: observing which internal state was left in the specimen would provide information about which electron channel was produced.

Why different final specimen states add as probabilities.

The general derivation was given in Derivation A.13. Here we specialize it to the scattering channels of Eq. (5.1386).

Suppose that we detect the imaging electron at position $\mathbf{r}$, but do not observe the final internal state of the specimen.

The electron wave associated with channel m is

$$\psi_m(\mathbf{r}) = {}_{\mathrm{e}}\langle \mathbf{r} | \psi_m \rangle_{\mathrm{e}}. \tag{5.1387}$$

To see what the detector records, first consider the complete probability amplitude involving two possible specimen channels m and m'. Squaring Eq. (5.1386) produces terms proportional to

$$\psi_m(\mathbf{r}) \psi^*_{m'}(\mathbf{r})\, {}_{\mathrm{s}}\langle \Xi_{m'} | \Xi_m \rangle_{\mathrm{s}}. \tag{5.1388}$$

The internal eigenstates of the specimen may be chosen orthonormal, so that

$$\boxed{{}_{\mathrm{s}}\langle \Xi_{m'} | \Xi_m \rangle_{\mathrm{s}} = \delta_{m'm}.} \tag{5.1389}$$

Consequently, every cross term with

$$m' \neq m \tag{5.1390}$$

vanishes. The electron probability density observed without measuring the specimen state is therefore

$$\boxed{P_{\mathrm{e}}(\mathbf{r}) = \sum_m |\psi_m(\mathbf{r})|^2.} \tag{5.1391}$$

This is precisely the result obtained in the more general quantum discussion in Eq. (A.668).

Thus, electron waves correlated with *different orthogonal final states of the specimen do not interfere with one another when the specimen*

state is unobserved. Their probabilities add:

$$\boxed{\begin{array}{c} |\psi_m\rangle \otimes |\Xi_m\rangle \\ \text{and} \\ |\psi_{m'}\rangle \otimes |\Xi_{m'}\rangle, \qquad m \neq m' \\ \Downarrow \\ |\psi_m|^2 + |\psi_{m'}|^2, \quad \text{not } |\psi_m + \psi_{m'}|^2. \end{array}} \tag{5.1392}$$

This explains microscopically why the sum over final states in the dynamic structure factor is a sum of terms $|\mathcal{F}_{mn}|^2$. Different final internal states of the specimen represent distinguishable quantum channels, even when the microscope does not actually perform a measurement that identifies them.

But each inelastic channel is still a complex electron wave. Equation (5.1391) must not be interpreted as saying that the electron wave associated with each inelastic channel has lost its phase.

Consider one particular final specimen state $|\Xi_m\rangle_{\mathrm{s}}$. There may be several quantum-mechanical alternatives, which we label by α, that lead to this *same* complete final channel. The corresponding electron state can be written as

$$\boxed{|\psi_m\rangle = \sum_\alpha |\psi_{m,\alpha}\rangle.} \tag{5.1393}$$

For example, different scattering positions or different sequences of elastic and inelastic interactions may lead to the same final specimen state and the same outgoing electron channel.

The probability associated with this single final specimen state is then

$$|\psi_m(\mathbf{r})|^2 = \left|\sum_\alpha \psi_{m,\alpha}(\mathbf{r})\right|^2 \tag{5.1394}$$

$$= \sum_\alpha |\psi_{m,\alpha}(\mathbf{r})|^2 + \sum_{\alpha\neq\beta} \psi_{m,\alpha}(\mathbf{r})\psi^*_{m,\beta}(\mathbf{r}). \tag{5.1395}$$

The second sum contains interference terms. Therefore,

$$\boxed{\text{amplitudes leading to the same final specimen state can interfere.}} \tag{5.1396}$$

This is exactly the general quantum-mechanical rule established in Eq. (A.93) and summarized together with the orthogonal-channel case in Eq. (A.670).

The distinction is therefore not

$$\text{elastic} \quad \text{versus} \quad \text{inelastic}, \tag{5.1397}$$

but rather

$$\boxed{\begin{array}{rcl} \text{same complete final quantum channel} & \Longrightarrow & \text{amplitudes may interfere,} \\ \text{different orthogonal final specimen states} & \Longrightarrow & \text{probabilities add.} \end{array}} \tag{5.1398}$$

What, then, is lost in an inelastic interaction? An inelastic interaction correlates the imaging electron with internal degrees of freedom of the specimen. If different electron components become correlated with different orthogonal specimen states, the relative phase between those *channels* is no longer accessible from a measurement of the imaging electron alone. This is the decoherence mechanism discussed generally in Sec. A.13.

But this does not erase the complex phase structure of $|\psi_m\rangle$ within an individual channel. An electron that has lost energy can still emerge as a coherent complex wave, propagate through the electron-optical system, acquire aberration phases, and interfere with other amplitudes belonging to the same final channel.

The consequences of this distinction for image formation are important: an electron that has suffered an inelastic energy loss is not necessarily devoid of phase information. We shall return to this point after completing the microscopic description of inelastic scattering, when we discuss how energy-loss electrons propagate through the electron-optical system and contribute to the recorded image.

A particularly intuitive example is provided by two alternatives:

$$\begin{array}{rcl} \text{elastic scattering} & \longrightarrow & \text{inelastic excitation,} \\ \text{inelastic excitation} & \longrightarrow & \text{elastic scattering.} \end{array} \tag{5.1399}$$

If both alternatives leave the specimen in the same final internal state and produce the same outgoing electron channel, quantum mechanics requires their amplitudes to be added before the squared modulus is taken. Their relative phase can therefore affect the detected signal.

What is often called "incoherent inelastic scattering" should thus be understood more precisely as an *incoherent sum over mutually orthogonal unobserved specimen channels*. It does not imply that every electron wave within each of those channels is itself devoid of phase.

Partial coherence as a more general statement. Equation (5.1389) represents the limiting case in which the corresponding specimen states are exactly orthogonal. The more general expression Eq. (5.1388) shows that the interference term between two electron channels is weighted by

$$\boxed{{}_{\mathrm{s}}\langle \Xi_{m'} | \Xi_m \rangle_{\mathrm{s}}.} \tag{5.1400}$$

Thus the specimen itself carries the information that can distinguish the two alternatives:

$$\begin{aligned} |\langle \Xi_{m'} | \Xi_m \rangle| = 1 \quad &\Longrightarrow \quad \text{no distinguishability from the specimen,} \\ 0 < |\langle \Xi_{m'} | \Xi_m \rangle| < 1 \quad &\Longrightarrow \quad \text{partial coherence,} \\ \langle \Xi_{m'} | \Xi_m \rangle = 0 \quad &\Longrightarrow \quad \text{no interference between the channels.} \end{aligned} \tag{5.1401}$$

For the distinct orthogonal internal eigenstates summed over in the dynamic structure factor, the last case applies. This is why their contributions enter as separate probabilities.

5.12.8 Small momentum transfer and the characteristic inelastic angle

The double differential cross section derived in Eq. (5.1376) contains the characteristic Coulomb factor

$$\frac{1}{R^4}. \tag{5.1402}$$

Taken by itself, this factor would appear to produce an extremely strong divergence as the momentum transfer tends to zero.

However, the Coulomb factor cannot be considered independently of the transition form factor. We showed in Eq. (5.1340) that for two different orthogonal internal states

$$F_{mn}(\mathbf{0}) = 0, \qquad m \neq n. \tag{5.1403}$$

The transition itself therefore suppresses the zero-momentum-transfer limit.

To determine the resulting angular dependence, we need to answer two questions. First, how rapidly does the transition form factor approach zero when $R \to 0$? Second, how is the three-dimensional spatial-frequency transfer R related to the scattering angle θ when the electron has lost a finite energy?

Small-R expansion of the transition form factor. The transition form factor was defined in Eq. (5.1316) as

$$F_{mn}(\mathbf{R}) = \int \rho_{mn}(\mathbf{r}) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}}\, d^3r. \tag{5.1404}$$

At sufficiently small spatial-frequency transfer, the exponential can be expanded systematically.

Dipole limit of the transition form factor.

For

$$2\pi|\mathbf{R}\cdot\mathbf{r}| \ll 1 \tag{5.1405}$$

over the region in which the transition density is appreciable,

$$e^{-j2\pi\mathbf{R}\cdot\mathbf{r}} = 1 - j2\pi\mathbf{R}\cdot\mathbf{r} - 2\pi^2(\mathbf{R}\cdot\mathbf{r})^2 + O(R^3). \tag{5.1406}$$

Substitution into Eq. (5.1316) gives

$$\begin{aligned} F_{mn}(\mathbf{R}) = &\int \rho_{mn}(\mathbf{r})\, d^3r \\ &- j2\pi\mathbf{R}\cdot\int \mathbf{r}\rho_{mn}(\mathbf{r})\, d^3r \\ &- 2\pi^2\int (\mathbf{R}\cdot\mathbf{r})^2\rho_{mn}(\mathbf{r})\, d^3r + O(R^3). \end{aligned} \tag{5.1407}$$

The first term is precisely $F_{mn}(\mathbf{0})$, which vanishes for an inelastic transition by Eq. (5.1340).

Define the electric-dipole operator of the specimen by

$$\boxed{\hat{\mathbf{d}}_{\mathrm{s}} \equiv \int \mathbf{r}\,\hat{\rho}_{\mathrm{s}}(\mathbf{r})\, d^3r,} \tag{5.1408}$$

and its transition matrix element by

$$\boxed{\mathbf{d}_{mn} \equiv {}_{\mathrm{s}}\langle\Xi_m|\hat{\mathbf{d}}_{\mathrm{s}}|\Xi_n\rangle_{\mathrm{s}}.} \tag{5.1409}$$

Using Eq. (5.1300),

$$\mathbf{d}_{mn} = \int \mathbf{r}\rho_{mn}(\mathbf{r})\, d^3r. \tag{5.1410}$$

Therefore, for an inelastic transition, Eq. (5.1407) begins as

$$\boxed{F_{mn}(\mathbf{R}) = -j2\pi\mathbf{R}\cdot\mathbf{d}_{mn} + O(R^2).} \tag{5.1411}$$

If

$$\mathbf{d}_{mn} \neq \mathbf{0}, \tag{5.1412}$$

the transition is *dipole allowed*, and its transition form factor is therefore linear in the momentum transfer:

$$\boxed{F_{mn}(\mathbf{R}) = O(R).} \tag{5.1413}$$

This result explains why the zero of the transition form factor found in Eq. (5.1340) is important. For a dipole-allowed transition, the form factor vanishes only linearly with R.

Using the dimensionless transition form factor defined in Eq. (5.1362), Eq. (5.1411) becomes

$$\boxed{\mathcal{F}_{mn}(\mathbf{R}) \simeq -j\frac{2\pi}{e}\mathbf{R}\cdot\mathbf{d}_{mn}.} \tag{5.1414}$$

The first-Born scattering amplitude contains one factor $1/R^2$ from the Coulomb interaction, as shown in Derivation 5.12.5. Consequently,

$$f_{mn}^{(1)} \propto \frac{F_{mn}(\mathbf{R})}{R^2} \sim \frac{R}{R^2} = \frac{1}{R} \tag{5.1415}$$

for a dipole-allowed transition, and therefore

$$\boxed{|f_{mn}^{(1)}|^2 \propto \frac{1}{R^2}.} \tag{5.1416}$$

Thus the $1/R^4$ dependence of the squared Coulomb kernel is partly cancelled by the R^2 dependence of the squared transition form factor. This is the microscopic origin of the much milder small-angle behavior characteristic of dipole-dominated inelastic scattering.

Energy loss requires longitudinal momentum transfer. We now relate R to the observable scattering angle.

The spatial-frequency transfer was defined in Eq. (5.341) as

$$\mathbf{R} = \frac{\mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}}{2\pi}. \tag{5.1417}$$

Using the relation between wavevector and momentum, Eq. (4.671), the change in momentum of the imaging electron is

$$\mathbf{p}_{\mathrm{f}} - \mathbf{p}_{\mathrm{i}} = h\mathbf{R}, \tag{5.1418}$$

as already obtained for the inelastic case in Eq. (5.1257).

Unlike elastic scattering, an inelastic event changes not only the direction but also the magnitude of the electron momentum. Even an electron that emerges at zero scattering angle must therefore transfer a finite longitudinal momentum to the specimen.

The characteristic energy-loss angle.

Let the incident electron propagate along the optical axis,

$$\mathbf{p}_{\mathrm{i}} = p\,\hat{\mathbf{z}}. \tag{5.1419}$$

Suppose that it loses an energy ϵ and that its momentum magnitude decreases by

$$\Delta p = p_{\mathrm{i}} - p_{\mathrm{f}} > 0. \tag{5.1420}$$

If the outgoing electron is deflected through a small angle θ, we may write

$$\mathbf{p}_{\mathrm{f}} = (p - \Delta p)\left(\sin\theta\,\hat{\mathbf{e}}_{\perp} + \cos\theta\,\hat{\mathbf{z}}\right), \tag{5.1421}$$

where $\hat{\mathbf{e}}_{\perp}$ is a unit vector in the transverse scattering direction.

Since

$$h\mathbf{R} = \mathbf{p}_{\mathrm{f}} - \mathbf{p}_{\mathrm{i}}, \tag{5.1422}$$

the magnitude of the transverse component is

$$hR_{\perp} = (p - \Delta p)\sin\theta \tag{5.1423}$$

$$\simeq p\theta, \tag{5.1424}$$

and therefore

$$\boxed{R_{\perp} \simeq \frac{p}{h}\theta = \frac{\theta}{\lambda}.} \tag{5.1425}$$

Here we used $p = h/\lambda$, which follows from Eq. (4.671) together with $k = 2\pi/\lambda$.

The longitudinal component is

$$hR_z = (p - \Delta p)\cos\theta - p \tag{5.1426}$$

$$\simeq -\Delta p - \frac{p\theta^2}{2}. \tag{5.1427}$$

To the order required for calculating R^2, the second term may be neglected: when R_z is squared, it produces terms of order θ^4 and $(\Delta p/p)\theta^2$. Hence

$$\boxed{R_z \simeq -\frac{\Delta p}{h}.} \tag{5.1428}$$

It is useful to define the positive longitudinal spatial-frequency scale

$$\boxed{R_E \equiv \frac{\Delta p}{h}.} \tag{5.1429}$$

Then

$$R_z \simeq -R_E. \tag{5.1430}$$

The complete three-dimensional spatial-frequency transfer therefore satisfies

$$R^2 = R_\perp^2 + R_z^2 \tag{5.1431}$$

$$\simeq \frac{\theta^2}{\lambda^2} + \frac{(\Delta p)^2}{h^2}. \tag{5.1432}$$

Define the *characteristic energy-loss angle*

$$\boxed{\theta_E \equiv \frac{\Delta p}{p}.} \tag{5.1433}$$

Since $p/h = 1/\lambda$,

$$R_E = \frac{\theta_E}{\lambda}, \tag{5.1434}$$

and hence

$$\boxed{R^2 \simeq \frac{\theta^2 + \theta_E^2}{\lambda^2}.} \tag{5.1435}$$

It remains to relate θ_E to the measured energy loss. The relativistic energy–momentum relation of Eq. (2.68) is

$$E^2 = p^2c^2 + m_0^2c^4. \tag{5.1436}$$

Differentiating with respect to p gives

$$\frac{dE}{dp} = \frac{pc^2}{E} = v. \tag{5.1437}$$

For a small energy loss,

$$\epsilon = E_{\mathrm{i}} - E_{\mathrm{f}} \simeq v\Delta p, \tag{5.1438}$$

and therefore

$$\boxed{\theta_E \simeq \frac{\epsilon}{pv}.} \tag{5.1439}$$

Equivalently, Eqs. (5.1429) and (5.1438) give

$$\boxed{R_E \simeq \frac{\epsilon}{hv}.} \tag{5.1440}$$

The characteristic angle therefore has a simple interpretation. Ordinary angular scattering produces the transverse spatial-frequency transfer

$$R_\perp \simeq \frac{\theta}{\lambda}, \tag{5.1441}$$

whereas the energy loss produces an unavoidable longitudinal transfer

$$|R_z| \simeq \frac{\theta_E}{\lambda}. \tag{5.1442}$$

Even at

$$\theta = 0, \tag{5.1443}$$

an inelastic event therefore has

$$R \simeq \frac{\theta_E}{\lambda} \neq 0. \tag{5.1444}$$

This is the physical reason why the apparent forward singularity of inelastic Coulomb scattering is regularized by the finite energy loss.

Relativistic expression for the characteristic angle. It is useful to express Eq. (5.1439) in terms of the incident kinetic energy E_k. From the relativistic relations

$$E = E_k + m_0 c^2, \tag{5.1445}$$

and

$$p^2 c^2 = E_k \left(E_k + 2m_0 c^2 \right), \tag{5.1446}$$

we have

$$v = \frac{pc^2}{E_k + m_0 c^2}. \tag{5.1447}$$

Consequently,

$$pv = \frac{E_k \left(E_k + 2m_0 c^2 \right)}{E_k + m_0 c^2}. \tag{5.1448}$$

Equation (5.1439) may therefore be written

$$\boxed{\theta_E \simeq \epsilon \frac{E_k + m_0 c^2}{E_k \left(E_k + 2m_0 c^2 \right)}.} \tag{5.1449}$$

In the non-relativistic limit, $E_k \ll m_0 c^2$, this reduces to

$$\theta_E \simeq \frac{\epsilon}{2E_k}. \tag{5.1450}$$

At the electron energies used in modern cryo-EM, however, the relativistic form Eq. (5.1449) should be used.

Numerical scales at cryo-EM electron energies. The characteristic angles involved in inelastic scattering are very small. This is worth quantifying.

Using the relativistic electron wavelength derived in Eq. (2.103),

$$\lambda_{200\,\mathrm{keV}} \simeq 0.02508\,\text{Å}, \qquad \lambda_{300\,\mathrm{keV}} \simeq 0.01969\,\text{Å}. \tag{5.1451}$$

Equation (5.1449) then gives the particularly useful numerical approximations

$$\boxed{\theta_E \simeq 2.91\,\mu\mathrm{rad} \left(\frac{\epsilon}{1\,\mathrm{eV}} \right), \qquad 200\ \mathrm{keV},} \tag{5.1452}$$

and

$$\boxed{\theta_E \simeq 2.04\,\mu\mathrm{rad} \left(\frac{\epsilon}{1\,\mathrm{eV}} \right), \qquad 300\ \mathrm{keV}.} \tag{5.1453}$$

Thus even an energy loss of several tens of electronvolts corresponds to only a few tens of microradians.

The same energy loss also defines the longitudinal spatial-frequency transfer derived in Eq. (5.1440),

$$R_E = \frac{\theta_E}{\lambda} \simeq \frac{\epsilon}{hv}. \tag{5.1454}$$

At 300 keV,

$$\boxed{R_E \simeq 1.04 \times 10^{-4}\,\text{Å}^{-1} \left(\frac{\epsilon}{1\,\text{eV}}\right).} \tag{5.1455}$$

There is also a temporal frequency associated with the excitation energy itself:

$$\boxed{\nu_E = \frac{\epsilon}{h} \simeq 2.418 \times 10^{14}\,\text{Hz} \left(\frac{\epsilon}{1\,\text{eV}}\right).} \tag{5.1456}$$

Equivalently,

$$\omega_E = 2\pi\nu_E = \frac{\epsilon}{\hbar}. \tag{5.1457}$$

Some representative values are shown in Table 5.5.

Numerical example: a 20 eV loss at 300 keV.

Consider a 300 keV imaging electron that loses

$$\epsilon = 20 \text{ eV}. \tag{5.1458}$$

Equation (5.1453) gives

$$\theta_E \simeq 40.9\ \mu\text{rad} = 0.0409 \text{ mrad}. \tag{5.1459}$$

Using $\lambda \simeq 0.01969$ Å, the corresponding unavoidable longitudinal spatial-frequency transfer is

$$R_E = \frac{\theta_E}{\lambda} \simeq 2.08 \times 10^{-3}\ \text{Å}^{-1}. \tag{5.1460}$$

Suppose now that the electron is observed at a scattering angle

$$\theta = 0.10 \text{ mrad}. \tag{5.1461}$$

Table 5.5: Representative characteristic angles, longitudinal spatial-frequency transfers, and excitation frequencies for several energy-loss scales. The spatial-frequency values R_E correspond to 300 keV electrons. The physical-process descriptions indicate approximate energy scales rather than sharp boundaries.

Energy-loss scale	ϵ (eV)	$\nu_E = \epsilon/h$	θ_E at 200 keV (mrad)	θ_E at 300 keV (mrad)
Molecular vibration	0.05	12.1 THz	1.45×10^{-4}	1.02×10^{-4}
Low-energy excitation	1	242 THz	2.91×10^{-3}	2.04×10^{-3}
Valence excitation	10	2.42 PHz	2.91×10^{-2}	2.04×10^{-2}
Typical low-loss / plasmon scale	20	4.84 PHz	5.82×10^{-2}	4.09×10^{-2}
Higher energy loss	100	24.2 PHz	0.291	0.204
Core-loss scale	300	72.5 PHz	0.873	0.613

ϵ (eV)	R_E at 300 keV (Å^{-1})	$1/R_E$
0.05	5.19×10^{-6}	19.3 μm
1	1.04×10^{-4}	0.963 μm
10	1.04×10^{-3}	96.3 nm
20	2.08×10^{-3}	48.1 nm
100	1.04×10^{-2}	9.63 nm
300	3.12×10^{-2}	3.21 nm

From Eq. (5.1435),

$$R \simeq \frac{\sqrt{\theta^2 + \theta_E^2}}{\lambda} \tag{5.1462}$$

$$= \frac{\sqrt{(0.10 \text{ mrad})^2 + (0.0409 \text{ mrad})^2}}{0.01969 \text{ Å}} \tag{5.1463}$$

$$\simeq 5.49 \times 10^{-3} \text{ Å}^{-1}. \tag{5.1464}$$

At zero observed scattering angle, by contrast,

$$R(\theta = 0) = R_E \simeq 2.08 \times 10^{-3} \text{ Å}^{-1}. \tag{5.1465}$$

Thus an inelastic event with finite energy loss never reaches $\mathbf{R} = \mathbf{0}$, even when the outgoing electron is observed exactly along the incident-beam direction.

For example, for a 20 eV loss at 300 keV, the relativistic expression gives

$$\theta_E \simeq 0.0409 \text{ mrad}. \tag{5.1466}$$

The non-relativistic approximation would instead give

$$\theta_E^{(\mathrm{nr})} = \frac{20}{2(300000)} \simeq 0.0333 \text{ mrad}, \tag{5.1467}$$

about 18.5% smaller. At 200 keV the corresponding underestimate is approximately 14%. This is sufficiently large that the relativistic expression should be retained for quantitative cryo-EM calculations.

The low-loss angular distribution. We can now combine the microscopic transition physics with the kinematics.

For one particular transition $n \to m$, integration of the double differential cross section Eq. (5.1376) over the narrow energy interval containing that transition gives

$$\frac{d\sigma_{n\to m}}{d\Omega} = \frac{\gamma^2}{4\pi^4 a_0^2 R^4} \frac{k_\mathrm{f}}{k_\mathrm{i}} \left|\mathcal{F}_{mn}(\mathbf{R})\right|^2 \tag{5.1468}$$

within the semi-relativistic plane-wave Born approximation introduced in the preceding subsection.

For a dipole-allowed transition, Eq. (5.1414) gives

$$\left|\mathcal{F}_{mn}(\mathbf{R})\right|^2 \simeq \frac{4\pi^2}{e^2} \left|\mathbf{R} \cdot \mathbf{d}_{mn}\right|^2. \tag{5.1469}$$

For a particular oriented transition, the factor $|\mathbf{R}\cdot\mathbf{d}_{mn}|^2$ retains information about the direction of the transition dipole. The simplest Lorentzian-like angular law emerges after averaging over isotropically distributed orientations, as is often appropriate for randomly oriented biological particles or for an isotropic set of transition directions.

Angular distribution of an isotropic dipole transition.

For an isotropic distribution of transition-dipole directions,

$$\boxed{\left\langle |\mathbf{R}\cdot\mathbf{d}_{mn}|^2 \right\rangle_{\text{orient}} = \frac{1}{3}R^2|\mathbf{d}_{mn}|^2.} \tag{5.1470}$$

This follows because the three Cartesian directions contribute equally under isotropic averaging.

Using Eq. (5.1469),

$$\left\langle |\mathcal{F}_{mn}(\mathbf{R})|^2 \right\rangle_{\text{orient}} \simeq \frac{4\pi^2}{3e^2}R^2|\mathbf{d}_{mn}|^2. \tag{5.1471}$$

Substituting this result into Eq. (5.1468) yields

$$\boxed{\left\langle \frac{d\sigma_{n\to m}}{d\Omega} \right\rangle_{\text{orient}} \simeq \frac{\gamma^2}{3\pi^2 a_0^2}\frac{k_{\text{f}}}{k_{\text{i}}}\frac{|\mathbf{d}_{mn}|^2}{e^2}\frac{1}{R^2}.} \tag{5.1472}$$

Finally, using the kinematic relation Eq. (5.1435),

$$\frac{1}{R^2} \simeq \frac{\lambda^2}{\theta^2+\theta_E^2}, \tag{5.1473}$$

we obtain

$$\boxed{\left\langle \frac{d\sigma_{n\to m}}{d\Omega} \right\rangle_{\text{orient}} \simeq \frac{\gamma^2\lambda^2}{3\pi^2 a_0^2}\frac{k_{\text{f}}}{k_{\text{i}}}\frac{|\mathbf{d}_{mn}|^2}{e^2}\frac{1}{\theta^2+\theta_E^2}.} \tag{5.1474}$$

Thus, apart from the transition-dependent prefactor,

$$\boxed{\frac{d\sigma_{\text{in}}}{d\Omega} \propto \frac{1}{\theta^2+\theta_E^2}.} \tag{5.1475}$$

Equation (5.1475) is the familiar strongly forward-peaked angular distribution of low-loss, dipole-dominated inelastic electron scattering. The characteristic angle θ_E prevents a true divergence at $\theta = 0$ and sets the angular scale over which the distribution turns over.

What is universal and what is not. The derivation above identifies the origin of the two factors in Eq. (5.1475):

$$\boxed{\begin{aligned} 1/R^4 &\longleftarrow \text{squared Coulomb interaction,} \\ R^2 &\longleftarrow \text{dipole transition form factor,} \\ R^2 \simeq \lambda^{-2}(\theta^2 + \theta_E^2) &\longleftarrow \text{inelastic scattering kinematics.} \end{aligned}} \tag{5.1476}$$

The resulting Lorentzian-like law is therefore not a universal form for every possible inelastic process. If the dipole matrix element vanishes,

$$\mathbf{d}_{mn} = \mathbf{0}, \tag{5.1477}$$

the linear term in Eq. (5.1407) disappears and the leading contribution comes from a higher multipole. Such dipole-forbidden transitions have a different small-angle dependence.

Likewise, for an oriented transition, Eq. (5.1469) retains the directional factor

$$|\mathbf{R} \cdot \mathbf{d}_{mn}|^2, \tag{5.1478}$$

and the scattering need not be isotropic about the incident beam.

Finally, a biological specimen has a broad spectrum of possible energy losses rather than a single value of ϵ. Since Eq. (5.1439) gives a different θ_E for every energy loss, the energy-integrated angular distribution is more accurately regarded as a weighted superposition, schematically,

$$\boxed{\frac{d\sigma_{\text{in}}}{d\Omega} \simeq \int \frac{A(\epsilon)}{\theta^2 + \theta_E^2(\epsilon)}\, d\epsilon,} \tag{5.1479}$$

where $A(\epsilon)$ contains the strength of the corresponding dipole-allowed excitations.

A single effective value of θ_E is therefore an approximation appropriate to a particular energy loss, a sufficiently narrow energy-loss interval, or a simplified description of the complete low-loss spectrum.

5.12.9 From microscopic cross section to the inelastic mean free path

The preceding sections described inelastic scattering at the level of individual quantum transitions. For a specified initial specimen state $|\Xi_n\rangle_{\mathrm{s}}$, the double differential cross section derived in Eq. (5.1376) gives the probability of scattering into a solid-angle element $d\Omega$ while transferring an energy in the interval $[\epsilon, \epsilon + d\epsilon]$:

$$\frac{d^2\sigma_{\mathrm{in}}}{d\Omega\, d\epsilon}. \tag{5.1480}$$

This quantity contains detailed information about both the direction of the scattered electron and the internal excitation produced in the specimen. For many questions concerning specimen thickness, however, we need only a simpler quantity: the probability that an electron undergoes *any* inelastic interaction while travelling through the material.

The first step is therefore to integrate over all possible outgoing directions and all allowed positive energy losses.

The total inelastic cross section. For a particular transition $n \to m$, the angular differential cross section was obtained in Eq. (5.1355). Its total cross section is

$$\boxed{\sigma_{n\to m} = \int_{4\pi} \frac{d\sigma_{n\to m}}{d\Omega}\, d\Omega.} \tag{5.1481}$$

The total inelastic cross section for an initial state n is obtained by summing over all final states different from the initial state:

$$\boxed{\sigma_{\mathrm{in},n} = \sum_{m\neq n} \sigma_{n\to m}.} \tag{5.1482}$$

The same result can be expressed directly in terms of the double differential cross section. Since positive energy loss corresponds to inelastic scattering,

$$\boxed{\sigma_{\mathrm{in},n} = \int_0^\infty d\epsilon \int_{4\pi} d\Omega\, \frac{d^2\sigma_{\mathrm{in}}}{d\Omega\, d\epsilon}.} \tag{5.1483}$$

Equation (5.1483) is simply the continuum counterpart of the channel sum in Eq. (5.1482). The integration over energy automatically collects all energetically accessible final states through the dynamic structure factor.

Using Eq. (5.1376), the semi-relativistic plane-wave Born approximation gives

$$\boxed{\sigma_{\mathrm{in},n} = \frac{\gamma^2}{4\pi^4 a_0^2} \int_0^\infty d\epsilon \, \frac{k_{\mathrm{f}}}{k_{\mathrm{i}}} \int_{4\pi} \frac{S_n(\mathbf{R}, \epsilon)}{R^4} \, d\Omega.} \tag{5.1484}$$

Here $\mathbf{R}$ is not an independent integration variable: for each energy loss ϵ and scattering direction Ω, it is fixed by the inelastic kinematics derived in Sec. 5.12.8.

Equation (5.1484) makes explicit that the total inelastic cross section depends both on the incident electron energy and on the excitation spectrum of the material.

Units of the integrated inelastic cross section.

The double differential cross section has units

$$\left[\frac{d^2\sigma}{d\Omega \, d\epsilon}\right] = \mathrm{m}^2 \, \mathrm{J}^{-1}, \tag{5.1485}$$

because the solid angle is dimensionless in SI.

Integration over energy therefore gives

$$\left[\int \frac{d^2\sigma}{d\Omega \, d\epsilon} d\epsilon\right] = \mathrm{m}^2, \tag{5.1486}$$

per unit solid angle, and the subsequent angular integration gives

$$\boxed{[\sigma_{\mathrm{in}}] = \mathrm{m}^2.} \tag{5.1487}$$

Thus σ_{in} is an effective area: it quantifies the strength of all microscopic inelastic channels associated with one chosen scattering entity.

From one scatterer to a macroscopic material. A cross section characterizes one scattering entity. A biological specimen contains a macroscopic number of such entities along the electron trajectory. To convert the microscopic cross section into a probability per unit path length, we must therefore introduce their number density.

Let n denote the number of scattering entities per unit volume and let each have a total inelastic cross section σ_{in}. Consider an electron traversing an infinitesimal specimen thickness dz.

From the microscopic cross section to the mean free path.

Let n denote the *number density of scattering entities*, that is, the number of effective scatterers per unit volume:

$$\boxed{n = \frac{N}{V}.} \tag{5.1488}$$

Its SI units are therefore

$$[n] = \mathrm{m}^{-3}. \tag{5.1489}$$

Now consider a beam crossing a transverse area A and an infinitesimal specimen thickness dz. The corresponding slab volume is

$$dV = A\,dz. \tag{5.1490}$$

Since n scatterers are present per unit volume, the expected number of scattering entities contained in this slab is

$$\boxed{dN = n\,dV = nA\,dz.} \tag{5.1491}$$

Here the meaning of "scattering entity" depends on the microscopic description being used. For an independent-atom model it may be an atom or molecule; for a multicomponent material there may be several species with different number densities and cross sections. For collective excitations, such as plasmons, it is often more natural to work directly with a cross section or response normalized per unit volume rather than to assign the excitation to an individual scatterer.

Each entity presents an effective inelastic scattering area σ_{in}. In the dilute, independent-scattering limit, the total effective inelastic area presented by the entities in the thin slab is therefore

$$dA_{\mathrm{in}} = dN\,\sigma_{\mathrm{in}} = nA\,\sigma_{\mathrm{in}}\,dz. \tag{5.1492}$$

The probability that an electron crossing the area A undergoes an inelastic interaction within this infinitesimal slab is the fraction of the

beam area occupied by this effective scattering area:

$$dP_{\text{in}} = \frac{dA_{\text{in}}}{A} \tag{5.1493}$$

$$= n\sigma_{\text{in}}\, dz. \tag{5.1494}$$

Thus,

$$\boxed{\frac{dP_{\text{in}}}{dz} = n\sigma_{\text{in}}} \tag{5.1495}$$

to first order in dz.

The inverse of this probability per unit distance defines the *inelastic mean free path*:

$$\boxed{\Lambda_{\text{in}} \equiv \frac{1}{n\sigma_{\text{in}}}.} \tag{5.1496}$$

Equivalently,

$$\boxed{\Lambda_{\text{in}}^{-1} = n\sigma_{\text{in}}.} \tag{5.1497}$$

The units provide an immediate check:

$$[n] = \text{m}^{-3}, \qquad [\sigma_{\text{in}}] = \text{m}^2, \tag{5.1498}$$

and hence

$$[n\sigma_{\text{in}}] = \text{m}^{-1}, \qquad [\Lambda_{\text{in}}] = \text{m}. \tag{5.1499}$$

The physical interpretation of Eq. (5.1496) is therefore simple: Λ_{in} is the characteristic distance over which the accumulated probability of an inelastic interaction becomes of order unity.

The differential inverse mean free path. There is a useful intermediate quantity between the microscopic double differential cross section and the total mean free path. Multiplying the cross section by the number density gives

$$\boxed{\frac{d^2\Lambda_{\text{in}}^{-1}}{d\Omega\, d\epsilon} \equiv n\frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon}.} \tag{5.1500}$$

This is the *double differential inverse inelastic mean free path.* It gives the probability per unit electron path length, per unit solid angle, and per unit energy loss.

Integrating it gives

$$\Lambda_{\rm in}^{-1} = \int_0^\infty d\epsilon \int_{4\pi} d\Omega \, \frac{d^2\Lambda_{\rm in}^{-1}}{d\Omega \, d\epsilon} \tag{5.1501}$$

$$= n \int_0^\infty d\epsilon \int_{4\pi} d\Omega \, \frac{d^2\sigma_{\rm in}}{d\Omega \, d\epsilon} \tag{5.1502}$$

$$= n\sigma_{\rm in}, \tag{5.1503}$$

recovering Eq. (5.1497).

This representation is particularly natural for a bulk material, because it expresses directly how much probability of each type of energy- and angle-resolved excitation is accumulated per unit distance travelled by the imaging electron.

Mixtures and biological specimens. The simple expression Eq. (5.1496) assumes one type of independent scattering entity. A biological specimen contains many chemical species and many kinds of excitation. If species a has number density n_a and total inelastic cross section $\sigma_{{\rm in},a}$, independent contributions add at the level of the inverse mean free path:

$$\boxed{\Lambda_{\rm in}^{-1} = \sum_a n_a \sigma_{{\rm in},a}.} \tag{5.1504}$$

This form is often more meaningful for heterogeneous matter than trying to assign a single microscopic "scatterer" to the entire material.

Collective excitations require a similar conceptual caution. A plasmon, for example, is not naturally associated with one isolated atom. In a bulk description the dynamic structure factor may already be normalized per unit volume, in which case the inverse mean free path is obtained directly from that macroscopic response rather than by introducing an additional atomic number density. The physical quantity is nevertheless the same: the integrated probability of inelastic scattering per unit path length.

Finite angular and energy acceptance. Equation (5.1483) defines the physical total inelastic cross section by integrating over all outgoing directions and all positive energy losses. An experiment, however, may collect only part of this domain.

For an angular acceptance region $\mathcal{A}$ and an energy-loss window $\mathcal{E}$, define

$$\boxed{\sigma_{\rm in}^{(\mathcal{A},\mathcal{E})} = \int_{\mathcal{E}} d\epsilon \int_{\mathcal{A}} d\Omega \, \frac{d^2\sigma_{\rm in}}{d\Omega \, d\epsilon}.} \tag{5.1505}$$

The corresponding experimentally selected mean free path is

$$\boxed{\Lambda_{\text{in}}^{(\mathcal{A},\mathcal{E})} = \frac{1}{n\sigma_{\text{in}}^{(\mathcal{A},\mathcal{E})}}.} \tag{5.1506}$$

Thus a mean free path inferred from an energy-loss experiment is not specified completely by the material and beam energy alone: the angular and energy windows used to define an "inelastic event" must also be stated when they do not include the full scattering distribution.

Why the collection angle matters especially for low-loss scattering.

For dipole-dominated low-loss scattering, Eq. (5.1475) gives

$$\frac{d\sigma_{\text{in}}}{d\Omega} \propto \frac{1}{\theta^2 + \theta_E^2}. \tag{5.1507}$$

For small angles,

$$d\Omega \simeq 2\pi\theta\, d\theta. \tag{5.1508}$$

If electrons are collected up to a semi-angle β, the contribution within that aperture is therefore

$$\sigma_{\text{in}}(\beta) \propto 2\pi \int_0^\beta \frac{\theta}{\theta^2 + \theta_E^2}\, d\theta \tag{5.1509}$$

$$= \pi \ln\left(\frac{\beta^2 + \theta_E^2}{\theta_E^2}\right). \tag{5.1510}$$

Hence,

$$\boxed{\sigma_{\text{in}}(\beta) \propto \pi \ln\left[1 + \left(\frac{\beta}{\theta_E}\right)^2\right].} \tag{5.1511}$$

The low-loss inelastic cross section therefore increases only logarithmically once the collection angle becomes much larger than θ_E.

Equation (5.1511) must not be extrapolated to arbitrarily large angles. The dipole and small-angle approximations used to derive Eq. (5.1475) eventually fail, and the full transition form factor must then be retained.

From mean free path to event statistics. We have now reached a macroscopic quantity starting entirely from the microscopic electron–specimen

interaction:

$$\boxed{\begin{array}{c} \widehat{H}_{\text{int}} \\ \Downarrow \\ F_{mn}(\mathbf{R}) \\ \Downarrow \\ M_{mn} \\ \Downarrow \\ \dfrac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon} \\ \Downarrow \\ \sigma_{\text{in}} \\ \Downarrow \\ \Lambda_{\text{in}} = \dfrac{1}{n\sigma_{\text{in}}}. \end{array}} \tag{5.1512}$$

Once Λ_{in} is known, the microscopic details of the individual excitation channels can be temporarily compressed into a single event rate per unit path length,

$$\boxed{\frac{1}{\Lambda_{\text{in}}}.} \tag{5.1513}$$

For a homogeneous specimen of thickness t, the corresponding mean number of inelastic interactions is therefore

$$\boxed{\mu_{\text{in}} = \frac{t}{\Lambda_{\text{in}}} = nt\sigma_{\text{in}}.} \tag{5.1514}$$

If successive interactions can be treated as independent, this quantity becomes the parameter of the Poisson distribution governing zero-, single-, and multiple-loss electrons. We shall return to that transport-level description when discussing the consequences of inelastic scattering for image formation. The purpose of the present section has been to establish where its fundamental parameter, Λ_{in}, comes from microscopically.

Numerical values for vitrified biological specimens. The inelastic mean free path is a macroscopic quantity, but its magnitude can be measured experimentally. For vitrified ice, values of a few hundred nanometres are typical at the electron energies used in cryo-EM.

The precise value is not universal. It depends on the incident electron energy, the composition and density of the specimen, and, for an experimentally determined mean free path, on the angular and energy acceptance

used to define an inelastic event. This distinction between total and partial mean free paths is important.

For example, at 120 keV, Grimm et al. measured

$$\Lambda_{\text{in}} = 203 \pm 33 \text{ nm} \tag{5.1515}$$

for vitrified ice (Grimm et al., 1996). Feja and Aebi obtained a total inelastic mean free path

$$\Lambda_{\text{in}} = 161 \text{ nm}, \tag{5.1516}$$

while the partial mean free path corresponding to an acceptance semi-angle of 4.2 mrad was

$$\Lambda_{\text{in}}^{(4.2\,\text{mrad})} = 232 \text{ nm} \tag{5.1517}$$

(Feja and Aebi, 1999). The larger partial mean free path follows because only a subset of all inelastic events contributes to the selected experimental cross section.

At 300 keV, values reported for vitrified ice are typically around 300–400 nm, depending on the experimental definition. Rice et al. measured an apparent value

$$\Lambda_{\text{in}}^{\text{app}} = 395 \pm 11 \text{ nm} \tag{5.1518}$$

for their Titan Krios/BioQuantum configuration (Rice et al., 2018). Other analyses give values close to 310–330 nm (Peet et al., 2019; Yesibolati et al., 2020). A recent analysis by Yu et al. obtained

$$\Lambda_{\text{in}} \simeq 303 \text{ nm} \tag{5.1519}$$

for vitreous ice at 300 keV after accounting for high-angle losses (Yu et al., 2025).

Thus, for order-of-magnitude calculations in cryo-EM, it is useful to remember

$$\boxed{\Lambda_{\text{in}}^{\text{ice}} \sim 300 \text{ nm} \qquad \text{at 300 keV.}} \tag{5.1520}$$

For quantitative thickness measurements, however, the value appropriate to the particular microscope geometry and energy-selection conditions should be used.

What does a 300 nm mean free path imply microscopically?.

As a numerical example, consider vitrified ice at 300 keV. Its density

is approximately

$$\rho_{\text{ice}} \simeq 0.94\ \text{g}\,\text{cm}^{-3} = 940\ \text{kg}\,\text{m}^{-3} \tag{5.1521}$$

(Yesibolati et al., 2020).

If one water molecule is taken as the effective scattering entity, its number density is

$$n = \frac{\rho_{\text{ice}} N_A}{M_{\text{H}_2\text{O}}}, \tag{5.1522}$$

where

$$M_{\text{H}_2\text{O}} = 18.015 \times 10^{-3}\ \text{kg}\,\text{mol}^{-1}. \tag{5.1523}$$

Hence

$$\boxed{n \simeq 3.14 \times 10^{28}\ \text{m}^{-3}.} \tag{5.1524}$$

Taking the representative value

$$\Lambda_{\text{in}} \simeq 310\ \text{nm} = 3.10 \times 10^{-7}\ \text{m}, \tag{5.1525}$$

Eq. (5.1496) gives

$$\sigma_{\text{in}} = \frac{1}{n\Lambda_{\text{in}}} \tag{5.1526}$$

$$\simeq \frac{1}{(3.14 \times 10^{28})(3.10 \times 10^{-7})} \tag{5.1527}$$

$$\simeq 1.03 \times 10^{-22}\ \text{m}^2. \tag{5.1528}$$

Since

$$1\ \text{Å}^2 = 10^{-20}\ \text{m}^2, \tag{5.1529}$$

this corresponds to

$$\boxed{\sigma_{\text{in}} \simeq 0.010\ \text{Å}^2} \tag{5.1530}$$

per water molecule.

This value should be regarded as an *effective* molecular cross section inferred from the macroscopic mean free path. The decomposition of the bulk response into independent water-molecule scatterers is an approximation, especially for collective electronic excitations such as plasmons.

5.12.10 Inelastic mean free path, zero-loss electrons, and effective absorption

The preceding microscopic treatment led from the individual electron–specimen transition amplitudes to the total inelastic cross section and finally to the inelastic mean free path,

$$\Lambda_{\text{in}} = \frac{1}{n\sigma_{\text{in}}}, \tag{5.1531}$$

as derived in Eq. (5.1496). Equivalently, the probability of an inelastic event per unit distance is

$$\Lambda_{\text{in}}^{-1}. \tag{5.1532}$$

We can now turn this microscopic result into a practical description of electrons traversing a specimen of finite thickness.

There are three populations that will be particularly useful:

$$\begin{aligned} \text{zero-loss electrons} &: \text{no inelastic interaction,} \\ \text{singly inelastically scattered electrons} &: \text{one inelastic interaction,} \\ \text{plural-scattered electrons} &: \text{two or more inelastic interactions.} \end{aligned} \tag{5.1533}$$

Poisson statistics of inelastic events. For a homogeneous specimen, the mean number of inelastic events along a path of thickness t was obtained in Eq. (5.1514):

$$\boxed{\mu_{\text{in}} = \frac{t}{\Lambda_{\text{in}}}.} \tag{5.1534}$$

If successive inelastic events are statistically independent and the event probability per unit path length is constant, their number follows a Poisson distribution.

Zero-, single-, and plural-loss probabilities.

A Poisson process with mean $\mu_{\text{in}} = t/\Lambda_{\text{in}}$ gives the probability of exactly N inelastic interactions as

$$\boxed{P_N(t) = \frac{1}{N!}\left(\frac{t}{\Lambda_{\text{in}}}\right)^N \exp\left(-\frac{t}{\Lambda_{\text{in}}}\right).} \tag{5.1535}$$

For $N = 0$,

$$\boxed{P_0(t) = \exp\left(-\frac{t}{\Lambda_{\text{in}}}\right).} \tag{5.1536}$$

This is the fraction of electrons that traverse the specimen without undergoing any inelastic interaction.

For exactly one inelastic event,

$$\boxed{P_1(t) = \frac{t}{\Lambda_{\text{in}}} \exp\left(-\frac{t}{\Lambda_{\text{in}}}\right).} \tag{5.1537}$$

Finally, the probability of plural inelastic scattering, that is, two or more events, is

$$P_{\geq 2}(t) = 1 - P_0(t) - P_1(t) \tag{5.1538}$$

$$= \boxed{1 - \left(1 + \frac{t}{\Lambda_{\text{in}}}\right) \exp\left(-\frac{t}{\Lambda_{\text{in}}}\right).} \tag{5.1539}$$

The complementary probability of suffering at least one inelastic interaction is

$$\boxed{P_{\geq 1}(t) = 1 - \exp\left(-\frac{t}{\Lambda_{\text{in}}}\right).} \tag{5.1540}$$

The zero-loss fraction therefore decreases exponentially with specimen thickness. This does not mean that all remaining electrons have been removed from the microscope. It means only that they have left the zero-inelastic-event channel.

Numerical example for vitrified ice. The scale is substantial at thicknesses routinely encountered in cryo-EM. Using, for illustration, an inelastic mean free path

$$\Lambda_{\text{in}} = 310 \text{ nm} \tag{5.1541}$$

for vitrified ice at 300 keV, representative of the values discussed in Sec. 5.12.9, we obtain

Thus, for 100 nm of ice, approximately 72% of the electrons have suffered no inelastic interaction, whereas for 300 nm of ice this fraction has fallen to approximately 38%. At the latter thickness, plural inelastic scattering is already a substantial population.

Table 5.6: Poisson probabilities for zero, exactly one, and two or more inelastic events for several ice thicknesses, using $\Lambda_{\text{in}} = 310$ nm.

t (nm)	t/Λ_{in}	P_0	P_1	$P_{\geq 2}$
50	0.161	0.851	0.137	0.012
100	0.323	0.724	0.234	0.042
200	0.645	0.525	0.338	0.137
300	0.968	0.380	0.368	0.252

What plural scattering means. Plural scattering does not represent a new elementary interaction. It means simply that the same imaging electron has undergone more than one of the microscopic transitions described in the preceding subsections.

If the individual energy losses are

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_N, \tag{5.1542}$$

the total energy lost by the electron is

$$\boxed{\epsilon_{\text{tot}} = \sum_{\ell=1}^{N} \epsilon_\ell.} \tag{5.1543}$$

Likewise, the successive momentum transfers modify the final direction of the electron.

Consequently, plural scattering broadens both the energy-loss and angular distributions. In an energy-loss spectrum, the distribution produced by two independent losses is the convolution of the single-loss distribution with itself; higher scattering orders produce successive convolutions. In an image, the corresponding electrons have experienced a broader range of energies, angles, and scattering histories than singly scattered electrons.

This is one reason why increasingly thick specimens become difficult to describe with a first-order single-scattering image model even when many transmitted electrons still reach the detector.

Zero-loss filtering. An energy filter placed after the specimen disperses electrons according to their energy and transmits only a selected energy interval. In *zero-loss imaging*, the energy-selecting slit is centred on the zero-loss peak.

In the idealized limit of infinitesimal energy resolution, a zero-loss filter would transmit only the $N = 0$ population of Eq. (5.1536). The transmitted

intensity would therefore satisfy

$$\boxed{\frac{I_{\mathrm{ZL}}(t)}{I_{\mathrm{inc}}} = P_0(t) = \exp\left(-\frac{t}{\Lambda_{\mathrm{in}}}\right).} \tag{5.1544}$$

A real energy filter has a finite slit width and finite energy resolution. Experimentally, "zero loss" therefore means an energy window around the elastic peak rather than the mathematical condition $\epsilon = 0$ exactly. Very small energy losses may consequently remain inside the selected window.

The logarithmic form of Eq. (5.1544),

$$\boxed{\frac{t}{\Lambda_{\mathrm{in}}} = \ln\left(\frac{I_{\mathrm{inc}}}{I_{\mathrm{ZL}}}\right),} \tag{5.1545}$$

also explains the basis of the familiar log-ratio approach to thickness estimation. In practice, the experimentally appropriate mean free path and the finite angular and energy acceptance must be used.

How an electron energy filter is constructed. The energy filter is a piece of electron optics placed *after* the specimen. Its purpose is to convert a difference in electron energy into a transverse displacement, so that electrons within a chosen energy interval can be selected physically by a slit.

The basic principle is the energy dependence of the trajectory of a charged particle in a magnetic field. For an electron moving perpendicularly to a uniform magnetic field, the relativistic cyclotron radius derived in Eq. (4.20) reduces to

$$\rho = \frac{p}{eB}, \tag{5.1546}$$

or equivalently

$$\boxed{p = eB\rho.} \tag{5.1547}$$

An electron that loses energy in the specimen also loses momentum. For a sufficiently small energy loss ϵ, we derived in Eq. (5.1438) that

$$\epsilon \simeq v\,\Delta p. \tag{5.1548}$$

Consequently, electrons with different energy losses follow slightly different trajectories through a magnetic dispersive element.

Electron lenses following the dispersive element are arranged so that, at a particular plane, electron position varies approximately linearly with energy loss:

$$\boxed{x(\epsilon) \simeq x_0 + D_E\epsilon.} \tag{5.1549}$$

The quantity

$$D_E = \frac{\partial x}{\partial \epsilon} \tag{5.1550}$$

is the *energy dispersion* of the filter, commonly quoted in units such as μm/eV.

An energy-selecting slit of physical width w is placed in this dispersive plane. In the first-order approximation of Eq. (5.1549), its corresponding energy width is

$$\boxed{\Delta\epsilon_{\text{slit}} \simeq \frac{w}{|D_E|}.} \tag{5.1551}$$

For zero-loss imaging the slit is centred on the zero-loss peak. Electrons whose dispersed trajectories fall inside the slit continue through the remaining electron-optical system, which removes the dispersion and reforms an energy-filtered image.

Two main geometries are used in transmission electron microscopy. In a *post-column imaging filter*, a magnetic prism located below the main microscope column bends the electrons and produces the energy-dispersed spectrum. Additional multipole electron optics focus the spectrum onto the selecting slit and then reconstruct the image (Gubbens et al., 2010). In an *in-column* filter, such as the Ω filter, several magnetic sectors are incorporated within the projector system. The electron trajectory is bent into an approximately Ω-shaped path and ultimately returned to the microscope optical axis (Reimer, 1995).

In both designs the physical principle is therefore the same:

$$\boxed{\begin{array}{c}\text{energy difference}\\ \Downarrow\\ \text{different momentum}\\ \Downarrow\\ \text{different trajectory in the dispersive optics}\\ \Downarrow\\ \text{different position at the energy slit.}\end{array}} \tag{5.1552}$$

The multipole elements of a modern imaging filter are important because a device intended for imaging must do more than disperse energies: after the desired energy interval has been selected, it must reconstruct the two-dimensional specimen image with sufficiently small geometrical and chromatic distortions.

A real energy filter is not an ideal projector onto zero loss. The idealized expression Eq. (5.1544),

$$\frac{I_{\mathrm{ZL}}}{I_{\mathrm{inc}}} = P_0(t), \tag{5.1553}$$

implicitly assumes a perfect measurement that accepts every electron with exactly zero energy loss and rejects every electron with $\epsilon > 0$.

A physical energy filter cannot make such an infinitely sharp selection. Even before considering instrumental imperfections, the finite slit width corresponds to a finite energy window. An idealized slit transmission function may be written

$$\boxed{T_{\mathrm{slit}}(\epsilon) = \begin{cases} 1, & |\epsilon - \epsilon_0| < \dfrac{\Delta\epsilon_{\mathrm{slit}}}{2}, \\ 0, & \text{otherwise}, \end{cases}} \tag{5.1554}$$

where $\epsilon_0 = 0$ for zero-loss filtering.

Thus, even an otherwise perfect filter with a finite slit does not strictly select

$$\epsilon = 0. \tag{5.1555}$$

It selects an *interval* of energies around zero.

The energy of an individual electron is also not measured with infinite resolution. Let

$$G(\epsilon_{\mathrm{meas}} - \epsilon) \tag{5.1556}$$

denote the normalized energy-response function of the complete source–spectrometer system: an electron whose true energy loss is ϵ is registered at the dispersed coordinate corresponding to ϵ_{meas} with probability density $G(\epsilon_{\mathrm{meas}} - \epsilon)$.

The effective probability that an electron with true energy loss ϵ passes the slit is therefore

$$\boxed{T_{\mathrm{eff}}(\epsilon) = \int_{-\infty}^{\infty} T_{\mathrm{slit}}(\epsilon') G(\epsilon' - \epsilon)\, d\epsilon'.} \tag{5.1557}$$

For a symmetric instrumental response, this is simply the convolution of the geometrical slit function with the energy-resolution function:

$$T_{\mathrm{eff}} = T_{\mathrm{slit}} * G. \tag{5.1558}$$

The sharp edges of Eq. (5.1554) are therefore rounded in a real instrument. Some true zero-loss electrons may fall outside the accepted interval, while some sufficiently small energy-loss electrons may fall inside it.

What a finite energy slit actually selects.

Let $L_1(\epsilon)$ be the normalized probability density for the energy lost in one inelastic interaction:

$$\int_0^\infty L_1(\epsilon)\, d\epsilon = 1. \tag{5.1559}$$

If the electron undergoes two statistically independent inelastic events, its total energy loss is the sum of the two losses, as discussed in Eq. (5.1543). Consequently, the two-event energy distribution is the convolution

$$L_2 = L_1 * L_1. \tag{5.1560}$$

More generally,

$$L_N = L_1^{*N}, \tag{5.1561}$$

where L_1^{*N} denotes the N-fold convolution. For zero inelastic events,

$$L_0(\epsilon) = \delta(\epsilon). \tag{5.1562}$$

Weighting the different scattering orders with the Poisson probabilities of Eq. (5.1535), the complete ideal energy-loss distribution after traversing thickness t is

$$\boxed{L_t(\epsilon) = e^{-\mu_{\text{in}}} \sum_{N=0}^{\infty} \frac{\mu_{\text{in}}^N}{N!} L_1^{*N}(\epsilon), \qquad \mu_{\text{in}} = \frac{t}{\Lambda_{\text{in}}}.} \tag{5.1563}$$

Here

$$L_1^{*0}(\epsilon) \equiv \delta(\epsilon). \tag{5.1564}$$

The coefficient of the zero-loss delta function is therefore

$$e^{-\mu_{\text{in}}} = e^{-t/\Lambda_{\text{in}}}, \tag{5.1565}$$

recovering Eq. (5.1536).

For a real energy filter, however, the fraction of electrons that pass the slit is not simply this coefficient. Instead,

$$\boxed{\eta_{\text{filter}}(t) = \int_{-\infty}^{\infty} T_{\text{eff}}(\epsilon) L_t(\epsilon)\, d\epsilon.} \tag{5.1566}$$

Only in the limiting case in which T_{eff} transmits the zero-loss term completely and rejects all $N \geq 1$ terms does Eq. (5.1566) reduce to

$$\eta_{\text{filter}}(t) = P_0(t) = e^{-t/\Lambda_{\text{in}}}. \tag{5.1567}$$

There are further departures from the one-dimensional model above. The transmission of an imaging energy filter can depend weakly on electron angle and on position within the image field as well as on energy. Geometrical and chromatic aberrations of the dispersive optics can make the energy selected by a fixed slit vary across the field, an effect usually referred to as imperfect *isochromaticity*. Modern imaging filters therefore employ multipole correction and careful alignment of the dispersive and achromatic planes (Gubbens et al., 2010).

At the most general level, the acceptance should consequently be regarded as a function

$$T = T(\epsilon, \Omega; x, y), \tag{5.1568}$$

rather than as a function of energy alone. For many cryo-EM applications the simpler approximation

$$T \simeq T_{\text{eff}}(\epsilon) \tag{5.1569}$$

is adequate, but the more general form clarifies that an experimental "zero-loss image" is an energy- and angle-selected image produced by a real electron-optical system, not a mathematical projection onto the single state $\epsilon = 0$.

An inelastically scattered electron has not disappeared. It is essential not to interpret Eq. (5.1544) as a loss of electrons from the complete physical system.

The channel structure derived in Sec. 5.12.7 showed that after the interaction the electron and specimen may occupy a state of the form

$$|\Psi_{\text{out}}\rangle = \sum_m |\psi_m\rangle_{\text{e}} \otimes |\Xi_m\rangle_{\text{s}}. \tag{5.1570}$$

The $m = n$ sector contains electrons for which the specimen has remained in the original internal state, whereas the $m \neq n$ sectors contain energy-loss electrons correlated with excited specimen states.

When the final specimen states are orthogonal, their probabilities add, as derived in Eq. (5.1391). Probability that leaves the zero-loss channel therefore appears in the inelastic channels:

$$\boxed{\text{attenuation of the zero-loss channel} \neq \text{destruction of the electron.}} \tag{5.1571}$$

An inelastically scattered electron can continue down the microscope column. If its angular deflection remains within the accepted aperture, it may pass through the objective lens and reach the detector. This is observed experimentally in unfiltered cryo-EM images (Langmore and Smith, 1992).

What differs is its state. It has a different energy and, generally, a different propagation direction. The angular distribution was derived in Sec. 5.12.8, while the electron energy after a specified loss was determined in Eq. (5.1248).

The different energy is particularly important for image formation. Chromatic aberration converts an energy change into a change of effective focus according to Eq. (4.1745). Therefore, even when an energy-loss electron reaches an unfiltered detector, its structural information is generally transferred differently from that of a zero-loss electron.

After averaging over many energy losses, scattering angles, and unobserved specimen channels, much of this intensity can appear as a diffuse or chromatically blurred background. This should not be confused with the stronger statement that every inelastically scattered electron has intrinsically lost its phase. As discussed in Sec. 5.12.7, an individual inelastic channel can still carry phase information.

Effective absorption and the optical potential. There is a compact way of representing the loss of probability from a selected channel without explicitly propagating every inelastic final state. One introduces a complex effective potential, often called an *optical potential*,

$$\boxed{V_{\text{eff}}(\mathbf{r}) = V_{\text{el}}(\mathbf{r}) + jV_{\text{abs}}(\mathbf{r}),} \tag{5.1572}$$

where, with the sign convention used in this book,

$$V_{\text{abs}}(\mathbf{r}) \geq 0 \tag{5.1573}$$

produces attenuation.

This type of effective complex potential has a long history in high-energy electron scattering (Peng et al., 1996). The terminology "absorptive" must, however, be interpreted carefully.

The full electron–specimen interaction is not being replaced by a fundamentally non-conserving theory. The complete quantum state belongs to the product Hilbert space

$$\boxed{\mathcal{H}_{\text{tot}} = \mathcal{H}_{\text{e}} \otimes \mathcal{H}_{\text{s}},} \tag{5.1574}$$

where $\mathcal{H}_{\mathrm{e}}$ describes the imaging electron and $\mathcal{H}_{\mathrm{s}}$ the internal degrees of freedom of the specimen.

For a specimen initially in the state $|\Xi_n\rangle_{\mathrm{s}}$, the elastic channel corresponds to the subspace

$$\boxed{\mathcal{H}_{\mathrm{el}} = \mathcal{H}_{\mathrm{e}} \otimes \mathrm{span}\left\{|\Xi_n\rangle_{\mathrm{s}}\right\}.} \tag{5.1575}$$

The complementary subspace contains the channels in which the specimen has been left in a different internal state,

$$|\Xi_m\rangle_{\mathrm{s}}, \qquad m \neq n. \tag{5.1576}$$

The complex optical potential is an effective description obtained when only the elastic, or more generally some selected, channel is propagated explicitly while the other channels are not retained. Probability transferred from the selected subspace into the omitted inelastic subspace then appears mathematically as attenuation of the selected wave.

The complete evolution in $\mathcal{H}_{\mathrm{tot}}$ still conserves probability. The apparent loss occurs only in the reduced description in which the omitted channels are no longer represented explicitly.

From the optical potential to exponential attenuation.

The fast-electron propagation equation was derived in Eq. (5.486):

$$\frac{\partial \psi}{\partial z} = \frac{j\lambda}{4\pi}\nabla_{\perp}^2\psi + j\sigma V(\mathbf{r})\psi, \tag{5.1577}$$

where the interaction constant σ was defined in Eq. (5.487).

Replace the real electrostatic potential by the effective complex potential of Eq. (5.1572):

$$\frac{\partial \psi}{\partial z} = \frac{j\lambda}{4\pi}\nabla_{\perp}^2\psi + j\sigma\left(V_{\mathrm{el}} + jV_{\mathrm{abs}}\right)\psi \tag{5.1578}$$

$$= \frac{j\lambda}{4\pi}\nabla_{\perp}^2\psi + j\sigma V_{\mathrm{el}}\psi - \sigma V_{\mathrm{abs}}\psi. \tag{5.1579}$$

The last term is real and negative. To isolate its meaning, neglect transverse propagation temporarily and consider a fixed ray through the

specimen. Equation (5.1579) then integrates to

$$\boxed{\psi(t) = \psi(0) \exp\left[j\sigma \int_0^t V_{\mathrm{el}}(z)\, dz\right] \exp\left[-\sigma \int_0^t V_{\mathrm{abs}}(z)\, dz\right].} \tag{5.1580}$$

The real part of the potential generates a phase shift, whereas the imaginary part attenuates the wave amplitude.

The corresponding intensity in this selected channel is

$$\boxed{\frac{I(t)}{I(0)} = \exp\left[-2\sigma \int_0^t V_{\mathrm{abs}}(z)\, dz\right].} \tag{5.1581}$$

For a homogeneous specimen, V_{abs} is constant. If the optical potential is chosen to represent removal from the zero-loss channel, Eq. (5.1581) must agree with the Poisson zero-loss probability Eq. (5.1536):

$$e^{-2\sigma V_{\mathrm{abs}} t} = e^{-t/\Lambda_{\mathrm{in}}}. \tag{5.1582}$$

Therefore,

$$\boxed{2\sigma V_{\mathrm{abs}} = \frac{1}{\Lambda_{\mathrm{in}}},} \tag{5.1583}$$

or

$$\boxed{V_{\mathrm{abs}} = \frac{1}{2\sigma \Lambda_{\mathrm{in}}}.} \tag{5.1584}$$

Thus the imaginary optical potential is another representation of the same microscopic event rate encoded by the inelastic mean free path.

The factor of two in Eq. (5.1583) is important. The mean free path describes attenuation of *intensity*, whereas the imaginary potential acts first on the complex wave *amplitude*. Thus,

$$\boxed{|\psi_{\mathrm{ZL}}(t)| = |\psi_{\mathrm{ZL}}(0)| \exp\left(-\frac{t}{2\Lambda_{\mathrm{in}}}\right),} \tag{5.1585}$$

while

$$\frac{I_{\mathrm{ZL}}(t)}{I_{\mathrm{inc}}} = \exp\left(-\frac{t}{\Lambda_{\mathrm{in}}}\right). \tag{5.1586}$$

Other sign conventions for the wave equation may write the optical potential as $V_{\mathrm{el}} - jW$ rather than $V_{\mathrm{el}} + jV_{\mathrm{abs}}$. What is physically invariant is that the imaginary contribution chosen to represent excluded channels must attenuate, rather than amplify, the selected wave.

Why "absorption" is only an effective description. The word "absorption" can be misleading in electron microscopy. The optical potential of Eq. (5.1572) does not imply

$$\text{electron} \longrightarrow \text{nothing}. \tag{5.1587}$$

Instead, it represents

$$\boxed{\begin{array}{c}\text{electron in the selected elastic/zero-loss channel}\\ \Downarrow \\ \text{interaction with unobserved specimen degrees of freedom}\\ \Downarrow \\ \text{electron in another scattering channel.}\end{array}} \tag{5.1588}$$

If we retain all those channels explicitly, the complete electron–specimen evolution conserves probability. If we project the problem onto only one channel and do not keep track of the others, its effective evolution is no longer norm preserving. The imaginary potential is a compact representation of that reduced description.

This is why an absorptive potential is useful computationally while remaining fully compatible with the fact that energy-loss electrons may still emerge from the specimen and reach the detector.

Microscopic origin of amplitude contrast. The optical-potential picture also closes a conceptual loop with the amplitude-contrast model introduced earlier in Sec. 4.12.

There the specimen exit wave was written, for a weak object, as

$$\psi_i(\mathbf{r}) \simeq 1 + a(\mathbf{r}) + j\varphi(\mathbf{r}) \tag{5.1589}$$

in Eq. (4.1767), where $a(\mathbf{r})$ is a real fractional amplitude modulation. At that point the microscopic origin of this term was deliberately left unspecified.

The complex potential now provides one such microscopic origin.

From an absorptive potential to the weak amplitude term.

For a thin specimen, define the projected elastic and absorptive potentials

$$V_{\text{el}}^{\text{proj}}(x, y) = \int V_{\text{el}}(x, y, z)\, dz, \tag{5.1590}$$

and

$$V_{\text{abs}}^{\text{proj}}(x,y) = \int V_{\text{abs}}(x,y,z)\,dz. \tag{5.1591}$$

Neglecting propagation within the specimen, the transmission function following from Eq. (5.1580) is

$$\boxed{t_{\text{eff}}(x,y) = \exp\left[-\sigma V_{\text{abs}}^{\text{proj}}(x,y)\right]\exp\left[j\sigma V_{\text{el}}^{\text{proj}}(x,y)\right].} \tag{5.1592}$$

For compactness, define the dimensionless projected absorptive and elastic interactions

$$\boxed{\alpha(x,y) \equiv \sigma V_{\text{abs}}^{\text{proj}}(x,y), \qquad \varphi(x,y) \equiv \sigma V_{\text{el}}^{\text{proj}}(x,y).} \tag{5.1593}$$

Equation (5.1592) then becomes

$$\boxed{t_{\text{eff}}(x,y) = e^{-\alpha(x,y)}e^{j\varphi(x,y)}.} \tag{5.1594}$$

For a weak specimen,

$$|\alpha(x,y)| \ll 1, \qquad |\varphi(x,y)| \ll 1. \tag{5.1595}$$

We now expand both exponential factors explicitly:

$$e^{-\alpha} = 1 - \alpha + \frac{\alpha^2}{2} + O(\alpha^3), \tag{5.1596}$$

$$e^{j\varphi} = 1 + j\varphi - \frac{\varphi^2}{2} + O(\varphi^3). \tag{5.1597}$$

Multiplying Eqs. (5.1596) and (5.1597) gives

$$t_{\text{eff}} = \left(1 - \alpha + \frac{\alpha^2}{2} + \cdots\right)\left(1 + j\varphi - \frac{\varphi^2}{2} + \cdots\right) \tag{5.1598}$$

$$= 1 - \alpha + j\varphi + \frac{\alpha^2}{2} - \frac{\varphi^2}{2} - j\alpha\varphi + O(3), \tag{5.1599}$$

where $O(3)$ denotes terms whose total order in α and φ is three or higher.

Thus the first-order terms are

$$-\alpha + j\varphi, \tag{5.1600}$$

whereas the terms

$$\frac{\alpha^2}{2}, \qquad -\frac{\varphi^2}{2}, \qquad -j\alpha\varphi \tag{5.1601}$$

are all second order.

The *weak complex-object approximation* consists precisely in discarding these quadratic and higher-order terms. Therefore,

$$\boxed{t_{\text{eff}}(x,y) \simeq 1 - \alpha(x,y) + j\varphi(x,y),} \tag{5.1602}$$

or, restoring the projected potentials,

$$\boxed{t_{\text{eff}}(x,y) \simeq 1 - \sigma V_{\text{abs}}^{\text{proj}}(x,y) + j\sigma V_{\text{el}}^{\text{proj}}(x,y).} \tag{5.1603}$$

The discarded terms also clarify what the first-order approximation is doing. The term

$$\frac{\alpha^2}{2} \tag{5.1604}$$

is the next correction to absorptive attenuation, while

$$-\frac{\varphi^2}{2} \tag{5.1605}$$

is the second-order correction to a pure phase object. The mixed term

$$-j\alpha\varphi \tag{5.1606}$$

couples absorption and phase.

In particular, the term $-\varphi^2/2$ should not be interpreted as physical absorption. If $\alpha = 0$, then

$$t_{\text{eff}} = e^{j\varphi} \tag{5.1607}$$

has exactly unit modulus,

$$|t_{\text{eff}}|^2 = 1. \tag{5.1608}$$

The real second-order term $-\varphi^2/2$ is simply required by the Taylor expansion of a unit-modulus complex exponential.

Comparing the first-order result Eq. (5.1603) with the weak amplitude–phase object introduced in Eq. (4.1772),

$$t(\mathbf{r}) \simeq 1 + a(\mathbf{r}) + j\varphi(\mathbf{r}), \tag{5.1609}$$

we identify

$$\boxed{a(x,y) \simeq -\sigma V_{\mathrm{abs}}^{\mathrm{proj}}(x,y),} \tag{5.1610}$$

and

$$\boxed{\varphi(x,y) \simeq \sigma V_{\mathrm{el}}^{\mathrm{proj}}(x,y).} \tag{5.1611}$$

Equation (5.1610) provides the microscopic justification for one of the amplitude-loss mechanisms anticipated in Sec. 4.12: regions with stronger inelastic scattering remove a larger fraction of the wave from the zero-loss coherent channel and therefore produce a negative amplitude modulation.

This is not the only source of amplitude contrast. Elastic electrons scattered to angles rejected by the objective aperture can also be removed from the image-forming wave, as discussed in Sec. 4.12. The optical potential introduced here specifically describes effective channel loss associated with inelastic processes, although more general absorptive models may incorporate other unresolved scattering mechanisms.

Filtered and unfiltered images are physically different. The distinction between actual electron loss and effective channel loss becomes particularly clear when comparing energy-filtered and unfiltered imaging.

With an ideal zero-loss filter,

$$\text{inelastic channel} \longrightarrow \text{rejected before detection.} \tag{5.1612}$$

The attenuation described by Eq. (5.1544) therefore corresponds directly to electrons missing from the recorded zero-loss image.

Without an energy filter, however,

$$\text{inelastic channel} \not\Longrightarrow \text{electron missing from detector.} \tag{5.1613}$$

Many energy-loss electrons remain within the angular acceptance of the microscope and are detected (Langmore and Smith, 1992). Their contribution is instead modified by their energy and angular distributions.

Schematically,

$$\boxed{\begin{array}{rcl} \text{zero-loss electron} & \longrightarrow & \text{nominal electron-optical transfer,} \\ \text{energy-loss electron} & \longrightarrow & \text{energy-dependent focus and phase transfer,} \\ \text{angularly deflected electron} & \longrightarrow & \text{modified spatial coherence / angular transfer,} \\ \text{electron rejected by filter or aperture} & \longrightarrow & \text{amplitude loss in the selected image.} \end{array}} \tag{5.1614}$$

A zero-loss energy filter therefore does more than select electrons that have "survived" the specimen. It selects a particular scattering channel whose electron-optical transfer is much more homogeneous. Energy-loss electrons that would otherwise contribute a broad, chromatically dispersed background are excluded from the detected image.

This distinction becomes increasingly important with increasing specimen thickness because Eq. (5.1540) shows that the population of energy-loss electrons grows rapidly as t/Λ_{in} increases.

A useful hierarchy of descriptions. We can now distinguish three levels of description of the same physical process:

$$\boxed{\begin{array}{rcl} \text{microscopic} & : & |\Xi_n\rangle \to |\Xi_m\rangle, \quad F_{mn}(\mathbf{R}), \\ \text{transport} & : & \sigma_{\text{in}}, \quad \Lambda_{\text{in}}, \quad P_N(t), \\ \text{effective wave optics} & : & V_{\text{eff}} = V_{\text{el}} + jV_{\text{abs}}. \end{array}} \tag{5.1615}$$

These are not competing theories. The cross section and mean free path summarize the microscopic transition probabilities, while the absorptive optical potential summarizes the removal of amplitude from a selected coherent channel.

The crucial qualification is therefore

$$\boxed{\text{effective absorption} \neq \text{physical disappearance of the electron.}} \tag{5.1616}$$

The electron may instead continue through the microscope in an energy-loss channel. Whether that electron contributes useful structural information, diffuse background, or no detected signal at all depends on the subsequent energy and angular selection and on the electron-optical transfer of that channel.

5.12.11 Quantitative balance of elastic and inelastic scattering

The preceding sections established the microscopic distinction between elastic and inelastic scattering and related their cross sections to the corresponding mean free paths. We now put numerical scales on these quantities and examine how they depend on electron energy and specimen thickness.

In this subsection, "radiation damage" will be used only in the restricted experimental sense of the loss of high-resolution diffraction intensity under irradiation. The microscopic chemical and structural mechanisms responsible for radiation damage will be considered separately later.

Energy dependence of elastic and inelastic scattering. Electron-scattering cross sections are commonly expressed in barns,

$$\boxed{1 \text{ barn} = 10^{-28} \text{ m}^2 = 10^{-8} \text{ Å}^2,} \tag{5.1617}$$

or in megabarns,

$$\boxed{1 \text{ Mb} = 10^{6} \text{ barn} = 10^{-22} \text{ m}^2 = 10^{-2} \text{ Å}^2.} \tag{5.1618}$$

A scattering cross section is an effective interaction area, not the literal geometrical area of the atom.

For carbon, a useful approximation to the total elastic cross section over the cryo-EM energy range is

$$\boxed{\sigma_{\mathrm{e}} \simeq \frac{2.1 \times 10^5}{\beta^2} \left[1 - \frac{1.6}{137\beta}\right] \text{ barn},} \tag{5.1619}$$

where

$$\beta = \frac{v}{c}. \tag{5.1620}$$

The dominant behavior is therefore approximately $\sigma_{\mathrm{e}} \propto 1/\beta^2$.

Representative values are

Table 5.7: Approximate total elastic cross section of carbon calculated from Eq. (5.1619).

E_k	β	σ_{e} (barn)	σ_{e} (Mb)
100 keV	0.548	6.84×10^5	0.684
200 keV	0.695	4.27×10^5	0.427
300 keV	0.777	3.43×10^5	0.343

Direct measurements between 100 and 300 keV give

$$\boxed{\frac{\sigma_{\mathrm{e}}(100 \text{ keV})}{\sigma_{\mathrm{e}}(300 \text{ keV})} = 2.01 \pm 0.05,} \tag{5.1621}$$

whereas the corresponding inelastic-scattering ratio is

$$\boxed{\frac{\sigma_{\mathrm{in}}(100 \text{ keV})}{\sigma_{\mathrm{in}}(300 \text{ keV})} = 1.51 \pm 0.10.} \tag{5.1622}$$

Thus both elastic and inelastic scattering become stronger as the electron energy is reduced, but elastic scattering increases more rapidly (Peet et al., 2019).

Connection with diffraction damage. The fading of diffraction from radiation-sensitive organic and biological crystals provides a simple operational measure of damage. If N_e denotes the electron fluence required for a diffraction intensity to fall to $1/e$ of its initial value, we may define

$$D \propto \frac{1}{N_e}. \tag{5.1623}$$

Experimentally,

$$\boxed{\frac{D(100\text{ keV})}{D(300\text{ keV})} = 1.57 \pm 0.03.} \tag{5.1624}$$

This scaling is close to the inelastic-scattering ratio 1.51 ± 0.10 and markedly different from the elastic-scattering ratio 2.01 ± 0.05. Within the conditions tested, the loss of high-resolution diffraction therefore follows the energy dependence of inelastic scattering much more closely than that of elastic scattering (Peet et al., 2019).

A useful consequence is obtained by comparing useful elastic scattering with this measured damage rate:

$$\frac{\sigma_{\mathrm{e}}(100)/D(100)}{\sigma_{\mathrm{e}}(300)/D(300)} \simeq \frac{2.01}{1.57} \simeq 1.28. \tag{5.1625}$$

Thus, from the interaction physics alone, 100 keV provides roughly 25–30% more elastic scattering per unit measured diffraction damage than 300 keV.

This does not by itself determine the optimum microscope voltage: detector performance, aberrations, specimen thickness, and multiple scattering must also be considered.

Specimen thickness converts cross sections into populations. The effect of a cross section becomes experimentally important through the mean free path. For inelastic scattering, Eq. (5.1496) gave

$$\Lambda_{\mathrm{in}} = \frac{1}{n\sigma_{\mathrm{in}}}, \tag{5.1626}$$

and the probability of at least one inelastic event after traversing a thickness t is, from Eq. (5.1540),

$$\boxed{P_{\geq 1}^{\mathrm{in}}(t) = 1 - \exp\left(-\frac{t}{\Lambda_{\mathrm{in}}}\right).} \tag{5.1627}$$

A particularly useful experimental example is amorphous carbon at 300 keV, for which an effective inelastic mean free path of approximately

$$\boxed{\Lambda_{\rm in}^{\rm C} \simeq 2400 \text{ Å}} \tag{5.1628}$$

was obtained.

Equation (5.1627) then predicts

$$P_{\geq 1}^{\rm in}(500 \text{ Å}) \simeq 1 - e^{-500/2400} = 0.188, \tag{5.1629}$$

$$P_{\geq 1}^{\rm in}(2000 \text{ Å}) \simeq 1 - e^{-2000/2400} = 0.565. \tag{5.1630}$$

The measured fractions of electrons with energy losses between 6 and 140 eV were approximately 19% and 56%, respectively (Dickerson et al., 2022). The agreement shows very directly how the mean-free-path description translates into observable electron populations.

For amorphous ice at 300 keV, a representative value is

$$\boxed{\Lambda_{\rm in}^{\rm ice} \simeq 3140 \text{ Å} = 314 \text{ nm}.} \tag{5.1631}$$

The corresponding populations are

Table 5.8: Expected zero-loss and inelastically scattered electron fractions in amorphous ice at 300 keV, using $\Lambda_{\rm in} = 3140$ Å.

Thickness	$P_0 = e^{-t/\Lambda_{\rm in}}$	$P_{\geq 1} = 1 - P_0$
300 Å	0.909	0.091
1000 Å	0.727	0.273
2000 Å	0.529	0.471
3000 Å	0.385	0.615
5000 Å	0.203	0.797

Thus in a thin single-particle specimen most electrons remain in the zero-loss population, whereas in several-thousand-Å cellular specimens inelastically scattered electrons can become the majority of the transmitted beam.

Typical energy loss and angular spread. At 300 keV, the low-loss spectrum of amorphous carbon has a broad maximum near

$$\boxed{\epsilon \simeq 23 \text{ eV}.} \tag{5.1632}$$

Using the characteristic-angle relation derived in Eq. (5.1439), this corresponds to approximately

$$\boxed{\theta_E \simeq 0.047 \text{ mrad.}} \tag{5.1633}$$

Although small on the scale of conventional objective apertures, this angle is much larger than the illumination angular spread of a high-coherence field-emission source under low-dose conditions. Consequently, for energy-loss electrons the specimen itself can become the dominant source of angular broadening.

The loss of phase information is gradual rather than abrupt. At 2.35 Å resolution, experimentally measured defocus ranges over which phase contrast survives decrease systematically with energy loss:

Table 5.9: Measured FWHM of the phase-contrast signal as a function of defocus for selected energy-loss electrons at 300 keV (Dickerson et al., 2022).

Energy loss	Defocus FWHM
9 eV	5600 ± 800 Å
15 eV	3400 ± 500 Å
23 eV	2600 ± 200 Å
30 eV	1800 ± 300 Å

This is consistent with the discussion in Sec. 5.12.8: larger energy losses imply larger momentum transfers and broader angular distributions. An inelastically scattered electron is therefore not automatically incoherent, but its useful phase contrast becomes progressively more restricted as the energy loss increases.

Elastic signal, inelastic loss, and specimen thickness. A compact way of illustrating the competition between these effects is to consider

$$\boxed{\zeta(E_k, t) = T(E_k, t)\frac{\sigma_{\mathrm{e}}(E_k)}{\sigma_{\mathrm{in}}(E_k)},} \tag{5.1634}$$

with

$$T(E_k, t) \simeq \exp\left[-\frac{t}{\Lambda_{\mathrm{tot}}(E_k)}\right]. \tag{5.1635}$$

This "information coefficient" is a simple interaction-based figure of merit rather than a general information-theoretic quantity (Peet et al., 2019).

Lower electron energies provide a larger $\sigma_{\mathrm{e}}/\sigma_{\mathrm{in}}$, which favours useful elastic scattering per inelastic event. At the same time, both cross sections increase, so transmission through a finite-thickness specimen falls more rapidly. The preferred electron energy from the point of view of beam–specimen interaction therefore depends on specimen thickness.

For thin specimens the improved elastic-to-inelastic ratio at lower energy dominates. As thickness increases, the growing probability of multiple elastic and inelastic scattering progressively shifts the balance toward higher electron energies.

5.12.12 Contribution of inelastic scattering to image formation and the power spectrum

We can now return to the image-formation model developed in the electron-optics chapter and ask where the inelastically scattered electrons appear in a recorded micrograph.

The empirical power-spectrum model introduced in Eq. (4.2162) was

$$\boxed{S_{\mathrm{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})\,|\mathrm{CTF}(\mathbf{R})|^2\,.} \tag{5.1636}$$

There, Eq. (4.2161) identified $A(\mathbf{R})$ with the slowly varying structural power produced by the particle, vitreous ice, support film, and other elastic specimen structure, after detector transfer.

The question is whether inelastically scattered electrons belong to the same CTF-modulated term or to the additive background $B(\mathbf{R})$.

The answer is more subtle than either identification. An inelastically scattered electron may retain phase information, but it generally belongs to a different quantum channel and experiences a different electron-optical transfer function.

Elastic and inelastic channels do not form one coherent wave. Suppose that the specimen begins in the internal state $|\Xi_n\rangle_{\mathrm{s}}$. After scattering, the combined electron–specimen state has the channel structure discussed in Sec. 5.12.7,

$$|\Psi_{\mathrm{out}}\rangle = |\psi_0\rangle_{\mathrm{e}} \otimes |\Xi_n\rangle_{\mathrm{s}} + \sum_{m\neq n} |\psi_m\rangle_{\mathrm{e}} \otimes |\Xi_m\rangle_{\mathrm{s}}, \tag{5.1637}$$

where the first term denotes the zero-loss elastic channel and the remaining terms denote inelastic channels.

The objective lens acts on the electron coordinate but does not erase the state of the specimen. Consequently, the different channels remain distinguishable through the orthogonal specimen states to which they are correlated.

Why elastic and inelastic image channels add in intensity.

Let $\widehat{\mathcal{L}}_m$ denote the electron-optical propagation operator appropriate to channel m. After passage through the microscope,

$$|\Psi_{\text{det}}\rangle = \widehat{\mathcal{L}}_0|\psi_0\rangle \otimes |\Xi_n\rangle + \sum_{m\neq n} \widehat{\mathcal{L}}_m|\psi_m\rangle \otimes |\Xi_m\rangle. \tag{5.1638}$$

The probability density for detecting the imaging electron at $\mathbf{r}$ without measuring the final specimen state is obtained by summing over that unobserved state. Cross terms between two different channels contain factors of the form

$$\langle \Xi_{m'}|\Xi_m\rangle. \tag{5.1639}$$

Since the final internal states are orthogonal,

$$\langle \Xi_{m'}|\Xi_m\rangle = \delta_{m'm}, \tag{5.1640}$$

all terms with $m \neq m'$ vanish.

Thus the detected intensity has the structure

$$\boxed{I_{\text{det}}(\mathbf{r}) = I_0(\mathbf{r}) + \sum_{m\neq n} I_m(\mathbf{r}).} \tag{5.1641}$$

The absence of the cross terms is a quantum-mechanical statement: elastic and mutually orthogonal inelastic specimen channels do not interfere with one another.

It does *not* imply that $I_m(\mathbf{r})$ contains no phase contrast. Within a fixed channel m, amplitudes leading to the same final specimen state may still interfere, as discussed in Sec. 5.12.7.

This distinction is important. We should therefore not describe an unfiltered image by adding an elastic CTF and an inelastic CTF at the *amplitude* level. The physically natural decomposition is instead a sum of channel intensities.

Each energy-loss channel has its own CTF. The complete elastic contrast transfer function was derived in Eq. (4.2117). Let us denote the zero-loss version by

$$\mathrm{CTF}_0(\mathbf{R}). \tag{5.1642}$$

An electron that has lost energy ϵ has a slightly different wavevector and therefore a different wavelength. In addition, chromatic aberration changes its effective defocus. The dependence of defocus on electron energy was derived in Eq. (4.1745).

We therefore denote the transfer function of an energy-loss channel by

$$\boxed{\mathrm{CTF}_\epsilon(\mathbf{R}) = \mathrm{CTF}\left(\mathbf{R}; \lambda_\epsilon, \Delta f + \delta f_\epsilon, C_s, \ldots\right).} \tag{5.1643}$$

For the small fractional energy losses normally encountered in cryo-EM, the change in wavelength is usually much less important than the chromatic defocus shift. Nevertheless,

$$\mathrm{CTF}_\epsilon \neq \mathrm{CTF}_0 \tag{5.1644}$$

in general.

There is a second difference. The angular spread produced by the inelastic interaction, derived in Sec. 5.12.8, causes additional specimen-induced loss of reproducible phase contrast. We represent this by an energy-dependent envelope

$$E_{\mathrm{SID}}(\mathbf{R}; \epsilon), \tag{5.1645}$$

where "SID" denotes specimen-induced decoherence.

A convenient channel transfer function is therefore

$$\boxed{G_\epsilon(\mathbf{R}) \equiv E_{\mathrm{SID}}(\mathbf{R}; \epsilon)\, \mathrm{CTF}_\epsilon(\mathbf{R}).} \tag{5.1646}$$

Elastic and inelastic contributions to the PSD. Let

$$S_0(\mathbf{R}) \tag{5.1647}$$

denote the structural power associated with the zero-loss channel before electron-optical transfer. Within the notation introduced in Sec. 4.16.3, let $S_s(\mathbf{R})$ denote the power spectral density of the reproducible specimen signal $s(\mathbf{r})$, for example the projected contrast of the particle of interest, and let $S_{\mathrm{el}}(\mathbf{R})$ denote the power spectral density of the elastic structural background $b_{\mathrm{el}}(\mathbf{r})$ arising from vitreous ice, support film, buffer, contaminants, and other specimen components. Here $\mathbf{R}$ is the two-dimensional spatial-frequency vector in the image plane.

If these two elastic contributions are treated as uncorrelated random fields, their power spectra add, as derived in Eq. (4.2157):

$$\boxed{S_0(\mathbf{R}) \equiv S_{\text{spec}}(\mathbf{R}) = S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R}).} \tag{5.1648}$$

Similarly, let

$$S_{\text{in}}(\mathbf{R}, \epsilon) \tag{5.1649}$$

denote the structural power carried by inelastic channels with energy loss ϵ, before their electron-optical transfer.

For each individual channel, the PSD propagation rule derived in Eq. (4.2133) gives

$$S_{\epsilon,\text{image}}(\mathbf{R}) = |G_\epsilon(\mathbf{R})|^2 \, S_{\text{in}}(\mathbf{R}, \epsilon). \tag{5.1650}$$

If the centered elastic and inelastic image contributions are treated as statistically uncorrelated, their power spectra add according to Eq. (4.2149). The structural part of the detected PSD is therefore, schematically,

$$\boxed{\begin{aligned} S_{\text{struct,det}}(\mathbf{R}) = \; & |\text{MTF}(\mathbf{R})|^2 \Bigg\{ |\text{CTF}_0(\mathbf{R})|^2 \, S_0(\mathbf{R}) \\ & + \int_0^\infty |E_{\text{SID}}(\mathbf{R}; \epsilon)|^2 \, |\text{CTF}_\epsilon(\mathbf{R})|^2 \, S_{\text{in}}(\mathbf{R}, \epsilon) \, d\epsilon \Bigg\}. \end{aligned}} \tag{5.1651}$$

Equation (5.1651) is the natural extension of the purely elastic expression Eq. (4.2159).

Strictly speaking, statistical independence of the *image intensity fields* is an additional approximation. Quantum orthogonality guarantees the absence of interference between the electron *amplitudes* associated with different specimen final states, but it does not by itself require all classical spatial correlations between their image intensities to vanish. If such correlations are important, cross-power spectra must be added to Eq. (5.1651). The independent-channel approximation provides the simplest useful PSD model.

Can inelastic scattering support the Thon rings? Equation (5.1651) answers this question rather directly.

In the empirical model of Eq. (4.2162), the smooth function $A(\mathbf{R})$ multiplies one particular oscillatory function,

$$|\text{CTF}_0(\mathbf{R})|^2 . \tag{5.1652}$$

It therefore acts as the smooth *support* of the zero-loss Thon rings.

The elastic term in Eq. (5.1651) has exactly this form:

$$\boxed{S_{\mathrm{el,det}}(\mathbf{R}) = A_{\mathrm{el}}(\mathbf{R})\,|\mathrm{CTF}_0(\mathbf{R})|^2\,,} \tag{5.1653}$$

with

$$A_{\mathrm{el}}(\mathbf{R}) = |\mathrm{MTF}(\mathbf{R})|^2\,S_0(\mathbf{R}). \tag{5.1654}$$

The inelastic contribution has instead the form

$$\boxed{\begin{aligned} S_{\mathrm{in,det}}(\mathbf{R}) = {} & |\mathrm{MTF}(\mathbf{R})|^2 \\ & \times \int_0^\infty |E_{\mathrm{SID}}(\mathbf{R};\epsilon)|^2\,|\mathrm{CTF}_\epsilon(\mathbf{R})|^2\,S_{\mathrm{in}}(\mathbf{R},\epsilon)\,d\epsilon. \end{aligned}} \tag{5.1655}$$

In general, Eq. (5.1655) cannot be written as

$$A_{\mathrm{in}}(\mathbf{R})\,|\mathrm{CTF}_0(\mathbf{R})|^2 \tag{5.1656}$$

with a slowly varying $A_{\mathrm{in}}(\mathbf{R})$.

The reason is that different values of ϵ have different CTF zeros.

Why broad inelastic scattering fills the Thon-ring minima.

Let $\mathbf{R}_0$ be a zero of the zero-loss CTF:

$$\mathrm{CTF}_0(\mathbf{R}_0) = 0. \tag{5.1657}$$

The elastic structural contribution then vanishes at that frequency,

$$S_{\mathrm{el,det}}(\mathbf{R}_0) = 0, \tag{5.1658}$$

apart from the additive background already present in the image.

For an energy-loss electron, however, chromatic aberration changes the effective defocus according to Eq. (4.1745). Consequently,

$$\mathrm{CTF}_\epsilon(\mathbf{R}_0) \neq 0 \tag{5.1659}$$

for a generic non-zero energy loss ϵ.

Equation (5.1655) therefore gives

$$\boxed{S_{\mathrm{in,det}}(\mathbf{R}_0) = |\,\mathrm{MTF}(\mathbf{R}_0)|^2 \int_0^\infty |E_{\mathrm{SID}}(\mathbf{R}_0;\epsilon)|^2 |\,\mathrm{CTF}_\epsilon(\mathbf{R}_0)|^2 S_{\mathrm{in}}(\mathbf{R}_0,\epsilon)\,d\epsilon,} \tag{5.1660}$$

which is generally non-zero.

Thus energy-loss channels fill the zeros of the zero-loss CTF rather than sharing them exactly.

When a broad distribution of energy losses is integrated, the oscillations of the different $|\,\mathrm{CTF}_\epsilon\,|^2$ functions are progressively out of register. Their sum is therefore much smoother than any individual channel.

This gives a microscopic interpretation of the empirical background introduced in Sec. 4.16. In a conventional microscope with appreciable chromatic aberration, much of the inelastic power appears as a slowly varying contribution underneath the zero-loss Thon rings, rather than as part of their common support.

The practical PSD model for a conventional microscope. It is therefore useful to write

$$\boxed{S_{\mathrm{obs}}(\mathbf{R}) \simeq B_{\mathrm{other}}(\mathbf{R}) + S_{\mathrm{in,det}}(\mathbf{R}) + A_{\mathrm{el}}(\mathbf{R})|\,\mathrm{CTF}_0(\mathbf{R})|^2,} \tag{5.1661}$$

where B_{other} contains shot noise, detector noise, and other contributions not represented explicitly.

For a sufficiently broad energy-loss distribution and conventional C_c,

$$S_{\mathrm{in,det}}(\mathbf{R}) \tag{5.1662}$$

varies much more slowly than $|\,\mathrm{CTF}_0\,|^2$. It can then be absorbed into an effective background,

$$\boxed{B_{\mathrm{eff}}(\mathbf{R}) = B_{\mathrm{other}}(\mathbf{R}) + S_{\mathrm{in,det}}(\mathbf{R}),} \tag{5.1663}$$

so that

$$\boxed{S_{\mathrm{obs}}(\mathbf{R}) \simeq B_{\mathrm{eff}}(\mathbf{R}) + A_{\mathrm{el}}(\mathbf{R})|\,\mathrm{CTF}_0(\mathbf{R})|^2.} \tag{5.1664}$$

This is precisely the empirical structure already introduced in Eq. (4.2162).

The physical interpretation is now sharper:

$$\boxed{\begin{array}{rcl} A_{\mathrm{el}}(\mathbf{R}) & : & \text{elastic structural power supporting the Thon rings,} \\ S_{\mathrm{in,det}}(\mathbf{R}) & : & \text{energy-loss structural power, usually strongly smeared,} \\ B_{\mathrm{other}}(\mathbf{R}) & : & \text{counting, detector, and other non-structural backgrounds.} \end{array}} \tag{5.1665}$$

The important exception: energy-selected or C_c-corrected imaging. The conclusion above is not a statement that inelastic electrons are intrinsically unable to produce Thon rings or phase contrast.

Suppose that only a narrow energy-loss interval is selected, or that chromatic aberration is corrected sufficiently well that

$$\mathrm{CTF}_\epsilon(\mathbf{R}) \simeq \mathrm{CTF}_0(\mathbf{R}) \tag{5.1666}$$

over the relevant range of energy losses. Then Eq. (5.1655) becomes approximately

$$\begin{aligned} S_{\mathrm{in,det}}(\mathbf{R}) \simeq\, & |\,\mathrm{CTF}_0(\mathbf{R})|^2 |\,\mathrm{MTF}(\mathbf{R})|^2 \\ & \times \int |E_{\mathrm{SID}}(\mathbf{R};\epsilon)|^2 S_{\mathrm{in}}(\mathbf{R},\epsilon)\, d\epsilon. \end{aligned} \tag{5.1667}$$

The inelastic contribution can then be factored by the same CTF and therefore becomes part of a generalized Thon-ring support:

$$\boxed{A_{\mathrm{eff}}(\mathbf{R}) = A_{\mathrm{el}}(\mathbf{R}) + A_{\mathrm{in}}(\mathbf{R}).} \tag{5.1668}$$

Energy-selected experiments directly demonstrate that inelastically scattered electrons can retain phase contrast (Dickerson et al., 2022). The contribution becomes progressively weaker with increasing energy loss because the angular distribution of the inelastic electrons broadens, producing the specimen-induced decoherence represented by E_{SID}.

Thus the correct statement is not

$$\text{inelastic scattering} \Longrightarrow \text{non-CTF background.} \tag{5.1669}$$

Rather,

$$\boxed{\begin{array}{c} \text{inelastic channel with reproducible phase transfer} \\ \Longrightarrow \text{CTF-modulated structural power,} \\[1em] \text{average over strongly different energies, angles, or unresolved channels} \\ \Longrightarrow \text{smooth power that fills the Thon-ring minima.} \end{array}} \tag{5.1670}$$

What does zero-loss filtering do to the power spectrum? A zero-loss energy filter removes most of $S_{\mathrm{in,det}}(\mathbf{R})$ before detection. The recorded structural PSD therefore approaches

$$\boxed{S_{\mathrm{struct,ZL}}(\mathbf{R}) \simeq |\,\mathrm{MTF}(\mathbf{R})|^2 |\,\mathrm{CTF}_0(\mathbf{R})|^2 S_0(\mathbf{R}).} \tag{5.1671}$$

The filter consequently makes the separation between the CTF-modulated elastic structural power and the smooth background much cleaner. This is one reason why Thon rings can become easier to interpret in energy-filtered images, particularly as specimen thickness and the fraction of energy-loss electrons increase.

An unfiltered detector, by contrast, still counts many inelastically scattered electrons. They contribute both the structural term of Eq. (5.1655), to the extent that phase information survives, and counting fluctuations associated with their detection. The latter belong to the shot-noise contribution rather than to the structural PSD itself.

5.13 Radiation damage and beam-induced specimen evolution

5.13.1 Irradiation changes the specimen while we are imaging it

Up to this point, the specimen has been described by a prescribed electrostatic potential

$$V(\mathbf{r}), \tag{5.1672}$$

and the main question has been how the imaging electron is scattered by that potential. This viewpoint is essential for image formation, but it hides an equally important fact: the same electron beam that forms the image also modifies the specimen while the image is being recorded.

It is therefore useful to make a conceptual shift:

$$\boxed{V(\mathbf{r}) \longrightarrow V(\mathbf{r}, \Phi),} \tag{5.1673}$$

where

$$\Phi \tag{5.1674}$$

denotes the accumulated electron exposure. The specimen potential is no longer strictly time independent in the experimental sense. Rather, it evolves as irradiation proceeds.

The three classes summarized in Table 5.10 can all degrade the recorded image, but they do so through different physical mechanisms. Structural damage changes the specimen itself; mechanical evolution changes its position and shape during acquisition; and charging modifies the electron wave through additional electrostatic fields.

Table 5.10: Main classes of beam-induced specimen changes relevant to cryo-EM. Although they may occur simultaneously, they affect image formation through different physical mechanisms.

Effect	**Examples and consequences**
Structural or chemical damage	Bond breaking, ionization, chemical modification, radical formation, mass loss, and progressive alteration of molecular and ice structure.
Beam-induced mechanical evolution	Motion of particles or larger specimen regions, doming of suspended ice films, creep, stress relaxation, and radiation-induced changes in mechanical stress.
Beam-induced electron-optical effects	Specimen charging, electrostatic lensing, beam deflection, and changes in the effective defocus or other low-order aberrations.

The previous sections on inelastic scattering have already identified the main microscopic route by which the beam deposits energy in the specimen. In particular, an inelastic event transfers energy

$$\epsilon \tag{5.1675}$$

and momentum

$$h\mathbf{R} \tag{5.1676}$$

to the specimen, as discussed in Sec. 5.12.3. This energy deposition, however, is only the beginning of a longer cascade. Electronic excitation and ionization may generate secondary electrons, radicals, local chemical rearrangements, and eventually visible structural change. A distinct but related pathway arises from sufficiently strong elastic collisions, which can transfer momentum directly to atomic nuclei and produce recoil or displacement.

Thus the phrase *radiation damage* should not be understood as if a primary electron caused one single elementary event. What the beam actually

initiates is a chain of processes:

$$\boxed{\begin{array}{c}\text{primary electron}\\ \Downarrow\\ \text{inelastic excitation/ionization} \quad \text{or} \quad \text{elastic nuclear recoil}\\ \Downarrow\\ \text{secondary electrons, radicals, atomic displacement}\\ \Downarrow\\ \text{chemistry, mass loss, stress, motion, charging}\\ \Downarrow\\ \text{loss or distortion of image information.}\end{array}} \tag{5.1677}$$

An important point for cryogenic biological specimens is that not only the macromolecule but also the embedding vitreous ice is radiation sensitive. The image therefore records a specimen environment that is itself evolving under irradiation. Later in this section we will analyze the main branches of this evolution separately: radiolysis and chemical damage, beam-induced motion, specimen charging, low-dose acquisition, and their consequences for the image formation model (Glaeser, 2016; Egerton, 2013).

5.13.2 Electron exposure, absorbed dose, and dose rate

Before discussing radiation damage quantitatively, it is important to distinguish several quantities that are often referred to loosely as "dose" in the cryo-EM literature.

The quantity most commonly controlled and reported in cryo-EM is the number of incident electrons per unit specimen area. More precisely, this is an *electron fluence*, although in cryo-EM it is usually called the *electron exposure*. We shall denote it by

$$\boxed{\Phi \equiv \frac{N_{\mathrm{e}}}{A},} \tag{5.1678}$$

where N_{e} is the number of incident electrons crossing an area A. The customary cryo-EM unit is

$$[\Phi] = \mathrm{e}^{-}\ \text{Å}^{-2}. \tag{5.1679}$$

The rate at which this exposure is delivered is the *electron fluence rate*, or *exposure rate*,

$$\boxed{\dot{\Phi} \equiv \frac{d\Phi}{dt},} \tag{5.1680}$$

with customary units

$$[\dot{\Phi}] = \mathrm{e}^- \,\text{Å}^{-2}\,\mathrm{s}^{-1}. \tag{5.1681}$$

For a constant exposure rate over an acquisition time T,

$$\boxed{\Phi = \dot{\Phi}\,T.} \tag{5.1682}$$

Neither Φ nor $\dot{\Phi}$ is, strictly speaking, a radiation dose. A physical radiation dose measures *energy deposited per unit mass*. The absorbed dose is

$$\boxed{D_{\mathrm{abs}} \equiv \frac{E_{\mathrm{deposited}}}{m},} \tag{5.1683}$$

whose SI unit is the gray,

$$\boxed{1\ \mathrm{Gy} = 1\ \mathrm{J\,kg^{-1}}.} \tag{5.1684}$$

Its time derivative,

$$\boxed{\dot{D}_{\mathrm{abs}} = \frac{dD_{\mathrm{abs}}}{dt},} \tag{5.1685}$$

is the physical *absorbed dose rate*, measured in $\mathrm{Gy\,s^{-1}}$.

The distinction is summarized in Table 5.11.

Table 5.11: Quantities used to characterize electron irradiation. In cryo-EM, "dose" is frequently used informally for electron exposure Φ, although the two quantities have different physical dimensions.

Quantity	**Symbol**	**Definition**	**Typical unit**
Electron fluence / exposure	Φ	N_{e}/A	$\mathrm{e}^-\,\text{Å}^{-2}$
Electron fluence rate / exposure rate	$\dot{\Phi}$	$d\Phi/dt$	$\mathrm{e}^-\,\text{Å}^{-2}\,\mathrm{s}^{-1}$
Absorbed dose	D_{abs}	$E_{\mathrm{deposited}}/m$	$\mathrm{Gy} = \mathrm{J\,kg^{-1}}$
Absorbed dose rate	$\dot{D}_{\mathrm{abs}}$	dD_{abs}/dt	$\mathrm{Gy\,s^{-1}}$

Thus, when a cryo-EM experiment is said to have used a "dose" of, for example,

$$40\ \mathrm{e}^-\,\text{Å}^{-2}, \tag{5.1686}$$

the stated quantity is normally the incident electron *exposure*, not the absorbed energy per unit mass.

From the microscopic inelastic cross section to stopping power. The connection between electron exposure and absorbed dose begins with the inelastic interactions developed in the preceding sections.

For a single scattering entity, the double differential inelastic cross section introduced in Eq. (5.1376),

$$\frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon}, \tag{5.1687}$$

gives the cross section per unit solid angle and per unit energy loss. Thus,

$$\frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon} d\Omega\, d\epsilon \tag{5.1688}$$

is the cross section for transferring an energy in $[\epsilon, \epsilon + d\epsilon]$ while the imaging electron is scattered into the solid-angle element $d\Omega$.

Let n be the number density of scattering entities, as introduced in Sec. 5.12.9. We can now ask not only how frequently an inelastic event occurs, but how much energy the primary electron loses, on average, per unit distance travelled through the specimen.

From the inelastic cross section to the stopping power.

Consider an infinitesimal path length dz. The probability of an inelastic event with energy loss between ϵ and $\epsilon + d\epsilon$ and scattering direction within $d\Omega$ is

$$dP = n\, dz\, \frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon} d\Omega\, d\epsilon. \tag{5.1689}$$

Such an event removes an energy ϵ from the primary electron. The corresponding mean energy loss over the distance dz is therefore

$$-\, d\langle E\rangle = n\, dz \int_0^\infty d\epsilon \int_{4\pi} d\Omega\, \epsilon \frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon}. \tag{5.1690}$$

Dividing by dz gives

$$\boxed{\mathcal{S}_{\text{loss}} \equiv -\frac{d\langle E\rangle}{dz} = n \int_0^\infty d\epsilon \int_{4\pi} d\Omega\, \epsilon \frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon}.} \tag{5.1691}$$

The quantity $\mathcal{S}_{\text{loss}}$ has units of energy per unit length and is the *stopping power* associated with the inelastic interactions. In the radiation-damage literature, a closely related quantity is commonly referred to as the *linear energy transfer* (LET).

The distinction from the total inelastic cross section is instructive:

$$\sigma_{\text{in}} = \int d\epsilon \int d\Omega \, \frac{d^2\sigma_{\text{in}}}{d\Omega \, d\epsilon} \tag{5.1692}$$

counts inelastic events, whereas Eq. (5.1691) weights each event by the energy ϵ that it transfers.

Thus two specimens could, in principle, have similar inelastic mean free paths while differing in the amount of energy transferred per interaction.

An equivalent representation follows from the differential inverse mean free path introduced in Sec. 5.12.9. If

$$\frac{d\Lambda_{\text{in}}^{-1}}{d\epsilon} = n \int_{4\pi} \frac{d^2\sigma_{\text{in}}}{d\Omega \, d\epsilon} d\Omega, \tag{5.1693}$$

then

$$\boxed{\mathcal{S}_{\text{loss}} = \int_0^\infty \epsilon \frac{d\Lambda_{\text{in}}^{-1}}{d\epsilon} d\epsilon.} \tag{5.1694}$$

Equation (5.1694) provides a particularly direct link between the energy-loss spectrum and radiation damage: it is the *first moment* of the energy-loss distribution per unit path length.

Energy lost by the primary electron is not necessarily energy deposited locally. There is an important qualification. The quantity $\mathcal{S}_{\text{loss}}$ describes energy removed from the primary electron. Not all of that energy necessarily remains in the local specimen volume.

For example, a sufficiently energetic secondary electron may travel a significant distance or may escape from a thin specimen altogether. It is therefore useful to distinguish the stopping power $\mathcal{S}_{\text{loss}}$ from the locally deposited energy per unit length,

$$\mathcal{S}_{\text{dep}}. \tag{5.1695}$$

We may write schematically

$$\boxed{\mathcal{S}_{\text{dep}} = f_{\text{dep}} \mathcal{S}_{\text{loss}}, \qquad 0 \le f_{\text{dep}} \le 1,} \tag{5.1696}$$

where f_{dep} depends on specimen composition, thickness, geometry, and the ranges of the secondary electrons.

This distinction will become important when we discuss the cascade of secondary electrons and radiolysis.

From stopping power to absorbed dose. We can now connect the microscopic energy-loss physics to the physical absorbed dose.

Absorbed dose produced by an electron exposure.

Consider a homogeneous specimen of thickness t, mass density ρ, and illuminated area A. Its mass is

$$m = \rho A t. \tag{5.1697}$$

An electron exposure Φ corresponds to

$$N_{\mathrm{e}} = \Phi A \tag{5.1698}$$

incident electrons.

If the deposited stopping power $\mathcal{S}_{\mathrm{dep}}$ is approximately constant through the thin specimen, the mean energy deposited by one primary electron is

$$E_{\mathrm{dep}}^{(1)} \simeq t\,\mathcal{S}_{\mathrm{dep}}. \tag{5.1699}$$

The total deposited energy is therefore

$$E_{\mathrm{deposited}} \simeq N_{\mathrm{e}} t \mathcal{S}_{\mathrm{dep}} \tag{5.1700}$$

$$= \Phi A t \mathcal{S}_{\mathrm{dep}}. \tag{5.1701}$$

Dividing by the specimen mass gives

$$D_{\mathrm{abs}} = \frac{E_{\mathrm{deposited}}}{\rho A t} \tag{5.1702}$$

$$\simeq \frac{\Phi}{\rho} \mathcal{S}_{\mathrm{dep}}. \tag{5.1703}$$

Hence

$$\boxed{D_{\mathrm{abs}} \simeq \frac{\Phi}{\rho} \frac{d\langle E_{\mathrm{dep}} \rangle}{dz}.} \tag{5.1704}$$

The cancellation of the specimen thickness in Eq. (5.1704) is worth noting. For a homogeneous material and constant local stopping power, a fixed electron exposure deposits approximately the same energy *per unit mass*, independently of thickness.

This statement is only approximate for very thin specimens, because the

fraction of energy carried away by escaping secondary electrons can depend strongly on thickness. In that case f_{dep} in Eq. (5.1696) is itself thickness dependent.

The enormous physical dose used in cryo-EM. The numerical scale is striking. For 300 keV electrons in vitreous ice, a representative mass stopping power is approximately

$$\boxed{\frac{\mathcal{S}_{\mathrm{loss}}}{\rho} \simeq 2.4\ \mathrm{MeV\,cm^2\,g^{-1}},} \tag{5.1705}$$

as discussed by Glaeser (2016).

Using

$$1\ \mathrm{e^-\,\mathring{A}^{-2}} = 10^{20}\ \mathrm{e^-\,m^{-2}}, \tag{5.1706}$$

and

$$1\ \mathrm{eV} = 1.602 \times 10^{-19}\ \mathrm{J}, \tag{5.1707}$$

Eq. (5.1704) gives, if all of the lost energy were deposited locally,

$$\boxed{D_{\mathrm{abs}} \simeq 3.8 \times 10^6\ \mathrm{Gy}\left(\frac{\Phi}{1\ \mathrm{e^-\,\mathring{A}^{-2}}}\right).} \tag{5.1708}$$

Thus an exposure of

$$\Phi = 10\ \mathrm{e^-\,\mathring{A}^{-2}} \tag{5.1709}$$

corresponds to the enormous scale

$$\boxed{D_{\mathrm{abs}} \simeq 3.8 \times 10^7\ \mathrm{Gy} = 38\ \mathrm{MGy},} \tag{5.1710}$$

before correcting for energy that escapes from the specimen in secondary electrons.

In the older radiation unit,

$$1\ \mathrm{rad} = 0.01\ \mathrm{Gy}, \tag{5.1711}$$

this is

$$D_{\mathrm{abs}} \simeq 3.8 \times 10^9\ \mathrm{rad}. \tag{5.1712}$$

The numerical magnitude explains why biological cryo-EM is intrinsically a radiation-limited experiment. Cryogenic preservation greatly slows the structural consequences of energy deposition, but it does not prevent the primary ionization and excitation events from occurring.

Exposure rate and absorbed dose rate. If the deposited stopping power is constant in time, Eq. (5.1704) can be differentiated to give

$$\boxed{\dot{D}_{\mathrm{abs}} \simeq \frac{\dot{\Phi}}{\rho}\mathcal{S}_{\mathrm{dep}}.} \tag{5.1713}$$

This equation makes an important distinction explicit:

$$\boxed{\begin{array}{lll} \dot{\Phi} & \text{is an electron exposure rate,} & \mathrm{e^{-}\,\text{Å}^{-2}\,s^{-1}}, \\ \dot{D}_{\mathrm{abs}} & \text{is an absorbed energy rate per unit mass,} & \mathrm{Gy\,s^{-1}}. \end{array}} \tag{5.1714}$$

Cryo-EM practitioners frequently call $\dot{\Phi}$ the "dose rate". We shall retain this terminology where it reflects common experimental usage, but will call it the *exposure rate* whenever the distinction matters.

The exposure rate is also related to the electrical beam-current density. If e denotes the magnitude of the electron charge,

$$\boxed{J_{\mathrm{beam}} = e\dot{\Phi}.} \tag{5.1715}$$

For example,

$$1\ \mathrm{e^{-}\,\text{Å}^{-2}\,s^{-1}} \tag{5.1716}$$

corresponds to

$$J_{\mathrm{beam}} \simeq 16.0\ \mathrm{A\,m^{-2}} = 1.60\ \mathrm{mA\,cm^{-2}}. \tag{5.1717}$$

Why the same accumulated exposure need not produce the same specimen state. If every effect of irradiation depended only on independent primary electron–specimen interactions, the accumulated exposure Φ would be sufficient to characterize the state of the specimen. Delivering

$$20\ \mathrm{e^{-}\,\text{Å}^{-2}} \tag{5.1718}$$

rapidly or slowly would then produce the same result.

For many aspects of primary chemical radiation damage, this is a useful first approximation at the relatively low current densities used in cryo-EM. It is not, however, generally true for all beam-induced specimen changes.

Charging, radical diffusion and recombination, stress relaxation, mechanical motion, heat transport, and electrical leakage all introduce their own characteristic time scales. Whenever such a time scale is comparable to the acquisition time, the instantaneous exposure rate can matter in addition to the accumulated exposure.

Why a dynamical beam-induced process can depend on exposure rate.

Consider a generic beam-induced quantity $q(t)$, such as a net charge, that is generated at a rate proportional to the incident electron exposure rate and relaxes with a characteristic time τ:

$$\boxed{\frac{dq}{dt} = \alpha\dot{\Phi} - \frac{q}{\tau}.} \tag{5.1719}$$

Here α measures how strongly each incident electron generates q.

For constant $\dot{\Phi}$ and initial condition $q(0) = 0$, the solution is

$$\boxed{q(t) = \alpha\tau\dot{\Phi}\left(1 - e^{-t/\tau}\right).} \tag{5.1720}$$

Since

$$t = \frac{\Phi}{\dot{\Phi}}, \tag{5.1721}$$

the same result at a fixed accumulated exposure Φ becomes

$$\boxed{q(\Phi; \dot{\Phi}) = \alpha\tau\dot{\Phi}\left[1 - \exp\left(-\frac{\Phi}{\dot{\Phi}\tau}\right)\right].} \tag{5.1722}$$

Thus two experiments with the same accumulated exposure Φ but different exposure rates $\dot{\Phi}$ need not leave the specimen in the same dynamical state.

Equation (5.1722) is only a minimal phenomenological example; specimen charging and mechanical motion are considerably more complicated. Its purpose is to show why a rate dependence is physically possible even when the same total number of electrons ultimately reaches the specimen.

This distinction has been observed directly in cryo-EM. The charging of vitreous ice evolves rapidly during the initial irradiation, and both the accumulated exposure required to approach a quasi-steady charge state and the resulting charge depend on the exposure rate. The charging is accompanied by physical movement of the ice film, and both effects modify the apparent defocus through electrostatic lensing and changes in specimen height (Schreiber et al., 2023a).

We shall therefore maintain the following distinction throughout this

chapter:

$$\boxed{\begin{array}{rcl} \Phi & \longrightarrow & \text{how many electrons have irradiated the specimen,} \\ D_{\text{abs}} & \longrightarrow & \text{how much energy has been deposited per unit mass,} \\ \dot{\Phi} & \longrightarrow & \text{how rapidly the irradiation has been delivered.} \end{array}} \tag{5.1723}$$

This distinction will be essential when we later discuss *low-dose acquisition*. Low dose means primarily that the finite available electron exposure is used economically; it does not, by itself, specify how rapidly that exposure is delivered.

5.13.3 Radiolysis: from an inelastic event to chemical damage

The preceding sections described an inelastic event as a quantum transition of the specimen,

$$|\Xi_n\rangle \longrightarrow |\Xi_m\rangle, \tag{5.1724}$$

accompanied by an electron energy loss

$$\epsilon_{mn} = E_m - E_n, \tag{5.1725}$$

as defined in Eq. (5.1240).

That description answers the scattering question: *how much energy and momentum did the imaging electron transfer?* It does not yet answer the radiation-damage question: *what subsequently happens to that energy inside the specimen?*

For biological materials, the dominant chemical damage produced by the electron beam is generally associated with *radiolysis*: structural change initiated by electronic excitation and ionization following inelastic scattering (Glaeser, 2016; Egerton, 2013).

The sequence is not usually

$$\text{one primary electron} \longrightarrow \text{one broken bond.} \tag{5.1726}$$

Instead, a single energy-loss event can initiate a cascade,

$$\boxed{\begin{array}{c}\text{valence excitation or collective electronic excitation}\\ \Downarrow\\ \text{secondary electrons and electron–hole excitations}\\ \Downarrow\\ \text{additional ionization and excitation}\\ \Downarrow\\ \text{radicals and altered bonding}\\ \Downarrow\\ \text{molecular fragmentation, rearrangement, or cross-linking.}\end{array}} \tag{5.1727}$$

The different stages of this cascade occur on very different spatial and temporal scales. The initial electronic excitation is essentially instantaneous on the scale of atomic motion, whereas secondary-electron thermalization, radical formation, chemical reaction, and eventual structural rearrangement occur progressively later.

Electronic excitation does not immediately imply structural damage. An inelastic transition first changes the *electronic* state of the specimen. For a molecule M, two elementary possibilities may be written schematically as

$$\boxed{M + e_{\mathrm{p}}^{-} \longrightarrow M^{*} + e_{\mathrm{p}}^{-},} \tag{5.1728}$$

for excitation without ionization, and

$$\boxed{M + e_{\mathrm{p}}^{-} \longrightarrow M^{+\bullet} + e_{\mathrm{p}}^{-} + e_{\mathrm{s}}^{-},} \tag{5.1729}$$

for ionization.

Here

$$e_{\mathrm{p}}^{-} \tag{5.1730}$$

denotes the fast *primary* imaging electron and

$$e_{\mathrm{s}}^{-} \tag{5.1731}$$

a *secondary electron* released within the specimen.

The difference in their characteristic energies is enormous. In modern cryo-EM, the primary electron typically enters the specimen with kinetic energy

$$\boxed{E_{\mathrm{p}} \simeq 200\ \mathrm{keV} \quad \text{or} \quad 300\ \mathrm{keV}.} \tag{5.1732}$$

A typical valence or collective inelastic event, however, transfers only of order

$$\boxed{\epsilon \sim 20\text{–}30 \text{ eV}} \tag{5.1733}$$

in organic matter (Egerton, 2013). Thus a 300 keV primary electron that loses, for example, 25 eV leaves the interaction with kinetic energy

$$300000 \text{ eV} - 25 \text{ eV} = 299975 \text{ eV}. \tag{5.1734}$$

Its fractional energy loss is only

$$\boxed{\frac{25 \text{ eV}}{300 \text{ keV}} \simeq 8.3 \times 10^{-5}.} \tag{5.1735}$$

The primary electron therefore remains a very fast electron even though the transferred energy can be chemically very important.

The secondary electrons produced by the interaction have much lower energies. Their distribution is broad rather than monoenergetic. For radiolytic processes in organic matter it is useful to distinguish roughly

$$\boxed{\begin{array}{rcl} E_{\mathrm{s}} \lesssim 10 \text{ eV} & : & \text{slow secondary electrons,} \\ E_{\mathrm{s}} \sim 10\text{–}50 \text{ eV} & : & \text{common low-energy secondaries,} \\ E_{\mathrm{s}} \gtrsim 50 \text{ eV} & : & \text{fast secondary electrons,} \end{array}} \tag{5.1736}$$

with secondary-electron kinetic energies extending to approximately 100 eV in typical electron-stimulated processes (Egerton et al., 2004; Hettler et al., 2018a).

Much more energetic secondary electrons can be produced by inner-shell ionization. For example, carbon K-shell ionization can generate an electron with kinetic energy of several hundred eV, and the subsequent Auger decay produces an electron with energy of approximately

$$\boxed{E_{\mathrm{Auger}} \simeq 270 \text{ eV}.} \tag{5.1737}$$

Such events are much less frequent than valence excitations, but each can initiate a substantial secondary-electron cascade (Egerton, 2013).

The relevant energy hierarchy is therefore approximately

$$\boxed{300 \text{ keV} \gg 20\text{–}30 \text{ eV} \gtrsim \text{a few-eV chemical bond energy.}} \tag{5.1738}$$

The imaging electron loses only a tiny fraction of its kinetic energy, yet that small loss is already several times the energy scale associated with an ordinary covalent bond.

The notation $M^{+\bullet}$ emphasizes that ionization generally leaves behind both positive charge and an unpaired electron configuration. Such a state is chemically reactive, but structural damage occurs only after the electronic excitation has been converted into changes of bonding or nuclear coordinates.

This distinction is important. The electron-scattering calculation describes the transition

$$|\Xi_n\rangle \to |\Xi_m\rangle, \tag{5.1739}$$

whereas radiolysis concerns the subsequent evolution of $|\Xi_m\rangle$ through a much larger network of electronic, vibrational, and chemical states.

Valence excitation and collective electronic losses. The most frequent energy losses in biological specimens involve valence electrons.

At relatively small energy losses, individual electronic transitions may be excited. At somewhat higher losses, many valence electrons can participate collectively. In conducting materials such a collective mode is normally called a *plasmon*. The same terminology is also commonly used for the broad collective or dielectric-loss feature in insulating biological materials, although "collective dielectric excitation" is physically more general.

Representative energy-loss scales are summarized in Table 5.12.

Table 5.12: Representative electronic energy-loss regimes in biological materials. The boundaries are approximate and depend on composition.

Process	**Typical energy scale**	**Immediate consequence**
Low-energy electronic excitation	few eV	Excitation of molecular electronic states without necessarily producing ionization.
Valence / collective dielectric loss	$\sim$ 10–100 eV	Collective or individual valence-electron excitation followed by decay into lower-energy electronic excitations and secondary electrons.
Inner-shell ionization	hundreds of eV for C, N, and O	Creation of a core hole followed predominantly, for low-Z elements, by Auger-electron emission and further electronic excitation.

The broad low-loss feature of organic and biological matter generally lies in the tens-of-electronvolts range (Glaeser, 2016; Egerton, 2013). This scale

should be compared with ordinary covalent-bond energies, which are only a few electronvolts.

Thus a typical primary inelastic event can transfer enough energy to alter several chemical bonds. It does not follow, however, that all of this energy is deposited into one bond. The energy is redistributed among many possible degrees of freedom.

Energy bookkeeping after one inelastic event.

Suppose that the primary imaging electron loses an energy ϵ. Energy conservation requires that this energy appear elsewhere in the combined specimen and secondary-electron system.

Schematically, we may write

$$\boxed{\epsilon = E_{\text{electronic}} + E_{\text{secondary}} + E_{\text{vibrational}} + E_{\text{chemical}} + E_{\text{escape}}.} \tag{5.1740}$$

Here:

- $E_{\text{electronic}}$ is energy remaining in electronic excitations and ionized states;
- $E_{\text{secondary}}$ is kinetic energy carried by secondary electrons while they remain within the specimen;
- $E_{\text{vibrational}}$ is energy transferred to molecular vibrations and lattice excitations;
- E_{chemical} is energy stored in altered bonding, radicals, fragments, and other chemically modified states;
- E_{escape} is energy carried away from the local specimen volume, for example by a secondary electron escaping into vacuum.

Equation (5.1740) is an energy balance, not a claim that these contributions are independent or uniquely separable at every instant. Energy may flow successively from one category to another as the radiation cascade develops.

Consequently,

$$\boxed{\text{energy deposited} \neq \text{one localized broken bond.}} \tag{5.1741}$$

This distinction also explains why the stopping power derived in Eq. (5.1691) cannot by itself predict the precise structural damage. Stopping power quantifies how rapidly energy is transferred from the primary electron. Radiation chemistry determines how that energy is subsequently redistributed.

Secondary electrons multiply the consequences of a primary event. The secondary electron in Eq. (5.1729) can itself undergo inelastic interactions.

If its kinetic energy is sufficiently high, it can ionize or excite another molecule,

$$e_{\mathrm{s}}^{-} + M' \longrightarrow e_{\mathrm{s}'}^{-} + M'^{*} \tag{5.1742}$$

or

$$e_{\mathrm{s}}^{-} + M' \longrightarrow e_{\mathrm{s}'}^{-} + e_{\mathrm{s}''}^{-} + M'^{+\bullet}. \tag{5.1743}$$

The secondary electrons produced in this way may undergo further collisions until their energies become too small to generate additional ionizations. The result is a branching cascade:

$$\boxed{\text{1 primary inelastic event} \quad \longrightarrow \quad \text{many lower-energy electronic events.}} \tag{5.1744}$$

For organic matter, typical valence energy losses are of order 20–30 eV, whereas only a few electronvolts are needed to alter a covalent bond. Secondary-electron production is therefore an important route by which the energy transferred in one primary inelastic event becomes distributed over several nearby molecular sites (Egerton, 2013).

This is also why radiation damage need not be localized exactly at the position at which the fast primary electron underwent its original inelastic interaction. Secondary electrons can travel away from that position before transferring their energy.

Direct damage to the macromolecule. A radiation event is called *direct* when the relevant electronic excitation or ionization occurs in the macromolecule itself.

For example,

$$P + e_{\mathrm{p}}^{-} \longrightarrow P^{+\bullet} + e_{\mathrm{p}}^{-} + e_{\mathrm{s}}^{-}, \tag{5.1745}$$

where P denotes the protein or other biological macromolecule.

The resulting molecular ion may relax through several competing pathways. Schematically,

$$\boxed{P^{+\bullet} \longrightarrow \begin{cases} \text{radical formation,} \\ \text{bond cleavage,} \\ \text{fragmentation,} \\ \text{cross-linking,} \\ \text{local conformational rearrangement.} \end{cases}} \tag{5.1746}$$

Which pathway occurs depends strongly on the local chemical environment. For this reason radiation damage is not spatially or chemically uniform, a point to which we shall return when discussing site-specific radiation sensitivity.

Indirect damage through the surrounding water. The macromolecule is not irradiated in isolation. In single-particle cryo-EM it is embedded in vitreous water, and water itself undergoes radiolysis.

A primary ionization event may be written schematically as

$$\mathrm{H_2O} + e^-_{\mathrm{p}} \longrightarrow \mathrm{H_2O}^{+\bullet} + e^-_{\mathrm{p}} + e^-_{\mathrm{s}}. \tag{5.1747}$$

The water radical cation is highly unstable and can transfer a proton to a neighbouring water molecule,

$$\boxed{\mathrm{H_2O}^{+\bullet} + \mathrm{H_2O} \longrightarrow \mathrm{H_3O}^+ + \mathrm{OH}^\bullet.} \tag{5.1748}$$

The released electron loses energy through further interactions and, in liquid water, can ultimately form a solvated electron,

$$e^-_{\mathrm{aq}}. \tag{5.1749}$$

Other pathways generate hydrogen radicals,

$$\mathrm{H}^\bullet, \tag{5.1750}$$

and subsequent reactions produce species such as

$$\mathrm{H_2}, \qquad \mathrm{H_2O_2}. \tag{5.1751}$$

The early radiation chemistry of water may therefore be summarized schematically as

$$\boxed{\mathrm{H_2O} \overset{\text{ionizing radiation}}{\longrightarrow} e^-_{\mathrm{aq}},\ \mathrm{H}^\bullet,\ \mathrm{OH}^\bullet,\ \mathrm{H_3O}^+,\ \mathrm{OH}^-,\ \mathrm{H_2},\ \mathrm{H_2O_2}, \ldots} \tag{5.1752}$$

Reactive intermediates are produced extremely rapidly; in liquid water, the principal early radical products emerge within approximately the first picosecond following ionization (Glaeser, 2016).

If such a reactive product subsequently attacks the macromolecule,

$$\mathrm{OH}^{\bullet} + P \longrightarrow \text{chemically modified } P, \tag{5.1753}$$

the resulting structural damage is called *indirect radiation damage*: the energy was initially deposited in the solvent rather than in the macromolecule itself.

Thus,

$$\boxed{\begin{array}{rcl} \text{direct damage} & : & \text{energy deposited in the macromolecule,} \\ \text{indirect damage} & : & \text{energy deposited elsewhere, followed by chemical attack.} \end{array}} \tag{5.1754}$$

This distinction is especially important when comparing irradiation in liquid water with irradiation in vitreous ice. At cryogenic temperature, the primary ionization and radical-production processes still occur, but the diffusion and secondary chemistry of many reactive species are strongly suppressed. We shall examine this cryoprotective "cage effect" in the next subsection. In frozen specimens, direct damage to the macromolecule therefore becomes relatively more important than it would be in a mobile aqueous environment (Glaeser, 2016).

Collective excitation is an intermediate, not the final damage. The strong low-loss feature of biological materials is often described as a plasmon or collective dielectric excitation. It is important not to interpret this collective mode itself as the chemical lesion.

Rather,

$$\boxed{\begin{aligned} &\text{collective excitation} \longrightarrow \text{electronic relaxation} \\ &\qquad\qquad\qquad\qquad \longrightarrow \text{ionization and secondary electrons} \longrightarrow \text{chemistry.} \end{aligned}} \tag{5.1755}$$

The collective excitation is therefore an efficient mechanism for placing energy into the electronic degrees of freedom of the specimen. The subsequent degradation of that electronic excitation into localized electron–hole excitations, secondary electrons, radicals, and molecular rearrangements is what converts the energy loss into structural damage.

This explains why the broad valence-loss region is so important for biological cryo-EM even though the individual imaging electron has lost only a tiny fraction of its initial kinetic energy.

Inner-shell ionization: rarer but more energetic events. A different branch of the inelastic spectrum occurs when the incident electron ejects a tightly bound inner-shell electron.

For a low-Z atom such as carbon, nitrogen, or oxygen, the result is a core-ionized state,

$$A \longrightarrow A^{+\bullet}_{\text{core}} + e^{-}_{\text{s}}. \tag{5.1756}$$

The core vacancy is unstable. A valence electron may fill it, releasing the excess energy through the emission of an Auger electron:

$$\boxed{A^{+\bullet}_{\text{core}} \longrightarrow A^{2+} + e^{-}_{\text{Auger}}.} \tag{5.1757}$$

The Auger electron can carry hundreds of electronvolts and can itself produce a secondary ionization cascade.

Core-loss events are much less frequent than valence and collective losses in light-element biological matter, but each individual event transfers substantially more energy and can produce several subsequent ionizations (Glaeser, 2016; Egerton, 2013). They therefore illustrate particularly clearly why the amount of chemical damage cannot always be inferred simply by counting the number of primary inelastic events.

Why deposited energy alone does not specify the chemical outcome. The absorbed dose introduced in Sec. 5.13.2 measures the amount of energy deposited per unit mass. It is indispensable for quantifying irradiation, but it compresses a complicated microscopic process into one number.

Two deposits of the same total energy need not produce exactly the same chemical changes. The outcome depends on, among other factors,

- whether the excitation is localized or collective;
- whether ionization occurs in the macromolecule or in the surrounding water;
- the energy and range of the secondary electrons;
- the chemical group that is ionized;
- whether a core hole or a valence hole is produced;
- whether radicals can diffuse, recombine, or remain trapped.

Radiation damage should therefore be viewed as a stochastic branching process rather than a deterministic conversion of deposited energy into broken bonds.

Radiation-chemical yield. A useful way of quantifying this branching is the *radiation-chemical yield*, traditionally called the G-value. For a chemical species or event X,

$$\boxed{G_X \equiv \frac{N_X}{E_{\text{abs}}/(100\ \text{eV})},} \tag{5.1758}$$

where N_X is the number of molecules of X produced or destroyed and E_{abs} is the absorbed energy.

The conventional units of G_X are therefore

$$\text{molecules per 100 eV}. \tag{5.1759}$$

For a deposited energy E_{abs}, the expected number of events of type X is

$$\boxed{N_X = G_X \frac{E_{\text{abs}}}{100\ \text{eV}}.} \tag{5.1760}$$

The G-value is not universal: it depends on the material, the chemical species being monitored, temperature, and experimental conditions. Its importance here is conceptual. It makes explicit that radiation chemistry is described by a *yield per deposited energy*, rather than by the assumption that one energy-loss event produces one specific molecular lesion (Glaeser, 2016).

A hierarchy of time scales. The complete radiation-damage cascade spans many orders of magnitude in time. A useful schematic hierarchy is shown in Table 5.13.

This separation of time scales is one of the reasons cryogenic cooling can strongly affect the eventual structural consequences of radiation without appreciably altering the probability of the initial inelastic scattering event.

From radiation chemistry back to the image-formation model. The image-formation model ultimately depends on the specimen electrostatic potential. Radiolysis makes that potential exposure dependent,

$$V(\mathbf{r}) \longrightarrow V(\mathbf{r}, \Phi), \tag{5.1761}$$

as anticipated in Eq. (5.8).

A useful schematic decomposition is

$$\boxed{V(\mathbf{r}, \Phi) = V_0(\mathbf{r}) + \delta V_{\text{chem}}(\mathbf{r}, \Phi),} \tag{5.1762}$$

Table 5.13: Approximate hierarchy of processes following an inelastic interaction. The times are indicative rather than universal and depend strongly on material and temperature.

Stage	**Approximate scale**	**Processes**
Primary electronic event	fs and below	Valence excitation, ionization, collective electronic excitation.
Electronic cascade	fs–ps	Secondary-electron production, further excitation and ionization, Auger decay, electron thermalization.
Early radiation chemistry	ps and beyond	Radical formation, proton transfer, altered bonding, early solvent radiolysis.
Structural evolution	longer and strongly temperature dependent	Fragment displacement, radical reactions, molecular rearrangement, mass loss, stress generation, and eventually visible specimen damage.

where V_0 is the potential of the initially vitrified specimen and δV_{chem} contains the accumulated changes caused by ionization, bond modification, radical formation, fragmentation, and other radiolytic processes.

The important point is that δV_{chem} is generally neither uniform nor a simple global attenuation of V_0. Radiation damage changes different chemical groups and different spatial frequencies at different rates. Later subsections will make this statement quantitative by examining site-specific damage and the exposure-dependent fading of Fourier components.

For now, the microscopic chain may be summarized as

$$\boxed{\begin{array}{c} \dfrac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon} \\ \Downarrow \\ \text{electronic excitation and ionization} \\ \Downarrow \\ \text{secondary-electron cascade and radicals} \\ \Downarrow \\ \text{chemical and structural modification} \\ \Downarrow \\ V_0(\mathbf{r}) \longrightarrow V(\mathbf{r}, \Phi). \end{array}} \tag{5.1763}$$

Radiolysis is therefore the physical bridge between the microscopic inelastic-scattering theory developed earlier and the progressive loss of structural information observed during cryo-EM acquisition.

5.13.4 Why cooling helps: radical trapping and the cage effect

Cryogenic cooling is one of the essential ingredients of biological electron microscopy for two distinct reasons.

First, it makes it possible to preserve a hydrated biological specimen inside the high vacuum of the electron-microscope column. At ordinary temperatures, liquid water cannot remain in equilibrium under such low pressure: it would rapidly evaporate, dehydrating and destroying the native environment of the macromolecule. By rapidly vitrifying the sample and maintaining it at cryogenic temperature, the water is immobilized as vitreous ice and its vapour pressure becomes sufficiently low for the specimen to remain stable in the microscope vacuum.

Second, cryogenic cooling substantially reduces the structural consequences of irradiation. It is tempting to interpret this protective effect as simply preventing the electron beam from heating the specimen. That interpretation is misleading. The fast imaging electron still undergoes essentially the same primary electronic interactions with the specimen, and ionization and excitation described in Sec. 5.13.3 still occur. Cooling acts mainly on what happens *after* these primary events: diffusion, radical migration,

fragment displacement, and secondary chemical reactions are strongly suppressed (Glaeser, 2016; Bammes et al., 2010).

Thus cryogenic operation serves two fundamentally different purposes:

$$\boxed{\begin{aligned}\text{vacuum compatibility} &\longrightarrow \text{preservation of the hydrated vitreous specimen,}\\ \text{radiation protection} &\longrightarrow \text{suppression of post-ionization molecular rearrangement.}\end{aligned}} \tag{5.1764}$$

The primary electronic event survives cooling. Consider again the ionization process introduced in Eq. (5.1729),

$$M + e_{\mathrm{p}}^{-} \longrightarrow M^{+\bullet} + e_{\mathrm{p}}^{-} + e_{\mathrm{s}}^{-}. \tag{5.1765}$$

The energy scales controlling this interaction are electronic energies of order electronvolts to tens of electronvolts, whereas the thermal energy available to the specimen is

$$\boxed{E_{\mathrm{th}} \sim k_{\mathrm{B}}T.} \tag{5.1766}$$

For reference,

$$k_{\mathrm{B}}T \simeq 25.9 \text{ meV} \qquad \text{at 300 K}, \tag{5.1767}$$

whereas

$$k_{\mathrm{B}}T \simeq 8.6 \text{ meV} \qquad \text{at 100 K}, \tag{5.1768}$$

and only

$$k_{\mathrm{B}}T \simeq 0.34 \text{ meV} \qquad \text{at 4 K}. \tag{5.1769}$$

These thermal energies are tiny compared with the tens of electronvolts transferred in a typical valence or collective inelastic event, Eq. (5.1733). Reducing the specimen temperature therefore does not remove the ability of a 200–300 keV electron to ionize the specimen.

The protection must arise at a later stage.

Thermally activated motion is strongly suppressed. After ionization or bond cleavage, a molecular fragment or radical can only produce substantial structural change if it is able to move, reorient, or react with its surroundings.

Many of the post-irradiation processes that can convert a local electronic excitation into a larger structural rearrangement require crossing an activation barrier E_a. Examples include the hopping of a radical to a neighbouring site, reorientation of a molecular fragment, or a local chemical rearrangement.

For such a thermally activated microscopic event, the corresponding rate constant may be written approximately as

$$\boxed{k(T) = k_0 \exp\left(-\frac{E_a}{k_\mathrm{B}T}\right),} \tag{5.1770}$$

where $k(T)$ is the probability per unit time for that event to occur, k_0 is the attempt frequency, and E_a is the activation barrier. Thus

$$[k] = [k_0] = \mathrm{s}^{-1}. \tag{5.1771}$$

The important feature of Eq. (5.1770) is its exponential dependence on inverse temperature.

Why cooling can suppress molecular rearrangement so strongly.

Consider the same activated process at two temperatures T_1 and T_2. From Eq. (5.1770),

$$\frac{k(T_2)}{k(T_1)} = \exp\left[-\frac{E_a}{k_\mathrm{B}}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)\right]. \tag{5.1772}$$

For cooling from room temperature to approximately liquid-nitrogen temperature,

$$T_1 = 300\ \mathrm{K}, \qquad T_2 = 100\ \mathrm{K}. \tag{5.1773}$$

Even for the relatively small illustrative barrier

$$E_a = 0.1\ \mathrm{eV}, \tag{5.1774}$$

one obtains

$$\frac{k(100\ \mathrm{K})}{k(300\ \mathrm{K})} \simeq 4.4 \times 10^{-4}. \tag{5.1775}$$

Thus a process that requires thermal activation can become thousands of times slower upon cooling, even though the primary electronic excitation that created the reactive species is essentially unaffected.

The numerical barrier used here is only illustrative; real radiation chemistry contains many different reaction and migration barriers.

This is the basic physical reason that low temperature suppresses diffusion, conformational relaxation, and many secondary chemical reactions.

The cage effect. A particularly important consequence of this reduced molecular mobility is the *cage effect*.

Suppose that irradiation breaks a bond,

$$A - B \longrightarrow A^{\bullet} + B^{\bullet}. \tag{5.1776}$$

At room temperature, the fragments may separate, rotate, diffuse, or react with neighbouring molecules. Their nuclear coordinates can therefore depart substantially from the original molecular structure.

At cryogenic temperature, the surrounding vitreous water and nearby parts of the macromolecule form a mechanically restrictive environment. The newly created fragments may be unable to escape from the local volume originally occupied by the parent molecule. Schematically,

$$\boxed{A - B \xrightarrow{\text{ionizing radiation}} \underbrace{A^{\bullet} \quad B^{\bullet}}_{\text{local molecular cage}}.} \tag{5.1777}$$

The chemical bond may therefore already be broken while the two atoms remain close to their original positions, possibly still at approximately van der Waals contact (Glaeser, 2016).

This leads to an important distinction:

$$\boxed{\text{chemical damage} \quad \not\Rightarrow \quad \text{immediate large structural displacement.}} \tag{5.1778}$$

The specimen may therefore have changed chemically before the change becomes equally conspicuous in its electron-scattering pattern.

Why caging protects structural information. The connection to image formation can be seen directly in Fourier space.

Suppose that irradiation displaces an atom from $\mathbf{r}_a$ to

$$\mathbf{r}_a + \mathbf{u}_a. \tag{5.1779}$$

As derived previously in Sec. 5.11.1, its contribution to the structure factor acquires the phase factor

$$\exp\left(-j2\pi\mathbf{R}\cdot\mathbf{u}_a\right), \tag{5.1780}$$

see Eq. (5.1052).

If the cage keeps

$$\|\mathbf{u}_a\| \tag{5.1781}$$

small, the phase perturbation

$$2\pi\mathbf{R}\cdot\mathbf{u}_a \tag{5.1782}$$

also remains relatively small. If the fragment instead diffuses or rearranges over a much larger distance, its phase relative to the undamaged structure changes substantially.

Thus the cage effect protects diffraction and image information not because the original chemical structure remains perfectly intact, but because it restricts the *positional consequences* of that damage.

This point is subtle but important:

$$\boxed{\begin{array}{r}\text{cooling does not necessarily preserve every bond;}\\ \text{it helps preserve the spatial organization of the damaged material.}\end{array}} \tag{5.1783}$$

In the language of the disorder treatment in Sec. 5.11, cooling reduces the growth of radiation-induced positional disorder and therefore slows the attenuation of coherent structural information.

Radical trapping and the physical origin of cryoprotection. The same restriction of molecular mobility applies to radicals. The radiolysis processes described in Sec. 5.13.3 generate reactive species both inside the macromolecule and in the surrounding water. In a liquid, such species can diffuse away from the site at which the original energy was deposited and participate in secondary chemical reactions.

At cryogenic temperature, this mobility is strongly reduced. Radicals can remain trapped in approximately the same local environment in which they were created, while molecular fragments remain confined by the surrounding vitreous ice and neighbouring parts of the macromolecule. In irradiated organic crystals, radicals may even remain trapped with well-defined orientations relative to their surroundings (Glaeser, 2016).

Consequently, the indirect pathway

$$\text{solvent radiolysis} \longrightarrow \text{mobile radical} \longrightarrow \text{attack on the macromolecule} \tag{5.1784}$$

is strongly suppressed. The distinction between direct and indirect damage introduced in Eq. (5.1754) therefore becomes temperature dependent: in

liquid water, mobile solvent-derived radicals can make an important contribution, whereas in vitreous ice near liquid-nitrogen temperature their diffusion is strongly restricted and the direct effects of irradiation on the macromolecule become relatively more important (Glaeser, 2016).

This also clarifies a common misconception. The main protective effect of cryogenic operation is not that it prevents the electron beam from heating the specimen. Rather,

$$\boxed{\begin{aligned} \text{cooling} &\longrightarrow \text{reduced molecular and radical mobility} \\ &\longrightarrow \text{fragment caging and radical trapping} \\ &\longrightarrow \text{slower secondary chemistry and rearrangement} \\ &\longrightarrow \text{slower loss of structural information.} \end{aligned}} \tag{5.1785}$$

The actual temperature rise caused by the electron beam will be considered separately later. Under the low-current conditions typical of cryo-EM imaging, beam heating is generally not the dominant explanation either for radiation damage itself or for the protective effect of cooling (Egerton, 2013).

Why the cage effect eventually fails. Cryogenic trapping is not permanent protection.

As exposure increases, the number of radicals and chemically modified sites also increases. If the concentration of trapped radicals becomes large enough, a newly generated radical is increasingly likely to be created close to an already existing one.

Two radicals in sufficiently close contact may react without requiring the thermally activated diffusion that cooling suppresses:

$$R_1^\bullet + R_2^\bullet \longrightarrow R_1 - R_2 + \text{released energy}. \tag{5.1786}$$

Such reactions may release enough local energy to activate rotations, translations, and further chemical rearrangements. Thus accumulated damage can eventually overcome the protection produced by radical trapping and molecular caging (Glaeser, 2016).

The qualitative progression is therefore

$$\boxed{\begin{aligned} \text{low exposure:} &\quad \text{isolated trapped radicals and caged fragments,} \\ \text{higher exposure:} &\quad \text{increasing overlap of damaged regions,} \\ \text{still higher exposure:} &\quad \text{radical reactions and structural rearrangement.} \end{aligned}} \tag{5.1787}$$

This explains why cooling increases radiation tolerance but can never make a biological specimen radiation resistant.

Does colder always mean better? One might expect that, if reduced molecular mobility provides cryoprotection, lowering the temperature from approximately liquid-nitrogen temperature to liquid-helium temperature should always preserve substantially more structural information. Historically, however, this improvement was difficult to demonstrate in cryo-EM images.

Bammes and coworkers examined frozen-hydrated catalase crystals at

$$T = 4,\ 25,\ 42,\ \text{and } 100\ \text{K} \tag{5.1788}$$

using exposure series and measured the evolution of Bragg-peak amplitudes and signal quality over a broad range of spatial frequencies (Bammes et al., 2010).

At high spatial frequencies, particularly in the 4–20 Å range, they found no statistically significant difference in exposure tolerance among the four temperatures. At lower spatial frequencies, intermediate temperatures of 25–42 K sometimes provided modestly improved tolerance relative to 100 K, whereas specimens at 4 K displayed additional abnormal changes of Fourier amplitudes, anisotropy, and beam-induced specimen motion.

The natural interpretation at the time was that cooling below liquid-nitrogen temperature provided little additional practical advantage for high-resolution cryo-EM, even though lower temperatures should further suppress radical diffusion and molecular rearrangement.

More recent experiments have refined this conclusion substantially. Dickerson and coworkers showed that the apparent lack of benefit at liquid-helium temperature was not primarily due to the absence of additional cryoprotection, but to additional *mechanical instabilities* of the specimen and support (Dickerson et al., 2025).

At approximately

$$T \simeq 13\ \text{K}, \tag{5.1789}$$

they observed substantial beam-induced motion when frozen water was suspended in holes of several hundred nanometres diameter. Part of this motion originated from stress in the support foil caused by differential thermal contraction between the gold foil and the frozen water. Upon irradiation, the stressed system relaxed, producing displacement of the foil.

In addition, the frozen water itself changed mechanically during the first few

$$\text{e}^{-}\ \text{Å}^{-2} \tag{5.1790}$$

of exposure. Rather than collapsing into a denser amorphous phase, as had previously been suggested, the water expanded under irradiation at liquid-helium temperature. The expansion was attributed to the accumulation of

radiolytic products that are less able to diffuse, recombine, or escape at such low temperature (Dickerson et al., 2025).

The resulting expansion generates stress in the suspended ice film and can eventually produce buckling or doming. Thus the very same reduction in molecular mobility that improves cryoprotection can also create a new mechanical problem:

$$\boxed{\begin{aligned} \text{lower temperature} &\longrightarrow \text{stronger trapping of radiolytic products} \\ &\longrightarrow \text{less chemical rearrangement} \\ &\quad\ \text{but also}\quad \text{greater accumulation of products} \\ &\longrightarrow \text{expansion, stress, and beam-induced motion.} \end{aligned}} \tag{5.1791}$$

This resolves the apparent contradiction between the earlier observations of Bammes et al. and the expected radiation-damage advantage of lower temperature. The intrinsic radiation chemistry may indeed be more strongly suppressed at liquid-helium temperature, but this benefit can be hidden in the recorded image by mechanical motion.

Recovering the radiation-damage advantage at liquid-helium temperature. Dickerson and coworkers reduced these mechanical effects by using all-gold specimen supports with holes only

$$100 \text{ nm} \tag{5.1792}$$

in diameter and by restricting the illuminated area to a beam of only a few hundred nanometres diameter (Dickerson et al., 2025).

The small holes prevented the expanding frozen-water film from reaching the strain required for substantial buckling, while the restricted beam reduced irradiation-induced relaxation of the surrounding support foil.

Under these conditions, specimen motion at 13 K was largely eliminated and the expected advantage of lower temperature became visible directly in single-particle cryo-EM.

For per-frame reconstructions, the radiation-damage contribution to the image B-factor increased approximately as

$$B_{\text{rad}}(\Phi) \simeq b_{\text{rad}}\Phi, \tag{5.1793}$$

with measured slopes of approximately

$$\boxed{\begin{aligned} b_{\text{rad}}(13\text{ K}) &\simeq 3.2\ \text{Å}^2/(\text{e}^-\ \text{Å}^{-2}), \\ b_{\text{rad}}(81\text{ K}) &\simeq 4.8\ \text{Å}^2/(\text{e}^-\ \text{Å}^{-2}). \end{aligned}} \tag{5.1794}$$

Thus the radiation-induced loss of high-resolution information was approximately

$$\frac{4.8}{3.2} \simeq 1.5 \tag{5.1795}$$

times slower at 13 K than at 81 K (Dickerson et al., 2025).

This improvement is consistent with earlier diffraction measurements, which had suggested that liquid-helium temperatures should reduce the effects of radiation damage by approximately a factor of 1.2–1.8 relative to liquid-nitrogen temperatures.

The important conclusion is therefore not that cooling below 100 K provides no additional protection, but rather that *the radiation-damage advantage and the mechanical stability of the specimen must be considered separately*:

$$\boxed{\begin{array}{rl} \text{intrinsic radiation damage} & \text{decreases at lower temperature,} \\ \text{beam-induced mechanical motion} & \text{may increase if the specimen geometry is unsuitable.} \end{array}} \tag{5.1796}$$

The earlier catalase experiments of Bammes et al. (2010) were therefore not inconsistent with additional cryoprotection at liquid-helium temperature. Rather, they revealed that the practical information content of an image is controlled by more than radiation chemistry alone. The later experiments of Dickerson et al. (2025) showed that once the dominant mechanical instability is suppressed, the additional radiation-damage resistance at liquid-helium temperature becomes measurable in single-particle images.

A revised picture of temperature and cryoprotection. The combined experimental evidence suggests a more nuanced picture than the simple statement that "colder is better":

$$\boxed{\begin{array}{rl} \text{room temperature} \longrightarrow & \text{rapid diffusion and secondary chemistry,} \\ \text{near 80–100 K} \longrightarrow & \text{strong radical trapping and fragment caging,} \\ \text{near 10–20 K} \longrightarrow & \text{still slower radiation-induced structural disorder,} \\ & \text{but potentially stronger mechanical instabilities.} \end{array}} \tag{5.1797}$$

Cooling therefore does not prevent the primary radiation chemistry. Ionization still occurs, bonds can still be broken, and radicals still accumulate. Its principal protective effect is to delay the conversion of these microscopic

chemical events into larger positional and structural changes:

$$\boxed{\begin{aligned} \text{ionization} &\not\Rightarrow \text{immediate fragment escape,} \\ \text{bond cleavage} &\not\Rightarrow \text{immediate large displacement,} \\ \text{radical formation} &\not\Rightarrow \text{immediate long-range chemical attack.} \end{aligned}} \tag{5.1798}$$

This distinction is fundamental for cryo-EM because the recorded image depends primarily on the spatial distribution of the specimen scattering potential. A specimen may therefore already be chemically modified while still retaining useful structural information, provided that the products of radiation damage remain sufficiently close to their original positions.

Lowering the temperature can extend this structural preservation, but only if the specimen remains mechanically stable. Radiation damage progressively changes the scattering potential itself, whereas beam-induced motion displaces that potential during acquisition. The experiments discussed above show that lowering the temperature can reduce the first effect while increasing the second unless the specimen support and illumination geometry are designed appropriately.

Eventually, accumulated chemical damage produces sufficient positional disorder that the coherent Fourier components of the specimen begin to fade. The rate of this fading depends strongly on spatial frequency and will be examined quantitatively in a later subsection.

5.13.5 Damage is chemically and spatially selective

The discussion so far might suggest that irradiation gradually converts the specimen into a uniformly more disordered version of itself. Such a picture is useful at low resolution, but it becomes increasingly inadequate as atomic detail is resolved.

Radiation damage is *chemically selective*. Different chemical groups have different susceptibilities to ionization, electron capture, radical formation, bond cleavage, and subsequent rearrangement. Consequently, some parts of a macromolecule can change substantially while others remain comparatively close to their initial structure.

Among the groups known to be particularly sensitive are

- disulfide bonds;
- carboxylate groups of acidic side chains, particularly aspartate and glutamate;

- metal-binding sites and coordinated ions;
- some catalytically important or otherwise chemically reactive groups.

These trends have long been recognized in radiation studies of protein crystals and have also been observed directly with electron irradiation (Glaeser, 2016; Hattne et al., 2018).

Thus radiation damage is not only exposure dependent,

$$V(\mathbf{r}) \longrightarrow V(\mathbf{r}, \Phi), \tag{5.1799}$$

as introduced in Eq. (5.8), but the dependence on Φ is also strongly position and chemistry dependent.

Global disorder and local chemical damage are different concepts. A common description of radiation damage is to represent the damaged specimen as a blurred version of the original one,

$$V(\mathbf{r}, \Phi) = V(\mathbf{r}, 0) * g_\Phi(\mathbf{r}), \tag{5.1800}$$

where g_Φ is an exposure-dependent blurring kernel.

In Fourier space,

$$\widehat{V}(\mathbf{R}, \Phi) = \widehat{V}(\mathbf{R}, 0)\widehat{g}_\Phi(\mathbf{R}). \tag{5.1801}$$

If g_Φ is Gaussian, this gives a Debye–Waller- or B-factor-like attenuation of high spatial frequencies. Such a model describes the *global loss of ordered structural information* reasonably well in many circumstances.

It does not, however, describe site-specific chemical damage.

For example, breaking one disulfide bond or removing the carboxyl group of one glutamate does not correspond to convolving the entire molecule with a common kernel. The electrostatic potential changes locally and chemically:

$$\boxed{V(\mathbf{r}, \Phi) \neq V(\mathbf{r}, 0) * g_\Phi(\mathbf{r}) \qquad \text{in general.}} \tag{5.1802}$$

A global B factor therefore describes one important consequence of damage—loss of high-frequency coherence—but not the full microscopic process that produced it.

A simple model of site-specific damage. To make this distinction more precise, consider a chemically sensitive site a. Let

$$V_a^{(0)}(\mathbf{r}) \tag{5.1803}$$

be its contribution to the specimen potential in the undamaged state, and let

$$V_a^{(1)}(\mathbf{r}) \tag{5.1804}$$

represent the potential after a particular radiation-induced chemical transformation.

Examples would be the intact and broken forms of a disulfide bond, or the intact and decarboxylated forms of an acidic side chain.

Let

$$p_a(\Phi) \tag{5.1805}$$

be the fraction of molecules in which site a remains in its original state after exposure Φ. The ensemble-averaged contribution of this site is then

$$\boxed{\langle V_a(\mathbf{r},\Phi)\rangle = p_a(\Phi)V_a^{(0)}(\mathbf{r}) + [1-p_a(\Phi)]\,V_a^{(1)}(\mathbf{r}).} \tag{5.1806}$$

For a simple independent-event model with a constant effective damage cross section $\sigma_{d,a}$,

$$\boxed{p_a(\Phi) = \exp\left(-\sigma_{d,a}\Phi\right).} \tag{5.1807}$$

Since Φ has dimensions of inverse area, $\sigma_{d,a}$ has dimensions of area.

The corresponding characteristic exposure is

$$\boxed{\Phi_{d,a} = \frac{1}{\sigma_{d,a}},} \tag{5.1808}$$

so that

$$p_a(\Phi_{d,a}) = e^{-1}. \tag{5.1809}$$

Different chemical groups generally have different values of $\sigma_{d,a}$ and therefore different characteristic exposures.

Why site-specific damage cannot generally be represented by one envelope.

Separate the specimen potential into a comparatively stable part and

a set of chemically sensitive sites:

$$V(\mathbf{r},0) = V_{\text{stable}}(\mathbf{r}) + \sum_a V_a^{(0)}(\mathbf{r}). \tag{5.1810}$$

After irradiation, Eq. (5.1806) gives

$$\begin{aligned}\langle V(\mathbf{r},\Phi)\rangle &= V_{\text{stable}}(\mathbf{r}) \\ &\quad + \sum_a \left[p_a(\Phi)V_a^{(0)}(\mathbf{r}) + (1-p_a(\Phi))V_a^{(1)}(\mathbf{r})\right].\end{aligned} \tag{5.1811}$$

Subtracting the undamaged potential,

$$\boxed{\delta V(\mathbf{r},\Phi) = \sum_a [1-p_a(\Phi)]\left[V_a^{(1)}(\mathbf{r}) - V_a^{(0)}(\mathbf{r})\right].} \tag{5.1812}$$

Taking the Fourier transform,

$$\boxed{\delta \widehat{V}(\mathbf{R},\Phi) = \sum_a [1-p_a(\Phi)]\left[\widehat{V}_a^{(1)}(\mathbf{R}) - \widehat{V}_a^{(0)}(\mathbf{R})\right].} \tag{5.1813}$$

There is no reason for this expression to be proportional to $\widehat{V}(\mathbf{R},0)$. Therefore, in general there is no single function $E_{\text{rad}}(\mathbf{R},\Phi)$ such that

$$\widehat{V}(\mathbf{R},\Phi) = E_{\text{rad}}(\mathbf{R},\Phi)\widehat{V}(\mathbf{R},0) \tag{5.1814}$$

for every spatial frequency.

A radiation-damage envelope is therefore an effective statistical description of global information loss, rather than a complete microscopic model of radiation chemistry.

Site-specific damage can precede large global information loss. This distinction is observed experimentally.

In MicroED measurements of proteinase K and a zinc-containing heptapeptide, Hattne et al. (2018) followed both the global decay of diffraction intensities and local changes in the reconstructed electrostatic potential as exposure accumulated.

Localized changes were detectable at exposures much smaller than those required to destroy the diffraction pattern globally. In particular, site-specific changes were observed at exposures below

$$1\ \text{e}^-\ \text{Å}^{-2}, \tag{5.1815}$$

and in some analyses changes were detectable from approximately

$$0.1\ \mathrm{e}^{-}\ \text{Å}^{-2}. \tag{5.1816}$$

The approximate experimental scales are summarized in Table 5.14.

Table 5.14: Examples of site-specific electron-radiation damage observed by Hattne et al. (2018). The numerical exposures refer to their MicroED measurements and should not be interpreted as universal thresholds for single-particle cryo-EM.

Sensitive site	**Approximate exposure**	**Observed change**
Disulfide bonds	$< 0.9\ \mathrm{e}^{-}\ \text{Å}^{-2}$	Difference density around sulfur atoms indicated the onset of disulfide-bond breakage.
Zn site	~ 0.2–$0.8\ \mathrm{e}^{-}\ \text{Å}^{-2}$	The atomic displacement parameter of the Zn ion increased early in the exposure series; pronounced site-specific perturbation was evident at higher exposure.
Aspartate and glutamate carboxylates	$\sim 2\ \mathrm{e}^{-}\ \text{Å}^{-2}$	Density associated with the acidic side chains began to deteriorate.
Aspartate and glutamate carboxylates	$\sim 5\ \mathrm{e}^{-}\ \text{Å}^{-2}$	The corresponding side-chain density was largely lost.

These changes occurred while substantial global diffraction remained. For proteinase K, after approximately

$$1\ \mathrm{e}^{-}\ \text{Å}^{-2}, \tag{5.1817}$$

the mean intensity of the measured reflections was still approximately 73% of its extrapolated zero-exposure value (Hattne et al., 2018).

Thus, local chemical damage can be detectable before global structural information is lost. This is precisely why a single global B factor cannot fully describe radiation damage.

Why some chemical groups are damaged earlier than others. The ordering of site-specific damage reflects the radiation chemistry described in Sec. 5.13.3. Different groups respond differently to electron removal, electron capture, radical formation, proton transfer, and changes in local electrostatic environment.

A disulfide bond,

$$\mathrm{R-S-S-R}, \tag{5.1818}$$

for example, may undergo reduction or cleavage after electron capture and subsequent rearrangement. The resulting sulfur positions and bonding potential differ strongly from those of the intact bridge.

For acidic side chains, loss of the carboxyl group may be represented schematically as

$$\boxed{\mathrm{R-COO^- \longrightarrow R^\bullet + CO_2},} \tag{5.1819}$$

where the equation represents the net structural outcome rather than a complete elementary reaction mechanism.

The change in electron-scattering potential is especially conspicuous because several atoms that contributed to the original local density are no longer present in the same configuration.

Metal sites can be affected in another way. Ionization of coordinating groups or changes in their charge state can alter metal coordination without requiring the metal atom itself to undergo the same chemistry. Thus the metal may shift, become more disordered, or eventually lose occupancy.

Active sites are particularly important conceptually because a small local chemical change may abolish biological function while producing only a negligible change in the global mass or low-resolution shape of the molecule (Glaeser, 2016).

The irradiated specimen is an ensemble of chemical states. There is an important consequence for single-particle cryo-EM. At a given accumulated exposure Φ, not every copy of the molecule has undergone the same radiation event. One particle may retain a particular disulfide bridge, another may have lost it, and another may have suffered damage at a different site altogether.

The specimen at exposure Φ is therefore better represented by an ensemble of chemically distinct states

$$\{V_c(\mathbf{r})\}, \tag{5.1820}$$

with probabilities

$$P_c(\Phi). \tag{5.1821}$$

The ensemble-averaged specimen potential is then

$$\boxed{\overline{V}(\mathbf{r},\Phi) = \sum_c P_c(\Phi) V_c(\mathbf{r}), \qquad \sum_c P_c(\Phi) = 1.} \tag{5.1822}$$

Equation (5.1822) generalizes the two-state description of Eq. (5.1806). It also clarifies what a reconstructed cryo-EM map represents. A map obtained from electrons accumulated over a finite exposure does not necessarily correspond to one molecular structure at one instant; at sufficiently high resolution, it may represent an average over molecules occupying different radiation-induced chemical states.

At lower resolution, many such local changes are not individually resolved and their combined effect can often be approximated by a smooth exposure-dependent attenuation or an effective B factor. At higher resolution, however, the identities and positions of individual atoms contribute directly to the map, and chemically selective damage can no longer be represented adequately as a uniform blur:

$$\boxed{\begin{array}{lcl} \text{lower resolution} & \longrightarrow & \text{damage often resembles global attenuation,} \\ \text{higher resolution} & \longrightarrow & \text{different chemical states become individually relevant.} \end{array}} \tag{5.1823}$$

The experiments of Hattne et al. (2018) show that site-specific changes become detectable at resolutions better than approximately 3 Å and at exposures of order 1 $e^- \text{Å}^{-2}$ or below.

The implication for image formation is fundamental: radiation damage changes the object, not merely the contrast with which the object is observed.

The exposure-dependent potential $V(\mathbf{r},\Phi)$ introduced in Eq. (5.8) should therefore be understood as a chemically evolving, spatially heterogeneous, and ensemble-valued quantity, rather than simply as the pristine potential multiplied by a global radiation-damage envelope.

5.13.6 Resolution-dependent fading of structural information

The site-specific damage discussed in Sec. 5.13.5 shows that radiation damage is not microscopically equivalent to applying a uniform blur to the specimen. Nevertheless, after averaging over many molecules, damage events, and Fourier components within a resolution shell, the loss of *coherent structural*

information can often be described remarkably well by a simple exponential law.

This effective description reveals one of the most important properties of radiation damage in cryo-EM:

$$\boxed{\text{high-resolution structural information fades faster than low-resolution information.}} \tag{5.1824}$$

Let

$$R = \|\mathbf{R}\| \tag{5.1825}$$

denote the magnitude of the spatial frequency and let Φ be the accumulated electron exposure, defined in Eq. (5.1678).

What quantity is fading? The site-specific picture developed in Sec. 5.13.5 implies that, at accumulated exposure Φ, different copies of the molecule may occupy different radiation-induced chemical states. Let these states have potentials $V_c(\mathbf{r})$ and probabilities $P_c(\Phi)$, as introduced in Eq. (5.1822).

The ensemble-averaged Fourier amplitude is therefore

$$\boxed{\overline{\widehat{V}}(\mathbf{R}, \Phi) = \sum_c P_c(\Phi) \widehat{V}_c(\mathbf{R}).} \tag{5.1826}$$

This is the part of the specimen Fourier transform that is reproducible from one particle to another and can therefore add coherently during single-particle averaging.

We define its shell-averaged coherent structural power as

$$\boxed{P_{\text{coh}}(R, \Phi) \equiv \left\langle \left| \overline{\widehat{V}}(\mathbf{R}, \Phi) \right|^2 \right\rangle_{\|\mathbf{R}\|=R},} \tag{5.1827}$$

where the brackets denote averaging over Fourier directions with $\|\mathbf{R}\| = R$.

This quantity must be distinguished from the ensemble-averaged *total* Fourier power,

$$\boxed{P_{\text{tot}}(R, \Phi) = \left\langle \sum_c P_c(\Phi) \left| \widehat{V}_c(\mathbf{R}) \right|^2 \right\rangle_{\|\mathbf{R}\|=R}.} \tag{5.1828}$$

In general,

$$P_{\text{coh}}(R, \Phi) \leq P_{\text{tot}}(R, \Phi). \tag{5.1829}$$

Their difference,

$$\boxed{P_{\text{het}}(R,\Phi) = P_{\text{tot}}(R,\Phi) - P_{\text{coh}}(R,\Phi),} \tag{5.1830}$$

measures Fourier power associated with particle-to-particle differences between radiation-induced states.

Thus radiation damage need not simply remove Fourier power. It can convert reproducible structural power into heterogeneous, non-reproducible power.

Critical exposure and the empirical exponential law. We first define the normalized fading of coherent structural power by

$$\boxed{F_{\text{rad}}(R,\Phi) \equiv \frac{P_{\text{coh}}(R,\Phi)}{P_{\text{coh}}(R,0)}.} \tag{5.1831}$$

No particular functional form for $F_{\text{rad}}(R,\Phi)$ is implied by this definition.

If the decay is monotonic, the *critical exposure* $\Phi_c(R)$ may be defined operationally as the exposure at which the coherent power has fallen to $1/e$ of its initial value:

$$\boxed{F_{\text{rad}}\left(R,\Phi_c(R)\right) = e^{-1}.} \tag{5.1832}$$

Experimentally, over a substantial range of exposures, F_{rad} is often well approximated by a single exponential,

$$\boxed{F_{\text{rad}}(R,\Phi) \simeq \exp\left[-\frac{\Phi}{\Phi_c(R)}\right].} \tag{5.1833}$$

Equivalently,

$$\boxed{P_{\text{coh}}(R,\Phi) \simeq P_{\text{coh}}(R,0)\exp\left[-\frac{\Phi}{\Phi_c(R)}\right].} \tag{5.1834}$$

When does exposure-dependent damage produce an exponential decay?.

Consider the coherent structural power at one spatial frequency R. Suppose that, during a small additional exposure $d\Phi$, there is a constant probability per incident electron of destroying the coherence of a small fraction of the remaining structural information.

Equivalently, assume that the fractional loss of coherent power is

proportional to the coherent power that is still present:

$$dP_{\text{coh}} = -\kappa(R) P_{\text{coh}} d\Phi, \tag{5.1835}$$

where $\kappa(R)$ is an effective damage rate per unit exposure.

Dividing by $d\Phi$,

$$\boxed{\frac{dP_{\text{coh}}(R,\Phi)}{d\Phi} = -\kappa(R) P_{\text{coh}}(R,\Phi).} \tag{5.1836}$$

This is the same mathematical structure as any first-order survival process. Integrating,

$$\frac{dP_{\text{coh}}}{P_{\text{coh}}} = -\kappa(R)\, d\Phi, \tag{5.1837}$$

$$\ln \frac{P_{\text{coh}}(R,\Phi)}{P_{\text{coh}}(R,0)} = -\kappa(R)\Phi, \tag{5.1838}$$

and therefore

$$P_{\text{coh}}(R,\Phi) = P_{\text{coh}}(R,0) e^{-\kappa(R)\Phi}. \tag{5.1839}$$

Identifying

$$\boxed{\Phi_c(R) = \frac{1}{\kappa(R)},} \tag{5.1840}$$

gives Eq. (5.1834).

The derivation above does *not* prove that radiation damage must follow a single exponential. It shows the physical assumptions under which an exponential follows: damaging events must be approximately independent, the effective damage probability per unit exposure must remain approximately constant, and the fractional loss must be proportional to the coherent information that remains.

The single-exponential law should therefore be regarded as an effective phenomenological model rather than a fundamental consequence of radiation chemistry. In single-particle measurements, ln(SNR) is often approximately linear in accumulated exposure over experimentally useful ranges, supporting a dominant single-exponential component (Grant and Grigorieff, 2015). Nevertheless, multi-stage chemical damage, site-specific sensitivity, radical accumulation, and mechanical rearrangement can all produce departures from this simple law (Heymann, 2022).

Power and amplitude differ by a factor of two in the exponent. The distinction between Fourier *power* and Fourier *amplitude* is essential when radiation damage is incorporated into an image-formation model.

From the critical exposure of power to the decay of Fourier amplitudes.

Let the effective coherent Fourier amplitude be written as

$$A_V(R,\Phi) = |\widehat{V}(R,\Phi)|. \tag{5.1841}$$

Its power is

$$P_V(R,\Phi) = A_V^2(R,\Phi). \tag{5.1842}$$

From Eq. (5.1833),

$$A_V^2(R,\Phi) = A_V^2(R,0)\exp\left[-\frac{\Phi}{\Phi_c(R)}\right]. \tag{5.1843}$$

Taking the positive square root gives

$$\boxed{A_V(R,\Phi) = A_V(R,0)\exp\left[-\frac{\Phi}{2\Phi_c(R)}\right].} \tag{5.1844}$$

It is therefore convenient to define the radiation-damage amplitude envelope

$$\boxed{E_{\mathrm{rad}}(R,\Phi) \equiv \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right].} \tag{5.1845}$$

The corresponding power envelope is

$$\boxed{|E_{\mathrm{rad}}(R,\Phi)|^2 = \exp\left[-\frac{\Phi}{\Phi_c(R)}\right].} \tag{5.1846}$$

The factor of two in Eq. (5.1844) is important. Image formation and coherent accumulation operate on Fourier amplitudes, whereas diffraction intensities, power spectra, and many SNR measures depend on squared amplitudes.

Why do high spatial frequencies disappear first? The resolution dependence of $\Phi_c(R)$ has a simple qualitative interpretation. Fine structural

detail depends strongly on the precise relative positions of atoms. A small radiation-induced displacement may therefore have little effect at low spatial frequency while producing a large phase error at high spatial frequency.

This can be made quantitative using the displacement model introduced previously in Sec. 5.11.

Critical exposure for a simple random-displacement model.

Suppose that radiation damage produces isotropic random atomic displacements with one-coordinate variance

$$U(\Phi), \tag{5.1847}$$

and suppose, as the simplest model, that this variance grows linearly with exposure:

$$\boxed{U(\Phi) = \alpha_{\mathrm{rad}}\Phi,} \tag{5.1848}$$

where α_{rad} has units of $\text{Å}^4/\mathrm{e}^-$ when Φ is expressed in $\mathrm{e}^-\,\text{Å}^{-2}$.

For isotropic Gaussian positional disorder, the Fourier amplitude attenuation derived in Sec. 5.11 is

$$E(R,\Phi) = \exp\left[-2\pi^2 U(\Phi) R^2\right]. \tag{5.1849}$$

Using Eq. (5.1848),

$$E(R,\Phi) = \exp\left[-2\pi^2 \alpha_{\mathrm{rad}}\Phi R^2\right]. \tag{5.1850}$$

The corresponding structural power therefore decays as

$$|E(R,\Phi)|^2 = \exp\left[-4\pi^2 \alpha_{\mathrm{rad}}\Phi R^2\right]. \tag{5.1851}$$

Comparison with Eq. (5.1833) gives

$$\boxed{\Phi_c(R) = \frac{1}{4\pi^2 \alpha_{\mathrm{rad}} R^2}.} \tag{5.1852}$$

Thus a simple independent-displacement model predicts

$$\boxed{\Phi_c(R) \propto R^{-2}.} \tag{5.1853}$$

Equation (5.1853) captures the basic reason that high-resolution information is more fragile: doubling the spatial frequency would, in this simple model, reduce the critical exposure by a factor of four.

The experimentally observed dependence is not universally exactly R^{-2}. This is expected. Radiation damage combines site-specific chemistry, fragment displacement, changes in molecular packing, secondary-electron cascades, and mechanical rearrangements rather than one pure diffusion process.

For example, the single-particle measurements of Grant and Grigorieff (2015) were fitted at 300 kV by the empirical relation

$$\boxed{\Phi_c(R) \simeq 0.245 \left(\frac{R}{1\ \text{Å}^{-1}}\right)^{-1.665} + 2.81 \qquad \text{e}^-\ \text{Å}^{-2},} \tag{5.1854}$$

over the range in which reliable measurements could be obtained.

More recent analyses of progressive spectral SNR also find a strong decrease of critical exposure with increasing spatial frequency, but the fitted power-law exponents vary among specimens and often lie between approximately -1 and -2 (Heymann, 2022). This variability reinforces the interpretation of $\Phi_c(R)$ as an empirical description of the combined consequences of several damage mechanisms rather than a universal material constant.

Crystal diffraction and single-particle measurements. The classical measurements of Baker et al. (2010) used thin catalase crystals imaged at 200 kV and liquid-nitrogen temperature. They found that the optimal exposure depended strongly on resolution. For example, their measurements gave approximately

$$\boxed{\begin{aligned} d \simeq 3\ \text{Å} \quad &: \quad \Phi_{\text{opt}} \simeq 10\ \text{e}^-\ \text{Å}^{-2}, \\ d \simeq 20\ \text{Å} \quad &: \quad \Phi_{\text{opt}} \simeq 20\ \text{e}^-\ \text{Å}^{-2}, \end{aligned}} \tag{5.1855}$$

where

$$d = \frac{1}{R}. \tag{5.1856}$$

The lower-resolution information was therefore able to tolerate substantially greater exposure before its SNR became suboptimal.

Single-particle measurements later showed an important difference. Using a 2.6 Å reconstruction of rotavirus VP6 recorded at 300 kV, Grant and

Grigorieff (2015) measured $\Phi_c(R)$ directly from movie data over approximately the 22–3.8 Å range. At high spatial frequencies, their critical exposures agreed reasonably well with the earlier crystalline measurements after accounting for accelerating voltage. At lower spatial frequencies, however, single-particle information faded more slowly than the crystal diffraction intensities.

One possible reason is that crystal diffraction is sensitive to two distinct forms of disorder:

$$\boxed{\begin{aligned}&\text{molecular damage} \longrightarrow \text{changes inside each repeating unit,}\\ &\quad\text{lattice disorder} \longrightarrow \text{loss of long-range crystalline registration.}\end{aligned}} \tag{5.1857}$$

Single-particle imaging does not require long-range crystalline order, and therefore does not suffer the second loss mechanism in the same way (Grant and Grigorieff, 2015; Glaeser, 2016).

This distinction is particularly important at lower spatial frequencies. It means that critical-exposure curves measured from protein crystals should not automatically be treated as universal single-particle limits.

There is an optimum exposure for an unweighted acquisition. The existence of radiation damage does not imply that the smallest possible exposure gives the best image. With too few electrons, the specimen is less damaged but counting noise is larger.

Consider one spatial frequency R. Suppose that at accumulated exposure ϕ the coherent Fourier amplitude contributed by the next small increment of exposure is attenuated by Eq. (5.1845):

$$E_{\text{rad}}(R, \phi) = \exp\left[-\frac{\phi}{2\Phi_c(R)}\right]. \tag{5.1858}$$

For an ordinary unweighted exposure, all electrons recorded between 0 and Φ are accumulated into the same image.

Optimal exposure at one spatial frequency.

Apart from factors independent of exposure, the accumulated coherent Fourier signal is

$$S(R, \Phi) \propto \int_0^{\Phi} \exp\left[-\frac{\phi}{2\Phi_c(R)}\right] d\phi. \tag{5.1859}$$

Performing the integral,

$$S(R,\Phi) \propto 2\Phi_c(R)\left[1-\exp\left(-\frac{\Phi}{2\Phi_c(R)}\right)\right]. \tag{5.1860}$$

For Poisson counting statistics, the noise variance is proportional to the total number of detected electrons and therefore to Φ:

$$\sigma_N^2 \propto \Phi. \tag{5.1861}$$

Hence

$$\mathrm{SNR}^2(R,\Phi) \propto \frac{\left[1-\exp\left(-\frac{\Phi}{2\Phi_c(R)}\right)\right]^2}{\Phi}. \tag{5.1862}$$

Introduce the dimensionless exposure

$$x \equiv \frac{\Phi}{\Phi_c(R)}. \tag{5.1863}$$

Apart from a multiplicative constant,

$$f(x) = \frac{\left(1-e^{-x/2}\right)^2}{x}. \tag{5.1864}$$

The optimum satisfies

$$\frac{df}{dx} = 0. \tag{5.1865}$$

Differentiating and removing the trivial solution at $x = 0$ gives

$$xe^{-x/2} = 1 - e^{-x/2}. \tag{5.1866}$$

Equivalently,

$$\boxed{e^{x/2} = x + 1.} \tag{5.1867}$$

The positive non-zero solution is

$$x \simeq 2.5129. \tag{5.1868}$$

Therefore,

$$\boxed{\Phi_{\mathrm{opt}}(R) \simeq 2.51\,\Phi_c(R).} \tag{5.1869}$$

This result, originally developed in the context of radiation-damaged diffraction data, explains the central experimental compromise (Hayward

and Glaeser, 1979; Baker et al., 2010):

$$\boxed{\begin{array}{c}\text{too little exposure} \longrightarrow \text{insufficient counting statistics,}\\ \text{too much exposure} \longrightarrow \text{additional electrons carry progressively less high-resolution signal,}\\ \text{intermediate exposure} \longrightarrow \text{maximum SNR at the chosen spatial frequency.}\end{array}} \tag{5.1870}$$

The optimum is resolution dependent because $\Phi_c(R)$ is resolution dependent. There is therefore no single exposure that is simultaneously optimal for all structural scales.

What exactly is meant by "optimal exposure"? There is an important qualification to Eq. (5.1869). A finite optimum arises for a particular measurement protocol: all electrons from an exposure are accumulated with equal weight while shot noise continues to increase.

With a direct electron detector, the acquisition is instead recorded as a movie. The electrons arriving early and late in the exposure remain separable. This changes the optimization problem.

Heymann has emphasized this distinction using the *progressive spectral signal-to-noise ratio* (PSSNR) (Heymann, 2022). Under that definition, the accumulated coherent information can increase monotonically with exposure and approach an asymptote rather than literally decrease after a particular exposure. The additional late electrons eventually contribute almost no new coherent information at a given spatial frequency, but they do not erase the information already recorded in the earlier frames.

For a simple exponential decay, the progressive coherent information has the generic form

$$\boxed{\mathrm{PSSNR}(R,\Phi) \propto \Phi_c(R)\left[1-\exp\left(-\frac{\Phi}{\Phi_c(R)}\right)\right],} \tag{5.1871}$$

and therefore approaches an asymptote as $\Phi \to \infty$.

For example,

$$1 - e^{-\Phi/\Phi_c} = 0.9 \tag{5.1872}$$

is reached at

$$\boxed{\Phi = -\Phi_c \ln(0.1) \simeq 2.30\,\Phi_c.} \tag{5.1873}$$

The numerical proximity between $2.30\Phi_c$ and the classical $2.51\Phi_c$ optimum should not obscure the conceptual distinction. The first describes approaching a saturation level of accumulated coherent information; the

second follows from maximizing the SNR of an equally weighted integrated exposure.

Thus, "optimal exposure" depends on what is being optimized and how the movie frames are combined.

Dose fractionation converts a fixed exposure limit into frequency-dependent weighting. The radiation-damage envelope Eq. (5.1845) provides the physical basis for exposure weighting of movie frames.

Let frame j correspond to accumulated exposure Φ_j. Its high-resolution structural amplitudes are expected to be attenuated by approximately

$$\boxed{q_j(R) = \exp\left[-\frac{\Phi_j}{2\Phi_c(R)}\right].} \tag{5.1874}$$

Because $\Phi_c(R)$ becomes smaller at high spatial frequency, Eq. (5.1874) has very different consequences at different resolutions:

$$\boxed{\begin{array}{l}\text{high } R \longrightarrow \text{late frames are strongly downweighted,}\\ \text{low } R \longrightarrow \text{late frames can remain useful much longer.}\end{array}} \tag{5.1875}$$

This is the essential physical idea behind the exposure-filtering scheme introduced by Grant and Grigorieff (2015). Their measurements showed that using a high total exposure need not require accepting all of the resulting high-resolution damage: early frames can provide the high-frequency information, while later frames continue to improve lower-frequency contrast and, in difficult cases, particle alignment.

Dose fractionation therefore does *not* prevent radiation damage. It preserves the time ordering of the damage so that information recorded before and after substantial irradiation can be treated differently.

Modern single-particle measurements. The practical consequences of this picture have been confirmed in single-particle datasets.

Grant and Grigorieff recorded rotavirus movies with a total exposure of approximately

$$100\ \mathrm{e}^{-}\ \text{Å}^{-2}, \tag{5.1876}$$

far above the exposure that would be used in a conventional single-image high-resolution acquisition. Because the data were recorded as movie frames, the high-resolution contribution of the later frames could be strongly attenuated while their lower-resolution information could still be retained (Grant and Grigorieff, 2015).

Their measurements also directly demonstrated the chemical selectivity discussed in Sec. 5.13.5. Carboxylate density, for example, faded substantially after only a few $e^- \text{Å}^{-2}$, whereas some aromatic side-chain and secondary-structure density remained visible at much larger accumulated exposures.

A recent systematic study of bacterial and human ribosomes provides a complementary perspective (Ważny et al., 2026). For both ribosomes, reconstructions improved rapidly as the total exposure increased toward approximately

$$\boxed{\Phi_{\text{tot}} \simeq 20\text{–}25\ e^- \text{Å}^{-2},} \tag{5.1877}$$

after which the gain largely saturated. Without exposure weighting, increasing the accumulated exposure beyond approximately 30 $e^- \text{Å}^{-2}$ progressively degraded the reconstructions. With exposure weighting, much of this degradation was suppressed.

The same study also examined *pre-exposure*. If a fixed set of frames was recorded only after the specimen had first received an additional exposure, the quality of the resulting reconstruction decreased systematically as the pre-exposure increased. This is a particularly direct demonstration that the early electrons contain structural information that cannot be recovered by simply increasing the exposure later.

The values in Eq. (5.1877) should not be interpreted as a universal optimum for SPA. They refer to large, high-contrast ribosomal particles, a particular microscope and detector, a particular reconstruction criterion, and frequency-dependent weighting. They nevertheless illustrate the general competition between electron statistics and radiation-induced information loss.

Critical exposure is specimen dependent. The critical exposure is not determined by spatial frequency alone. Heymann analyzed movie datasets containing several different biological specimens and found substantial specimen-to-specimen variation in $\Phi_c(R)$ even at the same accelerating voltage (Heymann, 2022).

This is physically reasonable. The effective fading measured from an image combines

- the chemical composition of the particle;
- the relative abundance of protein, nucleic acid, lipid, and solvent;
- site-specific radiation sensitivity;

- radiation-induced positional disorder;
- residual specimen motion;
- specimen thickness and surrounding ice;
- the microscope and detector configuration.

The critical-exposure curve is therefore best regarded as an empirical property of a particular specimen and imaging experiment, rather than as one universal function that applies identically to all biological materials.

Connection to the image-formation model. The result of this subsection can be incorporated directly into the image-formation model.

If the undamaged projected specimen contribution is

$$\widehat{s}_0(\mathbf{R}), \tag{5.1878}$$

then a simple shell-averaged model for the coherent contribution after accumulated exposure Φ is

$$\boxed{\widehat{s}(\mathbf{R}, \Phi) \simeq E_{\mathrm{rad}}(R, \Phi)\widehat{s}_0(\mathbf{R}),} \tag{5.1879}$$

with

$$E_{\mathrm{rad}}(R, \Phi) = \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right]. \tag{5.1880}$$

After passage through the microscope,

$$\boxed{\widehat{I}(\mathbf{R}, \Phi) \propto \mathrm{CTF}(\mathbf{R})E_{\mathrm{rad}}(R, \Phi)\widehat{s}_0(\mathbf{R}) + \widehat{n}(\mathbf{R}, \Phi),} \tag{5.1881}$$

where the notation suppresses the other envelopes and detector factors already developed in the electron-optics chapter.

At the power-spectrum level, the radiation term appears squared:

$$\boxed{S_{\mathrm{struct}}(\mathbf{R}, \Phi) \propto |\,\mathrm{CTF}(\mathbf{R})|^2 \exp\left[-\frac{\Phi}{\Phi_c(R)}\right] S_{\mathrm{struct}}(\mathbf{R}, 0).} \tag{5.1882}$$

Equation (5.1882) gives a compact description of why the high-frequency support of the CTF oscillations fades first as irradiation accumulates.

It must, however, be interpreted together with Sec. 5.13.5. The envelope E_{rad} describes the *average loss of coherent information*. The actual irradiated specimen is chemically heterogeneous and is not literally obtained by applying a single Gaussian blur to the pristine molecular structure.

The central picture is therefore

$$\boxed{\begin{aligned} \text{few electrons} &\longrightarrow \text{little damage but poor counting statistics,} \\ \text{intermediate exposure} &\longrightarrow \text{best balance of signal and damage,} \\ \text{large exposure} &\longrightarrow \text{low-resolution information may still accumulate,} \\ &\quad \text{while high-resolution information has largely saturated or faded.} \end{aligned}} \tag{5.1883}$$

5.13.7 The intrinsic spectral decay of the structural signal

The preceding subsection described how irradiation progressively reduces the coherent structural information that remains in the specimen. There is, however, another frequency dependence that must be separated from radiation damage: the structural signal is not spectrally white even before the first electron reaches the specimen.

In the notation of the electron-optics chapter, Eq. (4.2157) introduced $S_s(\mathbf{R})$ as the power spectral density (PSD) of the reproducible projected specimen signal. It is useful to characterize its broad spectral envelope through the radial average

$$\boxed{S_s^{\mathrm{rad}}(R) \equiv \langle S_s(\mathbf{R}) \rangle_{\|\mathbf{R}\|=R}, \qquad R = \|\mathbf{R}\|.} \tag{5.1884}$$

Empirically, over substantial frequency ranges, biological structures show a pronounced spectral falloff that can often be approximated by

$$\boxed{S_s^{\mathrm{rad}}(R) \propto R^{-\alpha}, \qquad \alpha \sim 2.} \tag{5.1885}$$

The value $\alpha = 2$ should not be interpreted as an exact or universal law of macromolecular structure. The local slope depends on the molecule, the frequency interval, its chemical composition and organization, and on whether a two-dimensional projected spectrum or a three-dimensional molecular spectrum is being considered. Rather, R^{-2} is a useful reference behavior around which real spectra vary (Vilas et al., 2020).

Why an inverse-square PSD is natural in two dimensions. There is a simple geometrical reason why an exponent close to two is natural for a two-dimensional signal containing structure over many spatial scales.

Consider an isotropic two-dimensional PSD $S(R)$. A Fourier annulus of radius R and thickness dR has area

$$dA_R = 2\pi R\, dR. \tag{5.1886}$$

The signal power contained in that annulus is therefore

$$dP = 2\pi R S(R)\, dR. \tag{5.1887}$$

Scale invariance and the inverse-square PSD.

Spatial scales are more naturally compared multiplicatively than additively. An octave, for example, extends from R to $2R$, irrespective of the absolute value of R. We therefore rewrite Eq. (5.1887) in terms of logarithmic frequency.

Since

$$dR = R\, d(\ln R), \tag{5.1888}$$

we obtain

$$dP = 2\pi R^2 S(R)\, d(\ln R), \tag{5.1889}$$

and consequently

$$\boxed{\frac{dP}{d\ln R} = 2\pi R^2 S(R).} \tag{5.1890}$$

If the signal is approximately scale balanced over some range, so that successive logarithmic frequency intervals contain comparable amounts of power,

$$\frac{dP}{d\ln R} \simeq \text{constant}. \tag{5.1891}$$

Equation (5.1890) then implies

$$R^2 S(R) \simeq \text{constant}, \tag{5.1892}$$

and therefore

$$\boxed{S(R) \propto R^{-2}.} \tag{5.1893}$$

Thus

$$\boxed{S(R) \propto R^{-2} \quad \Longleftrightarrow \quad \text{approximately equal power per logarithmic frequency interval.}} \tag{5.1894}$$

The exponent 2 in this argument is specifically associated with a *two-dimensional* PSD. More generally, equal power per logarithmic radial interval in d dimensions would require

$$S(R) \propto R^{-d}. \tag{5.1895}$$

This observation will become important below when we discuss Wilson statistics, which concerns a spherically averaged *three-dimensional* molecular spectrum.

A similar $R^{-\alpha}$ behavior with α frequently close to two is one of the classical statistical properties of natural images (Field, 1987; Ruderman and Bialek, 1994; van der Schaaf and van Hateren, 1996). The analogy should not be pushed too far—a macromolecular projection is not a photograph of a natural scene—but both contain organized structure over many spatial scales rather than a single characteristic length. Sharp interfaces provide another heuristic explanation: the Fourier amplitude of an ideal step decreases as R^{-1} in the direction normal to the edge, and its power therefore decreases as R^{-2}.

Macromolecular spectra have a robust intrinsic falloff. The decrease is also observed directly in spectra calculated from macromolecular structures. Using atomic models converted into Coulomb-potential maps with realistic electron atomic scattering factors, Vilas et al. (2020) found a pronounced non-flat spectral profile for several proteins, with a broad decay of approximately the same order as R^{-2} over part of the spectrum.

The overall falloff proved surprisingly robust. Replacing realistic atomic profiles by Gaussians, replacing carbon atoms by iron, binarizing the molecular potential, or randomly perturbing atomic coordinates by several angstroms changed details and absolute amplitudes but did not remove the broad spectral decrease. Specific molecular organization, on the other hand, produced localized deviations from this smooth envelope, associated with characteristic structural distances (Vilas et al., 2020).

This suggests a useful distinction:

$$\boxed{\begin{array}{l}\text{broad spectral envelope} \longleftarrow \text{overall geometry and multiscale organization,}\\ \text{local spectral deviations} \longleftarrow \text{specific molecular structure.}\end{array}} \tag{5.1896}$$

The power-law description is nevertheless only an empirical description of part of the spectrum. It cannot hold unchanged from zero frequency to atomic resolution. To understand why, it is useful to examine the three-dimensional molecular spectrum more closely.

Wilson statistics: from molecular shape to atomic scattering. We now temporarily switch from the two-dimensional projected signal $S_s(\mathbf{R})$

to the three-dimensional Coulomb potential $V(\mathbf{r})$ of a molecule. In this paragraph, $\mathbf{r} \in \mathbb{R}^3$ and $\mathbf{R} \in \mathbb{R}^3$.

Consider first the simplified model of N identical atoms with single-atom potential $f(\mathbf{r})$:

$$V(\mathbf{r}) = \sum_{a=1}^{N} f(\mathbf{r} - \mathbf{r}_a). \tag{5.1897}$$

Its Fourier transform is

$$\boxed{\widehat{V}(\mathbf{R}) = \widehat{f}(R) \sum_{a=1}^{N} \exp\left(-j2\pi \mathbf{R} \cdot \mathbf{r}_a\right).} \tag{5.1898}$$

The squared modulus contains two qualitatively different contributions:

$$|\widehat{V}(\mathbf{R})|^2 = |\widehat{f}(R)|^2 \left[N + \sum_{a \neq b} \exp\left(-j2\pi \mathbf{R} \cdot (\mathbf{r}_a - \mathbf{r}_b)\right) \right]. \tag{5.1899}$$

The first term is the sum of the individual atomic powers. The second contains interference between different atoms and therefore depends on their relative positions.

At low spatial frequency many of these phases are similar and the atoms scatter coherently. At higher frequencies the phases oscillate more rapidly, and radial or ensemble averaging progressively suppresses the positional cross terms. Wilson statistics formalizes this transition.

Singer provided a rigorous treatment using a random "bag of atoms" model (Singer, 2021). The first important point is that the molecular volume must increase with N. Otherwise, taking $N \to \infty$ while keeping the containing volume fixed would make the atomic density diverge.

At approximately constant atomic number density,

$$\text{volume} \propto N, \tag{5.1900}$$

so the molecular linear size scales as

$$\boxed{\text{linear size} \propto N^{1/3}.} \tag{5.1901}$$

Singer therefore writes

$$\mathbf{r}_a = N^{1/3} \mathbf{y}_a \tag{5.1902}$$

in dimensionless coordinates, where $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent and identically distributed random vectors with probability density $g(\mathbf{y})$.

Expected power in the random-bag-of-atoms model.

With

$$\mathbf{r}_a = N^{1/3}\mathbf{y}_a, \tag{5.1903}$$

Eq. (5.1898) becomes

$$\widehat{V}(\mathbf{R}) = \widehat{f}(R)\sum_{a=1}^{N}\exp\left(-j2\pi N^{1/3}\mathbf{R}\cdot\mathbf{y}_a\right). \tag{5.1904}$$

Its power is

$$|\widehat{V}(\mathbf{R})|^2 = |\widehat{f}(R)|^2\sum_{a=1}^{N}\sum_{b=1}^{N}\exp\left[-j2\pi N^{1/3}\mathbf{R}\cdot(\mathbf{y}_a-\mathbf{y}_b)\right]. \tag{5.1905}$$

For $a = b$, the exponential is one, giving N diagonal terms.
For $a \neq b$, define

$$\mathbf{Q} = N^{1/3}\mathbf{R}. \tag{5.1906}$$

Because $\mathbf{y}_a$ and $\mathbf{y}_b$ are independent,

$$\begin{aligned}\mathbb{E}\left\{e^{-j2\pi\mathbf{Q}\cdot(\mathbf{y}_a-\mathbf{y}_b)}\right\}\\ = \mathbb{E}\left\{e^{-j2\pi\mathbf{Q}\cdot\mathbf{y}_a}\right\}\mathbb{E}\left\{e^{+j2\pi\mathbf{Q}\cdot\mathbf{y}_b}\right\}.\end{aligned} \tag{5.1907}$$

The first factor is

$$\mathbb{E}\left\{e^{-j2\pi\mathbf{Q}\cdot\mathbf{y}_a}\right\} = \int_{\mathbb{R}^3} g(\mathbf{y})e^{-j2\pi\mathbf{Q}\cdot\mathbf{y}}d\mathbf{y} \tag{5.1908}$$

$$= \widehat{g}(\mathbf{Q}), \tag{5.1909}$$

while the second is

$$\widehat{g}(-\mathbf{Q}) = \widehat{g}^*(\mathbf{Q}), \tag{5.1910}$$

because g is real.
Consequently,

$$\boxed{\mathbb{E}\left\{e^{-j2\pi\mathbf{Q}\cdot(\mathbf{y}_a-\mathbf{y}_b)}\right\} = |\widehat{g}(\mathbf{Q})|^2.} \tag{5.1911}$$

There are $N(N-1)$ ordered pairs with $a \neq b$, and hence

$$\boxed{\mathbb{E}\left[|\widehat{V}(\mathbf{R})|^2\right] = |\widehat{f}(R)|^2 \left[N + N(N-1)\left|\widehat{g}(N^{1/3}\mathbf{R})\right|^2\right].} \tag{5.1912}$$

Equation (5.1912) gives a particularly clear decomposition:

$$\boxed{\begin{array}{lcl} N|\widehat{f}(R)|^2 & \longrightarrow & \text{individual-atom contribution,} \\ N(N-1)|\widehat{f}(R)|^2|\widehat{g}(N^{1/3}\mathbf{R})|^2 & \longrightarrow & \text{shape-dependent positional interference.} \end{array}} \tag{5.1913}$$

The Fourier transform $\widehat{g}$ describes the overall distribution of atoms. It is therefore concentrated at low spatial frequencies and decays as its argument increases. This is the mathematical mechanism by which the molecular-envelope term disappears.

There is also a useful geometrical interpretation. For two atoms separated by

$$d_{ab} = \|\mathbf{r}_a - \mathbf{r}_b\|, \tag{5.1914}$$

spherical averaging gives

$$\boxed{\frac{1}{4\pi}\int_{S^2} e^{-j2\pi R\boldsymbol{\omega}\cdot(\mathbf{r}_a-\mathbf{r}_b)}d\boldsymbol{\omega} = \frac{\sin(2\pi R d_{ab})}{2\pi R d_{ab}}.} \tag{5.1915}$$

At low R this factor is close to one; at larger R it oscillates and decays, producing cancellation among pairwise interference terms.

Where does the Wilson regime begin? Singer's analysis makes the phrase "high spatial frequency" quantitative. The essential issue is the rate at which the Fourier transform of the molecular envelope decays.

For a uniform distribution inside the unit ball,

$$g_B(\mathbf{y}) = \frac{3}{4\pi}\mathbf{1}_{\|\mathbf{y}\|\leq 1}, \tag{5.1916}$$

the Fourier transform can be calculated explicitly.

Molecular-envelope decay and the Wilson crossover.

Let

$$\kappa = \|\boldsymbol{\kappa}\|. \tag{5.1917}$$

Using spherical coordinates with the polar axis along $\boldsymbol{\kappa}$,

$$\widehat{g}_B(\kappa) = \frac{3}{4\pi}\int_{\|\mathbf{y}\|\le 1} e^{-j2\pi\boldsymbol{\kappa}\cdot\mathbf{y}} d^3\mathbf{y} \tag{5.1918}$$

$$= 3\int_0^1 r^2 \frac{\sin(2\pi\kappa r)}{2\pi\kappa r} dr. \tag{5.1919}$$

Performing the integral gives

$$\boxed{\widehat{g}_B(\kappa) = -\frac{3\cos(2\pi\kappa)}{4\pi^2\kappa^2} + \frac{3\sin(2\pi\kappa)}{8\pi^3\kappa^3}.} \tag{5.1920}$$

Therefore,

$$|\widehat{g}_B(\kappa)| = O(\kappa^{-2}), \tag{5.1921}$$

and

$$\boxed{|\widehat{g}_B(\kappa)|^2 = O(\kappa^{-4}).} \tag{5.1922}$$

For a broad class of three-dimensional convex or sufficiently smooth molecular envelopes, Singer shows that the *spherically averaged* shape power has the same asymptotic R^{-4} dependence. Thus, from Eq. (5.1912), the shape-dependent term scales as

$$N^2\left(N^{1/3}R\right)^{-4} = N^{2/3}R^{-4}. \tag{5.1923}$$

The Wilson term scales as N. The shape term becomes asymptotically smaller when

$$N^{2/3}R^{-4} \ll N, \tag{5.1924}$$

which gives

$$\boxed{R \gg N^{-1/12}} \tag{5.1925}$$

in the dimensionless coordinates of the model.

The very small exponent $1/12$ is therefore not empirical. It arises from the combination of the three-dimensional size scaling $N^{1/3}$ and the R^{-4} decay of the spherically averaged shape power.

For a quantitative estimate, Singer uses the bound

$$\boxed{|\widehat{g}_B(\kappa)| \lesssim \frac{C}{\kappa^2}, \qquad C \simeq 0.081,} \tag{5.1926}$$

and defines a practical crossover as the frequency at which the shape-dependent contribution has fallen to 30% of the Wilson term.

Let N_c be the number of carbon-atom equivalents and let a be the physical molecular radius. The resulting condition gives

$$\boxed{R_c = 0.3^{-1/4} C^{1/2} \frac{N_c^{1/4}}{a}.} \tag{5.1927}$$

Singer uses the approximate protein mass density

$$\rho_{\text{prot}} \simeq 0.8 \text{ Da Å}^{-3}, \tag{5.1928}$$

together with approximately 9.1 carbon-atom equivalents per 110 Da amino-acid mass:

$$N_c = M_W \frac{9.1}{110}, \tag{5.1929}$$

and, for a spherical protein,

$$M_W = \frac{4\pi}{3} a^3 \rho_{\text{prot}}. \tag{5.1930}$$

Substitution into Eq. (5.1927) gives

$$\boxed{R_c \simeq 0.279 \text{ Å}^{-1} \left(\frac{a}{1 \text{ Å}}\right)^{-1/4}.} \tag{5.1931}$$

Equivalently, eliminating a in favor of molecular weight gives

$$\boxed{R_c \simeq 0.309 \text{ Å}^{-1} \left(\frac{M_W}{1 \text{ Da}}\right)^{-1/12}.} \tag{5.1932}$$

Equations (5.1931) and (5.1932) explain a striking property of the Wilson crossover: it depends only very weakly on molecular size,

$$\boxed{R_c \propto a^{-1/4} \propto M_W^{-1/12}.} \tag{5.1933}$$

For example, changing the molecular weight by a factor of 100 changes the crossover frequency by only

$$100^{1/12} \simeq 1.47. \tag{5.1934}$$

Representative values are

$$\boxed{\begin{array}{llll} a \simeq 100\ \text{Å}, & M_W \simeq 3.3\ \text{MDa} & \longrightarrow & R_c \simeq 0.088\ \text{Å}^{-1}, \\ & & & d_c \simeq 11.3\ \text{Å}, \\ a \simeq 25\ \text{Å}, & M_W \simeq 52\ \text{kDa} & \longrightarrow & R_c \simeq 0.125\ \text{Å}^{-1}, \\ & & & d_c \simeq 8.0\ \text{Å}, \end{array}} \tag{5.1935}$$

where $d_c = 1/R_c$. Across molecular radii from roughly 20 to 150 Å, the predicted crossover remains within the rather narrow resolution range of approximately 7.5–12.5 Å (Singer, 2021).

What the Wilson limit means for real macromolecules. The Wilson crossover should not be interpreted as the frequency above which the physical electron-scattering spectrum becomes flat.

From Eq. (5.1912), once the shape-dependent term has become small,

$$\boxed{\mathbb{E}\left[|\widehat{V}(\mathbf{R})|^2\right] \simeq N|\widehat{f}(R)|^2.} \tag{5.1936}$$

A genuinely flat Wilson spectrum follows only for ideal point atoms, for which

$$f(\mathbf{r}) = \delta(\mathbf{r}), \qquad \widehat{f}(R) = 1. \tag{5.1937}$$

Real atoms have spatial extent. Their electron atomic scattering factors decrease with spatial frequency. For several atomic species, the corresponding Wilson-like limit is therefore

$$\boxed{P_V(R) \simeq \sum_a |f_{e,a}(R)|^2,} \tag{5.1938}$$

which retains an intrinsic frequency dependence (Sorzano et al., 2015; Vilas et al., 2020).

Furthermore, the random-bag-of-atoms model deliberately neglects the correlations that make a protein a protein. Covalent geometry, α-helices, β-sheets, tertiary packing, nucleic-acid helices, and other ordered structures produce systematic departures from the smooth Wilson envelope. For example, characteristic distances associated with secondary structure generate spectral features near approximately 10 Å and 4.5 Å (Singer, 2021).

Thus Wilson statistics describes the increasingly universal *radially averaged envelope*; the detailed amplitudes, their correlations, and their phases retain the specific molecular information.

Putting the spectral regimes together. The approximate $R^{-\alpha}$ behavior and Wilson statistics are therefore complementary descriptions, not competing universal laws.

The former is a useful empirical description of the broad decay of the two-dimensional projected structural PSD over part of its frequency range. Wilson statistics instead explains what happens to the three-dimensional spherically averaged molecular spectrum when long-range positional interference is progressively cancelled.

The resulting picture is summarized in Table 5.15.

Table 5.15: Schematic frequency regimes of the macromolecular structural spectrum. The transitions are gradual and the regimes overlap.

Frequency range	**Dominant origin**	**Spectral behavior**
Very low R	Overall molecular size and shape	Guinier regime; a pure power law cannot extend to $R = 0$.
Intermediate R	Multiscale organization of the projected structure	Broad power-law-like decay, $S_s^{\mathrm{rad}}(R) \sim R^{-\alpha}$, with α often of order 2.
Medium–high R	Progressive cancellation of interference between widely separated atoms	Transition toward the Wilson regime; dependence on global molecular shape becomes much weaker.
Highest R	Atomic scattering factors, local atomic correlations, positional disorder	Non-flat atomic envelope with superimposed structural features.

The table should not be interpreted as defining sharp universal boundaries. In particular, Singer's rigorous result concerns a *three-dimensional spherical average*, whereas Eq. (5.1885) describes the broad envelope of a *two-dimensional projected signal*. Wilson statistics therefore does not constitute a derivation of the two-dimensional R^{-2} law.

Rather than assigning one universal exponent to the entire projected spectrum, we may define its local logarithmic slope

$$\boxed{\alpha_{\mathrm{eff}}(R) \equiv -\frac{d \ln S_s^{\mathrm{rad}}(R)}{d \ln R}.} \tag{5.1939}$$

For a perfect power law, $\alpha_{\mathrm{eff}}(R) = \alpha$ is constant. For a real macro-

molecule it varies with R and contains superimposed deviations associated with particular structural length scales. Thus $\alpha \simeq 2$ should be regarded as a useful characteristic value over part of the projected spectrum, not as a fundamental constant of biological matter.

Intrinsic spectral decay and radiation damage are different effects. We can now return to the irradiated specimen.

Let

$$S_s^{(0)}(\mathbf{R}) \tag{5.1940}$$

denote the PSD of the undamaged projected structural signal. Its broad radial envelope may, over an appropriate intermediate range, be approximated as

$$\boxed{S_s^{(0),\mathrm{rad}}(R) \simeq A\left(\frac{R}{R_0}\right)^{-\alpha}, \qquad \alpha \sim 2,} \tag{5.1941}$$

where R_0 is an arbitrary reference frequency.

Radiation damage introduces a conceptually separate attenuation. Using the fading function defined in Eq. (5.1831),

$$\boxed{S_s(\mathbf{R},\Phi) \simeq S_s^{(0)}(\mathbf{R}) F_{\mathrm{rad}}(R,\Phi).} \tag{5.1942}$$

If the single-exponential approximation of Eq. (5.1833) is adequate,

$$\boxed{F_{\mathrm{rad}}(R,\Phi) \simeq \exp\left[-\frac{\Phi}{\Phi_c(R)}\right].} \tag{5.1943}$$

Hence the broad radial envelope is approximately

$$\boxed{S_s^{\mathrm{rad}}(R,\Phi) \simeq A\left(\frac{R}{R_0}\right)^{-\alpha} \exp\left[-\frac{\Phi}{\Phi_c(R)}\right].} \tag{5.1944}$$

The two factors describe different physics:

$$\boxed{\begin{array}{rcl} R^{-\alpha} & \longrightarrow & \text{intrinsic spectrum of the undamaged structure,} \\ \exp[-\Phi/\Phi_c(R)] & \longrightarrow & \text{additional loss produced by irradiation.} \end{array}} \tag{5.1945}$$

Finally, the microscope acts on this already non-white and exposure-dependent structural signal. Using the PSD image-formation relation derived in Eq. (4.2158),

$$\boxed{S_{\mathrm{image},s}(\mathbf{R},\Phi) \simeq |\,\mathrm{CTF}(\mathbf{R})|^2 S_s^{(0)}(\mathbf{R}) F_{\mathrm{rad}}(R,\Phi),} \tag{5.1946}$$

before including the detector MTF and the other envelopes introduced in the electron-optics chapter.

Thus three different mechanisms must not be conflated:

$$\boxed{\begin{aligned} \text{molecular structure} &\longrightarrow S_s^{(0)}(\mathbf{R}), \\ \text{irradiation} &\longrightarrow F_{\text{rad}}(R, \Phi), \\ \text{electron optics} &\longrightarrow |\,\text{CTF}(\mathbf{R})|^2. \end{aligned}} \tag{5.1947}$$

The envelope supporting the CTF oscillations therefore decreases with spatial frequency even for an undamaged specimen. Radiation damage progressively steepens that loss, particularly at high spatial frequencies, and the CTF finally modulates the remaining structural power into the oscillatory pattern observed in the micrograph power spectrum.

5.13.8 Radiation damage of vitreous ice itself

The vitreous ice surrounding a biological particle should not be regarded as an inert embedding medium. It is itself a major component of the specimen scattering potential, it undergoes both elastic and inelastic interaction with the imaging electrons, and its molecular structure changes progressively under irradiation.

This point is particularly important in cryo-EM because the volume occupied by the ice is often comparable to, or much larger than, the volume occupied by the macromolecule. The thickness and density of the water or ice layer therefore influence scattering, contrast, chromatic broadening, and the achievable resolution (de Jonge, 2018). More importantly for the present discussion, water is itself highly susceptible to ionizing radiation. Glaeser (2016) emphasizes that vitreous ice can be at least as radiation sensitive as the biological material embedded within it.

The consequences occur on several levels:

$$\boxed{\begin{aligned} \text{water radiolysis} &\longrightarrow \text{radicals and molecular products}, \\ \text{local rearrangement} &\longrightarrow \text{loss of high-resolution Fourier correlation}, \\ \text{larger accumulated exposure} &\longrightarrow \text{mass loss, thinning, and eventually gas formation.} \end{aligned}} \tag{5.1948}$$

Thus the specimen background against which the particle is imaged is itself changing while the image is being recorded.

Radiolysis of water continues in the vitreous state. The microscopic origin is the same inelastic energy deposition developed in Sec. 5.13.3. An imaging electron can excite or ionize a water molecule,

$$\mathrm{H_2O} \xrightarrow{\text{ionizing radiation}} \mathrm{H_2O^*}, \tag{5.1949}$$

or

$$\mathrm{H_2O} \xrightarrow{\text{ionization}} \mathrm{H_2O^{+\bullet}} + e_{\mathrm{s}}^{-}. \tag{5.1950}$$

The primary products then evolve through a network of reactions. Important intermediates and products include

$$\boxed{e_{\mathrm{aq}}^{-}, \quad \mathrm{H^\bullet}, \quad \mathrm{OH^\bullet}, \quad \mathrm{HO_2^\bullet}, \quad \mathrm{H_3O^+}, \quad \mathrm{OH^-}, \quad \mathrm{H_2}, \quad \mathrm{H_2O_2}.} \tag{5.1951}$$

For example, representative reactions include

$$\mathrm{H_2O^{+\bullet}} + \mathrm{H_2O} \longrightarrow \mathrm{H_3O^+} + \mathrm{OH^\bullet}, \tag{5.1952}$$

$$\mathrm{H_2O^*} \longrightarrow \mathrm{H^\bullet} + \mathrm{OH^\bullet}, \tag{5.1953}$$

$$2\mathrm{H^\bullet} \longrightarrow \mathrm{H_2}, \tag{5.1954}$$

$$2\mathrm{OH^\bullet} \longrightarrow \mathrm{H_2O_2}. \tag{5.1955}$$

The notation e_{aq}^{-} is conventional in water-radiolysis chemistry; in vitreous ice the electron may more properly be regarded as a localized or trapped excess electron rather than as an equilibrium hydrated electron in liquid water.

The characteristic sequence is extremely rapid. A useful approximate time hierarchy is summarized in Table 5.16.

Cooling strongly suppresses long-range diffusion, as discussed in Sec. 5.13.4, but it does not suppress the primary excitation and ionization events. The local hydrogen-bond network can therefore be chemically modified even when the resulting species remain trapped within approximately the same region.

In pure ice, many radiolytic products can subsequently recombine. Consequently, the specimen may show surprisingly little obvious macroscopic chemical change despite substantial microscopic radiolysis (Glaeser, 2016). Absence of a visible change should therefore not be interpreted as absence of radiation damage.

Vitreous ice has its own structural spectrum. Before considering its evolution under irradiation, recall that unirradiated amorphous ice is not structurally featureless.

Table 5.16: Approximate stages in the radiolysis of water. The times indicate orders of magnitude rather than sharply separated kinetic regimes (Glaeser, 2016).

Stage	Time scale	Main processes
Physical	$\sim 10^{-15}$ s	Electronic excitation and ionization of water.
Physicochemical	$\sim 10^{-12}$ s	Formation and localization of electrons, proton transfer, dissociation, and formation of $H^\bullet$ and $OH^\bullet$.
Chemical	up to $\sim 10^{-6}$ s	Secondary reactions among radicals and formation of products such as H_2 and H_2O_2.

Low-density amorphous ice, the form relevant to ordinary cryo-EM, has a density of approximately

$$\rho_{\mathrm{LDA}} \simeq 0.94\ \mathrm{g\,cm^{-3}}, \tag{5.1956}$$

and lacks long-range crystalline order while retaining strong short-range correlations between neighboring water molecules (Parkhurst et al., 2024).

These correlations generate a non-white structural PSD. This was developed explicitly in Sec. 4.16.4 of Chap. 4. In particular, amorphous ice contributes to the elastic structural background $S_{\mathrm{el}}(\mathbf{R})$ introduced in Eq. (4.2157) and is therefore transferred through the microscope according to Eq. (4.2158).

Different structural correlations appear at different characteristic distances. Atomistic simulations emphasize the nearest-neighbor oxygen–oxygen distance around

$$\boxed{d_{\mathrm{OO}} \simeq 2.88\ \text{Å},} \tag{5.1957}$$

which Parkhurst et al. (2024) use explicitly in their statistical model of the amorphous-ice PSD.

A second important feature occurs around

$$\boxed{d_{\mathrm{water}} \simeq 3.7\ \text{Å}, \qquad R_{\mathrm{water}} \simeq \frac{1}{3.7\ \text{Å}},} \tag{5.1958}$$

as already introduced in Eq. (4.2171). This strong "water ring" is associated with next-nearest-neighbor oxygen atoms in the tetrahedrally coordinated water network (McMullan et al., 2015).

Thus the vitreous ice supplies genuine elastic structural information over a broad frequency range. It is this structural power, rather than detector noise, that can support Thon-ring oscillations.

What exactly changes during irradiation? At this point an important distinction is required.

Let

$$\widehat{V}_{\text{ice}}(\mathbf{R}, \Phi) \tag{5.1959}$$

denote the Fourier transform of the instantaneous projected ice potential after accumulated exposure Φ.

One may define the instantaneous ensemble power

$$\boxed{S_{\text{ice}}^{\text{inst}}(R, \Phi) \equiv \left\langle \left| \widehat{V}_{\text{ice}}(\mathbf{R}, \Phi) \right|^2 \right\rangle_{\|\mathbf{R}\|=R}.} \tag{5.1960}$$

It might appear natural to assume that radiation damage simply causes this quantity to decrease monotonically. That is not necessarily what happens.

Irradiation can destroy one local arrangement of water molecules while creating another statistically similar arrangement. The radial distribution of intermolecular distances, and hence the instantaneous ensemble PSD, may therefore remain approximately similar even though the actual molecular configuration has changed substantially.

The quantity that is more directly sensitive to this rearrangement is the *cross-spectrum* between two stages of the exposure:

$$\boxed{C_{\text{ice}}(\mathbf{R}; \Phi_1, \Phi_2) \equiv \left\langle \widehat{V}_{\text{ice}}(\mathbf{R}, \Phi_1) \widehat{V}_{\text{ice}}^*(\mathbf{R}, \Phi_2) \right\rangle.} \tag{5.1961}$$

Thus

$$\boxed{\begin{array}{ccc} \text{equal-time PSD} & \longrightarrow & \text{how much structural power exists,} \\ \text{cross-time spectrum} & \longrightarrow & \text{whether it is still the same structure.} \end{array}} \tag{5.1962}$$

This distinction is the key to understanding the apparently paradoxical behavior of irradiated vitreous ice. The diffraction intensity can remain substantial even while successive high-resolution images of the ice are no longer mutually coherent.

Radiation-induced decorrelation of the ice structure. The dose-fractionated measurements of McMullan et al. (2015) provide a quantitative description of this phenomenon.

As derived in detail in Sec. 4.16.4, they can be described approximately by a Gaussian Markov displacement model in which the variance of the water-molecule displacement grows linearly with accumulated exposure. The resulting Fourier correlation coefficient is

$$\boxed{\rho_{\text{ice}}(R, \Delta\Phi) = \exp\left[-2\pi^2\sigma_0^2 R^2 \Delta\Phi\right],} \tag{5.1963}$$

which is the result derived previously in Eq. (4.2197).

Here σ_0^2 is the displacement variance in one Cartesian direction per unit exposure. The measurements at 300 keV gave approximately

$$\boxed{\sigma_0^2 \simeq 0.37\ \frac{\text{Å}^2}{\text{e}^-\ \text{Å}^{-2}}.} \tag{5.1964}$$

The most important feature of Eq. (5.1963) is the factor

$$R^2 \Delta\Phi. \tag{5.1965}$$

A given molecular displacement therefore destroys high-frequency correlation much more rapidly than low-frequency correlation.

Characteristic displacement and exposure scales for vitreous ice.

For isotropic independent displacements, Eq. (5.1964) implies

$$\left\langle \|\Delta\mathbf{r}\|^2 \right\rangle = 3\sigma_0^2 \Delta\Phi. \tag{5.1966}$$

Thus the three-dimensional RMS displacement after exposure Φ is

$$\boxed{u_{\text{rms}}(\Phi) = \sqrt{3\sigma_0^2 \Phi}.} \tag{5.1967}$$

For

$$\Phi = 25\ \text{e}^-\ \text{Å}^{-2}, \tag{5.1968}$$

and $\sigma_0^2 \simeq 0.37\ \text{Å}^2/(\text{e}^-\ \text{Å}^{-2})$,

$$u_{\text{rms}} \simeq 5.3\ \text{Å}, \tag{5.1969}$$

consistent with the estimate of McMullan et al. (2015).

We may also define the exposure separation at which the Fourier correlation at frequency R has fallen to $1/e$:

$$\rho_{\text{ice}}\left(R, \Phi_{\text{ice},c}\right) = e^{-1}. \tag{5.1970}$$

Using Eq. (5.1963),

$$\boxed{\Phi_{\text{ice},c}(R) = \frac{1}{2\pi^2\sigma_0^2 R^2}.} \tag{5.1971}$$

Thus,

$$\boxed{\Phi_{\text{ice},c}(R) \propto R^{-2}.} \tag{5.1972}$$

At the strong water-ring frequency $R = 1/(3.7\ \text{Å})$, Eq. (5.1971) gives approximately

$$\boxed{\Phi_{\text{ice},c} \simeq 1.9\ \text{e}^-\ \text{Å}^{-2}.} \tag{5.1973}$$

The value in Eq. (5.1973) is not the optimum exposure for measuring the water-ring Thon rings. It is the exposure separation required for the *correlation* between two configurations to fall to $1/e$. The optimum exposure also involves counting statistics and was shown in Eq. (4.2243) to be approximately

$$4\ \text{e}^-\ \text{Å}^{-2} \tag{5.1974}$$

for the 3.7 Å feature.

Why the high-resolution Thon rings of vitreous ice are weak. This framework resolves an observation that had long appeared surprising. Vitreous ice generally gives much weaker Thon rings than amorphous carbon.

At low and medium spatial frequencies there is already an intrinsic reason. Density fluctuations in water are relatively weak, and the mass density of vitreous ice is much lower than that of typical amorphous carbon films. The corresponding elastic structure factor is therefore weak (Glaeser, 2016).

At high spatial frequencies, however, this explanation is insufficient. Vitreous ice possesses the strong water-ring feature of Eq. (5.1958), so one might expect particularly prominent Thon rings around 3–4 Å.

The experiment of McMullan et al. (2015) demonstrated why this expectation fails. They compared two ways of combining a dose-fractionated movie:

1. sum the movie frames and then calculate one power spectrum;

2. calculate one power spectrum per frame and then sum those power spectra.

The distinction is fundamental. In the first procedure, Fourier *amplitudes* from different times are added and therefore require the specimen structure to remain correlated. In the second, the squared Fourier amplitudes are added and no inter-frame phase coherence is required.

The experiment gave opposite behavior for amorphous carbon and vitreous ice:

Table 5.17: Qualitative behavior observed by McMullan et al. (2015) in dose-fractionated images of stabilized amorphous carbon and vitreous ice.

Specimen	**Thon rings strongest in**	**Interpretation**
Pre-irradiated amorphous carbon	Power spectrum of the summed image	The structural configuration remains approximately fixed and Fourier amplitudes add coherently.
Vitreous ice	Sum of the individual-frame power spectra	The high-resolution molecular configuration changes between frames and inter-frame Fourier coherence is lost.

Thus radiation damage does not merely weaken the instantaneous water-ring power. It continuously replaces the particular structural features responsible for that power by new, statistically similar features.

In the terminology introduced above,

$$\boxed{S_{\text{ice}}^{\text{inst}}(R,\Phi) \quad \text{may change only modestly, while} \quad C_{\text{ice}}(R;\Phi_1,\Phi_2) \quad \text{decays rapidly.}} \tag{5.1975}$$

This is why a diffraction pattern or a short-frame power spectrum can retain the water ring while a long coherently integrated image suffers strong high-resolution blurring (Glaeser, 2016; McMullan et al., 2015).

The accumulated ice PSD is therefore exposure dependent. The cross-time decorrelation has a direct consequence for the PSD of a finite exposure.

The detailed derivation was already given in Sec. 4.16.4. For total exposure D, the resulting ice structural power is

$$\boxed{S_{\text{ice}}^{\text{acc}}(R;D) = D^2 S_{\text{ice},0}(R) G_{\text{ice}}\left(\alpha_R D\right),} \tag{5.1976}$$

with

$$\alpha_R = 2\pi^2 \sigma_0^2 R^2 \tag{5.1977}$$

from Eq. (4.2182), and G_{ice} defined in Eq. (4.2220).

For an exposure short enough that

$$\alpha_R D \ll 1, \tag{5.1978}$$

the configuration remains nearly fixed and

$$S_{\text{ice}}^{\text{acc}} \simeq D^2 S_{\text{ice},0}. \tag{5.1979}$$

The amplitudes from successive electrons add coherently.

When

$$\alpha_R D \gg 1, \tag{5.1980}$$

the configuration decorrelates during the exposure and the accumulated power grows only approximately linearly with D, as derived in Eq. (4.2228).

Therefore the "ice PSD" appearing in an experimentally integrated micrograph is not a fixed property of an immutable support medium. It depends on

$$\boxed{R, \qquad D, \qquad \text{and the temporal grouping of the movie frames.}} \tag{5.1981}$$

Connection to the CTF and the support of the Thon rings. This result connects directly to the power-spectrum image-formation model derived in Chap. 4.

For a static specimen, Eq. (4.2157) gave

$$S_{\text{spec}}(\mathbf{R}) = S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R}), \tag{5.1982}$$

where the surrounding vitreous ice is an important contribution to S_{el}.

For an irradiated cryo-EM movie, this expression should be understood more generally as

$$\boxed{\begin{aligned} S_{\text{spec}}(\mathbf{R};D) = S_s(\mathbf{R};D) + S_{\text{ice}}^{\text{acc}}(R;D) \\ + S_{\text{other,el}}(\mathbf{R};D), \end{aligned}} \tag{5.1983}$$

where $S_{\text{ice}}^{\text{acc}}$ includes the temporal decorrelation described by Eq. (5.1976).

The detected ice contribution is therefore, as already derived in Eq. (4.2229),

$$\boxed{\begin{aligned} S_{\text{ice,det}}(\mathbf{R}; D) = {} & |\,\text{MTF}(\mathbf{R})|^2 |\,\text{CTF}(\mathbf{R})|^2 \\ & \times D^2 S_{\text{ice},0}(R) G_{\text{ice}}(\alpha_R D). \end{aligned}} \tag{5.1984}$$

This equation is particularly important conceptually. The structural power that supports the visible Thon rings is not perfectly stationary throughout the exposure:

$$\boxed{\begin{aligned} S_{\text{ice},0}(R) &\longrightarrow \text{intrinsic short-range structure of the ice,} \\ G_{\text{ice}}(\alpha_R D) &\longrightarrow \text{loss of temporal coherence under irradiation,} \\ |\,\text{CTF}(\mathbf{R})|^2 &\longrightarrow \text{electron-optical Thon-ring modulation.} \end{aligned}} \tag{5.1985}$$

Since $\alpha_R \propto R^2$, the high-frequency part of the ice structural support loses coherence first. This provides a direct image-formation interpretation of the weak high-resolution Thon rings emphasized by Glaeser (2016).

A statistical model of the undamaged ice background. The recent simulations of Parkhurst et al. (2024) provide a useful complementary view. Instead of representing every water molecule explicitly, they showed that the amorphous-ice potential can be replaced, for image-simulation purposes, by a Gaussian random field having the same mean, variance, and PSD as an atomistic molecular-dynamics model.

Their model writes the broad radial PSD approximately as the sum of two Gaussian components,

$$\boxed{S_{\text{ice},0}(R) \simeq a_1 \exp\left(-\frac{R^2}{2s_1^2}\right) + a_2 \exp\left[-\frac{(R - R_{\text{OO}})^2}{2s_2^2}\right],} \tag{5.1986}$$

with

$$R_{\text{OO}} \simeq \frac{1}{2.88\ \text{Å}}. \tag{5.1987}$$

This particular parameterization should not be regarded as a universal analytical law for vitreous ice. Its significance here is that realistic cryo-EM images require a *structured* random ice potential rather than white noise or a spatially constant background.

Parkhurst and coworkers found that the exit-wave statistics and power spectra obtained from the Gaussian-random-field model were essentially the

same as those produced by a many-million-atom molecular model, and the simulated images also reproduced experimental statistics (Parkhurst et al., 2024).

There is an interesting connection with the radiation-damage picture. In their Gaussian-random-field implementation the exact microscopic water configuration need not remain fixed between independently generated samples. This would be inappropriate for a truly static solid, but is less problematic for irradiated vitreous ice because the high-resolution molecular configuration actually does change during electron exposure. For a quantitative simulation of a *dose-fractionated movie*, however, successive ice configurations should not be chosen independently: their correlation should follow a model such as Eq. (5.1963).

Radiation-induced thinning of the ice. Structural rearrangement is not the only way in which irradiation changes the ice. At sufficiently large exposure, water molecules and radiolytic fragments can leave the specimen surface through *electron-stimulated desorption*. The ice therefore becomes progressively thinner.

Measurements reviewed by Glaeser (2016) indicate that, for 100 keV electrons and temperatures below approximately 90 K, the mass-loss rate found by Heide (1984) was of order one molecular layer of water per

$$25\ \mathrm{e}^{-}\ \text{Å}^{-2}. \tag{5.1988}$$

The desorption rate rose strongly above approximately 100 K. For 300 keV electrons, Glaeser estimates a smaller rate, roughly scaling with the reduced inelastic-scattering cross section, although different experimental estimates of the absolute thinning rate do not agree closely.

At moderate exposure one may represent the effect phenomenologically as

$$t_{\mathrm{ice}} = t_{\mathrm{ice}}(\Phi), \qquad \frac{dt_{\mathrm{ice}}}{d\Phi} < 0. \tag{5.1989}$$

This introduces yet another time dependence into image formation. Changing t_{ice} changes not only the amount of elastic structural power but also the probabilities of elastic and inelastic scattering and, for sufficiently thick specimens, the angular and energy broadening analyzed for water and ice layers by de Jonge (2018).

Thus the ice evolves in two distinct ways:

$$\boxed{\begin{aligned} \text{internal rearrangement} &\longrightarrow \text{loss of Fourier correlation,} \\ \text{desorption} &\longrightarrow \text{loss of material and changing specimen thickness.} \end{aligned}} \tag{5.1990}$$

Molecular products and bubbling. Radiolysis generates stable molecular products in addition to the short-lived radicals discussed above. Molecular hydrogen is particularly important because it can accumulate in irradiated hydrated biological specimens. Electron energy-loss spectroscopy (EELS) measurements by Leapman and Sun (1995) identified the characteristic excitation of molecular hydrogen near

$$\Delta E_{\mathrm{H}_2} \simeq 13 \text{ eV}, \tag{5.1991}$$

and showed that visible hydrogen bubbles appear after accumulated electron fluences larger than approximately

$$\boxed{D_{\text{bubble}} \gtrsim 100 \ e^{-}\text{Å}^{-2}.} \tag{5.1992}$$

The molecular hydrogen contained in these bubbles can be highly compressed, with pressures exceeding

$$p_{\mathrm{H}_2} > 10^3 \text{ atm} \simeq 10^8 \text{ Pa} = 0.1 \text{ GPa}. \tag{5.1993}$$

At sufficiently high accumulated dose, radiation chemistry therefore ceases to be only a molecular-scale modification of the specimen: production and confinement of gaseous radiolysis products can generate mechanically significant pressures and ultimately macroscopic rearrangements of the ice and embedded specimen.

The situation in pure ice is more subtle. Radiolytic products can recombine or escape through the nearby ice–vacuum interfaces, and pure vitreous water at liquid-nitrogen temperature does not necessarily form visible bubbles even after substantial irradiation (Glaeser, 2016). Frozen solutions containing biological or other organic material bubble much more readily, presumably because both their radiation chemistry and the availability of nucleation sites differ from those of pure water. Visible bubbling must therefore be regarded as a relatively late and dramatic manifestation of radiation damage. The high-resolution structural rearrangements measured by McMullan et al. (2015), for example, occur after exposures of only a few $\mathrm{e}^{-}\text{Å}^{-2}$, far below the approximate bubbling threshold in Eq. (5.1992).

5.13.9 Elastic nuclear recoil: knock-on displacement and sputtering

The radiation-damage mechanisms discussed in the previous sections originate predominantly from *inelastic* scattering: the incident electron excites or ionizes the electronic system, and the subsequent cascade of secondary electrons, radicals, and chemical reactions modifies the specimen. There is, however, a second primary mechanism that does not pass through this electronic excitation cascade. An imaging electron can scatter elastically through a sufficiently large angle from an atomic nucleus and transfer enough momentum to move the nucleus itself.

This process is usually called *knock-on damage* or *displacement damage*. In a crystalline solid, the atom may be removed from its lattice site, leaving a vacancy and, if it remains inside the material, producing an interstitial defect. At a free surface, the same collision may eject the atom into vacuum; this surface analogue of knock-on displacement is called *sputtering* (Egerton et al., 2004).

Energy transferred to a nucleus. The kinematics of elastic recoil were introduced in the quantum appendix. In particular, Eq. (A.419) showed that a body of mass M, initially at rest and receiving momentum $\mathbf{q}$, acquires the recoil energy

$$T = \frac{q^2}{2M}. \tag{5.1994}$$

For radiation damage we apply the same argument not to the centre of mass of the complete specimen but to an individual atomic nucleus of mass M.

Let the incident electron have momentum

$$\mathbf{p}_i, \tag{5.1995}$$

and let the scattered electron have momentum

$$\mathbf{p}_f. \tag{5.1996}$$

We denote by θ the angle between the incident and scattered electron momenta,

$$\boxed{\cos\theta = \frac{\mathbf{p}_i \cdot \mathbf{p}_f}{|\mathbf{p}_i|, |\mathbf{p}_f|}.} \tag{5.1997}$$

Thus, $\theta = 0$ corresponds to forward scattering, whereas $\theta = \pi$ corresponds to complete backscattering.

Because the nucleus is initially at rest, conservation of momentum requires that the momentum gained by the nucleus be equal to the momentum lost by the electron. If $\mathbf{q}$ denotes the momentum transferred to the nucleus, then

$$\boxed{\mathbf{q} = \mathbf{p}_i - \mathbf{p}_f.} \tag{5.1998}$$

We now calculate the magnitude of this transferred momentum. Squaring Eq. (5.1998),

$$q^2 = |\mathbf{p}_i - \mathbf{p}_f|^2 \qquad = (\mathbf{p}_i - \mathbf{p}_f) \cdot (\mathbf{p}_i - \mathbf{p}_f) \tag{5.1999}$$

$$= p_i^2 + p_f^2 - 2\,\mathbf{p}_i \cdot \mathbf{p}_f. \tag{5.2000}$$

Using the definition of the scattering angle in Eq. (5.1997),

$$\mathbf{p}_i \cdot \mathbf{p}_f = p_i p_f \cos\theta, \tag{5.2001}$$

so that

$$q^2 = p_i^2 + p_f^2 - 2p_i p_f \cos\theta. \tag{5.2002}$$

In an elastic electron–nucleus collision, the nucleus receives only a very small fraction of the electron energy because

$$M \gg m_0. \tag{5.2003}$$

Consequently, the change in the *magnitude* of the electron momentum is very small, even though its direction may change substantially. We may therefore write

$$p_i \simeq p_f \simeq p. \tag{5.2004}$$

Equation (5.2002) then becomes

$$q^2 \simeq p^2 + p^2 - 2p^2 \cos\theta \tag{5.2005}$$

$$= 2p^2(1 - \cos\theta). \tag{5.2006}$$

Using the trigonometric identity

$$1 - \cos\theta = 2\sin^2\left(\frac{\theta}{2}\right), \tag{5.2007}$$

we finally obtain

$$\boxed{q = 2p\sin\left(\frac{\theta}{2}\right),} \tag{5.2008}$$

or, equivalently,

$$\boxed{q^2 = 4p^2 \sin^2\left(\frac{\theta}{2}\right).} \tag{5.2009}$$

The physical interpretation is immediate. For forward scattering, $\theta = 0$, the electron changes essentially no momentum and therefore

$$q = 0. \tag{5.2010}$$

As the scattering angle increases, the momentum transferred to the nucleus increases. Its largest possible value occurs for complete backscattering, $\theta = \pi$, for which

$$q_{\max} = 2p. \tag{5.2011}$$

This geometrical result will determine the maximum recoil energy that can be delivered to the nucleus.

Using Eq. (A.419), the energy transferred to the nucleus is

$$\boxed{T(\theta) = \frac{2p^2}{M} \sin^2\left(\frac{\theta}{2}\right).} \tag{5.2012}$$

The electron momentum must be treated relativistically at cryo-EM accelerating voltages. From the relativistic energy–momentum relation, Eq. (2.68), with

$$E = m_0 c^2 + E_{\mathrm{k}}, \tag{5.2013}$$

we obtain

$$p^2 c^2 = \left(m_0 c^2 + E_{\mathrm{k}}\right)^2 - m_0^2 c^4 \tag{5.2014}$$

$$= E_{\mathrm{k}}\left(E_{\mathrm{k}} + 2m_0 c^2\right). \tag{5.2015}$$

Equation (5.2012) can therefore be written as

$$\boxed{T(\theta) \simeq \frac{2E_{\mathrm{k}}\left(E_{\mathrm{k}} + 2m_0 c^2\right)}{Mc^2} \sin^2\left(\frac{\theta}{2}\right).} \tag{5.2016}$$

Maximum energy transferable to an atomic nucleus.

For fixed incident kinetic energy E_{k}, the transferred energy in Eq. (5.2016) is maximum when

$$\sin^2\left(\frac{\theta}{2}\right) = 1, \tag{5.2017}$$

which corresponds to electron backscattering,

$$\theta = \pi. \tag{5.2018}$$

Thus

$$\boxed{T_{\max}(E_{\mathrm{k}}, M) \simeq \frac{2E_{\mathrm{k}}\left(E_{\mathrm{k}} + 2m_0c^2\right)}{Mc^2}.} \tag{5.2019}$$

This expression assumes $M \gg m_0$ and treats the nuclear recoil non-relativistically. Both approximations are extremely accurate for atomic nuclei. The exact two-body relativistic result for a stationary target is

$$T_{\max}^{\mathrm{exact}} = \frac{2Mc^2E_{\mathrm{k}}\left(E_{\mathrm{k}} + 2m_0c^2\right)}{\left(Mc^2 + m_0c^2\right)^2 + 2Mc^2E_{\mathrm{k}}}. \tag{5.2020}$$

For atomic nuclei and conventional TEM energies this reduces to Eq. (5.2019) with negligible error.

Equation (5.2019) immediately reveals two important features. First, the maximum transferable energy increases strongly with accelerating voltage. Second,

$$T_{\max} \propto \frac{1}{M}, \tag{5.2021}$$

so light nuclei are much more susceptible to recoil than heavy ones.

For example, the kinematic upper limits obtained from Eq. (5.2019) are approximately The table illustrates why knock-on damage cannot be

Table 5.18: Maximum energy transferable in an elastic electron–nucleus collision for several elements abundant in frozen-hydrated biological specimens. The numbers are kinematic upper limits corresponding to 180° electron backscattering; they should not be interpreted as predicted damage energies.

E_{k} (keV)	H	C	N	O
	$T_{\max}$ (eV)			
100	240	20.0	17.2	15.0
200	524	43.7	37.4	32.8
300	850	70.9	60.7	53.2

dismissed merely because the electron–nucleus interaction is called “elastic”. A 300 keV electron can, in principle, transfer about 71 eV to a carbon nucleus and more than 800 eV to hydrogen. Only a small fraction of elastic scattering

events, however, occur at the very large angles required to approach these maximum values.

The displacement threshold. A nucleus is not displaced simply because it acquires non-zero recoil energy. It must receive enough energy to escape from its local bonding environment. We denote this minimum energy by the *displacement threshold energy* E_d. The elementary kinematic criterion for knock-on displacement is therefore

$$\boxed{T_{\max}(E_{\mathrm{k}}, M) > E_d \quad \Longrightarrow \quad \text{knock-on displacement is kinematically possible.}} \tag{5.2022}$$

Notice the wording "kinematically possible". Equation (5.2022) is a necessary condition, not a statement that every electron will cause displacement. The event still requires a sufficiently large-angle collision, and its probability is determined by the corresponding elastic-scattering cross section.

We can invert Eq. (5.2019) to find the incident electron energy at which displacement first becomes kinematically possible. Setting

$$T_{\max} = E_d \tag{5.2023}$$

gives

$$E_d M c^2 = 2E_{\mathrm{k}} \left(E_{\mathrm{k}} + 2m_0 c^2 \right), \tag{5.2024}$$

or

$$E_{\mathrm{k}}^2 + 2m_0 c^2 E_{\mathrm{k}} - \frac{E_d M c^2}{2} = 0. \tag{5.2025}$$

Keeping the positive solution,

$$\boxed{E_{\mathrm{k,th}} = \sqrt{m_0^2 c^4 + \frac{1}{2} E_d M c^2} - m_0 c^2.} \tag{5.2026}$$

The displacement threshold is not an atomic constant. It depends on chemical bonding, local structure, crystallographic direction, coordination, defects, and whether the atom lies in the bulk or near a surface. Representative values collected by Egerton et al. (2004) illustrate the scale. A displacement energy of approximately

$$E_d \simeq 30 \text{ eV} \tag{5.2027}$$

for graphite corresponds through Eq. (5.2026) to an incident-electron threshold of roughly

$$E_{\mathrm{k,th}} \simeq 140 \text{ keV}, \tag{5.2028}$$

whereas the much larger value

$$E_d \simeq 80 \text{ eV} \tag{5.2029}$$

historically quoted for diamond gives

$$E_{\mathrm{k,th}} \simeq 330 \text{ keV}. \tag{5.2030}$$

Thus two materials composed entirely of carbon can have very different knock-on thresholds because the relevant quantity is not the atomic number alone but the energy required to remove a particular atom from its particular bonding environment.

The sharp threshold of Eq. (5.2022) is itself an idealization. Atoms are not stationary: they possess zero-point and thermal motion. Close to threshold, the instantaneous velocity of the struck atom can increase or decrease the energy available for displacement. Measurements of single-layer graphene, in which displaced carbon atoms could be counted individually, demonstrated that a static-lattice model does not reproduce the measured displacement cross section near threshold, whereas including atomic vibrations gives substantially better agreement (Meyer et al., 2012). The concept of E_d therefore remains extremely useful, but the experimental onset of damage should not be imagined as an infinitely sharp voltage boundary.

Why knock-on damage is not the dominant mechanism in cryo-EM. The energies in Table 5.18 might suggest that knock-on displacement should be a major limitation in biological cryo-EM. In practice this is not the case. The maximum transferable energy tells us only what is possible in a rare, nearly backward elastic collision; it does not measure how frequently damaging events occur or how much subsequent damage each primary event generates.

For thin organic specimens, Egerton (2012) compared radiolysis and knock-on displacement quantitatively using both scattering cross sections and stopping power. When the secondary processes initiated by electronic excitation are included, radiolysis was found to dominate throughout the range

$$10 \lesssim E_{\mathrm{k}} \lesssim 300 \text{ keV}, \tag{5.2031}$$

even for materials containing hydrogen. This is consistent with the physical picture developed in the preceding sections and with the analysis of frozen-hydrated biological specimens by Glaeser (2016): most damaging interactions in ordinary cryo-EM originate from inelastic electronic excitation and the radiation chemistry that follows it.

This distinction is important. An inelastic ionization event can initiate several secondary processes and chemical reactions, whereas a knock-on event is essentially a direct momentum-transfer event. Consequently, even though Eq. (5.2019) shows that conventional 200–300 keV electrons are energetic enough to displace light nuclei under appropriate circumstances, the integrated damage to a frozen-hydrated organic specimen is normally governed by radiolysis rather than ballistic displacement (Egerton, 2012, 2013).

Cooling illustrates the difference between the two mechanisms. Lowering the specimen temperature can strongly modify diffusion, radical mobility, recombination, and therefore the secondary chemistry following ionization. It does not, however, remove the elementary momentum transfer described by Eq. (5.2016). Knock-on displacement is therefore much less amenable to suppression by cryogenic cooling, although atomic vibrations influence the detailed displacement cross section close to threshold (Egerton et al., 2004; Meyer et al., 2012).

Sputtering: displacement at a surface. The same nuclear-recoil kinematics applies to an atom at the surface of the specimen. The crucial difference is that a surface atom does not have to be pushed into an interstitial position inside the material. It only has to acquire enough energy in an appropriate direction to escape into vacuum. We denote this surface escape threshold by E_s. The corresponding criterion is

$$\boxed{T_{\max}(E_{\mathrm{k}}, M) > E_s \quad \Longrightarrow \quad \text{sputtering is kinematically possible.}} \tag{5.2032}$$

Typically,

$$E_s < E_d, \tag{5.2033}$$

so an atom that would remain bound if it were in the bulk may be removed if it is located at a surface. This is why electron-beam sputtering can occur at accelerating voltages below the threshold for bulk displacement (Egerton et al., 2004).

The probability of removal can be expressed through a sputtering cross section σ_s. For an atom exposed to an electron fluence D, the probability that it has undergone at least one sputtering event is approximately

$$P_{\mathrm{s}} = 1 - \exp(-\sigma_s D), \tag{5.2034}$$

and, in the low-dose limit,

$$\boxed{P_{\mathrm{s}} \simeq \sigma_s D, \qquad \sigma_s D \ll 1.} \tag{5.2035}$$

Thus the threshold determines whether sputtering is possible, whereas the cross section determines how rapidly material is actually removed.

Electron-induced sputtering is particularly relevant for exposed surfaces, ultrathin films, carbonaceous supports, graphene-like materials, and specimen edges. Egerton et al. (2004) notes that the lower surface-binding requirement makes light and medium-Z atoms susceptible to sputtering at typical TEM energies. The strong directionality of the momentum transfer also makes sputtering asymmetric between entrance and exit surfaces in sufficiently thin specimens.

For the frozen-hydrated biological specimens of primary interest here, however, sputtering and bulk knock-on displacement should be regarded as secondary contributors to the ordinary dose limitation. They are physically distinct and can certainly occur, especially at surfaces and for light atoms, but the experimentally relevant loss of high-resolution biological structure at conventional cryo-EM doses is normally reached through the inelastic-scattering and radiolysis pathway discussed in the preceding sections, long before extensive ballistic erosion of the specimen occurs (Egerton, 2012; Glaeser, 2016).

The useful conceptual separation is therefore

$$\boxed{\begin{array}{rcl} \text{inelastic electronic scattering} & \longrightarrow & \text{excitation/ionization} \longrightarrow \text{radiolysis}, \\ \text{large-angle elastic nuclear scattering} & \longrightarrow & \text{nuclear recoil} \longrightarrow \begin{cases} \text{bulk displacement}, \\ \text{sputtering at a surface}. \end{cases} \end{array}} \tag{5.2036}$$

Both mechanisms originate directly from interactions of the imaging electron with the specimen, but their physics and their relative importance in biological cryo-EM are fundamentally different.

5.13.10 Beam heating

Inelastic scattering removes energy from the imaging electrons and deposits it in the specimen. Some of this energy initially appears as electronic excitation, ionization, secondary electrons, and chemical potential energy, but a substantial fraction is eventually converted into heat. It is therefore natural to ask whether the radiation damage observed in cryo-EM might simply be caused by heating of the specimen.

For a thin specimen this question can be addressed quantitatively. We showed in Eq. (5.1534) that an electron crossing a homogeneous specimen

of thickness t undergoes, on average,

$$\langle n_{\mathrm{in}} \rangle = \mu_{\mathrm{in}} = \frac{t}{\Lambda_{\mathrm{in}}} \tag{5.2037}$$

inelastic interactions. Let

$$\bar{\epsilon}_{\mathrm{h}} \tag{5.2038}$$

denote the mean amount of energy that eventually appears as heat per inelastic interaction. For comparison with the low-loss spectrum discussed previously, a characteristic energy scale is a few tens of electronvolts; the broad plasmon feature in vitreous ice lies near 20 eV.

The important point is that energy deposition alone does not determine the specimen temperature. At steady state, the deposited power must be balanced by the rate at which heat is conducted away toward the surrounding support and ultimately the microscope grid (Egerton et al., 2004; Egerton, 2025).

Order-of-magnitude estimate of beam heating.

Consider a uniformly thick specimen of thickness t illuminated by a circular electron beam of radius a_{b}. Let the beam current be I. Since one electron carries charge e, the number of incident electrons per unit time is

$$\dot{N}_e = \frac{I}{e}. \tag{5.2039}$$

From Eq. (5.1534), each incident electron undergoes on average

$$\frac{t}{\Lambda_{\mathrm{in}}} \tag{5.2040}$$

inelastic interactions. If each of these deposits on average an energy $\bar{\epsilon}_{\mathrm{h}}$, expressed here in joules, the power deposited in the specimen is approximately

$$\boxed{P_{\mathrm{dep}} \simeq \frac{I}{e} \frac{t}{\Lambda_{\mathrm{in}}} \bar{\epsilon}_{\mathrm{h}}.} \tag{5.2041}$$

We now estimate how efficiently this energy is removed. Let κ be the thermal conductivity of the specimen and let r_{g} denote a characteristic distance from the illuminated region to a thermal reservoir, such as a grid bar, with

$$r_{\mathrm{g}} > a_{\mathrm{b}}. \tag{5.2042}$$

Outside the illuminated region, heat flows approximately radially through the thin specimen. At radius r, the cross-sectional area available for radial conduction is

$$A(r) = 2\pi r t. \tag{5.2043}$$

Fourier's law of heat conduction therefore gives

$$P_{\mathrm{dep}} = -\kappa\,(2\pi r t)\,\frac{dT}{dr}. \tag{5.2044}$$

Rearranging,

$$dT = -\frac{P_{\mathrm{dep}}}{2\pi\kappa t}\frac{dr}{r}. \tag{5.2045}$$

Integrating from the edge of the illuminated region, $r = a_{\mathrm{b}}$, to the thermal reservoir, $r = r_{\mathrm{g}}$,

$$T(a_{\mathrm{b}}) - T(r_{\mathrm{g}}) = \frac{P_{\mathrm{dep}}}{2\pi\kappa t}\int_{a_{\mathrm{b}}}^{r_{\mathrm{g}}}\frac{dr}{r} \qquad = \frac{P_{\mathrm{dep}}}{2\pi\kappa t}\ln\left(\frac{r_{\mathrm{g}}}{a_{\mathrm{b}}}\right). \tag{5.2046}$$

A more complete treatment also includes the temperature gradient inside the illuminated disk itself. For approximately uniform illumination this introduces only an order-unity additive correction, which Egerton writes as

$$\xi \simeq 0.3. \tag{5.2047}$$

The resulting temperature rise within the illuminated region is therefore approximately

$$\boxed{\Delta T \simeq \frac{P_{\mathrm{dep}}}{2\pi\kappa t}\left[\ln\left(\frac{r_{\mathrm{g}}}{a_{\mathrm{b}}}\right) + \xi\right].} \tag{5.2048}$$

Substituting Eq. (5.2041),

$$\boxed{\Delta T \simeq \frac{I\,\bar{\epsilon}_{\mathrm{h}}}{2\pi e\,\kappa\Lambda_{\mathrm{in}}}\left[\ln\left(\frac{r_{\mathrm{g}}}{a_{\mathrm{b}}}\right) + \xi\right].} \tag{5.2049}$$

An interesting consequence is that the specimen thickness t has cancelled. A thicker specimen absorbs proportionally more energy, according to Eq. (5.2041), but at the same time provides a proportionally larger cross section for lateral heat conduction. Within this simple thin-film model, these two effects compensate.

Equation (5.2049) makes several features of beam heating transparent.

The temperature rise is approximately linear in beam current,

$$\Delta T \propto I, \tag{5.2050}$$

and inversely proportional to thermal conductivity,

$$\Delta T \propto \frac{1}{\kappa}. \tag{5.2051}$$

By contrast, the dependence on beam radius is only logarithmic,

$$\Delta T \propto \ln\left(\frac{r_{\mathrm{g}}}{a_{\mathrm{b}}}\right). \tag{5.2052}$$

Thus a very small probe can have an enormous *current density* without necessarily producing a correspondingly enormous temperature rise. It is primarily the total beam current and the available thermal path to the grid that matter.

This somewhat counterintuitive result can be seen quantitatively in the calculations of Egerton et al. (2004). For a stationary 5 nA electron probe incident on a carbon film with thermal conductivity

$$\kappa \simeq 1.6\ \mathrm{W\,m^{-1}\,{}^{\circ}K^{-1}}, \tag{5.2053}$$

reducing the beam diameter from

$$1\ \mu\mathrm{m} \quad \text{to} \quad 1\ \mathrm{nm} \tag{5.2054}$$

increases the current density by a factor of 10^6, but the calculated temperature rise increases only from approximately

$$0.5\ {}^{\circ}\mathrm{K} \quad \text{to} \quad 1.4\ {}^{\circ}\mathrm{K}. \tag{5.2055}$$

More recent estimates give the same qualitative conclusion: for uniformly supported thin specimens and ordinary TEM beam currents, radial conduction is usually sufficiently efficient that the steady-state temperature rise remains modest (Egerton, 2025).

The geometry of a real cryo-EM grid is, of course, more complicated than the continuous thin sheet assumed in Eq. (5.2049). The vitreous ice is suspended across holes, the support foil has its own thermal conductivity, and thermal contact between foil, grid, cartridge, and cryo-holder may be imperfect. These additional thermal resistances can increase the actual temperature rise.

A useful cryo-EM-specific calculation was performed by Karuppasamy et al. (2011). They modelled a 150 nm-thick vitreous specimen irradiated at 120 kV with a 10 μm diameter beam. For an incident flux density of

$$5\ e^{-}\ \text{Å}^{-2}\, s^{-1} \tag{5.2056}$$

they estimated that approximately 48% of the incident electrons underwent inelastic scattering, depositing about 20 eV per event. Increasing the flux density by a factor of ten to

$$50\ e^{-}\ \text{Å}^{-2}\, s^{-1} \tag{5.2057}$$

could, in their deliberately poor-thermal-contact model, increase the specimen temperature from approximately

$$103\ ^{\circ}\text{K} \quad \text{to} \quad 140\ ^{\circ}\text{K} \tag{5.2058}$$

within milliseconds. Importantly, however, their conclusion was that such heating becomes significant only for very high flux densities combined with poor thermal contact between the grid and cryo-holder. At low and moderate flux densities with good thermal contact, beam heating was not expected to be a major limitation (Karuppasamy et al., 2011).

This distinction is important when interpreting the benefit of cooling in biological cryo-EM. Cryogenic temperatures do not primarily protect the specimen by preventing an otherwise enormous macroscopic temperature rise caused by the electron beam. Rather, as discussed in the preceding treatment of radiolysis, cooling strongly restricts the mobility of radicals and molecular fragments, cages reaction products, and suppresses many of the secondary chemical reactions that would proceed rapidly in a liquid or warmer specimen (Glaeser, 2016). Indeed, the cryo-protection described there is observed even though the local electronic excitations produced by individual inelastic events still involve energies of tens of electronvolts.

Beam heating is therefore a real physical effect, and it can become important when the beam current is high or the thermal path to the grid is poor. Under the low-dose conditions normally used for high-resolution cryo-EM, however, it is generally not the dominant explanation for radiation damage. The principal role of cryogenic cooling is to slow the *consequences* of energy deposition—the diffusion, reaction, and structural rearrangement of damaged molecules—rather than simply to remove a large amount of beam-generated heat.

5.13.11 Beam-induced mechanical motion

Radiation damage does not only change the chemical composition of the specimen. The specimen can also move while the image is being recorded. This distinction is important because mechanical motion enters the image-formation model directly: a translation changes Fourier phases, a rotation changes the particle orientation, and motion along the optical axis changes the local defocus.

It is useful to describe the mechanical state of the specimen by a displacement field

$$\boxed{\mathbf{u}(x, y, \Phi) = \begin{pmatrix} u_x(x, y, \Phi) \\ u_y(x, y, \Phi) \\ \Delta z(x, y, \Phi) \end{pmatrix},} \tag{5.2059}$$

where Φ is the accumulated electron exposure defined in Eq. (5.1678). The first two components describe motion parallel to the specimen plane and Δz describes motion along the optical axis.

Writing the motion as a field rather than as the trajectory of an individual particle is deliberate. One of the earliest direct measurements of beam-induced motion showed that neighboring particles move coherently over patches approximately

$$0.3\text{–}0.5\ \mu\text{m} \tag{5.2060}$$

across, with accompanying particle rotations of up to a few degrees (Brilot et al., 2012). The motion is therefore not adequately described as independent Brownian motion of individual macromolecules. A substantial part of it is the deformation of the ice film and support on which many particles are carried together.

Mechanical stress is already present before irradiation. The frozen specimen is not prepared in a mechanically relaxed state. During plunge freezing, the aqueous film, the thin support foil, and the much thicker metal grid cool on different time scales and undergo different thermal contractions. Furthermore, vitreous ice is a nonequilibrium amorphous solid whose density and specific volume change substantially during cooling.

There are therefore at least two related mechanisms by which stress can be frozen into the specimen. First, different materials have different thermal-expansion coefficients. This is the origin of the "cryocrinkling" historically observed, for example, when thin carbon foils are supported on metal grids. Replacing a carbon foil and metal grid by an all-gold system greatly reduces this source of mechanical instability (Russo and Passmore, 2014).

Second, stress can arise even when the foil and grid are made of the same material because they do not cool simultaneously. The thin support foil and the aqueous film near the centre of a grid square cool much more rapidly than the comparatively massive grid bars. A mechanical model proposed by Thorne (2020) describes the following sequence:

$$\boxed{\begin{array}{c}\text{foil and water cool rapidly}\\ \Downarrow\\ \text{foil develops transient tensile stress}\\ \Downarrow\\ \text{water vitrifies while the grid bars remain warmer}\\ \Downarrow\\ \text{grid bars subsequently cool and contract}\\ \Downarrow\\ \text{foil tension relaxes}\\ \Downarrow\\ \text{vitrified specimen is placed under compression.}\end{array}} \tag{5.2061}$$

The important point is that the mechanical state observed in the microscope need not be the equilibrium state corresponding simply to the final cryogenic temperature.

The magnitude of the volume change involved in vitrification is not small. The analysis of Naydenova et al. (2020) indicates that the largest specific-volume change experienced by water during rapid cooling below the homogeneous nucleation regime is of order

$$\left|\frac{\Delta V}{V}\right| \sim 6\%. \tag{5.2062}$$

The suspended film cannot accommodate an arbitrary fraction of this change as an in-plane compressive strain. Above a geometry-dependent critical stress it buckles.

There is also direct evidence that freezing can mechanically distort the embedded macromolecule itself. This should not be confused with ordinary beam-induced doming of vitreous ice, but it demonstrates that mechanical strain can be introduced before the first imaging electron arrives. Shi et al. (2023) quantified the compression produced when proteins were embedded in crystalline rather than vitreous ice. Under their lowest-cooling-rate condition, obtained using cryogen at approximately $-60\,^{\circ}\mathrm{C}$, reconstructed dimensions were reduced by 1.10% for apo-ferritin, 0.61% for glutamate dehydrogenase, 2.68% for empty Coxsackievirus A10, and 3.05% for EV71 virus-like particles. Under their fastest-freezing condition, using liquid ethane

at $-183\,^{\circ}\mathrm{C}$, the corresponding shrinkages were reduced to 0.00%, 0.24%, 0.85%, and 0.61%, respectively.

Interestingly, crystalline ice was not intrinsically detrimental once this deformation was sufficiently suppressed. At $-183\,^{\circ}\mathrm{C}$, Shi et al. (2023) reported B-factors changing from 83.0 to 62.7 $\mathrm{Å}^2$ for apo-ferritin, from 172.4 to 136.1 $\mathrm{Å}^2$ for glutamate dehydrogenase, and from 217.4 to 170.9 $\mathrm{Å}^2$ for aldolase when comparing vitreous with crystalline ice. The embedding ice must therefore be regarded as a mechanically active part of the specimen rather than as a passive background.

Irradiation allows pre-existing stress to relax. At cryogenic temperature, the stressed amorphous specimen can remain mechanically trapped for very long times in the absence of irradiation. Electron irradiation changes this situation. Atomic and molecular rearrangements induced by the beam give the amorphous material an effective mechanism for creep, allowing pre-existing stress to relax.

A phenomenological expression used in radiation-creep models is

$$\boxed{\epsilon_{\mathrm{c}}(\Phi) = a_{\mathrm{c}}\sigma\left[1 - \exp(-b_{\mathrm{c}}\Phi)\right] + c_{\mathrm{c}}\sigma\Phi,} \tag{5.2063}$$

where ϵ_{c} is the accumulated creep strain, σ is the pre-existing mechanical stress, and a_{c}, b_{c}, and c_{c} are material- and temperature-dependent coefficients (Thorne, 2020). Equation (5.2063) should not be regarded as a microscopic law for vitreous water. Its usefulness here is conceptual: irradiation can make an initially immobile stressed solid mechanically relax.

The first term in Eq. (5.2063) produces rapid motion at low exposure and subsequently saturates, whereas the second allows continued creep at larger exposure. This qualitative behavior is remarkably similar to a recurring experimental observation: beam-induced motion is usually largest during the first few movie frames and slows substantially as irradiation continues (Brilot et al., 2012; Campbell et al., 2012).

For a suspended film under in-plane compression, one way to release stress is to increase its surface area by moving perpendicular to the support plane. The result is the characteristic *doming* or "drum-head" deformation seen experimentally.

> **A very small in-plane strain can produce a large out-of-plane displacement.**

Consider a circular hole of radius a. Before deformation, the suspended film has approximately the area

$$A_0 = \pi a^2. \tag{5.2064}$$

Suppose that irradiation makes the film form approximately a spherical cap of height h. If R_s is the radius of the corresponding sphere, the geometry of a spherical cap gives

$$a^2 = 2R_s h - h^2. \tag{5.2065}$$

Its surface area is

$$A_{\text{cap}} = 2\pi R_s h. \tag{5.2066}$$

Eliminating R_s,

$$A_{\text{cap}} = \pi \left(a^2 + h^2\right). \tag{5.2067}$$

The fractional increase in area is consequently

$$\boxed{\frac{A_{\text{cap}} - A_0}{A_0} = \frac{h^2}{a^2}.} \tag{5.2068}$$

For a 1.2 μm diameter hole,

$$a = 0.6\ \mu\text{m} = 6000\ \text{Å}, \tag{5.2069}$$

and the approximately

$$h \simeq 150\ \text{Å} \tag{5.2070}$$

central displacement inferred from the measurements of Brilot et al. (2012) after an exposure of approximately 32 $e^-\ \text{Å}^{-2}$ gives

$$\frac{\Delta A}{A_0} = \left(\frac{150}{6000}\right)^2 \tag{5.2071}$$

$$\simeq 6.25 \times 10^{-4} \tag{5.2072}$$

$$\simeq 0.063\%. \tag{5.2073}$$

Thus an in-plane area relaxation of only about six parts in ten thousand can produce an axial displacement of 15 nm at the centre of the hole.

This calculation explains why apparently modest frozen-in stresses can produce displacements that are very large on the scale relevant to high-

resolution electron microscopy (Thorne, 2020).

Irradiation can also generate new mechanical stress. Relaxation of pre-existing stress is not the only possible origin of motion. The beam simultaneously changes the material itself. Radiolysis cleaves chemical bonds, produces radicals and stable molecular products, changes local density, and can remove material from the ice. The exposure-dependent change of ice thickness was discussed in Eq. (5.1989), while at much higher exposure molecular hydrogen can accumulate and eventually produce the high-pressure bubbles discussed in Eq. (5.1992).

Long before macroscopic bubbling becomes visible, radiation-induced changes of density, local specific volume, elastic modulus, and chemical composition can modify the internal stress field. Thus it is useful to distinguish conceptually

$$\boxed{\begin{array}{l}\text{pre-existing stress} \xrightarrow{\text{irradiation}} \text{stress relaxation and creep,}\\ \text{radiation chemistry} \rightarrow \text{new density and volume changes} \rightarrow \text{new stress.}\end{array}} \tag{5.2074}$$

The two processes occur simultaneously and need not be experimentally separable.

This coupling between radiation chemistry and mechanics becomes particularly clear at very low temperature. As discussed previously in Eq. (5.1791), trapping of radiolytic products at liquid-helium temperature can cause the frozen water to expand, generating additional stress and eventually buckling. Thus chemical cryoprotection and mechanical stability need not improve together.

The experimental signature is collective doming, not independent particle diffusion. The observations of Brilot et al. (2012) provide an important clue about the nature of the motion. Particle trajectories were correlated over patches approximately 0.3–0.5 μm in size, and particles underwent rotations of up to several degrees. Both observations are consistent with deformation of the entire suspended ice layer rather than independent translational diffusion of individual particles. Smaller foil holes and lower exposure rates reduced the observed motion, providing further evidence for a mechanical origin.

Dose-fractionated movies made the consequences directly measurable. In the experiments of Campbell et al. (2012), rotavirus particles were recorded

in 16 usable movie frames, each corresponding to approximately

$$2\ e^{-}\ \text{Å}^{-2}, \tag{5.2075}$$

for a total exposure of approximately

$$32\ e^{-}\ \text{Å}^{-2}. \tag{5.2076}$$

Between averages of the first and final groups of frames, the distribution of particle shifts had a maximum near

$$5\ \text{Å}, \tag{5.2077}$$

while the corresponding rotation distribution had a maximum near

$$0.3^{\circ}. \tag{5.2078}$$

Some particles moved much farther.

These numbers are already large enough to destroy substantial high-resolution information. Campbell et al. (2012) estimated that a particle trajectory of about 5 Å corresponds to a positional standard deviation of roughly

$$\sigma_{\rho} \simeq 1.5\ \text{Å}. \tag{5.2079}$$

Using the motion B-factor relation already derived in Eq. (5.1207),

$$B_{\text{motion}} = 8\pi^{2}\sigma_{\rho}^{2}, \tag{5.2080}$$

gives

$$B_{\text{motion}} \simeq 180\ \text{Å}^{2}, \tag{5.2081}$$

consistent with their estimate of approximately 200 Å^{2}. Particle rotations around 0.3° were estimated to contribute an additional effective B-factor of approximately 160 Å^{2} for the large rotavirus particles.

Correcting the movie frames reduced the reconstruction B-factor by approximately

$$160\ \text{Å}^{2}. \tag{5.2082}$$

For the rotavirus double-layered particle reconstruction, the reported resolution improved from approximately 7.5 Å without motion correction to 6.4 Å when translations were corrected and to 6.2 Å when rotations were also included. For the locally averaged VP6 reconstruction, the improvement was from 4.9 Å to 4.4 Å (Campbell et al., 2012).

The support geometry determines whether the film can buckle. The importance of mechanical stress is reinforced by experiments in which the support itself was changed. Russo and Passmore (2014) replaced conventional amorphous-carbon supports by a system in which both the foil and grid were made of gold. The resulting substrate motion was greatly reduced. For tracked ribosomal particles, the speed of the initial phase of motion decreased from approximately 8.8 to 5.0 $\text{Å}\,\text{s}^{-1}$, while the slower phase decreased from approximately 2.3 to 0.53 $\text{Å}\,\text{s}^{-1}$.

Removing support motion, however, does not by itself eliminate deformation of the ice suspended inside the holes. A later analysis by Naydenova et al. (2020) showed that buckling depends strongly on the aspect ratio of the suspended film. If $2a$ is the hole diameter and h_{ice} is the ice thickness, their mechanical treatment gives, as an approximate design criterion,

$$\boxed{\frac{2a}{h_{\text{ice}}} \lesssim 11 \qquad \Longrightarrow \qquad \text{buckling strongly suppressed.}} \tag{5.2083}$$

This number should be understood as a practical order-of-magnitude criterion rather than a universal material constant.

For example, if

$$h_{\text{ice}} = 300\ \text{Å}, \tag{5.2084}$$

Eq. (5.2083) suggests a maximum hole diameter of order

$$2a \lesssim 3300\ \text{Å} = 330\ \text{nm}. \tag{5.2085}$$

This explains why conventional micron-sized holes permit substantial doming whereas holes a few hundred nanometres across can remain mechanically stable.

Using this principle, Naydenova et al. (2020) developed an all-gold support with holes approximately 200–260 nm in diameter. Beam-induced particle movement was reduced to below approximately 1 Å. The experiment is especially important because it shows that beam-induced movement is not an unavoidable microscopic property of the protein: changing the mechanics of the macroscopic support can almost eliminate it.

How mechanical motion enters the image-formation model. The consequences of the displacement field in Eq. (5.2059) can now be stated quantitatively.

Consider first a sufficiently small region of the specimen that its motion can be approximated as rigid during one movie frame. Let

$$\mathbf{u}_{\parallel}(\Phi) = \begin{pmatrix} u_x(\Phi) \\ u_y(\Phi) \end{pmatrix} \tag{5.2086}$$

be its in-plane displacement.

Mechanical displacement as an image-formation phase.

Let $f(\mathbf{r}_\perp)$ denote any projected specimen quantity whose Fourier transform is defined according to Eq. (3.2). After an in-plane translation $\mathbf{u}_\parallel$,

$$f_{\mathbf{u}}(\mathbf{r}_\perp) = f\left(\mathbf{r}_\perp - \mathbf{u}_\parallel\right). \tag{5.2087}$$

Its Fourier transform is

$$\widehat{f}_{\mathbf{u}}(\mathbf{R}) = \int f\left(\mathbf{r}_\perp - \mathbf{u}_\parallel\right) e^{-j2\pi\mathbf{R}\cdot\mathbf{r}_\perp}\, d\mathbf{r}_\perp \tag{5.2088}$$

$$= e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_\parallel}\, \widehat{f}(\mathbf{R}). \tag{5.2089}$$

Thus the in-plane motion contributes the phase

$$\boxed{\delta\phi_\parallel = 2\pi\mathbf{R}\cdot\mathbf{u}_\parallel.} \tag{5.2090}$$

The axial component has a different effect. A local displacement along the optical axis changes the effective defocus. According to Eq. (4.1747), a defocus variation δf produces

$$\delta\chi(\mathbf{R}) = -\pi\lambda\delta f\, R^2. \tag{5.2091}$$

We define Δz with the geometrical sign convention such that an axial displacement Δz produces

$$\delta f = -\Delta z. \tag{5.2092}$$

Consequently,

$$\boxed{\delta\chi(\mathbf{R}) = \pi\lambda R^2\Delta z.} \tag{5.2093}$$

This is the same quadratic phase dependence derived for defocus in Eq. (4.1184), but here the defocus itself is changing during the exposure.

Combining the two contributions, the mechanical transfer factor for a locally rigid patch is

$$\boxed{M_{\text{motion}}(\mathbf{R},\Phi) = \exp\left[-j2\pi\mathbf{R}\cdot\mathbf{u}_{\parallel}(\Phi)\right]\exp\left[-j\pi\lambda R^2\Delta z(\Phi)\right].} \tag{5.2094}$$

If structural radiation damage is neglected temporarily so that we can isolate the mechanical effect, a Fourier component recorded over an exposure interval $[\Phi_1, \Phi_2]$ is proportional to

$$\boxed{\overline{M}_{\text{motion}}(\mathbf{R}) = \frac{1}{\Phi_2-\Phi_1}\int_{\Phi_1}^{\Phi_2} M_{\text{motion}}(\mathbf{R},\Phi)\,d\Phi.} \tag{5.2095}$$

Motion therefore attenuates a Fourier component whenever its phase changes substantially during the integration.

Equation (5.2094) makes an important distinction explicit:

$$\boxed{\begin{array}{rcl} \text{in-plane motion} & \longrightarrow & 2\pi\,\mathbf{R}\cdot\mathbf{u}_{\parallel}, \\ \text{out-of-plane motion} & \longrightarrow & \pi\lambda R^2\Delta z. \end{array}} \tag{5.2096}$$

A two-dimensional shift can therefore correct the first effect, but it does not in general correct the second.

The magnitude of the axial effect can be surprisingly large at high resolution. At 300 kV, Eq. (2.103) gives approximately

$$\lambda = 1.97 \text{ pm} = 0.0197 \text{ Å}. \tag{5.2097}$$

For the 150 Å central displacement discussed in Section 5.13.11, the additional phase at 3 Å resolution is

$$\delta\chi = \pi(0.0197)\left(\frac{1}{3}\right)^2(150) \tag{5.2098}$$

$$\simeq 1.0 \text{ rad}. \tag{5.2099}$$

At 2 Å resolution,

$$\delta\chi \simeq 2.3 \text{ rad}. \tag{5.2100}$$

Thus a doming displacement of only 15 nm changes the local defocus phase by order unity at near-atomic spatial frequencies.

This also reveals a difference in the frequency dependence of the two motion components. As derived previously in Sec. 5.11.5, a Gaussian distribution of in-plane translations produces the envelope

$$E_{\parallel}(R) = \exp\left(-2\pi^2\sigma_\rho^2 R^2\right), \tag{5.2101}$$

or equivalently the B-factor form of Eq. (5.1208).

If the axial displacement instead has an approximately Gaussian distribution with standard deviation σ_z, averaging Eq. (5.2093) gives

$$\boxed{E_z(R) = \exp\left[-\frac{1}{2}\left(\pi\lambda\sigma_z\right)^2 R^4\right].} \tag{5.2102}$$

This is mathematically identical to the focus-spread envelope derived for temporal coherence in Eq. (4.1749), although its physical origin is completely different. In-plane translational blur therefore produces an approximately e^{-cR^2} attenuation, whereas fluctuating out-of-plane doming produces an e^{-cR^4} attenuation.

The Gaussian models above should not be interpreted as implying Brownian particle motion. They merely describe the attenuation obtained when an exposure averages over a distribution of mechanical displacements. The actual field $\mathbf{u}(x, y, \Phi)$ can be strongly correlated in both space and exposure.

Particle rotations constitute a third effect. A rotation does not multiply the same projection by a simple phase factor: it changes the projection direction itself. This is why the rotations measured by Brilot et al. (2012) and Campbell et al. (2012) can cause an additional loss of high-resolution signal even after translational alignment.

Why motion correction must be local. If

$$\mathbf{u}(x, y, \Phi) \tag{5.2103}$$

were independent of (x, y), each movie frame could be corrected by one global translation. The correlated but non-rigid motion observed experimentally shows that this approximation is often insufficient. Different regions of the same micrograph can have different trajectories because they occupy different positions on the domed surface.

This is the physical reason why modern motion-correction algorithms estimate spatially varying displacement fields. MotionCor2, for example,

divides the image into local regions and estimates an anisotropic motion field rather than assuming one rigid translation for the complete micrograph (Zheng et al., 2017). Such methods correct much of the in-plane component $\mathbf{u}_{\parallel}(x, y, \Phi)$.

They cannot, however, turn the specimen back into the original three-dimensional object. Axial motion changes local defocus according to Eq. (5.2093), rotations change the particle projection direction, and radiation chemistry changes the specimen potential itself. Frame alignment is therefore a correction for one important consequence of specimen evolution, not an undoing of radiation damage.

The mechanical picture that emerges is consequently

$$\boxed{\begin{array}{c}\text{vitrification and cooling}\\ \Downarrow\\ \text{frozen-in stress and possible buckling}\\ \Downarrow\\ \text{electron irradiation}\\ \Downarrow\\ \text{radiation-induced creep and stress relaxation}\\ \text{+ new radiation-generated stress}\\ \Downarrow\\ \text{collective translation, rotation, and doming}\\ \Downarrow\\ \text{Fourier phase shifts}\\ \text{+ time-dependent local defocus}\\ \text{+ projection-direction changes}\\ \Downarrow\\ \text{loss of high-resolution image information.}\end{array}} \tag{5.2104}$$

Beam-induced mechanical motion is therefore not merely an experimental nuisance superimposed on the image-formation problem. It is itself a time-dependent part of that problem.

5.13.12 Charging: radiation-induced specimen modification or electron-optical artifact?

The inelastic interactions discussed in the preceding sections can remove electrons from the specimen. In a good electrical conductor the missing charge is rapidly replenished from ground. Vitreous ice, however, is an extremely poor conductor at cryogenic temperature, and the compensating

current can be much too small to maintain electrical neutrality. The irradiated specimen can therefore acquire a net electrical charge.

Charging occupies an unusual position among beam-induced effects. It is a genuine radiation-induced modification of the specimen: electrons have physically left one region and may have been recaptured elsewhere. At the same time, the resulting electrostatic field acts on subsequent imaging electrons. It can therefore change an image even if the macromolecule itself has neither moved nor undergone further chemical damage.

It is useful to distinguish four related effects:

$$\boxed{\begin{array}{rcl} \text{net charge accumulation} & \longrightarrow & \text{macroscopic electrostatic field} \\ \text{microscopic charge fluctuations} & \longrightarrow & \text{local beam deflections} \\ \text{electrostatic forces} & \longrightarrow & \text{mechanical deformation} \\ \text{smooth charge curvature} & \longrightarrow & \text{electrostatic lensing} \end{array}} \tag{5.2105}$$

These phenomena have a common origin but need not have the same magnitude or the same consequences for image formation.

How charge accumulates. High-energy primary electrons traversing the ice undergo ionization and other inelastic interactions. Some of the resulting secondary electrons have sufficient energy and a favorable direction to escape through one of the nearby ice–vacuum interfaces. Their escape leaves behind a positive charge. Other secondary electrons are recaptured by the irradiated region, by the surrounding support, or by nearby previously unirradiated regions. Consequently, the charge distribution need not consist of a uniformly positive disk (Glaeser, 2016; Russo and Henderson, 2018a).

Indeed, Russo and Henderson (2018a) observed both signs of charge. A region directly exposed to a focused electron beam accumulated positive charge, whereas surrounding initially unirradiated regions could accumulate negative charge produced by recaptured secondary electrons. The amount of negative charge decreased approximately with distance from the illuminated region; its measured magnitude fell by about one half over a distance of

$$3\ \mu\text{m}$$

, although a smaller fraction of secondary electrons traveled distances larger than

$$10\ \mu\text{m}$$

. Charging is therefore intrinsically non-local on the scale of a cryo-EM grid hole.

A minimal phenomenological description is obtained by writing the net charge Q in an illuminated region as

$$\frac{dQ}{dt} = \dot{Q}_{\text{loss}} - \dot{Q}_{\text{rec}} - \dot{Q}_{\text{ground}}. \tag{5.2106}$$

where $\dot{Q}_{\text{loss}}$ represents electrons escaping the region, $\dot{Q}_{\text{rec}}$ represents secondary electrons that return to it, and $\dot{Q}_{\text{ground}}$ represents charge neutralization through the support.

If the net production rate is approximately constant and the neutralization current is linear in Q, the simplest lumped model is

$$\frac{dQ}{dt} = S_q - \frac{Q}{\tau_q}, \tag{5.2107}$$

where S_q is the effective charging rate and τ_q is a charge relaxation time. Starting from an uncharged state, the solution of the differential equation above is

$$\boxed{Q(t) = Q_\infty \left[1 - \exp\left(-\frac{t}{\tau_q}\right)\right], \qquad Q_\infty = S_q \tau_q.} \tag{5.2108}$$

Equation (5.2108) is only a schematic model: real cryo-EM specimens have spatially distributed charge, secondary-electron recapture, nonlinear leakage currents, and dose-rate-dependent conductivity. Its main purpose is to show why a charged specimen can approach a dynamic equilibrium rather than charge indefinitely.

The experimentally observed approach to this equilibrium is extremely rapid. In the experiments of Russo and Henderson (2018a), most of the positive charging occurred within approximately

$$10^{-3}\ e^-\ \text{Å}^{-2}, \tag{5.2109}$$

with saturation occurring in the approximate range

$$\boxed{10^{-3} \lesssim \Phi_{\text{q}} \lesssim 10^{-1}\ e^-\ \text{Å}^{-2}.} \tag{5.2110}$$

For their irradiation conditions this corresponds to charging during approximately the first millisecond of a conventional exposure. They estimated a final electrostatic potential in the directly irradiated region of order

$$V_q \sim 1\ \text{V}. \tag{5.2111}$$

Thus charging can already be largely established before a conventional direct detector has accumulated enough electrons for a useful movie frame.

More recent time-resolved measurements by Schreiber et al. (2023b) show that the detailed dynamics depend strongly on illumination geometry and dose rate. They illuminated the ice within 1.2 μm Quantifoil holes using a 913 nm diameter beam, without illuminating the surrounding carbon, at fluence rates between

$$0.36 \quad \text{and} \quad 1.51\ e^{-}\ \text{Å}^{-2}\,\text{s}^{-1}. \tag{5.2112}$$

For fluence rates above a transition between approximately 0.36 and 0.48 $e^{-}\ \text{Å}^{-2}\,\text{s}^{-1}$, their Fresnel-diffraction signal approached a quasi-steady state after an accumulated exposure of approximately

$$\boxed{\Phi_{\text{q}} \simeq 0.4\ e^{-}\ \text{Å}^{-2}.} \tag{5.2113}$$

The most rapid change occurred within approximately the first

$$0.5\ e^{-}\ \text{Å}^{-2}. \tag{5.2114}$$

Fits of their electrostatic model gave potentials of approximately

$$7\text{–}30\ \text{V} \tag{5.2115}$$

near the center of the illuminated region and

$$4\text{–}17\ \text{V} \tag{5.2116}$$

near its edge, depending on dose rate (Schreiber et al., 2023b). These numerical values should not be compared directly with the approximately 1 V obtained by Russo and Henderson (2018a): the beam size, charge geometry, surrounding illumination, measurement procedure, and electrostatic model were substantially different. The robust conclusion from both experiments is instead that charging develops at extremely small exposures and rapidly evolves toward a dynamic state determined by secondary-electron emission, recapture, and charge transport.

A charged specimen produces an additional electron-optical phase. The electrostatic potential generated by charging is physically no different, from the point of view of the imaging electron, from any other electrostatic potential. We therefore denote it by

$$V_q(x, y, z, \Phi), \tag{5.2117}$$

where Φ is the accumulated electron exposure defined in Eq. (5.1678).

Unlike the atomic potential introduced previously, V_q need not be confined to the material itself. The electric field of a charged ice film extends into the vacuum above and below the specimen. Its projected potential must therefore include the complete region in which the charging field is appreciable:

$$\boxed{V_q^{\text{proj}}(x, y, \Phi) = \int_{-\infty}^{+\infty} V_q(x, y, z, \Phi), dz.} \tag{5.2118}$$

The electron interaction constant σ was defined in Eq. (5.487) and was written in the particularly transparent form

$$\sigma = \frac{e}{\hbar v} \tag{5.2119}$$

in Eq. (5.520). Applying the same phase-object argument that led to Eq. (5.548), the additional phase generated by charging is

$$\boxed{\phi_q(x, y, \Phi) = \sigma V_q^{\text{proj}}(x, y, \Phi) = \sigma \int V_q(x, y, z, \Phi), dz.} \tag{5.2120}$$

This equation is the direct connection between specimen charging and the electron-optical theory developed earlier in the book.

A smooth charge distribution acts as an electron lens.

Let

$$\boldsymbol{\rho} = \Big(x, y\Big) \tag{5.2121}$$

denote transverse position. Within a sufficiently small region, any smooth charging phase may be expanded as

$$\boxed{\phi_q(\boldsymbol{\rho}) \simeq \phi_0 + \mathbf{b}^{\mathsf{T}}\boldsymbol{\rho} + \frac{1}{2}\boldsymbol{\rho}^{\mathsf{T}}\mathbf{H}_q\boldsymbol{\rho} + \cdots ,} \tag{5.2122}$$

where $\mathbf{b}$ is the local phase gradient and $\mathbf{H}_q$ is the Hessian matrix of second derivatives. The three lowest-order terms have very different optical meanings. The constant term

$$\phi_0 \tag{5.2123}$$

adds only a spatially uniform phase and therefore has no direct effect on image intensity.

The transverse wavevector generated by charging is therefore

$$\delta \mathbf{k}_\perp = \nabla_\perp \phi_q. \tag{5.2124}$$

For a small angular deflection, using the paraxial relation derived in Eq. (3.624),

$$\boldsymbol{\theta}_q \simeq \frac{\delta \mathbf{k}_\perp}{k}. \tag{5.2125}$$

Hence,

$$\boxed{\boldsymbol{\theta}_q = \frac{1}{k} \nabla_\perp \phi_q.} \tag{5.2126}$$

Now consider the locally rotationally symmetric case

$$\boxed{\phi_q(r) \simeq \phi_0 + \alpha r^2, \qquad r^2 = x^2 + y^2.} \tag{5.2127}$$

Its transverse gradient is

$$\nabla_\perp \phi_q = 2\alpha \boldsymbol{\rho}, \tag{5.2128}$$

so that Eq. (5.2126) gives

$$\boldsymbol{\theta}_q = \frac{2\alpha}{k} \boldsymbol{\rho}. \tag{5.2129}$$

An ideal thin lens of focal length f produces the quadratic phase derived previously in Eq. (4.1097),

$$\varphi_l(\boldsymbol{\rho}) = -\frac{k}{2f} |\boldsymbol{\rho}|^2. \tag{5.2130}$$

Comparison with Eq. (5.2127) therefore identifies an effective charging-induced focal length f_q:

$$\alpha = -\frac{k}{2f_q}. \tag{5.2131}$$

Since

$$k = \frac{2\pi}{\lambda}, \tag{5.2132}$$

we obtain

$$\boxed{f_q = -\frac{\pi}{\lambda \alpha}.} \tag{5.2133}$$

Thus the quadratic component of the charging potential is literally an additional electron lens. Its sign determines whether the charge distribution is locally converging or diverging.

If $\mathbf{H}_q$ in Eq. (5.2122) is not proportional to the identity matrix, the curvatures in two orthogonal directions differ. The charge distribution then behaves not as a simple rotationally symmetric lens but as an astigmatic electrostatic element. Still higher spatial derivatives generate higher-order distortions.

The distinction between Eq. (5.2133) and a conventional objective-lens defocus is important. The coefficient α defines the local optical power of the charging field directly, but its conversion into a numerical shift of the objective defocus depends on where the charge-induced lens is located within the complete optical system and on the surrounding lens configuration. There is therefore no universal relation of the form

$$\Delta f_q = \Delta f_q(\alpha) \tag{5.2134}$$

that is independent of microscope geometry.

Experimentally, however, the consequence can often be represented as an *apparent* time-dependent defocus,

$$\boxed{\Delta f_{\text{app}}(\Phi) = \Delta f_0 + \Delta f_{\text{height}}(\Phi) + \Delta f_{\text{charge}}(\Phi),} \tag{5.2135}$$

where Δf_{height} is caused by actual motion of the ice along the optical axis and Δf_{charge} represents the electron-optical lensing effect.

Once the latter is expressed as an apparent CTF defocus change, the defocus-aberration formula already derived in Eq. (4.1191) gives

$$\boxed{\delta\chi_q(\mathbf{R}, \Phi) = -\pi\lambda\Delta f_{\text{charge}}(\Phi)R^2.} \tag{5.2136}$$

Thus two physically distinct processes—motion of the specimen in z and electrostatic lensing by accumulated charge—can both appear experimentally as a changing defocus.

This is precisely what was observed by Schreiber et al. (2023b). Their time-resolved Fresnel measurements showed simultaneous rapid changes in the inferred charge state and in the effective specimen-to-observation-plane distance. They concluded that the initial apparent defocus evolution contains contributions from both physical displacement of the ice film and electrostatic lensing.

Charging can also exert mechanical forces. The electrostatic field associated with specimen charging can act not only on the imaging electrons but also on the specimen itself. If

$$\mathbf{E}_q = -\nabla V_q, \tag{5.2137}$$

then charges on different parts of the film experience Coulomb forces. These forces can couple charging to the beam-induced mechanical motion described in the preceding subsection (Glaeser, 2016).

An order-of-magnitude measure of the mechanical importance of an electric field is the Maxwell pressure in vacuum,

$$\boxed{p_E \sim \frac{\epsilon_0 E_q^2}{2}.} \tag{5.2138}$$

For a dielectric film the precise stress depends on the geometry, dielectric constants, and field components, so Eq. (5.2138) should be regarded as a stress scale rather than as an exact pressure on the ice.

The charge-distribution model fitted by Schreiber et al. (2023b) gave fields at the edge of the illuminated region between approximately

$$E_q \simeq 2.0\text{–}8.4 \times 10^8 \ \mathrm{V\,m^{-1}}. \tag{5.2139}$$

Equation (5.2138) then corresponds to the stress scale

$$\boxed{p_E \sim 0.18\text{–}3.1 \ \mathrm{MPa}.} \tag{5.2140}$$

The largest fields occur close to the charge-gradient region and should not be interpreted as producing a uniform pressure over the entire ice film. Nevertheless, the resulting stress scale is mechanically significant. Values of order a few MPa are comparable to fracture stresses inferred for other vitrified aqueous materials, although these measurements cannot be transferred quantitatively to nanometre-thick vitreous water films. Electrostatic charging could therefore contribute to the initiation or propagation of cracks in an already stressed ice film, particularly near defects or ice–support interfaces. Charging cannot, however, explain cracks that are already present before electron irradiation; those must originate from specimen preparation and cooling. At present, beam-induced charging should therefore be regarded as a plausible contributor to fracture rather than as an established general mechanism for cracking in cryo-EM specimens.

Charging and beam-induced mechanical motion are therefore coupled:

$$\boxed{\begin{array}{c}\text{electron irradiation} \Downarrow \\ \text{charge accumulation} \\ \Downarrow \\ \begin{cases}\text{electrostatic lensing,} \\ \text{electrostatic mechanical stress,}\end{cases} \\ \Downarrow \\ \begin{cases}\text{apparent defocus change,} \\ \text{physical specimen displacement.}\end{cases}\end{array}} \tag{5.2141}$$

Microscopic charge fluctuations: the "bee-swarm" effect. Net or slowly varying charging must be distinguished from microscopic fluctuations of charge. At high defocus, cryogenic electron micrographs can display a rapidly fluctuating granular contrast often called the *bee-swarm effect*. It arises because small, spatially varying electric fields deflect different portions of the electron wave by slightly different angles (Russo and Henderson, 2018c).

This effect is visually striking, but its importance for modern high-resolution cryo-EM is much smaller than its appearance might suggest. Russo and Henderson (2018c) measured the 2.35 Å lattice reflection of gold particles embedded in amorphous ice. At the unusually large defocus

$$|\Delta f| = 10\ \mu\text{m}, \tag{5.2142}$$

microscopic charging reduced the amplitude of the 2.35 Å reflection by only approximately

$$7\%, \tag{5.2143}$$

corresponding to an intensity reduction of approximately

$$14\%. \tag{5.2144}$$

At the much smaller defoci of approximately

$$2\text{–}4\ \mu\text{m} \tag{5.2145}$$

and the spatial frequencies normally used in high-resolution cryo-EM, they concluded that the additional contrast loss caused by microscopic charge fluctuations is negligible when a field-emission gun is used (Russo and Henderson, 2018c).

This result resolves an apparent paradox. Charging is unquestionably present, and it can produce large low-frequency image distortions and time-dependent lensing, yet the *random microscopic charge fluctuations themselves* need not be an important direct source of high-resolution envelope damping. Beam-induced physical motion remains a much more serious limitation.

Charge gradients at the ice–support interface. A conducting support provides a path through which charge can be neutralized, but the situation is not equivalent to placing the ice inside an ideal grounded conductor. Charge transport across the ice–support interface may itself be limiting.

The measurements of Schreiber et al. (2023b) illustrate this clearly. When part of the surrounding carbon film was moved into the illuminated area, the rapid temporal evolution of charging was greatly reduced. Nevertheless, non-zero charge remained on both the ice and the carbon. Consequently, a charge gradient developed across their interface.

Moreover, amorphous carbon is not an ideal inert electrical ground. Hettler et al. (2018b) showed that thin carbon films can themselves develop weak negative charging under electron irradiation. In their experiments this behavior was associated with electron-stimulated desorption of adsorbed molecules, especially water, and the associated change of surface work function. Thus surface chemistry, contamination, temperature, and electrical conductivity can all influence the charging state.

The boundary between vitreous ice and its support should therefore be viewed as an electrostatic boundary condition of the imaging problem, not merely as a mechanical edge.

Mitigating charging with conductive supports. The most direct strategy for reducing charging is to provide an efficient electrical path between the irradiated region and ground. One traditional approach is therefore to illuminate a region slightly larger than the hole so that part of the conductive support foil lies inside the beam. The measurements of Schreiber et al. (2023b) directly confirm that this greatly reduces the rapid charging dynamics.

All-gold supports are attractive in this respect because gold remains highly conductive at cryogenic temperature and, unlike the usual combination of amorphous-carbon foil and metal grid, the foil and grid have matched thermal properties. Russo and Passmore (2016) measured more than an order-of-magnitude reduction in radiation-induced support motion with all-

gold substrates. For example, under one set of low-dose conditions the vertical displacement of the bare support was reduced from approximately

$$228 \text{ Å} \tag{5.2146}$$

for amorphous carbon to

$$3.8 \text{ Å} \tag{5.2147}$$

for gold, a factor of approximately 60. With vitreous ice present, the corresponding values were approximately

$$76 \text{ Å} \qquad \text{and} \qquad 1.9 \text{ Å}, \tag{5.2148}$$

respectively.

These measurements should not, however, be interpreted as proof that the improvement is caused primarily by reduced charging. Gold also greatly reduces differential thermal contraction and the associated mechanical stress. Mechanical stabilization and improved conductivity occur simultaneously.

Furthermore, a highly conducting foil surrounding a hole does not guarantee that the insulating ice at the center of the hole is equipotential with it. If the illumination remains entirely inside the ice, charge must still travel from the illuminated region to the foil. This distinction has been demonstrated directly in more recent experiments.

Zhang et al. (2023) placed a single layer of conducting graphene across conventional cryo-EM holes, thereby providing a conductive path directly beneath the vitrified specimen. Without graphene, an electron beam confined to the middle of a hole produced strong charging effects. With graphene present, charging was greatly reduced. Under data-collection conditions in which specimens without graphene failed to yield reconstructions better than approximately 3 Å, graphene-supported specimens produced a reconstruction at approximately

$$2 \text{ Å}. \tag{5.2149}$$

The same study found that the diffraction signature of charging could already be detected in their first recorded frame, corresponding to only

$$0.047\ e^{-}\ \text{Å}^{-2}, \tag{5.2150}$$

and approached a stable beam size near

$$1.5\ e^{-}\ \text{Å}^{-2}. \tag{5.2151}$$

These results reinforce the central lesson of this subsection. Specimen charging cannot be represented by a single scalar "amount of charge". Its image-forming consequences depend on

$$\boxed{\begin{array}{c}\text{how much charge is produced,}\\ \text{where positive and negative charge accumulate,}\\ \text{how rapidly charge can reach the support,}\\ \text{how the resulting potential varies across the beam,}\\ \text{and how that potential changes during the exposure.}\end{array}} \tag{5.2152}$$

For image formation, the especially important quantity is not the net charge itself but the spatially varying projected electrostatic potential in Eq. (5.2118). A constant charging potential contributes only a global phase, a gradient deflects the beam, a quadratic component behaves as an additional lens, and irregular higher-order components distort the image.

Charging is therefore simultaneously a radiation-induced specimen modification, a possible source of mechanical motion, and a time-dependent electron-optical element. In modern cryo-EM the direct high-resolution attenuation from microscopic charge fluctuations appears to be small, but charge-induced lensing, low-order distortion, and mechanical coupling—particularly during the earliest part of the exposure—remain physically important.

5.13.13 The low-dose acquisition principle

All of the preceding physics leads to a simple experimental principle: the specimen should not be irradiated before the electrons are needed for the measurement that will ultimately be used.

The conventional name for this procedure is *low-dose acquisition*. Strictly speaking, however, the quantity most directly controlled in cryo-EM is normally the electron exposure Φ defined in Eq. (5.1678), rather than the absorbed dose D_{abs}. We shall nevertheless retain the established term "low dose".

The central principle can be stated as

$$\boxed{\begin{array}{c}\text{Do not spend electrons obtaining information}\\ \text{that is not part of the final measurement.}\end{array}} \tag{5.2153}$$

The word *spend* is appropriate. Radiation damage makes the number of useful electrons that can interrogate a particular region of the specimen a finite resource. Once an electron has irradiated that region, the structural state of the specimen has changed, irrespective of whether that electron contributed to the final recorded image.

Pre-exposure consumes the high-resolution information budget. The quantitative reason for avoiding pre-irradiation follows directly from the radiation-damage envelope derived previously. At spatial frequency R, Eq. (5.1845) gives the coherent-amplitude attenuation

$$E_{\text{rad}}(R, \Phi) = \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right], \tag{5.2154}$$

where the critical exposure $\Phi_c(R)$ was defined in Eq. (5.1832).

Suppose that, before the actual high-resolution image is recorded, the target region has already received the exposure

$$\boxed{\Phi_{\text{pre}} = \Phi_{\text{search}} + \Phi_{\text{focus}} + \Phi_{\text{setup}},} \tag{5.2155}$$

where the terms represent any irradiation received by the target during searching, focusing and other microscope adjustments.

The acquisition itself then delivers an additional exposure Φ_{acq}.

The cost of exposing the specimen before acquisition.

Consider one spatial frequency R. As shown in Eq. (5.1853), the critical exposure decreases with increasing spatial frequency in the simple random-displacement model; experimentally, a similar but not exactly R^{-2} dependence was found in Eq. (5.1854).

If the specimen were pristine when acquisition began, its coherent contribution accumulated during an exposure Φ_{acq} would be proportional to

$$S_0(R) \propto \int_0^{\Phi_{\text{acq}}} \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right] d\Phi. \tag{5.2156}$$

Now suppose that the region has already received Φ_{pre} before recording starts. The same acquisition then samples the exposure interval

$$\Phi_{\text{pre}} \le \Phi \le \Phi_{\text{pre}} + \Phi_{\text{acq}}, \tag{5.2157}$$

and therefore

$$S_{\text{pre}}(R) \propto \int_{\Phi_{\text{pre}}}^{\Phi_{\text{pre}}+\Phi_{\text{acq}}} \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right] d\Phi. \tag{5.2158}$$

Introducing

$$u = \Phi - \Phi_{\text{pre}}, \tag{5.2159}$$

we obtain

$$S_{\mathrm{pre}}(R) \propto \exp\left[-\frac{\Phi_{\mathrm{pre}}}{2\Phi_c(R)}\right] \int_0^{\Phi_{\mathrm{acq}}} \exp\left[-\frac{u}{2\Phi_c(R)}\right] du \tag{5.2160}$$

$$= \exp\left[-\frac{\Phi_{\mathrm{pre}}}{2\Phi_c(R)}\right] S_0(R). \tag{5.2161}$$

Thus

$$\boxed{\frac{S_{\mathrm{pre}}(R)}{S_0(R)} = \exp\left[-\frac{\Phi_{\mathrm{pre}}}{2\Phi_c(R)}\right].} \tag{5.2162}$$

For the corresponding coherent power,

$$\boxed{\frac{P_{\mathrm{pre}}(R)}{P_0(R)} = \exp\left[-\frac{\Phi_{\mathrm{pre}}}{\Phi_c(R)}\right].} \tag{5.2163}$$

Equation (5.2163) gives a quantitative meaning to the low-dose principle. For example, if the target receives a pre-exposure equal to only

$$\Phi_{\mathrm{pre}} = 0.1\,\Phi_c(R), \tag{5.2164}$$

then the coherent power available when the real measurement begins has already been reduced to

$$e^{-0.1} \simeq 0.905. \tag{5.2165}$$

Approximately 9.5% of the coherent power at that spatial frequency has therefore been lost before the useful acquisition has even begun.

If instead

$$\Phi_{\mathrm{pre}} = \Phi_c(R), \tag{5.2166}$$

only

$$e^{-1} \simeq 0.37 \tag{5.2167}$$

of the original coherent power remains.

The penalty is especially severe at high spatial frequency because $\Phi_c(R)$ decreases as resolution increases. There is therefore no single universal "electron budget". The available exposure is resolution dependent, exactly as discussed in connection with Eq. (5.1870).

Importantly, increasing the subsequent acquisition exposure cannot restore the pristine high-resolution information that was destroyed during Φ_{pre}. Later electrons interrogate a different, more damaged specimen.

The classical search–focus–acquisition strategy. The practical low-dose procedure was developed precisely to make Φ_{pre} in Eq. (5.2155) as small as possible for the region from which the final high-resolution image will be recorded (Egerton, 2013; Grassucci et al., 2008).

Three distinct microscope states are traditionally used.

search	$\longrightarrow$	locate a suitable region using a broad field and the smallest practical exposure
focus	$\longrightarrow$	adjust focus and astigmatism on a sacrificial neighboring region
acquisition	$\longrightarrow$	illuminate the previously unexposed target only when recording begins.

(5.2168)

In the *search* state, a low magnification and weak, broadly spread illumination provide a large field of view with very little exposure per unit area. The purpose is to identify a suitable grid square, hole or specimen region without consuming an appreciable fraction of its high-resolution exposure budget.

In the *focus* state, the microscope is placed under optical conditions representative of those required for acquisition, but the beam is displaced to a nearby sacrificial region. Focus, astigmatism, and related adjustments can then be performed without irradiating the target itself.

Only after these adjustments are complete is the beam moved to the *acquisition* region and the final exposure begins. In the ideal low-dose experiment,

$$\boxed{\Phi_{\text{search,target}} \simeq \Phi_{\text{focus,target}} \simeq 0,} \tag{5.2169}$$

so that

$$\Phi_{\text{target}} \simeq \Phi_{\text{acq}}. \tag{5.2170}$$

Egerton describes this procedure explicitly: focusing is carried out on a nearby specimen region, after which the incident beam is deflected onto the region of interest only immediately before image acquisition (Egerton, 2013). A representative classical implementation described by Grassucci et al. (2008) used approximately 3000× magnification for search, approximately 115 000× for focusing on a neighboring region, and a separate exposure state for recording. These numerical settings are instrument-specific; the physically important feature is the spatial separation of the focus and acquisition regions.

The same principle applies to microscope adjustments other than focus. There is little advantage in obtaining a perfectly measured high-magnification diagnostic image of the target if obtaining that diagnostic image destroys part of the structure that the subsequent measurement was intended to record.

Low dose is not the same as low dose rate. A second distinction is essential. As already emphasized in Sec. 5.13.2, accumulated electron exposure and exposure rate are different physical quantities. In the notation of Eqs. (5.1678) and (5.1680),

$$\boxed{\begin{array}{ccc} \text{low dose} & & \text{low dose rate} \\ \Updownarrow & \neq & \Updownarrow \\ \text{small accumulated exposure } \Phi & & \text{small exposure rate } \dot{\Phi}. \end{array}} \tag{5.2171}$$

For constant exposure rate, Eq. (5.1682) gives

$$\Phi_{\text{acq}} = \dot{\Phi}\, T_{\text{acq}}, \tag{5.2172}$$

or

$$\boxed{T_{\text{acq}} = \frac{\Phi_{\text{acq}}}{\dot{\Phi}}.} \tag{5.2173}$$

Thus lowering the exposure rate while keeping the total exposure fixed does not save electrons. It simply delivers the same number of electrons over a longer time.

In the simplest exposure-dependent radiation-damage model developed above, two acquisitions with the same Φ would therefore consume the same structural-information budget. This is the sense in which *low dose* refers fundamentally to accumulated exposure.

The equality is not exact for a real cryo-EM specimen, however. Equation (5.1722) already showed that a specimen containing competing production and relaxation processes can reach different states at the same accumulated exposure when $\dot{\Phi}$ differs.

Several effects discussed in the preceding subsections provide experimental examples. Brilot et al. (2012) observed that lower exposure rates tended to reduce beam-induced specimen motion. Conversely, a lower rate requires a longer exposure time, which may make stage drift, environmental vibrations and other slow instabilities more important. The choice of exposure rate is therefore itself a compromise.

Charging provides an especially clear demonstration that total exposure and exposure rate cannot always be interchanged. The time-resolved measurements of Schreiber et al. (2023a) showed that both the charge state of vitreous ice and its physical movement depend on exposure rate. The accumulated exposure required to approach a quasi-steady charging state also changed with exposure rate.

Consequently,

$$\boxed{\begin{array}{ll} \Phi & \text{sets primarily how much of the finite structural information budget has been consumed,} \\ \dot{\Phi} & \text{can modify how the specimen evolves while that budget is being consumed.} \end{array}} \tag{5.2174}$$

This is why the two experimental decisions should be kept conceptually separate. Low-dose acquisition determines *where* the available electrons are spent and how many are spent on the final measurement. The exposure rate determines *how quickly* those electrons are delivered.

Dose fractionation does not abolish the low-dose principle. Modern direct detectors record the acquisition as a sequence of movie frames rather than as one indivisible exposure. This greatly improves how the available electrons can be used. Beam-induced translations can be estimated and corrected, and frames can be weighted according to the resolution-dependent radiation envelope, as described by Eq. (5.1874).

Dose fractionation does not, however, make pre-exposure harmless. It can preserve the temporal ordering of electrons that arrive *after recording has started*; it cannot reconstruct structural information that was destroyed before the detector began acquiring the useful measurement.

The modern form of the low-dose principle is therefore not simply "use as few electrons as possible". Too few electrons also give poor counting statistics, as expressed by the optimization discussed around Eq. (5.1869). Rather, the principle is

$$\boxed{\begin{array}{c} \text{preserve the pristine specimen until useful recording begins,} \\ \text{then extract as much information as possible from every electron that consumes the available exposure budget.} \end{array}} \tag{5.2175}$$

5.13.14 From low-dose images to dose-fractionated movies

The classical low-dose strategy of Sec. 5.13.13 determines *when* the valuable specimen region is first exposed and how the available electron exposure is spent. Modern direct electron detectors add another possibility: instead of integrating the complete acquisition into one image, they preserve information about *when during the exposure* the detected electrons arrived.

Suppose that the total acquisition exposure is

$$\Phi_{\text{tot}}. \tag{5.2176}$$

A dose-fractionated movie divides this acquisition into N_f temporal fractions,

$$\boxed{\Phi_{\text{tot}} = \sum_{j=1}^{N_f} \Delta\Phi_j,} \tag{5.2177}$$

where $\Delta\Phi_j$ is the electron exposure accumulated during frame j.

For equal fractions,

$$\Delta\Phi_j = \frac{\Phi_{\text{tot}}}{N_f}, \tag{5.2178}$$

but equality is not essential. What matters is that the temporal ordering of the acquisition has been retained.

The fundamental physical point is

$$\boxed{\begin{array}{l}\text{Dose fractionation preserves the temporal ordering}\\ \text{of radiation damage; it does not undo the damage.}\end{array}} \tag{5.2179}$$

Fractionating the detector recording does not reduce the primary exposure. The reason is straightforward. The specimen is upstream from the detector. Whether the electrons that emerge from the microscope are recorded in one image, 40 movie frames, or thousands of temporal fractions cannot alter the interactions that have already occurred inside the specimen.

This statement can be made quantitative.

Why dose fractionation does not reduce the number of primary inelastic interactions.

Consider an illuminated specimen area A. From the definition of electron exposure in Eq. (5.1678), the number of primary electrons incident

on this area during movie frame j is

$$N_{e,j} = A\Delta\Phi_j. \tag{5.2180}$$

We showed in Eq. (5.1534) that an electron traversing a specimen of thickness t undergoes on average

$$\mu_{\text{in}} = \frac{t}{\Lambda_{\text{in}}} \tag{5.2181}$$

inelastic interactions.

The expected number of primary inelastic interactions produced during frame j is therefore

$$\langle N_{\text{in},j} \rangle = A\Delta\Phi_j \frac{t}{\Lambda_{\text{in}}}. \tag{5.2182}$$

Summing over all movie frames,

$$\langle N_{\text{in,tot}} \rangle = \sum_{j=1}^{N_f} A\Delta\Phi_j \frac{t}{\Lambda_{\text{in}}} \tag{5.2183}$$

$$= A\frac{t}{\Lambda_{\text{in}}} \sum_{j=1}^{N_f} \Delta\Phi_j. \tag{5.2184}$$

Using Eq. (5.2177),

$$\boxed{\langle N_{\text{in,tot}} \rangle = A\Phi_{\text{tot}} \frac{t}{\Lambda_{\text{in}}}.} \tag{5.2185}$$

The number of fractions N_f has disappeared.

Thus, if genuine exposure-rate effects are negligible, a specimen receiving

$$40\ e^-\ \text{Å}^{-2} \tag{5.2186}$$

experiences essentially the same expected number of primary inelastic interactions whether those electrons are recorded as one integrated image or as 40, 400, or 4000 temporal fractions.

Dose fractionation does not protect the specimen from those interactions. Its advantage is that it prevents the detector from irreversibly mixing together electrons that encountered the specimen at very different stages of its evolution.

This conclusion must be qualified by the distinction between exposure and exposure rate discussed previously. If changing the acquisition scheme also changes $\dot{\Phi}$, then charging, mechanical motion, chemical relaxation, or other rate-dependent processes can change, as illustrated by Eq. (5.1722). The statement above concerns fractionating the *recording* while keeping the irradiation conditions unchanged.

Each frame observes a different specimen. Let

$$\Phi_j = \sum_{\ell=1}^{j} \Delta\Phi_\ell \tag{5.2187}$$

be the accumulated exposure at the end of frame j, with

$$\Phi_0 = 0. \tag{5.2188}$$

Frame j therefore records electrons that encounter the specimen between the two radiation states

$$\boxed{\Phi_{j-1} \leq \Phi \leq \Phi_j.} \tag{5.2189}$$

This is more than a bookkeeping distinction. The specimen potential is itself exposure dependent, as expressed in Eq. (5.1762). The first and last frames of a movie therefore do not constitute repeated measurements of exactly the same object.

For radiation damage alone, the coherent amplitude at spatial frequency R is attenuated according to Eq. (5.1845),

$$E_{\text{rad}}(R, \Phi) = \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right]. \tag{5.2190}$$

The average radiation envelope sampled during frame j is consequently

$$\overline{E}_j(R) = \frac{1}{\Delta\Phi_j} \int_{\Phi_{j-1}}^{\Phi_j} \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right] d\Phi \tag{5.2191}$$

$$= \boxed{\frac{2\Phi_c(R)}{\Delta\Phi_j}\left[\exp\left(-\frac{\Phi_{j-1}}{2\Phi_c(R)}\right) - \exp\left(-\frac{\Phi_j}{2\Phi_c(R)}\right)\right].} \tag{5.2192}$$

For a sufficiently small exposure per frame,

$$\Delta\Phi_j \ll \Phi_c(R), \tag{5.2193}$$

the envelope varies little across one frame and Eq. (5.2192) reduces approximately to the frame weighting introduced previously in Eq. (5.1874).

The dependence on R is crucial. The simple displacement model gave

$$\Phi_c(R) \propto R^{-2} \tag{5.2194}$$

in Eq. (5.1853), while the experimental dependence measured by Grant and Grigorieff (2015) was given in Eq. (5.1854). High spatial frequencies therefore lose useful coherent information much earlier than low spatial frequencies.

The useful part of the movie changes during the exposure. Radiation damage alone would suggest using the earliest possible electrons for high-resolution information. The preceding sections show why the real situation is more subtle.

Immediately after irradiation begins, the specimen is chemically closest to its pristine state. However, this is also commonly the period in which the largest mechanical relaxation and the fastest charging dynamics occur. The beam-induced motion discussed in Sec. 5.13.11 is often strongest during the first part of the exposure, while the charging state described in Sec. 5.13.12 can evolve extremely rapidly during initial irradiation.

At the other end of the movie, motion and charging may have approached a more slowly varying state, but radiation damage has progressively removed high-resolution structural information.

A useful qualitative picture is therefore

early frames	:	least chemically damaged, but often strongest motion and charging,
intermediate frames	:	reduced initial mechanical transients, with substantial high-frequency signal remaining,
late frames	:	progressively radiation damaged, with mainly lower-frequency information surviving.

(5.2195)

The boundaries between these regimes are not universal. They depend on the specimen, support, exposure rate, accelerating voltage, ice thickness, and spatial frequency under consideration.

The early-motion behavior is directly visible in dose-fractionated measurements. Campbell et al. (2012), for example, observed faster particle motion during the beginning of the exposure and slower motion toward its end. Recording the acquisition as a movie made it possible to distinguish those stages rather than averaging them irreversibly into one image.

Similarly, the time-resolved charging measurements of Schreiber et al. (2023a) reveal a rapid initial change of the electrostatic state of the ice. Thus the earliest electrons have the special advantage of interrogating the least chemically damaged specimen, but they do not necessarily provide the mechanically most stable image.

Temporal fractionation also limits irrecoverable intra-frame motion. Dose fractionation is useful even before any sophisticated image-processing method is considered.

Suppose that an illuminated patch undergoes an in-plane displacement $\mathbf{u}_{\parallel}(\Phi)$. As derived in Eq. (5.2090), its Fourier phase contains the factor

$$\exp\left[-j2\pi\mathbf{R}\cdot\mathbf{u}_{\parallel}(\Phi)\right]. \tag{5.2196}$$

A movie frame integrates this phase throughout its exposure interval:

$$\boxed{M_j(\mathbf{R}) = \frac{1}{\Delta\Phi_j}\int_{\Phi_{j-1}}^{\Phi_j} \exp\left[-j2\pi\mathbf{R}\cdot\mathbf{u}_{\parallel}(\Phi)\right] d\Phi.} \tag{5.2197}$$

If the specimen moves substantially during one frame, the different phases have already been averaged together inside that frame. A subsequent rigid translation of the complete frame can correct its mean position but cannot reconstruct information lost through this *intra-frame* averaging.

The phase excursion caused by a displacement change $\Delta\mathbf{u}$ is of order

$$\Delta\varphi \sim 2\pi\mathbf{R}\cdot\Delta\mathbf{u}. \tag{5.2198}$$

Thus a useful criterion for small intra-frame blur is

$$\boxed{2\pi R|\Delta\mathbf{u}| \ll 1.} \tag{5.2199}$$

The higher the desired spatial frequency, the finer the temporal sampling required to keep this condition satisfied.

This is the physical reason that increased temporal resolution in the detector can remain useful even when the final image will eventually be formed by summing many detected electrons.

From movie frames to electron-event representation. A conventional dose-fractionated movie still makes a decision about temporal resolution during data acquisition. The detector operates at some underlying physical frame rate, but successive raw camera frames are commonly

summed into larger *exposure fractions* before those fractions are stored. Once that summation has been performed, the temporal ordering of the electrons within one stored fraction has been lost.

Electron-event representation (EER) takes the idea of temporal fractionation one step further (Guo et al., 2020). Instead of storing a sequence of pixel-intensity images, the data are stored as a sparse stream of individual *detected electron events*. Conceptually, the EER data can be written as

$$\boxed{\mathcal{E} = \{(x_i, y_i, n_i)\}_{i=1}^{N_{\mathrm{det}}}\,,} \tag{5.2200}$$

where

$$(x_i, y_i) \tag{5.2201}$$

is the localized detector position of event i, and n_i identifies the raw detector frame in which the event occurred.

Thus the time coordinate is discrete rather than an infinitely precise continuous timestamp, but it retains the temporal resolution of the underlying detector readout. The impact position can also be localized at subpixel resolution in an electron-counting detector (Guo et al., 2020).

It is important to interpret Eq. (5.2200) correctly. An EER file is not a list of every primary electron that struck the specimen and certainly not a list of every scattering event inside it. It is a list of *electron-detection events at the camera*, after the electrons have passed through the specimen and microscope.

Once the event stream has been recorded, conventional movie frames can be constructed afterwards. If $\mathcal{B}_j$ denotes the set of raw camera-frame indices assigned to output fraction j, then conceptually

$$\boxed{I_j(x, y) = \sum_{i:\, n_i \in \mathcal{B}_j} \delta\,(x - x_i)\,\delta\,(y - y_i)\,,} \tag{5.2202}$$

followed in practice by the appropriate spatial binning and detector corrections.

The same recorded event stream can therefore be converted afterwards into, for example,

$$20, \qquad 40, \qquad 100, \tag{5.2203}$$

or many more exposure fractions, provided that the underlying detector time resolution is not exceeded.

This is an important conceptual change:

$$\boxed{\begin{array}{c}\text{conventional movie}\\ \Downarrow\\ \text{temporal binning is partly chosen}\\ \text{during acquisition}\\ \text{EER}\\ \Downarrow\\ \text{fine event timing is stored and}\\ \text{temporal binning can be chosen later}\end{array}} \tag{5.2204}$$

Guo et al. (2020) demonstrated this approach experimentally. Their EER format stores the position and raw-frame time of each detected electron while exploiting the sparsity of counting-mode data to keep the file size manageable. It preserves both the temporal information required for fine motion correction and the subpixel position information used for super-resolution imaging.

EER therefore does not increase the radiation tolerance of the specimen. Its benefit is informational: it postpones the decision about how the detected electrons should be grouped in time.

Motion correction and exposure weighting use different information in the same movie. Once temporal ordering has been retained, two distinct physical corrections become possible.

First, successive fractions contain information about the trajectory of the specimen. Their relative displacements can therefore be estimated and the frames brought into register. The direct benefit of this procedure was already demonstrated in the early movie experiments of Campbell et al. (2012): correcting beam-induced motion substantially improved both the measured B-factor and the final reconstruction resolution.

Second, different temporal fractions need not contribute equally at all spatial frequencies. The critical exposure is strongly resolution-dependent, as derived in Eq. (5.1853) and measured experimentally in Eq. (5.1854). The exposure-dependent weighting introduced previously in Eq. (5.1874) can therefore preserve the high-frequency contribution of early frames while allowing later frames to continue contributing at lower spatial frequencies.

Grant and Grigorieff demonstrated this principle directly using rotavirus movies acquired with a total exposure of approximately

$$100\ e^{-}\ \text{Å}^{-2}. \tag{5.2205}$$

They measured the optimal exposure as a function of spatial frequency and used that information to filter the movie frames before combining them, obtaining improved image contrast and higher-resolution reconstructions (Grant and Grigorieff, 2015).

The important point for the present chapter is not the details of the filtering algorithm. It is the physical reason that such an algorithm can work:

$$\boxed{\begin{array}{c}\text{the detector preserves when the electrons arrived}\\ \Downarrow\\ \text{motion can be estimated as a function of time}\\ \text{and}\\ \text{radiation damage can be weighted as a function of}\\ \text{both time and spatial frequency.}\end{array}} \tag{5.2206}$$

What dose fractionation can and cannot recover. Dose-fractionated acquisition has transformed cryo-EM, but its limitations should remain clear.

It *can* preserve information that would otherwise be lost by averaging together different specimen positions. It can also prevent late, strongly damaged electrons from being given the same high-frequency weight as early electrons.

It *cannot* reverse bond cleavage, restore atoms that have moved, remove radiolysis products, return a mechanically deformed specimen to its original three-dimensional state, or undo electrostatic changes that occurred before an electron was detected. Nor can motion correction reconstruct information that has already been averaged away within one temporal fraction.

The sequence is therefore

$$\boxed{\begin{array}{c}\text{pristine specimen}\\ \Downarrow\\ \text{continuous irradiation and continuous specimen evolution}\\ \Downarrow\\ \text{temporally resolved detection}\\ \Downarrow\\ \text{motion estimation}\\ +\ \text{resolution-dependent exposure weighting}\\ \Downarrow\\ \text{more efficient use of the information that survived,}\\ \text{not reversal of the underlying radiation damage.}\end{array}} \tag{5.2207}$$

In this sense, dose-fractionated movies and EER are the modern completion of the low-dose principle. Classical low-dose microscopy tries to ensure that the useful region remains pristine until the measurement begins. Dose-fractionated acquisition then tries to retain as much information as possible about *when* each part of that finite electron budget was recorded.

5.14 Summary

This chapter has followed the electron through the specimen, from the microscopic electrostatic potential of matter to the detected cryo-EM image. The central object introduced at the beginning of the chapter was the specimen scattering operator,

$$\boxed{\psi_{\text{exit}} = \mathcal{S}\,\psi_{\text{in}},} \tag{5.2208}$$

originally defined in Eq. (5.4). Everything developed in this chapter can be regarded as a progressively more realistic description of the operator $\mathcal{S}$.

At the simplest level, $\mathcal{S}$ is an elastic phase modulation produced by the projected electrostatic potential. At greater accuracy, it includes propagation through a three-dimensional potential, multiple elastic scattering, transitions of the specimen into inelastic final states, and finally the fact that the specimen itself changes while it is being irradiated.

The essential hierarchy is therefore

$$\boxed{\begin{array}{c}\text{charge distribution}\\ \Downarrow\\ \text{electrostatic potential}\\ \Downarrow\\ \text{elastic and inelastic electron scattering}\\ \Downarrow\\ \text{exit electron wave and energy-loss channels}\\ \Downarrow\\ \text{electron-optical image formation}\\ \Downarrow\\ \text{detected electrons,}\end{array}} \tag{5.2209}$$

with the important complication that irradiation progressively changes several elements of this chain.

From atomic charge to the specimen potential. For elastic scattering, the central material quantity is the electrostatic potential $V(\mathbf{r})$. It is

determined by the charge density through Poisson's equation,

$$\boxed{\nabla^2 V(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\varepsilon_0},} \tag{5.2210}$$

derived in Eq. (5.32).

The same atomic potential can be represented in reciprocal space by the electron scattering factor,

$$\boxed{f_e(\mathbf{R}) = \frac{m_0 e}{2\pi\hbar^2}\widehat{V}_a(\mathbf{R}),} \tag{5.2211}$$

as defined in Eq. (5.82). The equivalent relation

$$\rho(\mathbf{r}) \longleftrightarrow V(\mathbf{r}) \longleftrightarrow f_e(\mathbf{R}) \tag{5.2212}$$

was summarized in Eq. (5.30).

For biological specimens, atomic potentials are normally assembled using the independent-atom approximation. Chemical bonding modifies the true charge density, particularly at low spatial frequency, but the atomic scattering-factor representation provides the practical bridge between an atomic structural model and the electrostatic potential required for electron-wave calculations.

The Mott–Bethe relation, Eq. (5.113),

$$f_e(\mathbf{R}) = \frac{1}{2\pi^2 a_0}\frac{Z - f_X(\mathbf{R})}{R^2}, \tag{5.2213}$$

also makes clear why electrons are exceptionally sensitive to redistribution of charge: electron scattering responds to the complete electrostatic potential, not only to the electron density.

The fundamental propagation equation inside the specimen. For a fast electron propagating predominantly along the optical axis, the three-dimensional specimen interaction is described by the paraxial fast-electron equation,

$$\boxed{\frac{\partial\psi}{\partial z} = \frac{j\lambda}{4\pi}\nabla_\perp^2\psi + j\sigma V\psi,} \tag{5.2214}$$

derived in Eq. (5.486), where

$$\boxed{\sigma = \frac{e}{\hbar v}} \tag{5.2215}$$

is the electron interaction constant derived in Eq. (5.520).

Equation (5.2214) contains the two processes that govern coherent propagation through the specimen:

$$\boxed{\begin{aligned} \frac{j\lambda}{4\pi}\nabla_\perp^2\psi \quad &\longrightarrow \quad \text{transverse diffraction,} \\ j\sigma V\psi \quad &\longrightarrow \quad \text{interaction with the specimen potential.} \end{aligned}} \tag{5.2216}$$

This equation is the natural starting point for understanding both the phase-object approximation and multislice.

Phase object, projection approximation, and the standard SPA model. If transverse propagation while crossing the specimen can be neglected, the three-dimensional potential collapses into its projected potential,

$$V^{\text{proj}}(x,y) = \int V(x,y,z)\,dz, \tag{5.2217}$$

defined in Eq. (5.546), and the specimen acts as the phase transmission function

$$\boxed{t(x,y) = \exp\left[j\sigma V^{\text{proj}}(x,y)\right],} \tag{5.2218}$$

derived in Eq. (5.548).

The additional weak-phase-object approximation assumes

$$\left|\sigma V^{\text{proj}}\right| \ll 1 \tag{5.2219}$$

and linearizes Eq. (5.2218):

$$\boxed{\psi_{\text{exit}} \simeq 1 + j\sigma V^{\text{proj}}.} \tag{5.2220}$$

This result was derived in Eq. (5.569).

A central conceptual distinction of the chapter is that

$$\boxed{\text{first Born} \neq \text{projection approximation} \neq \text{weak-phase approximation.}} \tag{5.2221}$$

Their different assumptions were summarized explicitly in Eq. (5.579).

Combining the weak-phase specimen model with the electron optics developed in Chap. 4 gives the standard single-particle Fourier-space model,

$$\boxed{\widehat{y}_i(\mathbf{R}) = a_i\,\text{CTF}_i(\mathbf{R})e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i}\widehat{\mathcal{P}_{\mathcal{R}_i}V_0}(\mathbf{R}) + \widehat{n}_i(\mathbf{R}),} \tag{5.2222}$$

derived in Eq. (5.608).

This familiar equation is therefore not the starting point of the physics. It is the end result of several approximations to the complete three-dimensional electron–specimen interaction.

When the specimen is genuinely three dimensional. When propagation through the specimen thickness cannot be neglected, interaction and propagation must be alternated. Multislice implements this directly:

$$\boxed{\psi_{n+1} = \mathcal{P}_{\Delta z_n} \left\{ \mathcal{T}_n \psi_n \right\},} \tag{5.2223}$$

as derived in Eq. (5.667). The transmission operator $\mathcal{T}_n$ describes interaction with one slice and $\mathcal{P}_{\Delta z_n}$ propagates the modified electron wave to the next slice.

Multislice therefore retains the depth ordering of the specimen and naturally contains multiple elastic scattering. In the weak, single-scattering limit it also reveals the connection with the Ewald sphere.

The important geometrical result is that specimen depth, Fresnel propagation, Ewald-sphere curvature, and depth-dependent defocus are different descriptions of the same longitudinal phase. This equivalence was summarized in Eq. (5.1032).

For a specimen of thickness D, the useful criterion

$$\boxed{\pi \lambda D R_\perp^2 \ll 1} \tag{5.2224}$$

implies simultaneously that propagation through specimen depth is negligible, that one effective defocus is adequate, and that the Ewald sphere can be replaced by its tangent plane, as shown in Eq. (5.1040).

Thus the projection approximation is not simply synonymous with a "thin specimen". Its validity depends jointly on specimen thickness, electron wavelength, and the spatial frequency one wishes to recover.

Disorder and B-factors describe loss of coherent structure, not a unique physical mechanism. Atomic displacement, incomplete occupancy, conformational heterogeneity, beam-induced motion, radiation damage, alignment errors, and other sources of disorder can all reduce coherent Fourier amplitudes. When their net effect is approximately Gaussian in spatial frequency, it is conveniently summarized by

$$\boxed{F_{\text{obs}}(\mathbf{R}) \simeq F_{\text{ideal}}(\mathbf{R}) \exp\left(-\frac{B_{\text{overall}} R^2}{4} \right),} \tag{5.2225}$$

as discussed in Eq. (5.1210).

A B-factor is therefore a useful description of signal attenuation, but it does not by itself identify its physical origin. In particular, static conformational heterogeneity and correlated molecular motion need not obey an exact Debye–Waller law even when an empirical B-factor provides a useful approximation.

Inelastic scattering creates different specimen channels. Elastic scattering leaves the specimen in the same quantum state. Inelastic scattering leaves it in a different state and transfers both energy and momentum. The measurable double-differential cross section is governed by the dynamic structure factor,

$$\boxed{\frac{d^2\sigma_{\text{in}}}{d\Omega\, d\epsilon} = \frac{\gamma^2}{4\pi^4 a_0^2 R^4}\frac{k_{\text{f}}}{k_{\text{i}}} S_n(\mathbf{R}, \epsilon),} \tag{5.2226}$$

derived in Eq. (5.1376).

Different final specimen states are mutually incoherent after the unobserved specimen degrees of freedom are traced out. This does not mean that an inelastically scattered electron has lost all phase information: each individual energy-loss channel still propagates as a complex electron wave with its own angular distribution and energy-dependent CTF.

For a homogeneous specimen, the total inelastic cross section gives the inelastic mean free path

$$\boxed{\Lambda_{\text{in}} = \frac{1}{n\sigma_{\text{in}}},} \tag{5.2227}$$

from Eq. (5.1496). A specimen of thickness t consequently has

$$\mu_{\text{in}} = \frac{t}{\Lambda_{\text{in}}} \tag{5.2228}$$

mean inelastic events per transmitted electron [Eq. (5.1514)], and the number of inelastic interactions follows

$$\boxed{P_N(t) = \frac{1}{N!}\left(\frac{t}{\Lambda_{\text{in}}}\right)^N \exp\left(-\frac{t}{\Lambda_{\text{in}}}\right).} \tag{5.2229}$$

In particular,

$$P_0 = \exp\left(-\frac{t}{\Lambda_{\text{in}}}\right) \tag{5.2230}$$

is the zero-loss fraction [Eq. (5.1536)].

An energy filter therefore does not "repair" inelastically scattered electrons; it selects a different population of electron–specimen channels. Without energy filtering, the inelastic contribution to the power spectrum must in general be treated through its energy-dependent transfer functions, as expressed in Eq. (5.1655).

The imaging beam changes the specimen. Inelastic scattering is also the starting point of the dominant radiation-damage mechanism in frozen-hydrated biological specimens. Electronic excitation and ionization produce secondary electrons, radicals, bond cleavage, molecular products, density changes, and ultimately structural rearrangement.

The central consequence is that the specimen potential is no longer static:

$$\boxed{V(\mathbf{r}) \longrightarrow V(\mathbf{r}, \Phi).} \tag{5.2231}$$

More explicitly,

$$V(\mathbf{r}, \Phi) = V_0(\mathbf{r}) + \delta V_{\text{chem}}(\mathbf{r}, \Phi), \tag{5.2232}$$

as introduced in Eq. (5.1762).

Radiation damage is chemically selective. It cannot in general be represented as a simple convolution of an undamaged structure, Eq. (5.1802). The irradiated specimen is more fundamentally an exposure-dependent ensemble of chemical and conformational states, as represented in Eq. (5.1822).

For averaged coherent structural information, however, a very useful effective description is the radiation-damage envelope

$$\boxed{E_{\text{rad}}(R, \Phi) = \exp\left[-\frac{\Phi}{2\Phi_c(R)}\right].} \tag{5.2233}$$

This is Eq. (5.1845). The corresponding structural power decays as

$$\boxed{|E_{\text{rad}}(R, \Phi)|^2 = \exp\left[-\frac{\Phi}{\Phi_c(R)}\right],} \tag{5.2234}$$

as derived in Eq. (5.1846).

The critical exposure $\Phi_c(R)$ decreases toward high spatial frequency. The simple random-displacement model gives $\Phi_c \propto R^{-2}$ [Eq. (5.1853)], while the experimental dependence measured for cryo-EM specimens is given in Eq. (5.1854). High-resolution information is therefore consumed before low-resolution information.

This is one of the most important results of the chapter:

$$\boxed{\begin{array}{c}\text{radiation tolerance is not described by one exposure limit;}\\ \text{the useful electron budget depends on spatial frequency.}\end{array}} \tag{5.2235}$$

Exposure and exposure rate are different physical variables. The experimentally controlled quantity is normally the electron exposure

$$\Phi = \frac{N_e}{A}, \tag{5.2236}$$

defined in Eq. (5.1678), whereas the exposure rate is

$$\dot{\Phi} = \frac{d\Phi}{dt}. \tag{5.2237}$$

Absorbed dose is an energy-per-mass quantity and is related to exposure through the deposited stopping power,

$$D_{\text{abs}} \simeq \frac{\Phi}{\rho}\frac{d\langle E_{\text{dep}}\rangle}{dz}, \tag{5.2238}$$

derived in Eq. (5.1704).

The distinction between Φ and $\dot{\Phi}$ is physically essential:

$$\boxed{\begin{array}{c}\Phi \quad \text{primarily determines how much of the finite}\\ \text{structural-information budget has been consumed,}\\ \dot{\Phi} \quad \text{can determine how the specimen evolves while}\\ \text{that budget is being consumed.}\end{array}} \tag{5.2239}$$

This result was emphasized in Eq. (5.2174).

Cooling primarily suppresses diffusion, radical mobility and secondary chemical rearrangement. It does not eliminate the primary electronic excitation itself. This is the physical origin of the cryoprotective cage effect.

Other beam-induced processes also modify image formation. Radiolysis normally dominates damage in frozen-hydrated biological specimens, but it is not the only mechanism.

Elastic nuclear scattering can transfer the maximum recoil energy

$$\boxed{T_{\max}(E_{\text{k}}, M) \simeq \frac{2E_{\text{k}}\left(E_{\text{k}} + 2m_0c^2\right)}{Mc^2},} \tag{5.2240}$$

derived in Eq. (5.2019). Knock-on displacement is kinematically possible when

$$T_{\max} > E_d, \tag{5.2241}$$

as stated in Eq. (5.2022). Surface atoms can be removed at lower thresholds by sputtering.

Beam heating is real, but under ordinary low-dose cryo-EM conditions it is normally modest. The radial heat-conduction estimate derived in Eq. (5.2049) shows why total beam current and thermal conductivity matter more directly than current density alone.

Mechanically, the frozen specimen can already contain substantial stress before irradiation. Irradiation can relax this stress by creep and simultaneously generate new stress through chemical and density changes. The resulting motion is collective rather than independent particle diffusion.

For a locally rigid patch, the mechanical contribution to the image-forming wave can be summarized as

$$\boxed{M_{\text{motion}}(\mathbf{R}, \Phi) = \exp\left[-j2\pi\mathbf{R}\cdot\mathbf{u}_{\parallel}(\Phi)\right]\exp\left[-j\pi\lambda R^2\Delta z(\Phi)\right].} \tag{5.2242}$$

This is Eq. (5.2094). The first exponential is produced by in-plane motion. The second shows that axial doming is not merely another translation: it changes the local defocus during the exposure.

Charging introduces yet another exposure-dependent electrostatic potential,

$$\boxed{\phi_q(x, y, \Phi) = \sigma \int V_q(x, y, z, \Phi)\, dz,} \tag{5.2243}$$

as derived in Eq. (5.2120). A constant component is optically irrelevant, a linear phase deflects the beam, and a locally quadratic phase behaves as an additional electrostatic lens. Charging can also exert mechanical forces on the specimen.

Thus radiation chemistry, mechanical motion and charging are not separate from image formation: each can change the wave that reaches the objective lens.

Low-dose acquisition follows directly from this physics. Once the specimen is recognized as an evolving object, the classical low-dose principle becomes unavoidable:

$$\boxed{\text{Do not spend electrons obtaining information that is not part of the final measurement.}} \tag{5.2244}$$

This was introduced in Eq. (5.2153).

Searching and focusing should therefore avoid the region reserved for the high-resolution exposure. Electrons used before acquisition consume part of the same finite structural-information budget as electrons recorded during acquisition.

Low dose does not mean low exposure rate:

$$\boxed{\text{small accumulated } \Phi \neq \text{small } \dot{\Phi}.} \tag{5.2245}$$

The former controls how many electrons have interrogated the specimen; the latter can modify charging, mechanical relaxation, radical chemistry, heating, and other time-dependent processes.

Movies preserve the temporal ordering of an irreversible process. A dose-fractionated acquisition partitions the total exposure,

$$\boxed{\Phi_{\text{tot}} = \sum_{j=1}^{N_f} \Delta\Phi_j,} \tag{5.2246}$$

as introduced in Eq. (5.2177).

The partition does not reduce the number of primary electron–specimen interactions. Its importance is that it records different stages of an evolving specimen separately:

$$\boxed{\begin{array}{c}\text{dose fractionation preserves temporal information;}\\ \text{it does not reverse radiation damage.}\end{array}} \tag{5.2247}$$

Early frames contain the least chemically damaged structure but often the largest motion and charging transients. Later frames are often more mechanically stable but contain progressively less high-resolution structural information. Motion correction and resolution-dependent dose weighting exploit precisely this temporal separation.

Electron-event representation takes this idea further by storing individual detected events,

$$(x_i, y_i, n_i), \tag{5.2248}$$

as described in Eq. (5.2200). Temporal fractionation can then be selected after acquisition, subject to the underlying detector time resolution. EER preserves more information about the detection process; it does not alter the specimen exposure.

The image-formation model at the end of this chapter. The most important conceptual change introduced by this chapter is that the specimen cannot always be regarded as a fixed object followed by a fixed microscope.

For a short movie fraction centered near accumulated exposure $\overline{\Phi}_j$, a useful schematic expression for the coherent zero-loss contribution is

$$\boxed{\begin{aligned}\widehat{y}_j^{(0)}(\mathbf{R}) \simeq\ & \mathrm{CTF}_j(\mathbf{R})\, E_{\mathrm{rad}}\left(R, \overline{\Phi}_j\right) \\ & \times \exp\left[-j2\pi\mathbf{R}\cdot\mathbf{u}_{\parallel,j}\right] \widehat{s}_0(\mathbf{R}) + \widehat{n}_j(\mathbf{R}).\end{aligned}} \tag{5.2249}$$

This equation is deliberately schematic. Its factors summarize several different physical effects:

$$\boxed{\begin{aligned} \text{chemical radiation damage} &\longrightarrow E_{\mathrm{rad}}(R, \Phi), \\ \text{in-plane motion} &\longrightarrow e^{-j2\pi\mathbf{R}\cdot\mathbf{u}_\parallel}, \\ \text{doming} &\longrightarrow \text{time-dependent local defocus}, \\ \text{charging} &\longrightarrow \text{beam deflection and electrostatic lensing}, \\ \text{inelastic scattering} &\longrightarrow \text{additional energy-dependent image channels}. \end{aligned}} \tag{5.2250}$$

The frame-dependent CTF_j in Eq. (5.2249) can therefore differ from frame to frame even if the microscope lenses themselves are perfectly stable: specimen doming and charging can change the effective local electron optics.

Inelastic channels should not, in general, be added coherently to Eq. (5.2249). Their contributions are most naturally combined at the intensity or power-spectrum level, as developed in Eq. (5.1655).

The complete physical picture is therefore

$$\boxed{\begin{array}{c}\text{the electron beam probes the specimen} \\ \textit{and simultaneously changes it.}\end{array}} \tag{5.2251}$$

The exit wave, the specimen structure, the mechanical configuration, and the local electrostatic environment can all evolve during a single exposure. The standard SPA equation (5.2222) remains an extraordinarily useful forward model, but this chapter has established the physical approximations under which it is valid and the principal mechanisms by which a real cryo-EM experiment departs from it.

Chapter 6

Electron detection

6.1 From an arriving electron to a recorded measurement

The preceding chapters have followed the imaging electron through the successive stages of the microscope. The electron source determines the initial distribution of electron energies and illumination directions; the specimen transforms the incident electron wave through elastic and inelastic interactions; and the electron-optical system propagates the resulting wave to the image plane.

At the end of this sequence, however, there is still no digital image. The electron-optical system produces a spatial distribution of *electron probability flux* at the detector plane. The detector must convert the individual electrons arriving according to this distribution into measurable electrical signals and, ultimately, into numerical data.

The last part of the image-formation chain is therefore

$$\boxed{\begin{array}{c}\text{electron wave} \longrightarrow \text{probability flux} \longrightarrow \text{electron arrival} \longrightarrow \\ \text{interaction with detector} \longrightarrow \text{charge carriers} \longrightarrow \text{electronic readout.}\end{array}} \tag{6.1}$$

Several physically different operations are hidden in this apparently simple sequence. The wave-optical calculation determines the *probability distribution of the primary-electron arrivals.* The arrivals themselves are discrete and stochastic. Finally, each arriving electron interacts with the detector material and produces a spatially distributed and fluctuating response from which its arrival position must be inferred.

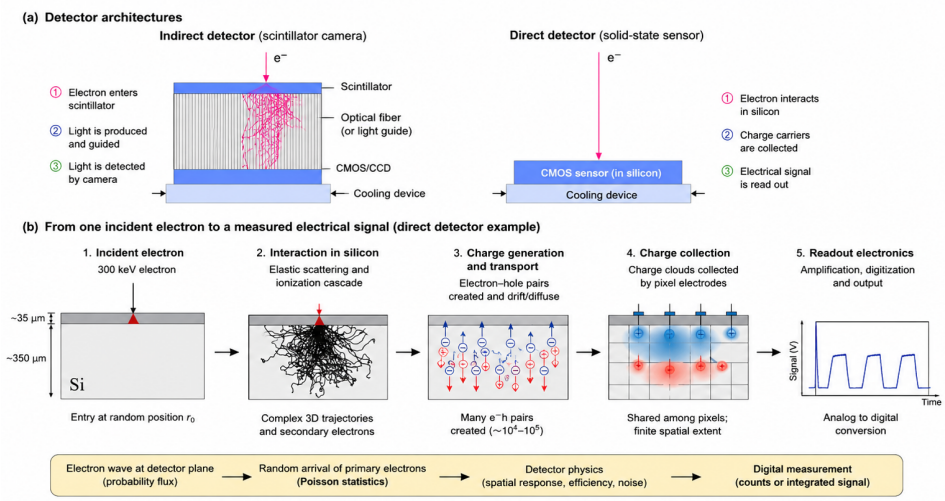


Figure 6.1: **From an incident electron to a recorded measurement.** The electron-optical system determines the probability flux at the detector plane and therefore the spatial distribution from which individual primary-electron arrivals are drawn. These arrivals are discrete stochastic events. In an indirect detector, the incident electron deposits energy in a scintillator, producing photons that are transported through an optical coupling system and detected by a CCD or CMOS sensor. In a direct electron detector, the primary electron interacts directly with the semiconductor sensor, where elastic and inelastic interactions generate a spatially distributed cascade of excitations and electron–hole pairs. The resulting charge cloud is transported and collected by the pixel electrodes and is subsequently amplified, digitized, and converted into recorded data. The lower panels emphasize that the interaction of a single high-energy electron with silicon extends over a finite volume rather than occurring at a mathematical point. Consequently, the position assigned to an event, $\hat{\mathbf{r}}$, is inferred from the measured charge distribution and need not coincide exactly with the true entrance position $\mathbf{r}$. This distinction provides the physical basis for the detector point-spread function, modulation transfer function, noise power spectrum, and detective quantum efficiency developed later in this chapter.

This distinction will be central throughout the chapter:

$$\boxed{\begin{array}{c}\text{electron optics determines where the primary electron arrives,}\\ \text{detector physics determines where we infer that it arrived.}\end{array}} \tag{6.2}$$

Probability flux and arrival statistics The relation between the electron wave and the flux arriving at a detector has already been established in Sec. 3.11 of Chap. 3. For the relativistic electron, the conserved probability-current density is given by Eq. (3.709),

$$\mathbf{J}(\mathbf{r},t) = c, \mathbf{\Psi}^\dagger(\mathbf{r},t)\boldsymbol{\alpha}\mathbf{\Psi}(\mathbf{r},t), \tag{6.3}$$

and satisfies the continuity equation Eq. (3.714).

For the narrow, paraxial, essentially monochromatic beam used for cryo-EM image formation, the longitudinal probability current is proportional to the squared modulus of the scalar spatial wave function. Consequently, as shown in Eq. (3.724),

$$I(\mathbf{r}) \propto |\psi_E(\mathbf{r})|^2. \tag{6.4}$$

We shall not derive this relation again here. The important point for the present chapter is that the wave-optical calculation determines an *expected electron flux*, not a deterministic arrangement of electron impacts.

Let the detector surface be denoted by Σ_d, with unit normal $\mathbf{n}$, and let $\mathbf{r}_\perp = (x, y)$ denote a position on that surface. After normalization to the physical beam current, it is convenient to define the incident electron-rate density

$$\boxed{\Phi_\mathrm{d}(\mathbf{r}_\perp, t) = \mathbf{J}(\mathbf{r}_\perp, t) \cdot \mathbf{n},} \tag{6.5}$$

with units of electrons per unit area and unit time.

For a stationary exposure, Φ_d is independent of time. The expected number of electrons reaching a region $A \subset \Sigma_\mathrm{d}$ during an exposure of duration T is therefore

$$\mu_A = T \int_A \Phi_\mathrm{d}(\mathbf{r}_\perp) dA, \tag{6.6}$$

which is precisely the result derived previously in Eq. (3.727).

The quantity μ_A is an expectation. Even an ideal detector does not record exactly μ_A electrons in every exposure. Individual electrons arrive as discrete events. Under the independent-arrival approximation used in the standard single-particle image-formation model, the number of primary electrons arriving in A is described by a Poisson random variable,

$$\boxed{N_A \sim \mathrm{Poisson}(\mu_A).} \tag{6.7}$$

This is the continuous-space counterpart of the pixelwise model introduced in Eq. (5.604) of Chap. 5.

Consequently,

$$\boxed{\mathbb{E}\{N_A\} = \mathrm{Var}\{N_A\} = \mu_A.} \tag{6.8}$$

This fluctuation is present even if every incident electron is detected perfectly and its position is measured without error. It is therefore important not to identify all fluctuations in a recorded image with detector imperfections. The finite number of primary electrons already introduces an irreducible *electron shot noise*. Its power spectrum will be derived in Sec. 6.8.1.

At this stage we can therefore distinguish two quantities:

$$\boxed{\begin{array}{rcl} \Phi_{\mathrm{d}}(\mathbf{r}_\perp) & : & \text{expected incident electron-rate density,} \\ \mathbf{r}_1, \ldots, \mathbf{r}_N & : & \text{one random realization of primary-electron arrivals.} \end{array}} \tag{6.9}$$

The first is determined by the wave-optical image-formation model. The second is the particular set of electrons that physically encounters the detector during one exposure.

The detector as the final interaction medium The detector surface marks an important physical boundary in the image-formation process. Before reaching it, the electron is propagated as an imaging electron whose wave function contains the structural information introduced by the specimen and transformed by the electron optics. Once the electron enters the detector material, the purpose of the interaction changes completely.

The specimen-interaction problem developed in Chap. 5 was concerned with how an energetic electron interacts with matter while retaining enough information to form an image. Both elastic and inelastic interactions were important, but excessive scattering and energy deposition were undesirable because they degraded the structural signal and damaged the specimen.

Inside the detector, by contrast, interaction with matter is precisely what makes detection possible. The kinetic energy carried by the incoming electron must ultimately be converted into a signal that can be amplified, read electronically, and represented by a number.

The detector can therefore be regarded as the *final interaction medium*

of the imaging electron:

specimen	detector
interaction encodes structure	interaction produces measurement
preserve useful outgoing electrons	convert incoming electron energy into signal.

(6.10)

An incident electron entering a solid detector does not, in general, deposit all of its energy at a single mathematical point. It may undergo elastic deflections and many inelastic interactions. Energy transferred to the electronic system of the detector can generate secondary electrons and, in a semiconductor, ultimately produce a large number of mobile electron–hole pairs. These charge carriers are then transported through the sensor and collected by the readout electronics.

The complete process is therefore more accurately represented as

$$\boxed{\begin{aligned} \text{primary electron} &\longrightarrow \text{trajectory through detector material} \\ &\longrightarrow \text{elastic and inelastic interactions} \\ &\longrightarrow \text{secondary electronic excitations} \\ &\longrightarrow \text{spatial distribution of charge carriers} \\ &\longrightarrow \text{charge collection} \\ &\longrightarrow \text{digital measurement.} \end{aligned}} \tag{6.11}$$

The microscopic physics of these steps will be developed in the following sections. For now, the essential observation is that the position at which a primary electron crosses the detector surface and the position at which its electrical signal is eventually recorded need not coincide exactly.

Let

$$\mathbf{r}_i \tag{6.12}$$

denote the true position at which primary electron i enters the active detector material, and let

$$\widehat{\mathbf{r}}_i \tag{6.13}$$

denote the position assigned to that event by the detector and its readout algorithm. In general,

$$\boxed{\widehat{\mathbf{r}}_i \neq \mathbf{r}_i.} \tag{6.14}$$

The difference is not merely an electronic measurement error. It may originate in the physical trajectory of the primary electron inside the sensor, the spatial extent of the secondary-electron cascade, transport and diffusion of the generated charge, the finite pixel size, electronic noise, and the algorithm used to estimate the event position.

The statistical distribution of $\widehat{\mathbf{r}}_i - \mathbf{r}_i$ will eventually determine the spatial response of the detector and therefore its point-spread function and modulation transfer function.

What information can a detector measure? It is useful to state explicitly what is meant by "detecting an electron". The detector does not directly measure the complex electron wave function

$$\psi(\mathbf{r}), \tag{6.15}$$

nor does it return its phase. The electron-optical system must first convert phase differences in the outgoing wave into variations of the arrival probability, as developed in Chap. 4. The detector then samples those arrival probabilities through discrete physical events.

For an individual incident electron, the quantities relevant at the detector entrance may be represented schematically as

$$\boxed{e_i = (\mathbf{r}_i, t_i, E_i)\,,} \tag{6.16}$$

where $\mathbf{r}_i$ is its arrival position, t_i its arrival time, and E_i its energy when it enters the detector. More detailed descriptions could additionally retain the incidence direction and spin, but these variables are normally not measured by a conventional cryo-EM imaging detector.

The detector converts this incident event into some measured response, which we may write schematically as

$$\boxed{e_i \xrightarrow{\mathcal{D}} y_i,} \tag{6.17}$$

where $\mathcal{D}$ denotes the physical detector and readout process. The quantity y_i need not be a single number. Depending on the detector architecture and acquisition mode, it may consist of a distributed signal over several pixels, an estimated electron position, a frame number or timestamp, and a measure related to the amount of deposited energy.

This immediately suggests three conceptually different measurement questions.

First, *was the electron detected at all?* A real detector may fail to produce a usable event for some incident electrons. Detection efficiency therefore concerns the probability that an incoming primary electron produces a measurable response.

Second, *where was the electron detected?* The signal produced by one primary electron generally extends over a finite region. The precision with

which $\mathbf{r}_i$ can be inferred from this response determines the spatial localization performance of the detector.

Third, *what signal amplitude was produced?* Different primary electrons need not deposit exactly the same energy or generate exactly the same amount of collected charge. An integrating detector retains these amplitude fluctuations, whereas an electron-counting detector attempts instead to identify the occurrence of one primary-electron event and assign it a standardized count.

These three questions already anticipate several detector quantities that will be developed later:

$$\boxed{\begin{array}{rcl} \text{spatial localization} & \longrightarrow & \text{PSF and MTF,} \\ \text{fluctuations of the measured response} & \longrightarrow & \text{noise power spectrum,} \\ \text{signal transfer relative to noise transfer} & \longrightarrow & \text{DQE.} \end{array}} \tag{6.18}$$

It is therefore useful to think of the complete final stage of cryo-EM image formation as two consecutive stochastic mappings:

$$\boxed{\psi_{\mathrm{d}} \longrightarrow \Phi_{\mathrm{d}} \longrightarrow \{(\mathbf{r}_i, t_i, E_i)\}_{i=1}^{N} \xrightarrow{\mathcal{D}} \text{recorded data.}} \tag{6.19}$$

The first mapping, from the detector-plane wave ψ_{d} to the incident flux Φ_{d}, belongs to electron optics and quantum mechanics. The second produces the random set of primary-electron arrivals and is responsible for electron shot noise. The third is the detector itself and introduces the spatial response, efficiency, additional fluctuations, and possible nonlinearities of the measurement process.

Keeping these three levels separate will be essential when we later analyze the power spectrum of a cryo-EM image. In particular, it will allow us to distinguish noise caused by the finite number of imaging electrons from noise and spatial correlations introduced by the detector.

The task of the following sections is therefore to understand the last arrow in Eq. (6.19): how a high-energy electron entering a detector material generates a measurable electrical event, and how accurately that event preserves the position and statistical information carried by the incident electron flux.

6.2 From photographic film to direct electron detection

The problem described in Sec. 6.1 is not specific to modern cryo-EM: an electron reaching the image plane must somehow be converted into a persistent, measurable record. What has changed profoundly over the history of electron microscopy is the number of intermediate physical conversions required to accomplish this task.

Three detector technologies are particularly useful for understanding this evolution: photographic film, scintillator-coupled electronic cameras, and direct electron detectors. Their essential difference can be summarized as

$$\boxed{\begin{aligned} \text{photographic film:} \quad & e^-_{\text{primary}} \longrightarrow \text{energy deposition in emulsion} \longrightarrow \text{latent image,} \\ \text{scintillator camera:} \quad & e^-_{\text{primary}} \longrightarrow \text{scintillator excitation} \longrightarrow \text{photons} \longrightarrow \text{CCD/CMOS signal,} \\ \text{direct detector:} \quad & e^-_{\text{primary}} \longrightarrow \text{electronic excitations in semiconductor} \longrightarrow \text{charge collection.} \end{aligned}} \tag{6.20}$$

The first and third of these processes convert the energy of the imaging electron within the recording medium itself. The important distinction between indirect and direct *electronic* detectors is that the former inserts an additional conversion into visible light before the signal reaches the electronic sensor.

Photographic film. For much of the history of transmission electron microscopy, photographic film was the standard recording medium. The imaging electrons interacted directly with a photographic emulsion containing radiation-sensitive silver-halide grains. Energy deposited by an incident electron produced microscopic changes in some of these grains, creating a *latent image*. At the microscopic level this latent image consists of stable modifications of sensitized silver-halide crystals, including the formation of small metallic-silver clusters. Subsequent chemical development amplifies these microscopic changes, converting the affected grains into metallic silver and producing a macroscopic distribution of optical density.

Schematically,

$$\boxed{\begin{aligned} e^-_{\text{primary}} &\longrightarrow \text{energy deposition} \longrightarrow \text{latent image in silver-halide grains} \\ &\longrightarrow \text{chemical development} \longrightarrow \text{optical density.} \end{aligned}} \tag{6.21}$$

Photographic film had several remarkable properties for electron microscopy. It provided a very large recording area, contained a very large number of small photosensitive grains, and could preserve fine spatial detail. Indeed, film remained for many years an exceptionally good detector for high-resolution, low-dose electron microscopy. Its combination of high spatial fidelity, good detective quantum efficiency at useful spatial frequencies, and large field of view was difficult to match with early electronic cameras (McMullan et al., 2009a).

Film does not possess an intrinsic regular pixel grid. Its effective sampling is imposed later by the scanner used to digitize the developed micrograph. Consequently, a single photographic exposure could contain a considerably larger usable image area than the electronic sensors available during the early development of digital cryo-EM.

Its principal disadvantages were therefore not primarily spatial. Rather, they were practical and quantitative. The film had to be removed, chemically developed, and subsequently digitized. The complete response depended on the emulsion, exposure, development conditions, scanner response, and the conversion between deposited electron energy and final optical density. Immediate inspection of an exposure was impossible, and large-scale automated data collection was cumbersome.

Thus, photographic film presented an interesting combination:

$$\boxed{\begin{array}{c}\text{excellent spatial recording medium}\\ \text{but}\\ \text{slow, indirect, and inconvenient digital workflow.}\end{array}} \tag{6.22}$$

Scintillator–CCD and scintillator–CMOS cameras. The transition to electronic imaging solved many of the practical limitations of photographic film. Conventional CCD and CMOS sensors, however, are designed primarily to detect visible photons rather than high-energy electrons. Electron microscopy cameras therefore commonly placed a *scintillator* in front of the electronic sensor, as illustrated in Fig. 6.1.

An imaging electron deposits energy in the scintillator and creates electronic excitations. Part of this deposited energy is subsequently emitted as visible photons. These photons are transported to a CCD or CMOS sensor, frequently through a fiber-optic plate, taper, or another optical coupling system. The electronic sensor then converts the photons into charge and

reads that charge pixel by pixel:

$$\boxed{\begin{aligned} e^-_{\text{primary}} &\longrightarrow \text{energy deposition in scintillator} \longrightarrow \text{visible photons} \\ &\longrightarrow \text{optical transport} \longrightarrow \text{photoelectrons in CCD/CMOS} \longrightarrow \text{digital signal.} \end{aligned}} \tag{6.23}$$

This architecture represented a major practical advance. The image became available immediately in digital form, enabling rapid inspection, automated microscope operation, and eventually the acquisition of the large datasets required for single-particle cryo-EM.

There was, however, an important price. The first generations of scintillator-coupled electronic cameras did not simply replace film by an equivalent digital recording medium. For high-energy electrons, particularly near the high-spatial-frequency end of the recorded band, their modulation transfer and detective quantum efficiency could be substantially poorer than those of photographic film (McMullan et al., 2009a). Their sensitive areas were also much smaller.

Thus, the historical transition from film to electronic cameras was, in one important sense, initially a trade:

$$\boxed{\begin{array}{c} \text{immediate digital readout and automation} \\ \text{were obtained partly at the expense of} \\ \text{high-frequency spatial fidelity and field of view.} \end{array}} \tag{6.24}$$

This point is important for understanding the later impact of direct detectors. Their importance was not merely that they were electronic. Scintillator–CCD cameras had already provided electronic imaging. Rather, direct detectors made it possible to combine electronic acquisition with spatial and statistical performance approaching, and eventually exceeding, that of film.

The reason for the loss of high-frequency information in an indirect camera is directly related to its physical construction. A primary electron does not create light at a single geometrical point. It travels through the scintillator, undergoes elastic and inelastic interactions, and deposits energy over a finite region. The generated photons then propagate through the scintillator and optical coupling before reaching the electronic sensor. Consequently, an electron entering at position $\mathbf{r}_i$ may generate signal over several detector pixels.

There are therefore at least two successive mechanisms of spatial spread-

ing,

$$\boxed{\text{electron spreading in scintillator} \quad + \quad \text{photon spreading during optical transport.}} \tag{6.25}$$

In addition, every conversion stage is stochastic: the deposited energy fluctuates, the number of generated photons fluctuates, only a fraction of those photons reach the sensor, and their conversion into electronic charge is itself statistical. These processes affect both the spatial resolution and the noise of the recorded image.

In the language developed later in this chapter, the successive spatial spreading mechanisms reduce the modulation transfer function, whereas the additional statistical fluctuations reduce the fraction of the input signal-to-noise ratio retained by the detector.

The scintillator had another important practical advantage. Because the high-energy electron beam was intercepted by the scintillator, the CCD or CMOS imaging sensor itself was not directly exposed to the full electron flux. The scintillator therefore acted not only as an electron-to-photon converter but also as a physical barrier protecting the underlying semiconductor electronics from direct irradiation.

Direct electron detectors. The central idea of a direct electron detector is to eliminate the scintillator and its optical coupling stage. The imaging electron enters a semiconductor sensor directly and deposits energy in the semiconductor itself, as illustrated in Fig. 6.1. The resulting electronic excitations generate mobile charge carriers that are transported and collected by the sensor electronics.

For a silicon detector, the conceptual chain is

$$\boxed{\begin{aligned} e^-_{\text{primary}} &\longrightarrow \text{elastic and inelastic interactions in Si} \\ &\longrightarrow \text{electron–hole pairs} \longrightarrow \text{charge transport} \\ &\longrightarrow \text{charge collection} \longrightarrow \text{electronic readout.} \end{aligned}} \tag{6.26}$$

The word *direct* requires some care. It does not mean that a 300 keV imaging electron simply enters one pixel and is collected as one electron of electrical charge. On the contrary, the incoming electron has an energy enormously larger than the characteristic electronic excitation energies of the semiconductor. It undergoes many elastic and inelastic interactions and creates a cascade of secondary excitations and electron–hole pairs. The detector measures the charge generated by this cascade.

The decisive simplification is instead

$$\boxed{\text{imaging electron} \longrightarrow \text{electronic charge,}} \tag{6.27}$$

without the intermediate sequence

$$\text{imaging electron} \longrightarrow \text{light} \longrightarrow \text{electronic charge.} \tag{6.28}$$

Although this signal chain is conceptually simpler, implementing it was technologically much more difficult. Directly exposing a semiconductor sensor and its associated electronics to a high-energy electron beam produces cumulative radiation damage. Energetic electrons can generate both ionizing damage in insulating structures and displacement damage in the semiconductor lattice, progressively changing the electronic properties of the device.

Practical direct electron cameras therefore required radiation-tolerant sensor architectures and readout electronics. An important development was the use of radiation-hardened monolithic active pixel sensors (MAPS), in which charge generation and pixel electronics could be integrated in a CMOS-compatible architecture specifically designed to survive the accumulated electron dose encountered in electron microscopy (McMullan et al., 2009a, 2014).

The delay between the apparently simple conceptual chain of Eq. (6.27) and its successful implementation in high-energy electron microscopy is therefore not surprising:

$$\boxed{\begin{array}{c}\text{removing the scintillator improves the information transfer,}\\ \text{but exposes the semiconductor directly to the electron beam.}\end{array}} \tag{6.29}$$

Why modern direct detectors are thin. Direct detection does not require the primary electron to stop inside the sensor. In fact, for the 200–300 keV electrons commonly used in cryo-EM, allowing the electron to travel through a thick silicon substrate would be undesirable.

As the primary electron propagates through silicon it undergoes repeated elastic and inelastic scattering. The longer the trajectory inside the sensor, the larger the possible lateral displacement of the primary electron and its secondary-electron cascade. Energy can therefore be deposited far from the original entrance position.

Modern direct electron sensors are consequently made thin, commonly by removing most of the silicon substrate from the back of the device. Back-thinning strongly suppresses the contribution from electrons that would otherwise propagate deep into the substrate and later scatter back toward the sensitive region (McMullan et al., 2009b).

For a 300 keV electron, the range in silicon is of order hundreds of micrometres, whereas back-thinned active sensors used for high-resolution electron detection are typically only of order tens of micrometres thick. Most primary electrons therefore traverse the active sensor and leave it rather than depositing their complete kinetic energy inside it.

This apparently inefficient use of the incident electron energy is actually a crucial imaging advantage. Even a partial energy loss of only order tens of kiloelectronvolts generates thousands of electron–hole pairs, providing a readily measurable electrical signal. At the same time, forcing the electron to leave the sensor after a relatively short path truncates its random walk and substantially reduces the lateral extent of the deposited-energy distribution.

The design principle is therefore not

$$\text{maximize the energy absorbed from each incident electron,} \tag{6.30}$$

but rather

$$\boxed{\begin{array}{c}\text{deposit enough energy for reliable detection}\\ \text{while minimizing the lateral spreading of that deposition.}\end{array}} \tag{6.31}$$

This design choice has a direct frequency-domain consequence: reducing the lateral extent of the single-electron response preserves modulation at higher spatial frequencies and therefore improves the detector MTF. The quantitative relation between interaction volume, point-spread function, and MTF will be developed later in this chapter.

Why direct detection changed cryo-EM. The historical progression can therefore be interpreted as a progressive shortening of the physical chain between the incident imaging electron and the recorded numerical value:

$$\boxed{\begin{array}{c}\text{film}\\ e^- \to \text{emulsion} \to \text{development} \to \text{scanner}\\ \\ \Downarrow\\ \text{scintillator camera}\\ e^- \to \text{scintillator} \to \text{photons} \to \text{CCD/CMOS}\\ \\ \Downarrow\\ \text{direct detector}\\ e^- \to \text{semiconductor charge} \to \text{electronic readout.}\end{array}} \tag{6.32}$$

The progression was therefore not simply

$$\text{film} \longrightarrow \text{better electronic detector.} \tag{6.33}$$

Historically, the first electronic cameras offered enormous practical advantages but initially sacrificed some of the spatial performance and recording area of film. Direct electron detectors removed the major optical conversion responsible for this loss and eventually combined the advantages of both approaches:

$$\boxed{\begin{array}{c}\text{high spatial fidelity} \;+\; \text{high DQE} \;+\; \text{immediate digital readout} \\ +\; \text{high frame rate} \;+\; \text{automated acquisition.}\end{array}} \tag{6.34}$$

Measurements at 300 keV have indeed shown that modern back-thinned direct electron detectors can attain DQE values exceeding those of photographic film (McMullan et al., 2014).

The physical reason is now clear. Every intermediate conversion can broaden the spatial response or introduce additional statistical fluctuations. Direct detection removes the scintillator and optical transport while retaining enough energy deposition in a thin semiconductor layer to identify the incident electron.

It does not, however, make the detector ideal. The primary electron still follows a finite and stochastic trajectory; secondary electrons distribute energy away from the entrance position; the generated charge carriers drift and diffuse; and the resulting signal may extend over several pixels. Thus the distinction introduced in Eq. (6.2) remains: the electron optics determines the true arrival position $\mathbf{r}_i$, whereas the detector must infer that position from a spatially distributed physical response.

Modern direct detectors exploit this response in two different ways. They may *integrate* the charge deposited by all electrons during a given time interval, or, when the incident rate is sufficiently low, identify individual electron events and *count* them. In counting mode, the spatially extended signal created by one incident electron can also be used to estimate its arrival position with subpixel precision.

To understand how accurately this can be done, we must now examine quantitatively what happens when a high-energy electron enters the semiconductor sensor: how much energy it loses, how many charge carriers it generates, how far the primary and secondary electrons travel, and how these processes determine the spatial extent of the single-electron response.

6.3 Interaction of the imaging electron with the detector

Once an imaging electron enters a direct detector, it encounters matter again. Microscopically, many of the processes are the same as those developed for the specimen in Chap. 5: elastic scattering changes the direction of propagation, whereas inelastic scattering transfers energy to the internal degrees of freedom of the material.

The role of these interactions is nevertheless very different. In the specimen, the objective was to preserve as much structural information as possible while depositing as little energy as possible:

$$\boxed{\text{specimen:} \qquad \text{minimal energy deposition while preserving structural information.}} \tag{6.35}$$

The detector instead needs the incident electron to produce a measurable signal. However, this does *not* imply that the complete kinetic energy of the electron should be absorbed. For a modern thin direct electron detector, a better statement is

$$\boxed{\begin{array}{c}\text{detector:} \qquad \text{sufficient energy deposition for reliable detection} \\ \text{while minimizing the spatial spreading of that deposition.}\end{array}} \tag{6.36}$$

This distinction is important. A thick semiconductor can absorb a larger fraction of the incident electron energy, but the electron then travels farther inside the detector, undergoes more elastic scattering, and may deposit energy tens or hundreds of micrometres away from its entrance position. A thin semiconductor deposits less energy but can preserve the arrival position much more accurately.

Figure 6.1 illustrates precisely this competition. The imaging electron enters at a well-defined position $\mathbf{r}_0$, but the energy subsequently deposited in the semiconductor occupies a finite three-dimensional interaction volume.

The inelastic-scattering theory developed in Sec. 5.12 already provides the microscopic description of electron energy loss. In particular, Eq. (5.1238) showed that the energy lost by the imaging electron appears as internal excitation of the material and, when relevant, recoil.

For a macroscopic solid such as silicon, it is convenient to compress the large number of possible microscopic channels into the *stopping power*,

$$\boxed{S(E) \equiv -\frac{dE}{ds},} \tag{6.37}$$

where s is distance measured along the actual electron trajectory. The SI unit of S is $\mathrm{J\,m^{-1}}$, although $\mathrm{keV}\,\mu\mathrm{m}^{-1}$ is more convenient at the dimensions relevant to electron detectors.

The stopping power introduced in Eq. (6.37) is conventionally separated into two contributions,

$$\boxed{S(E) = S_{\mathrm{col}}(E) + S_{\mathrm{rad}}(E).} \tag{6.38}$$

They correspond to two physically different ways in which kinetic energy can leave the primary electron. In the *collisional* contribution, energy is transferred to the electronic degrees of freedom of the material, producing excitation and ionization. In the *radiative* contribution, energy is emitted as electromagnetic radiation, predominantly bremsstrahlung. We consider these two mechanisms separately.

6.3.1 Collisional stopping power.

The dominant energy-loss mechanism for the electron energies relevant to cryo-EM is the Coulomb interaction of the incident electron with the electrons of the detector material. These interactions are *inelastic*: part of the kinetic energy of the incident electron is transferred to the electronic system of the material.

Two limiting outcomes are useful for developing physical intuition. An atomic electron may be promoted to a higher-energy bound state,

$$\boxed{e^-_{\mathrm{primary}} + \mathrm{Si} \longrightarrow {e^-_{\mathrm{primary}}}' + \mathrm{Si}^*,} \tag{6.39}$$

or sufficient energy may be transferred to remove an electron from its initial state,

$$\boxed{e^-_{\mathrm{primary}} + \mathrm{Si} \longrightarrow {e^-_{\mathrm{primary}}}' + e^-_{\mathrm{secondary}} + \mathrm{Si}^+.} \tag{6.40}$$

The first process is electronic excitation and the second is ionization. In a solid such as silicon, the corresponding states are more appropriately described in terms of electronic bands rather than isolated atomic levels, but the essential energy-transfer mechanism is the same.

If an individual inelastic interaction transfers an energy T from the primary electron to the material, then

$$E_{\mathrm{in}} = E_{\mathrm{out}} + T. \tag{6.41}$$

For excitation, T is stored initially as electronic excitation energy. For ionization, it includes the energy required to liberate the secondary electron together with the kinetic energy carried by that electron.

The collisional stopping power is obtained by averaging these energy transfers over all possible inelastic interactions. If $d\sigma_{\text{inel}}/dW$ denotes the differential cross section for transferring an energy W, then its physical structure can be written schematically as

$$\boxed{S_{\text{col}}(E) = n_{\text{Si}} \int W \, \frac{d\sigma_{\text{inel}}(E, W)}{dW} \, dW,} \tag{6.42}$$

where n_{Si} is the number of silicon atoms per unit volume and the cross section is understood to include all allowed electronic excitation and ionization channels.

Equation (6.42) is important conceptually: the stopping power is not the energy transferred in one collision. It is the *mean energy transferred per unit trajectory length* after averaging over the complete distribution of possible inelastic events. Individual electrons can therefore lose substantially more or less energy than predicted by this average, a point that will become important when we discuss energy straggling below.

The word *collisional* should also not be confused with elastic scattering by atomic nuclei. Elastic electron–nucleus interactions can strongly change the direction of the primary electron while transferring comparatively little kinetic energy to the much heavier nucleus. They are therefore crucial for the spatial trajectory of the electron but contribute little to its stopping power at cryo-EM energies. Their effect on trajectory spreading will be considered separately in Sec. 6.4.1.

At sufficiently high energies, Eq. (6.42) can be evaluated within the quantum-mechanical theory developed by Bethe. Rather than requiring the detailed excitation spectrum of every atom explicitly, the result summarizes much of the electronic structure of the material through a single characteristic quantity, the *mean excitation energy* I.

The relativistic quantities required below were introduced in Chap. 2, Sec. 2.3. In particular, Eq. (2.47) defines

$$\beta = \frac{v}{c}, \qquad \gamma = \frac{1}{\sqrt{1-\beta^2}}, \tag{6.43}$$

where v is the electron speed and γ is the Lorentz factor. The total and kinetic relativistic energies were given in Eqs. (2.51), (2.52), and Eq. (2.96)

$$E_{\text{tot}} = \gamma m_0 c^2, \qquad K = (\gamma - 1) m_0 c^2 = eV. \tag{6.44}$$

For the stopping-power expressions, it is convenient to introduce the

dimensionless kinetic energy

$$\boxed{\tau = \frac{K}{m_0 c^2}.} \tag{6.45}$$

Using Eq. (2.52),

$$\tau = \gamma - 1, \tag{6.46}$$

and therefore

$$\boxed{\gamma = 1 + \tau.} \tag{6.47}$$

Combining this relation with Eq. (2.47) gives

$$\boxed{\beta^2 = 1 - \frac{1}{(1+\tau)^2}.} \tag{6.48}$$

For an elemental material, the relativistic Bethe expression for the *mass collisional stopping power* of an electron can be written as

$$\boxed{\begin{aligned}\frac{S_{\text{col}}}{\rho} &= \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M} \\ &\quad \times \left\{ \ln\left[\frac{\tau^2(\tau+2)}{2\left(I/m_0c^2\right)^2}\right] + F^-(\tau) - \delta \right\}.\end{aligned}} \tag{6.49}$$

Here

$$\boxed{F^-(\tau) = 1 - \beta^2 + \frac{\tau^2/8 - (2\tau+1)\ln 2}{(\tau+1)^2},} \tag{6.50}$$

is the electron-specific kinematic and exchange correction (Berger et al., 2017).

In Eq. (6.49), r_e is the classical electron radius,

$$r_e = \frac{e^2}{4\pi\varepsilon_0 m_0 c^2}, \tag{6.51}$$

N_A is Avogadro's constant, Z is the atomic number of the stopping material, M is its molar mass, and ρ is its mass density. The quantity I is the mean excitation energy of the material, and δ is the density-effect correction.

For silicon,

$$\boxed{Z_{\text{Si}} = 14, \qquad M_{\text{Si}} \simeq 28.0855\ \text{g}\,\text{mol}^{-1}, \qquad I_{\text{Si}} \simeq 173\ \text{eV}.} \tag{6.52}$$

The value of I_{Si} used by ESTAR is 173 eV (Berger et al., 2017).

The structure of Eq. (6.49) is physically informative. The factor

$$\frac{Z}{M} \tag{6.53}$$

is proportional to the number of target electrons available per unit mass of material. The factor $1/\beta^2$ expresses the reduction of the Coulomb interaction time as the incident electron becomes faster. The logarithmic term results from integrating over the broad range of possible energy and momentum transfers in the inelastic collisions.

The mean excitation energy I sets the characteristic energy scale of the electronic response of the material. It should not be confused with the first ionization energy of an isolated silicon atom. Rather, it is an effective logarithmic average over the many excitation and ionization channels through which the fast electron can transfer energy to the material. For a condensed solid it therefore contains information about the collective electronic response as well as about individual atomic excitations.

The term $F^-(\tau)$ is specific to an incident electron. The projectile and the target electrons are identical fermions, and the kinematics of an electron–electron collision therefore differs from that of a heavy charged projectile colliding with an electron. The superscript "$-$" emphasizes that this is the correction appropriate to an incident electron rather than a positron.

Finally, the density-effect correction δ accounts for the polarization of the surrounding medium. At sufficiently high velocity, distant atoms do not respond independently to the rapidly varying electric field of the projectile. Collective dielectric polarization screens part of the long-range interaction and therefore reduces the collisional stopping power. This reduction appears through the subtraction of δ in Eq. (6.49).

Thus Bethe theory converts the microscopic picture

$$\boxed{\begin{array}{c}\text{many possible electronic excitations and ionizations}\\ \Downarrow \\ \text{distribution of transferred energies } W \\ \Downarrow \\ \text{mean collisional energy loss per unit path length}\end{array}} \tag{6.54}$$

into a quantitative stopping power that depends on the electron energy and on a small number of material parameters.

Collisional stopping power of a fast electron.

We now derive the electron collisional stopping-power expression in Eq. (6.49). The derivation is slightly more subtle than for a heavy charged projectile because the incident electron is indistinguishable from the electrons with which it collides.

The concept of scattering cross section was introduced in the quantum-mechanics appendix and developed for inelastic electron scattering in Sec. 5.12.6 of Chap. 5. In particular, Eq. (5.1359) introduced the double differential cross section

$$\frac{d^2\sigma_{\text{inel}}}{d\Omega\, dW}, \tag{6.55}$$

which resolves an inelastic event both by scattering direction and by the energy W transferred from the incident electron to the material.

For the stopping power we are not interested in the direction in which the electron is scattered, but only in how much energy it loses. Integrating over all outgoing directions therefore defines the energy-differential inelastic cross section

$$\boxed{\frac{d\sigma_{\text{inel}}}{dW} = \int \frac{d^2\sigma_{\text{inel}}}{d\Omega\, dW}\, d\Omega.} \tag{6.56}$$

Here the cross section is taken per target atom.

Let n_{at} denote the atomic number density of the material, that is, the number of target atoms per unit volume. If the incident electron travels an infinitesimal path length ds, the expected number of inelastic interactions transferring an energy between W and $W + dW$ is

$$d\langle N\rangle = n_{\text{at}}\, ds\, \frac{d\sigma_{\text{inel}}}{dW}\, dW. \tag{6.57}$$

Each such interaction removes an energy W from the primary electron. The corresponding mean energy loss is therefore

$$-\, dK = \int W\, d\langle N\rangle, \tag{6.58}$$

where K is the kinetic energy of the incident electron. Substituting Eq. (6.57) gives

$$-\,dK = n_{\mathrm{at}}\,ds \int W \frac{d\sigma_{\mathrm{inel}}}{dW}\,dW. \tag{6.59}$$

Dividing by ds, and using the definition of stopping power in Eq. (6.37), we obtain

$$\boxed{S_{\mathrm{col}}(K) = n_{\mathrm{at}} \int W \frac{d\sigma_{\mathrm{inel}}(K,W)}{dW}\,dW.} \tag{6.60}$$

Equation (6.60) shows that the collisional stopping power is the first moment of the distribution of inelastic energy transfers: each possible transfer W is weighted by the cross section with which it occurs.

For an elemental material of mass density ρ_{m} and molar mass M, the atomic number density is

$$\boxed{n_{\mathrm{at}} = \frac{\rho_{\mathrm{m}} N_A}{M},} \tag{6.61}$$

where N_A is Avogadro's constant. Hence the corresponding mass collisional stopping power is

$$\boxed{\frac{S_{\mathrm{col}}}{\rho_{\mathrm{m}}} = \frac{N_A}{M} \int W \frac{d\sigma_{\mathrm{inel}}}{dW}\,dW.} \tag{6.62}$$

Separation into soft and hard collisions. The energy transfers cover a very broad range. For small W, the binding and excitation spectrum of the target electrons is essential. For sufficiently large W, however, the transferred energy is much larger than the binding energy, and the target electron may be treated approximately as a free electron initially at rest.

We therefore introduce an intermediate energy W_c and write

$$S_{\mathrm{col}} = S_{\mathrm{soft}} + S_{\mathrm{hard}}, \tag{6.63}$$

where

$$W < W_c \tag{6.64}$$

defines the soft-collision region and

$$W > W_c \tag{6.65}$$

defines the hard-collision region.

The value of W_c is only a mathematical device. It is chosen large compared with characteristic binding energies but small compared with the kinetic energy of the incident electron,

$$\text{binding energies} \ll W_c \ll K. \tag{6.66}$$

Since the physical stopping power cannot depend on this arbitrary choice, W_c must disappear when the two contributions are added.

Soft collisions: the Bethe contribution. We begin with the small-energy-transfer part of the collisional stopping power. The quantum-mechanical description of such an event is not new. In Chap. 5 we developed inelastic electron scattering in the first Born approximation, Sec. 5.12.4. There, the incident electron induces a transition of the target from an initial state to a different internal state, and the corresponding scattering amplitude is given by Eq. (5.1282). For a Coulomb interaction, the internal structure of the transition enters through the transition form factor, Eq. (5.1332).

Bethe's calculation uses this same microscopic description for the purpose of computing an average energy loss. The new step is therefore not the scattering theory itself, but the summation over all possible electronic transitions and over the allowed momentum transfers.

We separate the collisions according to the energy W transferred from the primary electron to the material. The soft region is defined by

$$W < W_c, \tag{6.67}$$

where W_c is an intermediate energy chosen such that electronic binding cannot be neglected below it, while the target electrons can be treated as approximately free above it. The value of W_c is artificial and must disappear when the soft and hard contributions are eventually combined.

Consider one possible transition from the initial electronic state $|0\rangle$ to a final state $|j\rangle$. Its excitation energy is

$$\boxed{W_j = E_j - E_0.} \tag{6.68}$$

This is the same physical energy transfer that was denoted by ϵ_{mn} in Sec. 5.12.3; the simpler index j is convenient here because the initial state is kept fixed while we sum over all possible final states.

The corresponding transition strength is contained in the transition form factor introduced in Eq. (5.1316). Its dimensionless version, Eq. (5.1362), will be denoted here by

$$\mathcal{F}_{j0}(\mathbf{q}), \tag{6.69}$$

where we have expressed it using the conventional momentum-transfer wavevector $\mathbf{q}$. As established in Eq. (5.1378), this variable is related to the spatial-frequency transfer used elsewhere in this book by

$$\boxed{\mathbf{q} = -2\pi\mathbf{R}, \qquad q = 2\pi R.} \tag{6.70}$$

Bethe packages the strength of transition j at a given momentum transfer into the *generalized oscillator strength*, defined as

$$\boxed{f_j(\mathbf{q}) \equiv \frac{2m_0 W_j}{\hbar^2 q^2} \left|\mathcal{F}_{j0}(\mathbf{q})\right|^2 .} \tag{6.71}$$

The quantity $f_j(\mathbf{q})$ is dimensionless. The normalization in Eq. (6.71) is chosen so that its small-momentum-transfer limit reduces to the familiar optical oscillator strength.

To evaluate this limit, recall the dipole expansion already derived in Sec. 5.12.8. Equation (5.1411) gives

$$F_{j0}(\mathbf{R}) = -j2\pi\, \mathbf{R}\cdot\mathbf{d}_{j0} + O(R^2), \tag{6.72}$$

where $\mathbf{d}_{j0}$ is the electric transition dipole moment defined in Eq. (5.1409).

Using Eq. (6.70),

$$-2\pi\mathbf{R} = \mathbf{q}, \tag{6.73}$$

and therefore

$$F_{j0}(\mathbf{q}) = j\,\mathbf{q}\cdot\mathbf{d}_{j0} + O(q^2). \tag{6.74}$$

Since $\mathcal{F}_{j0} = F_{j0}/e$,

$$\boxed{\mathcal{F}_{j0}(\mathbf{q}) = \frac{j}{e}\mathbf{q}\cdot\mathbf{d}_{j0} + O(q^2).} \tag{6.75}$$

Substitution into Eq. (6.71) gives

$$f_j(\mathbf{q}) = \frac{2m_0 W_j}{\hbar^2 q^2} \frac{|\mathbf{q}\cdot\mathbf{d}_{j0}|^2}{e^2} + O(q) \tag{6.76}$$

$$= \frac{2m_0 W_j}{\hbar^2 e^2} |\widehat{\mathbf{q}}\cdot\mathbf{d}_{j0}|^2 + O(q), \tag{6.77}$$

where

$$\widehat{\mathbf{q}} = \frac{\mathbf{q}}{q} \tag{6.78}$$

is the unit vector in the direction of the momentum transfer.

For a fixed orientation, the result still depends on the angle between $\widehat{\mathbf{q}}$ and the transition dipole. For an isotropic material all orientations are equally probable. Choose the polar axis along $\mathbf{d}_{j0}$ and let θ be the angle between $\widehat{\mathbf{q}}$ and $\mathbf{d}_{j0}$. Then

$$|\widehat{\mathbf{q}}\cdot\mathbf{d}_{j0}|^2 = |\mathbf{d}_{j0}|^2 \cos^2\theta. \tag{6.79}$$

The uniform orientational average over the unit sphere is therefore

$$\left\langle |\widehat{\mathbf{q}}\cdot\mathbf{d}_{j0}|^2 \right\rangle_\Omega = \frac{|\mathbf{d}_{j0}|^2}{4\pi} \int_0^{2\pi}\int_0^{\pi} \cos^2\theta\, \sin\theta\, d\theta\, d\phi. \tag{6.80}$$

The azimuthal integral gives

$$\int_0^{2\pi} d\phi = 2\pi, \tag{6.81}$$

while, with $u = \cos\theta$,

$$\int_0^{\pi} \cos^2\theta \sin\theta\, d\theta = \int_{-1}^{1} u^2\, du \tag{6.82}$$

$$= \frac{2}{3}. \tag{6.83}$$

Consequently,

$$\boxed{\left\langle |\widehat{\mathbf{q}}\cdot\mathbf{d}_{j0}|^2 \right\rangle_\Omega = \frac{1}{3}|\mathbf{d}_{j0}|^2.} \tag{6.84}$$

Taking the small-q limit of Eq. (6.77) and averaging over orientations therefore gives

$$\boxed{f_j \equiv \lim_{q\to 0} \langle f_j(\mathbf{q})\rangle_\Omega = \frac{2m_0 W_j}{3\hbar^2 e^2}|\mathbf{d}_{j0}|^2.} \tag{6.85}$$

This is the ordinary, or optical, oscillator strength of transition j.

The factor e^2 appears because $\mathbf{d}_{j0}$ is the *electric* transition dipole moment defined in Eq. (5.1409); it already contains the electron charge. Thus

$$[\mathbf{d}_{j0}] = \mathrm{C\,m}, \tag{6.86}$$

and

$$\left[\frac{m_0 W_j}{\hbar^2 e^2}|\mathbf{d}_{j0}|^2\right] = 1, \tag{6.87}$$

confirming that f_j is dimensionless.

We can now return from the strength of one transition to the energy lost by the incident electron. The starting point is the inelastic first-Born cross section already derived in Chap. 5. In the conventional EELS notation, Eq. (5.1380) gives

$$\frac{d^2\sigma_{\mathrm{in}}}{d\Omega\, dW} = \frac{4\gamma^2}{a_0^2 q^4}\frac{k_{\mathrm{f}}}{k_{\mathrm{i}}} S(\mathbf{q}, W). \tag{6.88}$$

For a particular transition $|0\rangle \to |j\rangle$ of energy W_j, the corresponding contribution to the dynamic structure factor is, from Eq. (5.1364),

$$S_j(\mathbf{q}, W) = |\mathcal{F}_{j0}(\mathbf{q})|^2\, \delta(W - W_j). \tag{6.89}$$

Integration over W therefore gives

$$\frac{d\sigma_j}{d\Omega} = \frac{4\gamma^2}{a_0^2 q^4}\frac{k_{\mathrm{f}}}{k_{\mathrm{i}}}\, |\mathcal{F}_{j0}(\mathbf{q})|^2\,. \tag{6.90}$$

Using the definition of the generalized oscillator strength, Eq. (6.71),

$$|\mathcal{F}_{j0}(\mathbf{q})|^2 = \frac{\hbar^2 q^2}{2m_0 W_j} f_j(\mathbf{q}), \tag{6.91}$$

and hence

$$\boxed{W_j \frac{d\sigma_j}{d\Omega} = \frac{2\gamma^2\hbar^2}{m_0 a_0^2}\frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{f_j(\mathbf{q})}{q^2}.} \tag{6.92}$$

This equation already displays the important simplification introduced by the oscillator strength. The original Coulomb cross section contains q^{-4}, but the transition form factor contributes two powers of q in the dipole regime, leaving an energy-weighted cross section proportional to q^{-2}.

Integration over momentum transfer. To obtain the energy loss produced by transition j, we must integrate Eq. (6.92) over the allowed scattering directions. The magnitude of the momentum transfer is determined by the geometry of the incident and outgoing wavevectors. Since

$$\mathbf{q} = \mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}, \tag{6.93}$$

its squared magnitude is

$$q^2 = |\mathbf{k}_\mathrm{f} - \mathbf{k}_\mathrm{i}|^2 \tag{6.94}$$

$$= k_\mathrm{f}^2 + k_\mathrm{i}^2 - 2\mathbf{k}_\mathrm{i}\cdot\mathbf{k}_\mathrm{f}. \tag{6.95}$$

If θ is the angle between $\mathbf{k}_\mathrm{i}$ and $\mathbf{k}_\mathrm{f}$,

$$\mathbf{k}_\mathrm{i}\cdot\mathbf{k}_\mathrm{f} = k_\mathrm{i}k_\mathrm{f}\cos\theta, \tag{6.96}$$

so that

$$\boxed{q^2 = k_\mathrm{i}^2 + k_\mathrm{f}^2 - 2k_\mathrm{i}k_\mathrm{f}\cos\theta.} \tag{6.97}$$

For a particular transition j, the energy loss W_j is fixed. Consequently, the incident and outgoing energies are fixed, and so are the magnitudes k_i and k_f. When we vary the scattering direction, the only variable on the right-hand side of Eq. (6.97) is therefore θ, while q varies as a function of θ.

Taking the differential of both sides gives

$$d(q^2) = d\Big(k_\mathrm{i}^2 + k_\mathrm{f}^2 - 2k_\mathrm{i}k_\mathrm{f}\cos\theta\Big). \tag{6.98}$$

Since k_i and k_f are constant for the fixed transition,

$$d(k_\mathrm{i}^2) = d(k_\mathrm{f}^2) = 0, \tag{6.99}$$

and therefore

$$2q\,dq = -2k_\mathrm{i}k_\mathrm{f}\,d(\cos\theta) \tag{6.100}$$
$$= -2k_\mathrm{i}k_\mathrm{f}\,(-\sin\theta\,d\theta) \tag{6.101}$$
$$= 2k_\mathrm{i}k_\mathrm{f}\sin\theta\,d\theta. \tag{6.102}$$

Dividing by 2 gives

$$\boxed{q\,dq = k_\mathrm{i}k_\mathrm{f}\sin\theta\,d\theta.} \tag{6.103}$$

Because the scattering problem is azimuthally symmetric around the incident direction, the element of solid angle is

$$d\Omega = \sin\theta\,d\theta\,d\phi. \tag{6.104}$$

Integration over the azimuthal angle, $0 \le \phi < 2\pi$, gives

$$d\Omega = 2\pi\sin\theta\,d\theta. \tag{6.105}$$

Using Eq. (6.103),

$$\sin\theta\,d\theta = \frac{q\,dq}{k_\mathrm{i}k_\mathrm{f}}, \tag{6.106}$$

and hence

$$\boxed{d\Omega = \frac{2\pi q\,dq}{k_\mathrm{i}k_\mathrm{f}}.} \tag{6.107}$$

Multiplying Eq. (6.92) by Eq. (6.107) gives

$$W_j\,d\sigma_j = \frac{4\pi\gamma^2\hbar^2}{m_0a_0^2}\frac{1}{k_\mathrm{i}^2}f_j(\mathbf{q})\frac{dq}{q}. \tag{6.108}$$

Notice that k_f cancels exactly.

The incident momentum is

$$p = \hbar k_\mathrm{i} = \gamma m_0 v, \tag{6.109}$$

as derived in Chap. 2. Therefore,

$$k_\mathrm{i}^2 = \frac{\gamma^2 m_0^2 v^2}{\hbar^2}. \tag{6.110}$$

Using also

$$a_0 = \frac{4\pi\varepsilon_0\hbar^2}{m_0 e^2}, \tag{6.111}$$

and the classical electron radius

$$r_e = \frac{e^2}{4\pi\varepsilon_0 m_0 c^2}, \tag{6.112}$$

the prefactor reduces to

$$\frac{4\pi\gamma^2\hbar^2}{m_0 a_0^2 k_{\mathrm{i}}^2} = \frac{4\pi r_e^2 m_0 c^2}{\beta^2}, \tag{6.113}$$

where $\beta = v/c$. Thus

$$\boxed{W_j\, d\sigma_j = \frac{4\pi r_e^2 m_0 c^2}{\beta^2} f_j(\mathbf{q}) \frac{dq}{q}.} \tag{6.114}$$

Equation (6.114) explains the logarithm characteristic of Bethe stopping theory: in the soft, dipole-dominated region,

$$f_j(\mathbf{q}) \simeq f_j, \tag{6.115}$$

and therefore the momentum-transfer dependence is simply

$$\frac{dq}{q}. \tag{6.116}$$

Its integral is logarithmic.

Limits of the momentum-transfer integral. The lower limit is fixed by energy conservation. As shown in Eq. (5.1438), an energy loss W_j requires, even at zero scattering angle, a longitudinal momentum change

$$\Delta p \simeq \frac{W_j}{v}. \tag{6.117}$$

Since $\hbar q$ is the momentum transfer,

$$\boxed{q_{\mathrm{min},j} \simeq \frac{W_j}{\hbar v}.} \tag{6.118}$$

This is the same physical lower cutoff that regularized the apparent forward singularity of the inelastic cross section in Sec. 5.12.8.

At the opposite end, introduce an intermediate momentum-transfer scale q_c separating the region in which binding must be retained from the region in which the struck electron can be treated approximately as free. It is convenient to express this scale as an energy,

$$\boxed{W_c \equiv \frac{\hbar^2 q_c^2}{2m_0}.} \tag{6.119}$$

The precise value of W_c is arbitrary; it will disappear when the soft and hard contributions are added.

Integrating Eq. (6.114) from $q_{\min,j}$ to q_c gives, for the Coulomb part of a soft transition,

$$W_j \sigma_{j,\text{soft}}^{(\text{C})} = \frac{4\pi r_e^2 m_0 c^2}{\beta^2} f_j \int_{q_{\min,j}}^{q_c} \frac{dq}{q} \tag{6.120}$$

$$= \frac{2\pi r_e^2 m_0 c^2}{\beta^2} f_j \ln\left(\frac{q_c^2}{q_{\min,j}^2}\right). \tag{6.121}$$

Using Eqs. (6.118) and (6.119),

$$\frac{q_c^2}{q_{\min,j}^2} = \frac{2m_0 W_c/\hbar^2}{W_j^2/(\hbar^2 v^2)} \tag{6.122}$$

$$= \frac{2m_0 v^2 W_c}{W_j^2} \tag{6.123}$$

$$= \frac{2m_0 c^2 \beta^2 W_c}{W_j^2}. \tag{6.124}$$

Consequently,

$$\boxed{W_j \sigma_{j,\text{soft}}^{(\text{C})} = \frac{2\pi r_e^2 m_0 c^2}{\beta^2} f_j \ln\left[\frac{2m_0 c^2 \beta^2 W_c}{W_j^2}\right].} \tag{6.125}$$

Equation (6.125) is not yet a stopping power. For one target atom, $\sigma_{j,\text{soft}}^{(\text{C})}$ is the cross section for producing transition j, and each such event

removes an energy W_j from the primary electron. Therefore,

$$W_j \sigma_{j,\mathrm{soft}}^{(\mathrm{C})} \tag{6.126}$$

has units of energy times area.

To obtain the energy lost per unit trajectory length in the material, we must multiply by the atomic number density n_{at}. Indeed, for a path length ds, the expected number of transitions of type j is

$$d\langle N_j \rangle = n_{\mathrm{at}} \sigma_{j,\mathrm{soft}}^{(\mathrm{C})} ds. \tag{6.127}$$

The corresponding mean energy loss is therefore

$$-\, dK_j = W_j \, d\langle N_j \rangle = n_{\mathrm{at}} W_j \sigma_{j,\mathrm{soft}}^{(\mathrm{C})} ds. \tag{6.128}$$

Hence the stopping power associated with this single transition is

$$\boxed{S_{j,\mathrm{soft}}^{(C)} = n_{\mathrm{at}} W_j \sigma_{j,\mathrm{soft}}^{(\mathrm{C})}.} \tag{6.129}$$

The total soft collisional stopping power is obtained by summing over all possible electronic excitations and ionization channels:

$$\boxed{S_{\mathrm{soft}}^{(C)} = n_{\mathrm{at}} \sum_j W_j \sigma_{j,\mathrm{soft}}^{(\mathrm{C})}.} \tag{6.130}$$

Using the atomic number density from Eq. (6.61) and the single-transition result Eq. (6.125), the Coulomb part of the soft mass stopping power is

$$\begin{aligned} \frac{S_{\mathrm{soft}}^{(\mathrm{C})}}{\rho_{\mathrm{m}}} &= \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2 M} \\ &\quad \times \sum_j f_j \ln\left[\frac{2 m_0 c^2 \beta^2 W_c}{W_j^2}\right]. \end{aligned} \tag{6.131}$$

The apparent dependence on every individual transition energy W_j can be simplified considerably. The optical oscillator strengths defined in Eq. (6.85) satisfy the Thomas–Reiche–Kuhn sum rule,

$$\boxed{\sum_j f_j = Z,} \tag{6.132}$$

where the sum includes both discrete excitations and continuum ionization states, and Z is the number of electrons per neutral atom.

This sum rule motivates the definition of the *mean excitation energy* I by the oscillator-strength-weighted logarithmic mean

$$\boxed{\ln I \equiv \frac{1}{Z}\sum_j f_j \ln W_j.} \tag{6.133}$$

The logarithmic average appears naturally here because the individual transition energies enter Eq. (6.131) through logarithms.

To see the simplification explicitly, write

$$\begin{aligned}\sum_j f_j \ln\left[\frac{2m_0c^2\beta^2W_c}{W_j^2}\right] \\ = \sum_j f_j \ln\Big(2m_0c^2\beta^2W_c\Big) - 2\sum_j f_j \ln W_j.\end{aligned} \tag{6.134}$$

Using Eqs. (6.132) and (6.133),

$$\begin{aligned}\sum_j f_j \ln\left[\frac{2m_0c^2\beta^2W_c}{W_j^2}\right] \\ &= Z\ln\Big(2m_0c^2\beta^2W_c\Big) - 2Z\ln I \qquad (6.135)\\ &= \boxed{Z\ln\left[\frac{2m_0c^2\beta^2W_c}{I^2}\right].} \qquad (6.136)\end{aligned}$$

Equation (6.136) is the result of the longitudinal Coulomb calculation carried out in Eqs. (6.114)– (6.125). For a relativistic incident electron, however, the complete electromagnetic interaction also contains a transverse contribution. The relativistic completion of the soft term adds

$$\ln\gamma^2 - \beta^2 \tag{6.137}$$

per target electron. In a condensed medium, the long-range part of the interaction is additionally reduced by collective polarization; this is represented by the density-effect correction $-\delta$.

Including these two effects changes Eq. (6.136) into

$$\boxed{\begin{aligned}\frac{S_{\text{soft}}}{\rho_{\text{m}}} &= \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M} \\ &\quad\times\left[\ln\left(\frac{2m_0c^2\beta^2\gamma^2 W_c}{I^2}\right) - \beta^2 - \delta\right].\end{aligned}} \tag{6.138}$$

Hard collisions: the Møller contribution. Equation (6.138) describes the soft-collision contribution. In this region, the binding of the target electrons is retained explicitly through the transition energies W_j and the corresponding transition cross sections $\sigma_{j,\text{soft}}^{(\text{C})}$.

We now consider the complementary hard-collision region,

$$W \geq W_c, \tag{6.139}$$

where the transferred energy is sufficiently large compared with the characteristic binding energies that the struck electron can be approximated as free and initially at rest. The problem then reduces to relativistic scattering between two electrons.

Because the incident particle and the struck particle are identical fermions, the appropriate elementary process is Møller scattering (Møller, 1932). Its relativistic derivation is given in Sec. A.14 of the quantum-mechanics appendix. In particular, Sec. A.14.4 derives the two indistinguishable scattering amplitudes, Sec. A.14.5 evaluates their unpolarized spin average, Sec. A.14.6 converts the invariant amplitude into a differential cross section, and Sec. A.14.7 specializes the result to the laboratory frame in which the target electron is initially at rest. The final conversion to SI units is carried out in Sec. A.14.8.

For the detector calculation, we therefore take the resulting energy-transfer cross section directly from Eq. (A.1075). If the incident electron has kinetic energy K, and W denotes the kinetic energy carried by the less energetic outgoing electron, the differential cross section per target electron is

$$\boxed{\begin{aligned}\frac{d\sigma_{\text{M}}}{dW} &= \frac{2\pi r_e^2 m_0 c^2}{\beta^2 W^2}\left[1 + \frac{W^2}{(K-W)^2} + \frac{\tau^2}{(1+\tau)^2}\frac{W^2}{K^2}\right. \\ &\qquad\left. - \frac{2\tau+1}{(1+\tau)^2}\frac{W}{K-W}\right],\end{aligned}} \tag{6.140}$$

where

$$\tau = \frac{K}{m_0 c^2}, \tag{6.141}$$

and

$$\beta = \frac{v}{c} \tag{6.142}$$

is the incident-electron velocity in units of the speed of light. The classical electron radius r_e is the quantity defined in Eq. (6.51).

The four terms in Eq. (6.140) contain not only the relativistic electron–electron kinematics but also the exchange contribution associated with the indistinguishability of the two outgoing electrons. Their microscopic origin is derived explicitly in Sec. A.14; here our interest is in the energy deposited in the detector material.

A special feature of Møller scattering is that the two outgoing electrons are indistinguishable. The symmetric form of the Møller cross section derived in Eq. (A.1069) is unchanged under the exchange

$$W \longleftrightarrow K - W. \tag{6.143}$$

Thus a final state in which one outgoing electron has kinetic energy W and the other has kinetic energy $K - W$ is physically the same final state as the one obtained by interchanging those two labels.

If we integrated over the complete range of both possibilities, every physical final state would therefore be counted twice. To avoid this double counting, we define W to be the kinetic energy of the less energetic outgoing electron, conventionally called the *secondary* electron. Hence

$$W \leq K - W, \tag{6.144}$$

which immediately gives

$$\boxed{0 < W \leq \frac{K}{2}.} \tag{6.145}$$

The upper limit $K/2$ is therefore not a dynamical restriction on the energy that an individual labelled electron may carry. It arises from the indistinguishability of the two outgoing electrons and from the convention that W denotes the energy of the less energetic one.

We now convert the microscopic Møller cross section into a macroscopic stopping power.

Consider first how many target electrons are available in the material. If the material has mass density ρ_{m} and molar mass M, then the number of moles per unit volume is

$$\frac{\rho_{\mathrm{m}}}{M}. \tag{6.146}$$

Multiplying by Avogadro's constant N_A gives the number of atoms per unit volume,

$$\boxed{n_{\mathrm{at}} = \rho_{\mathrm{m}} \frac{N_A}{M}.} \tag{6.147}$$

If each atom contains Z electrons, the corresponding number density of target electrons is

$$\boxed{n_e = Z n_{\mathrm{at}} = \rho_{\mathrm{m}} \frac{N_A Z}{M}.} \tag{6.148}$$

We now follow one energetic electron as it travels an infinitesimal distance ds through the material. The Møller differential cross section

$$\frac{d\sigma_{\mathrm{M}}}{dW} \tag{6.149}$$

is defined so that

$$d\sigma_{\mathrm{M}} = \frac{d\sigma_{\mathrm{M}}}{dW}\, dW \tag{6.150}$$

is the cross-sectional area associated with collisions in which the transferred kinetic energy lies between W and $W + dW$.

During the path length ds, this cross-sectional area sweeps out the effective interaction volume

$$dV_{\mathrm{int}} = d\sigma_{\mathrm{M}}\, ds. \tag{6.151}$$

Since the material contains n_e target electrons per unit volume, the mean number of collisions in this energy-transfer interval is

$$\begin{aligned} d\langle N(W)\rangle &= n_e\, dV_{\mathrm{int}} && (6.152)\\ &= n_e\, ds\, d\sigma_{\mathrm{M}} && (6.153)\\ &= \boxed{n_e\, ds\, \frac{d\sigma_{\mathrm{M}}}{dW}\, dW.} && (6.154) \end{aligned}$$

Each collision belonging to this interval transfers approximately the energy W from the incident electron to the struck electron. Therefore,

the mean energy transferred in collisions with energies between W and $W + dW$ is

$$dE_{\rm tr}(W) = W\, d\langle N(W)\rangle. \tag{6.155}$$

Using Eq. (6.154),

$$\boxed{dE_{\rm tr}(W) = n_e\, ds\, W \frac{d\sigma_{\rm M}}{dW}\, dW.} \tag{6.156}$$

The hard-collision region extends from the matching energy W_c to the maximum transferred energy $K/2$ defined in Eq. (6.145). The total mean energy transferred in hard collisions while the incident electron travels the distance ds is therefore

$$dE_{\rm tr,hard} = \int_{W_c}^{K/2} dE_{\rm tr}(W) \tag{6.157}$$

$$= n_e\, ds \int_{W_c}^{K/2} W \frac{d\sigma_{\rm M}}{dW}\, dW. \tag{6.158}$$

This transferred energy is lost by the incident electron. If its kinetic energy changes by dK, then

$$dK = -dE_{\rm tr,hard}. \tag{6.159}$$

Substituting Eq. (6.158),

$$dK = -n_e\, ds \int_{W_c}^{K/2} W \frac{d\sigma_{\rm M}}{dW}\, dW. \tag{6.160}$$

The hard-collision stopping power is the mean kinetic-energy loss per unit path length,

$$\boxed{S_{\rm hard} \equiv -\frac{dK}{ds}.} \tag{6.161}$$

Dividing Eq. (6.160) by ds therefore gives

$$\boxed{S_{\rm hard} = n_e \int_{W_c}^{K/2} W \frac{d\sigma_{\rm M}}{dW}\, dW.} \tag{6.162}$$

Finally, we are interested in the *mass stopping power*, obtained by dividing the stopping power by the mass density $\rho_{\rm m}$. Substituting the

target-electron density from Eq. (6.148) into Eq. (6.162),

$$\frac{S_{\text{hard}}}{\rho_{\text{m}}} = \frac{1}{\rho_{\text{m}}}\left(\rho_{\text{m}}\frac{N_A Z}{M}\right)\int_{W_c}^{K/2} W\frac{d\sigma_{\text{M}}}{dW}\,dW \tag{6.163}$$

$$= \boxed{\frac{N_A Z}{M}\int_{W_c}^{K/2} W\frac{d\sigma_{\text{M}}}{dW}\,dW.} \tag{6.164}$$

Equation (6.164) establishes the bridge between the microscopic Møller scattering cross section and the macroscopic stopping power. The factor $d\sigma_{\text{M}}/dW$ determines the probability of a collision with a given transferred energy, the factor W weights that collision by the energy removed from the incident electron, and the factor $N_A Z/M$ gives the number of target electrons per unit mass of material.

Integration over the hard-collision region. To evaluate Eq. (6.164), introduce the dimensionless transferred-energy fraction

$$x = \frac{W}{K}, \qquad x_c = \frac{W_c}{K}. \tag{6.165}$$

Thus

$$W = Kx, \qquad dW = K\,dx, \tag{6.166}$$

and the integration limits become

$$x_c \le x \le \frac{1}{2}. \tag{6.167}$$

For compactness, define

$$A = \frac{\tau^2}{(1+\tau)^2}, \qquad B = \frac{2\tau+1}{(1+\tau)^2}. \tag{6.168}$$

Substituting Eq. (6.140) into Eq. (6.164) gives

$$\frac{S_{\text{hard}}}{\rho_{\text{m}}} = \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M}\int_{W_c}^{K/2}\frac{1}{W}\left[1 + \frac{W^2}{(K-W)^2} + A\frac{W^2}{K^2} - B\frac{W}{K-W}\right]dW. \tag{6.169}$$

Using Eq. (6.165), each term transforms according to

$$\begin{aligned}
\frac{dW}{W} &= \frac{dx}{x}, \\
\frac{W}{(K-W)^2}\,dW &= \frac{x}{(1-x)^2}\,dx, \\
\frac{W}{K^2}\,dW &= x\,dx, \\
\frac{dW}{K-W} &= \frac{dx}{1-x}.
\end{aligned} \tag{6.170}$$

Therefore,

$$\boxed{\begin{aligned}
\frac{S_{\text{hard}}}{\rho_{\text{m}}} &= \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M} \\
&\quad \times \int_{x_c}^{1/2}\left[\frac{1}{x} + \frac{x}{(1-x)^2} + Ax - \frac{B}{1-x}\right]dx.
\end{aligned}} \tag{6.171}$$

The four contributions can be integrated separately. In particular,

$$\frac{x}{(1-x)^2} = \frac{1}{(1-x)^2} - \frac{1}{1-x}, \tag{6.172}$$

so an antiderivative of the complete integrand is

$$\boxed{\mathcal{G}(x) = \ln x + \frac{1}{1-x} + (1+B)\ln(1-x) + \frac{A}{2}x^2.} \tag{6.173}$$

Indeed,

$$\frac{d\mathcal{G}}{dx} = \frac{1}{x} + \frac{x}{(1-x)^2} + Ax - \frac{B}{1-x}. \tag{6.174}$$

At the upper limit,

$$\mathcal{G}\left(\frac{1}{2}\right) = \ln\frac{1}{2} + 2 + (1+B)\ln\frac{1}{2} + \frac{A}{8} \tag{6.175}$$

$$= 2 - (2+B)\ln 2 + \frac{A}{8}. \tag{6.176}$$

The matching energy W_c separates atomic-scale soft collisions from hard collisions of a fast incident electron and is therefore much smaller than its kinetic energy K. Hence

$$x_c = \frac{W_c}{K} \ll 1. \tag{6.177}$$

For the lower limit,

$$\frac{1}{1-x_c} = 1 + \mathcal{O}(x_c), \tag{6.178}$$

$$\ln(1-x_c) = -x_c + \mathcal{O}(x_c^2), \tag{6.179}$$

and

$$\frac{A}{2}x_c^2 = \mathcal{O}(x_c^2). \tag{6.180}$$

Consequently,

$$\mathcal{G}(x_c) = \ln x_c + 1 + \mathcal{O}(x_c). \tag{6.181}$$

Subtracting Eq. (6.181) from Eq. (6.176), and neglecting corrections of relative order W_c/K, gives

$$\begin{aligned}\mathcal{G}\left(\frac{1}{2}\right) - \mathcal{G}(x_c) = & -\ln x_c + 1 - (2+B)\ln 2 + \frac{A}{8} \\ = & \ln\left(\frac{K}{4W_c}\right) + 1 - B\ln 2 + \frac{A}{8}.\end{aligned} \tag{6.182}$$

Finally, substituting the definitions of A and B from Eq. (6.168), we obtain

$$\boxed{\begin{aligned}\frac{S_{\text{hard}}}{\rho_{\text{m}}} = & \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M} \\ & \times\left[\ln\left(\frac{K}{4W_c}\right) + 1 + \frac{\tau^2}{8(1+\tau)^2} - \frac{2\tau+1}{(1+\tau)^2}\ln 2\right].\end{aligned}} \tag{6.183}$$

Combination of the soft and hard regions. We have now calculated the two contributions introduced in Eq. (6.63). The hard part has been reduced to the closed expression Eq. (6.183) and the soft part in Eq. (6.138).

We can now perform the addition promised in Eq. (6.63). Dividing that equation by ρ_{m},

$$\frac{S_{\mathrm{col}}}{\rho_{\mathrm{m}}} = \frac{S_{\mathrm{soft}}}{\rho_{\mathrm{m}}} + \frac{S_{\mathrm{hard}}}{\rho_{\mathrm{m}}}. \tag{6.184}$$

Substituting Eqs. (6.138) and (6.183), both contributions contain the common prefactor

$$\frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M}. \tag{6.185}$$

Therefore,

$$\begin{aligned}\frac{S_{\mathrm{col}}}{\rho_{\mathrm{m}}} = \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M}\Bigg\{ & \\ & \ln\left(\frac{2m_0c^2\beta^2\gamma^2 W_c}{I^2}\right) - \beta^2 - \delta \\ & + \ln\left(\frac{K}{4W_c}\right) + 1 + \frac{\tau^2}{8(1+\tau)^2} \\ & - \frac{2\tau+1}{(1+\tau)^2}\ln 2\Bigg\}.\end{aligned} \tag{6.186}$$

We now see explicitly why the intermediate energy W_c cannot have physical significance. The two logarithms in Eq. (6.186) combine as

$$\begin{aligned}\ln\left(\frac{2m_0c^2\beta^2\gamma^2 W_c}{I^2}\right) + \ln\left(\frac{K}{4W_c}\right) & \\ = \ln\left[\frac{2m_0c^2\beta^2\gamma^2 W_c}{I^2}\frac{K}{4W_c}\right] & \qquad (6.187) \\ = \ln\left[\frac{m_0c^2 K\beta^2\gamma^2}{2I^2}\right]. & \qquad (6.188)\end{aligned}$$

The factor W_c has cancelled exactly from the logarithmic terms.

This is the consistency condition anticipated when the arbitrary matching energy was introduced in Eq. (6.66). The physical collisional stopping power cannot depend on where we choose to separate the soft and hard descriptions.

It remains to express Eq. (6.188) in terms of the dimensionless kinetic energy τ. From Eq. (6.45),

$$K = \tau m_0 c^2. \tag{6.189}$$

Furthermore, Eqs. (6.47) and (6.48) give

$$\beta^2\gamma^2 = \gamma^2 - 1 \tag{6.190}$$
$$= (1+\tau)^2 - 1 \tag{6.191}$$
$$= \boxed{\tau(\tau+2).} \tag{6.192}$$

Consequently,

$$\frac{m_0 c^2 K \beta^2 \gamma^2}{2I^2} = \frac{m_0 c^2 \left(\tau m_0 c^2\right) \tau(\tau+2)}{2I^2} \tag{6.193}$$
$$= \frac{(m_0 c^2)^2 \tau^2 (\tau+2)}{2I^2} \tag{6.194}$$
$$= \frac{\tau^2(\tau+2)}{2\left(I/m_0 c^2\right)^2}. \tag{6.195}$$

Thus,

$$\boxed{\ln\left[\frac{m_0 c^2 K \beta^2 \gamma^2}{2I^2}\right] = \ln\left[\frac{\tau^2(\tau+2)}{2\left(I/m_0 c^2\right)^2}\right].} \tag{6.196}$$

The non-logarithmic kinematic terms in Eq. (6.186) are

$$-\beta^2 + 1 + \frac{\tau^2}{8(1+\tau)^2} - \frac{2\tau+1}{(1+\tau)^2}\ln 2. \tag{6.197}$$

Collecting the terms with the common denominator $(1+\tau)^2$,

$$-\beta^2 + 1 + \frac{\tau^2}{8(1+\tau)^2} - \frac{2\tau+1}{(1+\tau)^2}\ln 2$$
$$= 1 - \beta^2 + \frac{\tau^2/8 - (2\tau+1)\ln 2}{(1+\tau)^2} \tag{6.198}$$
$$= \boxed{F^-(\tau),} \tag{6.199}$$

where $F^-(\tau)$ is precisely the electron-specific correction defined in Eq. (6.50).

Substituting this result and Eq. (6.196) into Eq. (6.186) finally gives

$$\boxed{\begin{aligned}\frac{S_{\rm col}}{\rho_{\rm m}} &= \frac{2\pi r_e^2 m_0 c^2 N_A}{\beta^2}\frac{Z}{M} \\ &\quad\times\left\{\ln\left[\frac{\tau^2(\tau+2)}{2\left(I/m_0c^2\right)^2}\right] + F^-(\tau) - \delta\right\}.\end{aligned}} \tag{6.200}$$

Equation (6.200) is the collisional stopping-power expression stated in Eq. (6.49).

The derivation also explains why the division into soft and hard collisions is legitimate. The two approximations overlap in the intermediate region specified by Eq. (6.66): the soft calculation retains the bound-electron response, whereas the hard calculation uses free relativistic electron–electron scattering. The matching energy W_c only determines where we change from one description to the other. Its explicit logarithmic dependence cancels when the two contributions are added.

In deriving Eq. (6.183), terms of relative order W_c/K were neglected. The cancellation is therefore exact at the order retained in the present calculation; any residual dependence on the arbitrary matching point belongs to the higher-order terms omitted in constructing the soft–hard approximation.

6.3.2 Radiative stopping power.

The second contribution in Eq. (6.38) has a different physical origin. Instead of transferring energy directly to electronic excitations of the material, the incident electron emits part of its energy as electromagnetic radiation.

The radiative term describes *bremsstrahlung* (German for "braking radiation"). A fast electron passing through the Coulomb field of an atomic nucleus is deflected and therefore accelerated—a change in the direction of the velocity is already an acceleration, even if its speed changes little. An accelerated electric charge emits electromagnetic radiation, so the electron can lose part of its kinetic energy through the emission of an X-ray photon.

Schematically,

$$\boxed{e^- + Z \longrightarrow e^- + Z + \gamma,} \tag{6.201}$$

where the nucleus Z absorbs the momentum required by the interaction while acquiring only a very small recoil energy. Conservation of energy gives

$$E_{k,\rm in} = E_{k,\rm out} + h\nu_\gamma + E_{\rm recoil}. \tag{6.202}$$

Because different encounters can transfer different fractions of the electron kinetic energy, bremsstrahlung produces a continuous photon spectrum rather than a single characteristic photon energy.

Classically, the connection between acceleration and electromagnetic radiation can be made quantitative through the *Larmor formula*. For a nonrelativistic point charge undergoing an instantaneous acceleration $\mathbf{a}$, the total electromagnetic power radiated into the surrounding space is

$$\boxed{P = \frac{q^2 a^2}{6\pi\varepsilon_0 c^3}.} \tag{6.203}$$

Here

$$a = \|\mathbf{a}\| = \left\|\frac{d\mathbf{v}}{dt}\right\| \tag{6.204}$$

is the magnitude of the acceleration of the particle, q is its electric charge, ε_0 is the vacuum permittivity, and c is the speed of light in vacuum. The quantity P is the energy radiated per unit time and therefore has SI units of watts ($\mathrm{J\,s^{-1}}$).

For an electron,

$$q = -e, \tag{6.205}$$

where $e > 0$ is the elementary charge. Since the charge enters Eq. (6.203) quadratically, its sign is irrelevant, and the electron formula becomes

$$\boxed{P = \frac{e^2 a^2}{6\pi\varepsilon_0 c^3}.} \tag{6.206}$$

The important feature of Eq. (6.206) is the quadratic dependence

$$P \propto a^2. \tag{6.207}$$

Thus a rapidly changing velocity produces much more radiation than a slowly changing one. Notice that acceleration does not require a change in the *magnitude* of the velocity. A change in its direction also produces

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} \neq 0. \tag{6.208}$$

This is precisely what happens when a fast electron is deflected by the Coulomb field of an atomic nucleus.

Larmor radiation formula.

Consider a point charge q moving with velocity $\mathbf{v}(t)$ and acceleration

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt}. \tag{6.209}$$

We shall derive the electromagnetic power radiated by the charge in the non-relativistic limit

$$v \ll c. \tag{6.210}$$

The derivation proceeds in two stages. We first determine the part of the electromagnetic field generated by the acceleration of the charge. We then calculate the energy carried by this field through a distant surface.

From Maxwell's equations to a retarded field. For a stationary point charge, the electric field is simply the Coulomb field,

$$\mathbf{E}_{\mathrm{C}} = \frac{q}{4\pi\varepsilon_0}\frac{\mathbf{n}}{R^2}, \tag{6.211}$$

where R is the distance from the charge to the observation point and $\mathbf{n}$ is the corresponding unit vector.

Equation (6.211) contains no explicit speed of light because it describes a time-independent situation. Once the charge moves non-uniformly, however, changes in its electromagnetic field cannot be communicated instantaneously: Maxwell's equations imply that they propagate through vacuum at the finite speed c.

To make this propagation explicit, introduce the scalar and vector potentials,

$$\boxed{\mathbf{B} = \nabla \times \mathbf{A}, \qquad \mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}.} \tag{6.212}$$

The potentials are not unique. For any sufficiently smooth scalar function $\chi(\mathbf{r}, t)$, the transformation

$$\boxed{\begin{aligned} \mathbf{A}' &= \mathbf{A} + \nabla\chi, \\ \phi' &= \phi - \frac{\partial \chi}{\partial t} \end{aligned}} \tag{6.213}$$

leaves both **E** and **B** unchanged, because

$$\nabla \times \nabla \chi = 0 \tag{6.214}$$

and the two additional terms in **E** cancel.

This freedom allows us to impose a convenient additional condition on the potentials. For radiation problems we choose the *Lorenz gauge*,

$$\boxed{\nabla \cdot \mathbf{A} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0.} \tag{6.215}$$

We now use Maxwell's equations to determine the equations satisfied by the potentials ϕ and **A**. We shall do this separately for the scalar and vector potentials.

Recall from Eq. (6.212) that

$$\mathbf{E} = -\nabla \phi - \frac{\partial \mathbf{A}}{\partial t}, \tag{6.216}$$

and

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{6.217}$$

Equation for the scalar potential. We begin with Gauss's law,

$$\boxed{\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}.} \tag{6.218}$$

Substituting the expression for the electric field,

$$\nabla \cdot \left(-\nabla \phi - \frac{\partial \mathbf{A}}{\partial t}\right) = \frac{\rho}{\varepsilon_0}. \tag{6.219}$$

Using linearity of the divergence,

$$-\nabla \cdot \nabla \phi - \nabla \cdot \frac{\partial \mathbf{A}}{\partial t} = \frac{\rho}{\varepsilon_0}. \tag{6.220}$$

The first term is the Laplacian of the scalar potential,

$$\nabla \cdot \nabla \phi = \nabla^2 \phi, \tag{6.221}$$

while, for sufficiently smooth fields, the spatial and temporal derivatives can be interchanged,

$$\nabla \cdot \frac{\partial \mathbf{A}}{\partial t} = \frac{\partial}{\partial t} (\nabla \cdot \mathbf{A}) . \tag{6.222}$$

Therefore,

$$\boxed{-\nabla^2 \phi - \frac{\partial}{\partial t} (\nabla \cdot \mathbf{A}) = \frac{\rho}{\varepsilon_0}.} \tag{6.223}$$

At this point the scalar potential ϕ is still coupled to the vector potential $\mathbf{A}$ through the term $\nabla \cdot \mathbf{A}$. This is precisely where the Lorenz gauge, Eq. (6.215), becomes useful:

$$\nabla \cdot \mathbf{A} = -\frac{1}{c^2} \frac{\partial \phi}{\partial t}. \tag{6.224}$$

Differentiating with respect to time,

$$\frac{\partial}{\partial t} (\nabla \cdot \mathbf{A}) = -\frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2}. \tag{6.225}$$

Substitution into Eq. (6.223) gives

$$-\nabla^2 \phi + \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{\rho}{\varepsilon_0}. \tag{6.226}$$

Multiplying by -1,

$$\boxed{\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) \phi = -\frac{\rho}{\varepsilon_0}.} \tag{6.227}$$

Thus the Lorenz gauge has removed the coupling to $\mathbf{A}$, leaving a wave equation for ϕ alone.

Equation for the vector potential. We now start from Ampère–Maxwell's law,

$$\boxed{\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}.} \tag{6.228}$$

Substituting

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{6.229}$$

into the left-hand side gives

$$\nabla \times (\nabla \times \mathbf{A}) \,. \tag{6.230}$$

For the right-hand side, differentiating

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t} \tag{6.231}$$

with respect to time gives

$$\frac{\partial \mathbf{E}}{\partial t} = -\nabla\frac{\partial \phi}{\partial t} - \frac{\partial^2 \mathbf{A}}{\partial t^2}. \tag{6.232}$$

Therefore Ampère–Maxwell's law becomes

$$\begin{aligned} \nabla \times (\nabla \times \mathbf{A}) = \mu_0 \mathbf{J} - \mu_0 \varepsilon_0 \nabla \frac{\partial \phi}{\partial t} \\ - \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{A}}{\partial t^2}. \end{aligned} \tag{6.233}$$

We now use the vector identity

$$\boxed{\nabla \times (\nabla \times \mathbf{A}) = \nabla (\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}.} \tag{6.234}$$

Equation (6.233) therefore becomes

$$\begin{aligned} \nabla (\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{J} - \mu_0 \varepsilon_0 \nabla \frac{\partial \phi}{\partial t} \\ - \mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{A}}{\partial t^2}. \end{aligned} \tag{6.235}$$

Once again, $\mathbf{A}$ and ϕ are still coupled. The Lorenz gauge removes this coupling. From Eq. (6.215),

$$\nabla \cdot \mathbf{A} = -\frac{1}{c^2}\frac{\partial \phi}{\partial t}, \tag{6.236}$$

so

$$\nabla (\nabla \cdot \mathbf{A}) = -\frac{1}{c^2}\nabla\frac{\partial \phi}{\partial t}. \tag{6.237}$$

We now use the relation

$$\boxed{\mu_0\varepsilon_0 = \frac{1}{c^2}.} \tag{6.238}$$

With this substitution, Eq. (6.235) becomes

$$\begin{aligned} -\frac{1}{c^2}\nabla\frac{\partial\phi}{\partial t} - \nabla^2\mathbf{A} = \mu_0\mathbf{J} - \frac{1}{c^2}\nabla\frac{\partial\phi}{\partial t} \\ - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2}. \end{aligned} \tag{6.239}$$

The terms containing the scalar potential are identical on the two sides and cancel:

$$-\nabla^2\mathbf{A} = \mu_0\mathbf{J} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2}. \tag{6.240}$$

Rearranging,

$$\boxed{\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{A} = -\mu_0\mathbf{J}.} \tag{6.241}$$

The Lorenz gauge has therefore achieved the same simplification for the vector potential as for the scalar potential: $\mathbf{A}$ now satisfies a wave equation independently of ϕ.

Equations (6.227) and (6.241) are the key point: the electromagnetic potentials obey wave equations whose propagation velocity is c. Their causal solutions must therefore depend on the source at an earlier time.

The corresponding outgoing solutions are the *retarded potentials*,

$$\phi(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0}\int\frac{\rho(\mathbf{r}',t_\mathrm{r})}{|\mathbf{r}-\mathbf{r}'|}\,d^3r', \tag{6.242}$$

and

$$\mathbf{A}(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int\frac{\mathbf{J}(\mathbf{r}',t_\mathrm{r})}{|\mathbf{r}-\mathbf{r}'|}\,d^3r', \tag{6.243}$$

where

$$\boxed{t_\mathrm{r} = t - \frac{|\mathbf{r}-\mathbf{r}'|}{c}} \tag{6.244}$$

is the *retarded time*. Thus the field observed at time t depends on what the source was doing sufficiently earlier for the electromagnetic disturbance to have reached the observation point.

Field of an accelerated point charge. For a point charge following a trajectory $\mathbf{r}_q(t)$,

$$\rho(\mathbf{r},t) = q\,\delta\left(\mathbf{r}-\mathbf{r}_q(t)\right), \tag{6.245}$$

and

$$\mathbf{J}(\mathbf{r},t) = q\mathbf{v}(t)\,\delta\left(\mathbf{r}-\mathbf{r}_q(t)\right). \tag{6.246}$$

Evaluation of the retarded potentials for this moving point source gives the Liénard–Wiechert potentials. Differentiating them according to Eq. (6.212) gives the electric field

$$\boxed{\mathbf{E}(\mathbf{r},t) = \frac{q}{4\pi\varepsilon_0}\left[\frac{(1-\beta^2)(\mathbf{n}-\boldsymbol{\beta})}{(1-\mathbf{n}\cdot\boldsymbol{\beta})^3R^2} + \frac{\mathbf{n}\times\left[(\mathbf{n}-\boldsymbol{\beta})\times\dot{\boldsymbol{\beta}}\right]}{c(1-\mathbf{n}\cdot\boldsymbol{\beta})^3R}\right]_{\mathrm{r}},} \tag{6.247}$$

where

$$\boldsymbol{\beta} = \frac{\mathbf{v}}{c}, \tag{6.248}$$

and

$$\mathbf{R} = \mathbf{r}-\mathbf{r}_q(t_{\mathrm{r}}), \qquad R = \|\mathbf{R}\|, \qquad \mathbf{n} = \frac{\mathbf{R}}{R}. \tag{6.249}$$

The notation $[\cdots]_{\mathrm{r}}$ indicates evaluation at the retarded time.

The two terms in Eq. (6.247) have fundamentally different distance dependences:

$$\boxed{\begin{array}{rcl} \text{velocity field} & \propto & R^{-2}, \\ \text{acceleration field} & \propto & R^{-1}. \end{array}} \tag{6.250}$$

The first is the moving-charge generalization of the Coulomb field. The second exists only when the charge accelerates and is the radiation field.

In the non-relativistic limit,

$$\beta \ll 1, \tag{6.251}$$

we may take

$$\boldsymbol{\beta} \simeq 0, \qquad 1-\mathbf{n}\cdot\boldsymbol{\beta} \simeq 1, \tag{6.252}$$

while

$$\dot{\boldsymbol{\beta}} = \frac{\mathbf{a}}{c}. \tag{6.253}$$

Then, the radiation part of the Liénard–Wiechert field becomes

$$\boxed{\mathbf{E}_{\mathrm{rad}}(\mathbf{r},t) = \frac{q}{4\pi\varepsilon_0 c^2 R}\left[\mathbf{n}\times(\mathbf{n}\times\mathbf{a})\right]_{t_{\mathrm{r}}}.} \tag{6.254}$$

The notation

$$\left[\cdots\right]_{t_{\mathrm{r}}} \tag{6.255}$$

means that the quantities describing the moving charge inside the brackets must be evaluated at the *retarded time* t_{r}, rather than at the observation time t.

Suppose that the field is observed at position $\mathbf{r}$ and time t, while the charge follows the trajectory $\mathbf{r}_q(t)$. The retarded time is defined implicitly by

$$\boxed{t_{\mathrm{r}} = t - \frac{\|\mathbf{r}-\mathbf{r}_q(t_{\mathrm{r}})\|}{c}.} \tag{6.256}$$

Thus the electromagnetic signal observed at $(\mathbf{r},t)$ was emitted when the charge was at the earlier position

$$\mathbf{r}_q(t_{\mathrm{r}}). \tag{6.257}$$

We therefore define

$$\boxed{\mathbf{R} = \mathbf{r}-\mathbf{r}_q(t_{\mathrm{r}}), \qquad R = \|\mathbf{R}\|, \qquad \mathbf{n} = \frac{\mathbf{R}}{R}.} \tag{6.258}$$

The acceleration appearing in Eq. (6.254) is likewise

$$\mathbf{a}(t_{\mathrm{r}}) = \left.\frac{d\mathbf{v}}{dt}\right|_{t=t_{\mathrm{r}}}. \tag{6.259}$$

Consequently,

$$\left[\mathbf{n}\times(\mathbf{n}\times\mathbf{a})\right]_{t_{\mathrm{r}}} \tag{6.260}$$

is shorthand for

$$\boxed{\mathbf{n}(t_{\mathrm{r}})\times\left[\mathbf{n}(t_{\mathrm{r}})\times\mathbf{a}(t_{\mathrm{r}})\right],} \tag{6.261}$$

where $\mathbf{n}(t_{\mathrm{r}})$ points from the position occupied by the charge at the retarded time toward the observation point.

The distinction between t and t_{r} is essential. The observer does not see the radiation corresponding to the acceleration of the charge *now*;

it sees the radiation generated by the acceleration at the earlier time required for that electromagnetic disturbance to travel from the charge to the observer.

Schematically,

$$\boxed{\begin{array}{ccc} t_{\mathrm{r}} & & t \\ \text{charge accelerates} & \xrightarrow{\;c\;} & \text{radiation is observed.} \end{array}} \tag{6.262}$$

The factor c^{-2} now has a clear origin. One factor $1/c$ is present in the acceleration term of the Liénard–Wiechert field because changes in the source propagate at speed c, while the other arises from

$$\dot{\boldsymbol{\beta}} = \frac{\mathbf{a}}{c}. \tag{6.263}$$

The vector triple-product identity gives

$$\mathbf{n} \times (\mathbf{n} \times \mathbf{a}) = \mathbf{n}\,(\mathbf{n} \cdot \mathbf{a}) - \mathbf{a} = -\mathbf{a}_{\perp}, \tag{6.264}$$

where $\mathbf{a}_{\perp}$ is the component of the acceleration perpendicular to $\mathbf{n}$. If θ is the angle between $\mathbf{a}$ and $\mathbf{n}$,

$$a_{\perp} = a \sin\theta, \tag{6.265}$$

and hence

$$\boxed{E_{\mathrm{rad}} = \frac{|q| a \sin\theta}{4\pi\varepsilon_0 c^2 R}.} \tag{6.266}$$

Thus the radiation field is transverse,

$$\mathbf{n} \cdot \mathbf{E}_{\mathrm{rad}} = 0, \tag{6.267}$$

and its amplitude decreases only as $1/R$.

Why the $1/R$ field is the radiation field. The significance of the $1/R$ dependence can already be anticipated from energy conservation. As discussed in Sec. 3.11 of Chap. 3, the instantaneous electromagnetic energy-flux density is described by the Poynting vector, Eq. (3.696),

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}. \tag{6.268}$$

For monochromatic fields, its time-averaged value is given by Eq. (3.698). In particular, for a propagating electromagnetic wave the magnitude of the energy flux is proportional to the square of the field amplitude. Therefore,

$$E_{\text{rad}} \propto \frac{1}{R} \qquad \Longrightarrow \qquad S_{\text{rad}} \propto E_{\text{rad}}^2 \propto .\frac{1}{R^2}. \tag{6.269}$$

The area of a sphere of radius R grows as

$$A = 4\pi R^2, \tag{6.270}$$

so the total electromagnetic power crossing such a sphere behaves as

$$P \propto R^2 S_{\text{rad}} \propto R^2 \frac{1}{R^2} = \text{constant}. \tag{6.271}$$

Thus the $1/R$ field can transport a finite amount of energy to arbitrarily large distances. This is why the acceleration-dependent $1/R$ term of the Liénard–Wiechert field is identified as the *radiation field*.

We now calculate that power explicitly.

Spatial variation of the radiation field. We have obtained the electric component of the radiation field, Eq. (6.254). We now determine the corresponding magnetic field. For this purpose we use Faraday's law,

$$\boxed{\nabla \times \mathbf{E}_{\text{rad}} = -\frac{\partial \mathbf{B}_{\text{rad}}}{\partial t}.} \tag{6.272}$$

The curl on the left-hand side involves spatial derivatives. We must therefore determine how the radiation field changes when the observation position $\mathbf{r}$ is varied while the observation time t is held fixed.

From Eq. (6.254), the radiation field has the general far-field structure

$$\boxed{\mathbf{E}_{\text{rad}}(\mathbf{r}, t) = \frac{1}{R}\mathbf{F}(\mathbf{n}, t_{\text{r}}),} \tag{6.273}$$

where we have defined

$$\boxed{\mathbf{F}(\mathbf{n}, t_{\text{r}}) = \frac{q}{4\pi\varepsilon_0 c^2}\, \mathbf{n} \times [\mathbf{n} \times \mathbf{a}(t_{\text{r}})]\,.} \tag{6.274}$$

Thus the factor $1/R$ describes the geometrical decrease of the field with distance, whereas $\mathbf{F}$ contains its angular dependence and the time dependence inherited from the acceleration of the charge at the retarded time.

In the radiation zone, the dimensions of the source are much smaller than the observation distance. To leading order in $1/R$ we may therefore write

$$R \simeq \|\mathbf{r}\|, \qquad \mathbf{n} \simeq \frac{\mathbf{r}}{R}, \qquad t_{\mathrm{r}} \simeq t - \frac{R}{c}. \tag{6.275}$$

To see which spatial variations dominate, let E_j and F_j denote Cartesian components of $\mathbf{E}_{\mathrm{rad}}$ and $\mathbf{F}$. Then

$$E_j(\mathbf{r}, t) = \frac{1}{R} F_j(\mathbf{n}, t_{\mathrm{r}}). \tag{6.276}$$

Differentiating with respect to the observation coordinate x_i at fixed t, the product and chain rules give

$$\frac{\partial E_j}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{1}{R}\right) F_j + \frac{1}{R}\sum_k \frac{\partial F_j}{\partial n_k}\frac{\partial n_k}{\partial x_i} + \frac{1}{R}\frac{\partial F_j}{\partial t_{\mathrm{r}}}\frac{\partial t_{\mathrm{r}}}{\partial x_i}. \tag{6.277}$$

The three terms have a simple interpretation. They describe, respectively, the change in the geometrical factor $1/R$, the change in the observation direction $\mathbf{n}$, and the change in the retarded time.

Their dependence on R can be determined directly. Since

$$\frac{\partial R}{\partial x_i} = n_i, \tag{6.278}$$

we have

$$\frac{\partial}{\partial x_i}\left(\frac{1}{R}\right) = -\frac{n_i}{R^2}. \tag{6.279}$$

The first term of Eq. (6.277) is therefore of order R^{-2}.

Similarly, since

$$n_k = \frac{x_k}{R}, \tag{6.280}$$

we obtain

$$\boxed{\frac{\partial n_k}{\partial x_i} = \frac{\delta_{ik} - n_i n_k}{R},} \tag{6.281}$$

where δ_{ik} is the Kronecker delta. Because the second term in Eq. (6.277) already contains a factor $1/R$, it is also of order R^{-2}.

The third term behaves differently. From Eq. (6.275), at fixed observation time,

$$\frac{\partial t_{\mathrm{r}}}{\partial x_i} = -\frac{1}{c}\frac{\partial R}{\partial x_i} \tag{6.282}$$

$$= \boxed{-\frac{n_i}{c}.} \tag{6.283}$$

Consequently, the retarded-time contribution is

$$-\frac{n_i}{cR}\frac{\partial F_j}{\partial t_{\mathrm{r}}}, \tag{6.284}$$

which is of order R^{-1}.

We have therefore found the hierarchy

$$\boxed{\begin{array}{rcl} \text{variation of } 1/R & \sim & R^{-2}, \\ \text{variation of } \mathbf{n} & \sim & R^{-2}, \\ \text{variation through } t_{\mathrm{r}} & \sim & R^{-1}. \end{array}} \tag{6.285}$$

Far from the source, the R^{-1} term dominates. Retaining only this leading contribution, Eq. (6.277) becomes

$$\boxed{\frac{\partial E_j}{\partial x_i} \simeq -\frac{n_i}{cR}\frac{\partial F_j}{\partial t_{\mathrm{r}}}.} \tag{6.286}$$

We can now relate this spatial derivative to the temporal derivative of the same field. At fixed observation position, R and $\mathbf{n}$ are constant and

$$\frac{\partial t_{\mathrm{r}}}{\partial t} = 1. \tag{6.287}$$

Therefore, from Eq. (6.276),

$$\boxed{\frac{\partial E_j}{\partial t} = \frac{1}{R}\frac{\partial F_j}{\partial t_{\mathrm{r}}}.} \tag{6.288}$$

Comparison with Eq. (6.286) gives

$$\boxed{\frac{\partial E_j}{\partial x_i} \simeq -\frac{n_i}{c}\frac{\partial E_j}{\partial t}.} \tag{6.289}$$

The physical meaning of this relation is simple. If the observation point is displaced outward by dR, the radiation must travel an additional distance dR, corresponding to an additional propagation time dR/c. At the same observation time, the more distant point therefore samples the source at an earlier retarded time,

$$dt_{\mathrm{r}} = -\frac{dR}{c}. \tag{6.290}$$

Thus the dominant spatial variation of a far-field radiation wave is the variation of its retarded waveform, rather than the much slower variation of its $1/R$ amplitude.

Equation (6.289) can be written compactly as

$$\boxed{\nabla \simeq -\frac{\mathbf{n}}{c}\frac{\partial}{\partial t},} \tag{6.291}$$

with the important understanding that this is not a general identity between spatial and temporal derivatives. It is the leading-order relation obtained when the gradient acts on an outgoing radiation field of the form Eq. (6.273).

We can now return to Faraday's law. Using Eq. (6.291),

$$\nabla \times \mathbf{E}_{\mathrm{rad}} \simeq -\frac{1}{c}\mathbf{n} \times \frac{\partial \mathbf{E}_{\mathrm{rad}}}{\partial t}. \tag{6.292}$$

Substituting this result into Eq. (6.272) gives

$$-\frac{1}{c}\mathbf{n} \times \frac{\partial \mathbf{E}_{\mathrm{rad}}}{\partial t} = -\frac{\partial \mathbf{B}_{\mathrm{rad}}}{\partial t}, \tag{6.293}$$

or

$$\frac{\partial \mathbf{B}_{\mathrm{rad}}}{\partial t} = \frac{1}{c}\mathbf{n} \times \frac{\partial \mathbf{E}_{\mathrm{rad}}}{\partial t}. \tag{6.294}$$

Integrating with respect to time and retaining only the radiation field gives

$$\boxed{\mathbf{B}_{\mathrm{rad}} = \frac{1}{c}\mathbf{n} \times \mathbf{E}_{\mathrm{rad}}.} \tag{6.295}$$

Because $\mathbf{E}_{\text{rad}}$ is perpendicular to $\mathbf{n}$, Eq. (6.295) shows that $\mathbf{E}_{\text{rad}}$, $\mathbf{B}_{\text{rad}}$, and $\mathbf{n}$ are mutually perpendicular. Their magnitudes satisfy

$$\boxed{B_{\text{rad}} = \frac{E_{\text{rad}}}{c}.} \tag{6.296}$$

Thus, locally, the far-field radiation produced by the accelerated charge has the same relation between electric and magnetic fields as a propagating plane electromagnetic wave.

Energy flux and angular distribution. The electromagnetic energy flux is given by the Poynting vector,

$$\boxed{\mathbf{S} = \frac{1}{\mu_0}\mathbf{E} \times \mathbf{B}.} \tag{6.297}$$

Using Eq. (6.295),

$$\mathbf{S}_{\text{rad}} = \frac{1}{\mu_0 c}\mathbf{E}_{\text{rad}} \times (\mathbf{n} \times \mathbf{E}_{\text{rad}}). \tag{6.298}$$

Since $\mathbf{E}_{\text{rad}} \cdot \mathbf{n} = 0$, the vector triple-product identity gives

$$\mathbf{E}_{\text{rad}} \times (\mathbf{n} \times \mathbf{E}_{\text{rad}}) = E_{\text{rad}}^2 \mathbf{n}. \tag{6.299}$$

Therefore,

$$\mathbf{S}_{\text{rad}} = \frac{E_{\text{rad}}^2}{\mu_0 c}\mathbf{n}. \tag{6.300}$$

Using

$$\mu_0 \varepsilon_0 c^2 = 1, \tag{6.301}$$

we obtain

$$\boxed{\mathbf{S}_{\text{rad}} = \varepsilon_0 c E_{\text{rad}}^2 \mathbf{n}.} \tag{6.302}$$

Substituting Eq. (6.266),

$$\boxed{S_{\text{rad}} = \frac{q^2 a^2 \sin^2\theta}{16\pi^2 \varepsilon_0 c^3 R^2}.} \tag{6.303}$$

An element of area dA on a sphere of radius R subtends the solid angle

$$d\Omega = \frac{dA}{R^2}. \tag{6.304}$$

Since

$$dP = S_{\rm rad}\, dA, \tag{6.305}$$

we obtain

$$\frac{dP}{d\Omega} = R^2 S_{\rm rad}, \tag{6.306}$$

and therefore

$$\boxed{\frac{dP}{d\Omega} = \frac{q^2 a^2}{16\pi^2 \varepsilon_0 c^3} \sin^2\theta.} \tag{6.307}$$

The radiation therefore has the characteristic angular dependence

$$\frac{dP}{d\Omega} \propto \sin^2\theta. \tag{6.308}$$

It vanishes along the acceleration direction ($\theta = 0, \pi$) and is maximal perpendicular to it ($\theta = \pi/2$).

Total radiated power. The total radiated power is obtained by integrating over all directions,

$$P = \int_{4\pi} \frac{dP}{d\Omega}\, d\Omega. \tag{6.309}$$

Choosing the polar axis along $\mathbf{a}$,

$$d\Omega = \sin\theta\, d\theta\, d\phi, \tag{6.310}$$

and therefore

$$P = \frac{q^2 a^2}{16\pi^2 \varepsilon_0 c^3} \int_0^{2\pi} d\phi \int_0^{\pi} \sin^3\theta\, d\theta. \tag{6.311}$$

The two angular integrals are

$$\int_0^{2\pi} d\phi = 2\pi, \qquad \int_0^{\pi} \sin^3\theta\, d\theta = \frac{4}{3}. \tag{6.312}$$

Consequently,

$$P = \frac{q^2 a^2}{16\pi^2 \varepsilon_0 c^3} (2\pi) \left(\frac{4}{3}\right) \tag{6.313}$$

$$= \boxed{\frac{q^2 a^2}{6\pi \varepsilon_0 c^3}.} \tag{6.314}$$

This is the Larmor formula, Eq. (6.203).

The derivation shows the complete physical chain:

$$\boxed{\begin{array}{c} \text{acceleration of a charge} \\ \Downarrow \\ \text{transverse } 1/R \text{ electromagnetic field} \\ \Downarrow \\ \text{outward energy flux } \propto 1/R^2 \\ \Downarrow \\ P = \dfrac{q^2 a^2}{6\pi \varepsilon_0 c^3}. \end{array}} \tag{6.315}$$

For an electron deflected by the Coulomb field of an atomic nucleus, this radiated energy is bremsstrahlung. The stronger the acceleration, the larger the radiation field and, quadratically, the larger the radiated power:

$$\boxed{E_{\text{rad}} \propto a, \qquad P_{\text{rad}} \propto a^2.} \tag{6.316}$$

The derivation above assumes $v \ll c$. Electrons used in cryo-EM at 200–300 keV are relativistic, so Eq. (6.314) is not quantitatively adequate for them. The corresponding classical result is the Liénard generalization of the Larmor formula, and a complete microscopic description of bremsstrahlung requires quantum electrodynamics. Nevertheless, the Larmor derivation identifies the essential physical mechanism: deflection of a charged particle by the Coulomb field of matter produces acceleration, and acceleration produces electromagnetic radiation.

6.3.3 Total stopping power

The qualitative detector trade-off stated in Eq. (6.36) can now be made quantitative. Equation (6.38) separated the stopping power into collisional and radiative contributions, and the preceding derivation identified the phys-

ical origin of the radiative term. The next question is therefore how large these two contributions actually are for the electron energies and silicon thicknesses relevant to cryo-EM.

Stopping powers are commonly tabulated as *mass stopping powers*, obtained by dividing the linear stopping power by the mass density of the material. To avoid confusion with the charge density $\rho(\mathbf{r}, t)$ used in the electromagnetic derivation, we denote the mass density of silicon explicitly by

$$\boxed{\rho_{\mathrm{Si}} \simeq 2.33\ \mathrm{g\,cm^{-3}}.} \tag{6.317}$$

Thus the mass stopping power is written

$$\frac{S(E)}{\rho_{\mathrm{Si}}}, \tag{6.318}$$

and has units $\mathrm{MeV\,cm^2\,g^{-1}}$.

For a 300 keV electron in silicon, ESTAR gives approximately

$$\left(\frac{S_{\mathrm{col}}}{\rho_{\mathrm{Si}}}\right)_{300\,\mathrm{keV}} \simeq 1.892\ \mathrm{MeV\,cm^2\,g^{-1}}, \tag{6.319}$$

whereas the radiative contribution is only

$$\left(\frac{S_{\mathrm{rad}}}{\rho_{\mathrm{Si}}}\right)_{300\,\mathrm{keV}} \simeq 1.06 \times 10^{-2}\ \mathrm{MeV\,cm^2\,g^{-1}} (\textit{Bergeret al.},\ 2017). \tag{6.320}$$

Consequently,

$$\frac{S_{\mathrm{rad}}}{S_{\mathrm{col}} + S_{\mathrm{rad}}} \simeq \frac{1.06 \times 10^{-2}}{1.892 + 1.06 \times 10^{-2}} \tag{6.321}$$

$$\simeq \boxed{5.6 \times 10^{-3}.} \tag{6.322}$$

Thus only about 0.56% of the stopping power at 300 keV is radiative. Although bremsstrahlung is physically interesting and motivated the preceding derivation, it contributes very little to the energy transfer relevant to a silicon direct detector. The overwhelmingly dominant processes are electronic excitation and ionization, which will be connected to the generation of electron–hole pairs in Sec. 6.4.2.

The total mass stopping power is

$$\left(\frac{S}{\rho_{\mathrm{Si}}}\right)_{300\,\mathrm{keV}} \simeq 1.903\ \mathrm{MeV\,cm^2\,g^{-1}}. \tag{6.323}$$

The corresponding linear stopping power is obtained by multiplying by the silicon mass density,

$$\boxed{S(E) = \rho_{\mathrm{Si}} \frac{S(E)}{\rho_{\mathrm{Si}}}.} \tag{6.324}$$

At 300 keV,

$$S(300\ \mathrm{keV}) \simeq (2.33\ \mathrm{g\,cm^{-3}})(1.903\ \mathrm{MeV\,cm^2\,g^{-1}}) \tag{6.325}$$

$$\simeq 4.43\ \mathrm{MeV\,cm^{-1}} \tag{6.326}$$

$$\simeq \boxed{0.443\ \mathrm{keV}\,\mu\mathrm{m}^{-1}.} \tag{6.327}$$

The dimensional conversion is

$$\left(\mathrm{g\,cm^{-3}}\right)\left(\mathrm{MeV\,cm^2\,g^{-1}}\right) = \mathrm{MeV\,cm^{-1}}, \tag{6.328}$$

as required by the definition of stopping power in Eq. (6.37).

Equation (6.327) gives useful physical intuition. Near 300 keV, the primary electron loses on average only about 0.44 keV for every micrometre of actual trajectory through silicon. The characteristic distance over which such an electron would lose all of its kinetic energy is therefore much larger than the thickness of a modern direct detector.

This characteristic distance can be quantified through the *continuous-slowing-down approximation* (CSDA). From Eq. (6.37),

$$\frac{dE}{ds} = -S(E), \tag{6.329}$$

and therefore

$$ds = -\frac{dE}{S(E)}. \tag{6.330}$$

If the electron starts with kinetic energy E_0 and is followed until its kinetic energy is reduced to zero, integration gives

$$\boxed{R_{\mathrm{CSDA}}(E_0) = \int_0^{E_0} \frac{dE}{S(E)}.} \tag{6.331}$$

The CSDA replaces the discrete, stochastic sequence of individual energy-loss events by a continuous average loss governed by $S(E)$. Thus R_{CSDA} is the trajectory length that the electron would accumulate while slowing from E_0 to zero in this continuous-loss approximation.

Representative ESTAR values for electrons in silicon are (Berger et al., 2017)

Electron energy	Linear stopping power S [keV μm^{-1}]	CSDA range R_{CSDA} [μm]
100 keV	0.763	78
200 keV	0.523	242
300 keV	0.443	452

The CSDA range is not the depth reached below the detector surface. The coordinate s in Eq. (6.37) measures distance along the actual electron trajectory. Because elastic scattering changes the direction of propagation, the accumulated trajectory length can be considerably larger than its projection along the detector normal. Hence,

$$\boxed{\text{CSDA range} \neq \text{projected penetration depth.}} \tag{6.332}$$

In particular, $R_{\text{CSDA}} \simeq 452$ μm at 300 keV does not mean that a 300 keV electron normally travels 452 μm in a straight line into silicon. The consequences of elastic scattering for the actual trajectory will be examined in Sec. 6.4.1.

We can now quantify why the thin-detector regime emphasized in Eq. (6.36) is possible. If an electron follows a trajectory of length L through the sensitive silicon, its mean energy loss is approximately

$$\langle E_{\text{loss}} \rangle \simeq \int_0^L S(E(s))\, ds. \tag{6.333}$$

If the detector is sufficiently thin that the electron energy changes only modestly during its traversal, then $S(E) \simeq S(E_0)$. If the trajectory is also sufficiently close to straight that $L \simeq t$, where t is the sensor thickness, Eq. (6.333) reduces to

$$\boxed{\langle E_{\text{loss}} \rangle \simeq S(E_0) t.} \tag{6.334}$$

For example, for a 300 keV electron traversing $t = 35$ μm of silicon,

$$\langle E_{\text{loss}} \rangle \simeq (0.443 \text{ keV}\, \mu\text{m}^{-1})(35\ \mu\text{m}) \tag{6.335}$$

$$\simeq \boxed{15.5 \text{ keV}.} \tag{6.336}$$

This is only

$$\frac{15.5}{300} \simeq 5.2\% \tag{6.337}$$

of the incident kinetic energy. Thus the thin detector does not need to stop the primary electron. Depositing only a few percent of its kinetic energy

is already sufficient to generate a substantial electronic signal, as we shall quantify in Sec. 6.4.2.

The value in Eq. (6.336) is an *average*. Individual primary electrons do not all lose 15.5 keV while traversing the sensor. Energy transfer occurs through a stochastic sequence of inelastic interactions, and occasional collisions can transfer substantially more energy than typical ones. The resulting event-to-event fluctuation of the energy loss is known as *energy straggling*.

For a sufficiently thin absorber, the distribution of the energy loss is strongly asymmetric rather than Gaussian. It is commonly described, to first approximation, by a Landau or Landau–Vavilov distribution, with a relatively sharp maximum and a long tail toward large energy losses. Thus the most probable energy loss,

$$E_{\text{loss}}^{\text{mp}}, \tag{6.338}$$

is generally smaller than the mean,

$$E_{\text{loss}}^{\text{mp}} < \langle E_{\text{loss}} \rangle \,. \tag{6.339}$$

This behavior is directly relevant to thin silicon electron detectors. For example, simulations of 300 keV electrons traversing a 50 μm silicon sensitive layer produce a markedly asymmetric single-electron signal distribution that is well described by a Landau distribution (Dourki et al., 2017). Experimentally measured 300 keV single-electron responses of modern direct detectors show the same characteristic Landau-like shape (McMullan et al., 2014).

Energy straggling is important because the charge generated by one incident electron is approximately proportional to the energy deposited in the sensitive volume. Consequently, even before electronic readout noise is considered, identical incident electrons can produce signals of different amplitudes:

$$\boxed{\begin{aligned} \text{stochastic inelastic interactions} &\longrightarrow \text{fluctuating } E_{\text{dep}} \\ &\longrightarrow \text{fluctuating number of electron–hole pairs} \\ &\longrightarrow \text{fluctuating single-electron signal.} \end{aligned}} \tag{6.340}$$

This intrinsic variation of the single-electron response will later contribute to the noise statistics of an integrating detector and is one of the reasons why electron counting can improve detector performance.

One final distinction will be important in what follows. Equation (6.334) estimates the energy *lost by the primary electron*, but not all of that energy

must remain inside the active detector volume. Energetic secondary electrons may leave the sensitive layer, and the small fraction transferred to bremsstrahlung photons may be absorbed elsewhere or may escape from the detector.

We therefore distinguish

$$\boxed{E_{\text{loss}} = E_{\text{in}} - E_{\text{out,primary}},} \tag{6.341}$$

from the deposited energy,

$$\boxed{E_{\text{dep}} = \text{energy transferred to the active detector volume.}} \tag{6.342}$$

For an individual primary-electron event,

$$\boxed{E_{\text{dep}} \leq E_{\text{loss}},} \tag{6.343}$$

with equality only when all energy transferred from the primary electron is ultimately deposited within the active volume. It is E_{dep}, rather than E_{loss}, that determines the number of electron–hole pairs generated in the detector, which is the subject of the next section.

Why the detector requires a different interaction description. At this point it is useful to clarify why the interaction of the imaging electron with the detector has been treated in considerably more detail than its interaction with the specimen. The underlying microscopic physics is, of course, the same. In both cases a fast electron interacts with atomic nuclei and with bound electrons, can scatter elastically or inelastically, transfers energy to electronic excitations and ionizations, and can in principle emit bremsstrahlung.

The difference is not the existence of these processes, but the *physical question* being asked and the characteristic amount of matter traversed.

In the specimen-interaction chapter, our principal objective was to understand how the specimen modifies the electron wave and therefore encodes structural information in the emerging wave. The first Born approximation was introduced precisely as a single-interaction description, Eq. (5.434),

$$\text{incident wave} \longrightarrow \text{one interaction} \longrightarrow \text{scattered wave}. \tag{6.344}$$

Its usefulness is that the scattering amplitude can then be related directly to the Fourier components of the specimen potential, as expressed in Eq. (5.445). The limitations of this approximation and its distinction from propagation through the complete specimen were emphasized in Eq. (5.452).

This does not imply that an electron traversing a real specimen can undergo only one inelastic interaction. The connection between the microscopic inelastic cross section and the inelastic mean free path was derived in Sec. 5.12.9. In particular, Eq. (5.1496) gives

$$\Lambda_{\text{in}} = \frac{1}{n\sigma_{\text{in}}}. \tag{6.345}$$

The statistics of repeated inelastic interactions through a specimen of finite thickness were then developed in Sec. 5.12.10. For a specimen thickness t, the mean number of inelastic interactions is given by Eq. (5.1534),

$$\mu_{\text{in}} = \frac{t}{\Lambda_{\text{in}}}, \tag{6.346}$$

and Eq. (5.1535) gives the probability of exactly N inelastic interactions. In particular, Eq. (5.1539) gives the probability of two or more inelastic interactions.

For a conventional thin single-particle cryo-EM specimen, however, the central image-formation problem can still be organized naturally around the unscattered and singly scattered components. The specimen is typically only tens to a few hundreds of nanometres thick, and its purpose is not to absorb the incident electron but to perturb its wave just enough to encode structural information. Electrons that undergo additional inelastic or multiple scattering are then treated as separate energy and angular channels, or as contributions that progressively degrade the useful image information.

The detector has a fundamentally different task. Its purpose is precisely to convert the energy carried by the incident electron into a measurable signal. A silicon sensitive layer may be tens of micrometres thick, and the electron continues interacting with the solid throughout its trajectory. The quantity of interest is therefore no longer the amplitude or angular distribution associated with one particular scattering event, but the *total energy transferred through the ensemble of possible interactions*.

This description is directly connected with the inelastic-scattering theory developed in Sec. 5.12. There, an inelastic event was described as a transition of the specimen from one internal state to another. The first-Born transition amplitude was obtained in Eq. (5.1282), and for a Coulomb interaction its dependence on the internal excitation of the specimen was expressed through the transition form factor in Eq. (5.1332).

From these microscopic transition amplitudes we constructed the corresponding inelastic scattering probabilities. In particular, Eq. (5.1359) intro-

duced the double differential cross section

$$\frac{d^2\sigma_{\text{inel}}}{d\Omega\, dW}, \tag{6.347}$$

which answers the question

$$\boxed{\text{with what probability does the electron transfer energy } W \text{ and emerge inside the solid angle } d\Omega?} \tag{6.348}$$

That is precisely the information required for the specimen problem. The energy loss changes the wavelength and therefore the subsequent electron optics, while the angular deflection changes the direction in which the electron leaves the specimen. Consequently, in Sec. 5.12 it was useful to retain both variables W and Ω.

For the detector problem we ask a different question. We are primarily interested in how much kinetic energy the incident electron loses, irrespective of the direction in which a particular inelastic collision scatters it. We therefore first integrate the double differential cross section over all outgoing directions:

$$\boxed{\frac{d\sigma_{\text{inel}}}{dW} = \int \frac{d^2\sigma_{\text{inel}}}{d\Omega\, dW}\, d\Omega.} \tag{6.349}$$

Thus $d\sigma_{\text{inel}}/dW$ retains the energy-transfer information developed in Sec. 5.12, but no longer retains the scattering direction.

The second step is specific to the stopping-power question. A collision with energy transfer W removes that amount of kinetic energy from the incident electron. We must therefore weight the probability of each possible event by W. For a material containing n_{at} scattering centres per unit volume,

$$\boxed{\begin{aligned} S_{\text{col}}(K) &= -\frac{dK}{ds} \\ &= n_{\text{at}} \int W \frac{d\sigma_{\text{inel}}(K,W)}{dW}\, dW \\ &= n_{\text{at}} \int\int W \frac{d^2\sigma_{\text{inel}}(K,W)}{d\Omega\, dW}\, d\Omega\, dW. \end{aligned}} \tag{6.350}$$

Equation (6.350) provides the direct bridge between the two descriptions. The microscopic inelastic-scattering theory of Sec. 5.12 resolves the possible events by both energy loss and scattering direction. The stopping power is obtained from the same distribution by discarding the angular information and taking its first moment with respect to the transferred energy.

For discrete specimen transitions the connection is even more transparent. If transition j transfers the fixed energy W_j and has differential angular cross section $d\sigma_j/d\Omega$, then its contribution to the double differential cross section has the form

$$\frac{d^2\sigma_j}{d\Omega\, dW} = \frac{d\sigma_j}{d\Omega}\delta(W - W_j). \tag{6.351}$$

Substituting this expression into Eq. (6.350) gives

$$S_j = n_{\mathrm{at}} \int\int W \frac{d\sigma_j}{d\Omega}\delta(W - W_j)\, d\Omega\, dW \tag{6.352}$$

$$= n_{\mathrm{at}} W_j \int \frac{d\sigma_j}{d\Omega}\, d\Omega \tag{6.353}$$

$$= \boxed{n_{\mathrm{at}} W_j \sigma_j.} \tag{6.354}$$

This is exactly the structure obtained for the soft-collision contribution in Eq. (6.129).

The relation between the specimen and detector descriptions can therefore be summarized as

$$\boxed{\begin{array}{ccccc} \text{transition amplitude} & \longrightarrow & \dfrac{d^2\sigma}{d\Omega\, dW} & \longrightarrow & \text{different observables} \\ & & \downarrow & & \downarrow \\ & & \text{retain } (\Omega, W) & & n_{\mathrm{at}} \displaystyle\iint W(\cdots)\, d\Omega\, dW \\ & & \text{specimen image formation} & & S_{\mathrm{col}}. \end{array}} \tag{6.355}$$

Thus the specimen and detector calculations do not invoke different microscopic physics. They interrogate the same scattering distribution in different ways: for the specimen we need to know *where* the electron is scattered and *how much* energy it loses; for the detector stopping power we integrate over where it goes and ask for the mean energy transferred per unit trajectory length.

The two descriptions can therefore be summarized as

specimen	:	how does scattering modify the electron wave?
		individual amplitudes, cross sections, coherence, and the probability of plural scattering,
detector	:	how much energy is accumulated in the material?
		many collisions, stopping power, range, energy deposition, and straggling.

(6.356)

This distinction also explains why radiative losses were not developed in comparable detail in the specimen-interaction model. Bremsstrahlung is allowed in both specimen and detector materials, but at cryo-EM energies it is a very small contribution to the total stopping power. For silicon at 300 keV, Eq. (6.322) showed that the radiative fraction is only about 0.56%. It is therefore negligible for the usual description of image formation in a thin biological specimen. Here we have included it because the detector problem requires an explicit accounting of all channels by which the primary electron can lose energy before relating that energy loss to the measured signal.

6.4 Where did the electron hit?

The electron optics determines the position at which an imaging electron reaches the detector surface. The distinction was introduced in Eq. (6.2): the electron optics determines the physical arrival coordinate, whereas the detector must infer that coordinate from the signal generated inside the sensor.

We now examine the second part of this problem.

Let the detector surface define the plane $z = 0$, and write a position inside the sensor as

$$\mathbf{r} = (\boldsymbol{\rho}, z), \qquad \boldsymbol{\rho} = (x, y). \tag{6.357}$$

Suppose that one imaging electron reaches the detector at the exactly known entrance position

$$\boxed{\mathbf{r}_0 = (\boldsymbol{\rho}_0, 0).} \tag{6.358}$$

An ideal position detector would convert this event into a signal localized at precisely the same transverse coordinate. If Q_{event} denotes the total collected charge, its response would have the form

$$\mathcal{Q}_{\text{ideal}}(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0) = Q_{\text{event}} \delta^{(2)} (\boldsymbol{\rho} - \boldsymbol{\rho}_0) . \tag{6.359}$$

A real detector does not behave in this way. The incoming electron follows a stochastic trajectory through the silicon, transfers energy at different depths and lateral positions, creates energetic secondary electrons, and ultimately generates a spatial distribution of mobile charge carriers. We shall denote the collected charge-density distribution produced by one particular realization of this process by

$$\boxed{\mathcal{Q}(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega),} \tag{6.360}$$

where ω represents the particular random microscopic history of that electron.

Thus, even for fixed incident energy and exactly fixed entrance coordinate $\boldsymbol{\rho}_0$, different electrons produce different charge distributions.

Several mechanisms contribute to this loss of localization:

$$\boxed{\begin{array}{l}\text{elastic scattering of the primary electron,}\\ \text{transport of energetic secondary electrons,}\\ \text{depth-dependent and stochastic energy deposition,}\\ \text{drift and diffusion of the generated charge carriers,}\\ \text{pixel geometry, charge sharing, and event reconstruction.}\end{array}} \tag{6.361}$$

These mechanisms should not be conflated. The first three determine *where energy is deposited*; the fourth transforms the deposited energy distribution into a collected charge distribution; and the fifth determines how that continuous charge distribution is represented by the discrete detector readout.

This sequence refines the direct-detection chain introduced in Eq. (6.26):

$$\boxed{\begin{array}{l}\boldsymbol{\rho}_0 \longrightarrow \text{stochastic primary track} \longrightarrow \text{secondary-electron cascade}\\ \quad \longrightarrow \text{three-dimensional energy deposition} \longrightarrow \text{charge generation and transport}\\ \quad \longrightarrow \text{pixel charges} \longrightarrow \widehat{\boldsymbol{\rho}}_0.\end{array}} \tag{6.362}$$

Here $\widehat{\boldsymbol{\rho}}_0$ is the detector's estimate of the true entrance coordinate $\boldsymbol{\rho}_0$.

The central question of this section is therefore:

$$\boxed{\begin{array}{c}\text{If an electron reaches silicon at the exact coordinate } \boldsymbol{\rho}_0,\\ \text{how precisely can } \boldsymbol{\rho}_0 \text{ ultimately be reconstructed?}\end{array}} \tag{6.363}$$

6.4.1 Transport and spatial spreading of energetic electrons in silicon

The stopping-power treatment developed earlier describes how rapidly the kinetic energy of the incident electron decreases as it traverses the detector. It does not, however, determine the geometrical path followed by the electron or the positions at which the transferred energy is ultimately deposited.

Two effects must be considered. Elastic collisions change the direction of the primary electron and therefore displace its subsequent trajectory from the entrance coordinate. Inelastic collisions transfer energy away from that trajectory; sufficiently energetic transfers produce secondary electrons that can themselves propagate, scatter, and generate further secondaries. The energetic part of a single detector event is therefore not one straight track, but a stochastic branching set of electron trajectories.

These two mechanisms will be considered separately below, before we translate the resulting energetic-electron trajectories into a spatial distribution of deposited energy.

Primary-electron trajectory and elastic spreading. The continuous-slowing-down approximation introduced in Eq. (6.331) describes the accumulated path length required for an electron to lose its kinetic energy. As emphasized in Eq. (6.332), this path length should not be identified with the projected penetration depth, because the electron trajectory is not generally straight.

Let

$$\mathbf{r}_{\mathrm{p}}(s) \tag{6.364}$$

denote the position of the primary electron after it has travelled a path length s, and define its instantaneous propagation direction as

$$\widehat{\mathbf{u}}(s) = \frac{d\mathbf{r}_{\mathrm{p}}}{ds}. \tag{6.365}$$

Since s is arc length,

$$|\widehat{\mathbf{u}}(s)| = 1. \tag{6.366}$$

The mean kinetic-energy evolution along this path is governed by the stopping power introduced in Eq. (6.37),

$$\boxed{\frac{dK}{ds} = -S(K).} \tag{6.367}$$

Equation (6.367) specifies how the electron energy changes with accumulated path length, but it contains no equation for the direction $\widehat{\mathbf{u}}(s)$.

That direction changes predominantly through elastic scattering from the atomic potential. The microscopic relation between scattering amplitude and elastic differential cross section was derived in Eq. (5.408). In the specimen chapter, that cross section was used to describe how elastic scattering contributes to image formation. Here the same quantity plays a different role: it determines the stochastic angular deflections of an energetic electron propagating through the detector.

If n_{at} is the atomic number density, the mean number of elastic collisions occurring along an infinitesimal path ds and producing a deflection into $d\Omega$ is

$$\boxed{d\langle N_{\mathrm{el}}\rangle = n_{\mathrm{at}}\, ds\, \frac{d\sigma_{\mathrm{el}}}{d\Omega} d\Omega.} \tag{6.368}$$

The primary-electron direction therefore undergoes a stochastic sequence of angular changes whose statistics are determined by the elastic differential cross section.

At high electron energies, $d\sigma_{\mathrm{el}}/d\Omega$ is strongly concentrated toward small scattering angles. The total elastic cross section is consequently not always the most useful measure of spatial spreading. A very small deflection changes the subsequent trajectory only slightly, whereas a rarer large-angle collision can displace the later part of the track by many micrometres.

The total elastic cross section counts all elastic collisions equally, although not all collisions are equally important for the loss of the initial propagation direction. A more appropriate quantity is the *transport cross section*,

$$\boxed{\sigma_{\mathrm{tr}} = \int (1-\cos\theta)\, \frac{d\sigma_{\mathrm{el}}}{d\Omega}\, d\Omega.} \tag{6.369}$$

The origin of the angular weight $1-\cos\theta$ can be understood by following the component of the electron momentum along its original direction.

Why the transport cross section contains $1-\cos\theta$.

Consider an electron whose momentum immediately before an elastic collision is

$$\mathbf{p} = p\,\widehat{\mathbf{z}}. \tag{6.370}$$

Since the collision is elastic, and nuclear recoil is negligible at the level

considered here, the magnitude p is essentially unchanged. If the electron is deflected through an angle θ, the component of its final momentum along the original direction is

$$p'_{\parallel} = p\cos\theta. \tag{6.371}$$

The loss of forward momentum in that collision is therefore

$$\boxed{\Delta p_{\parallel} = p - p\cos\theta = p(1-\cos\theta).} \tag{6.372}$$

Equation (6.372) explains immediately why collisions through different angles should not be weighted equally. For a nearly forward event, $\theta \simeq 0$ and

$$1-\cos\theta \simeq 0, \tag{6.373}$$

so the electron retains almost all of its original forward momentum. For a 90° deflection,

$$1-\cos\theta = 1, \tag{6.374}$$

whereas for complete backscattering, $\theta = \pi$,

$$1-\cos\theta = 2. \tag{6.375}$$

Large-angle and backward events therefore contribute much more strongly to the loss of directional memory than nearly forward scattering.

During an infinitesimal path length ds, the mean number of collisions into the solid angle $d\Omega$ is, from Eq. (6.368),

$$d\langle N_{\rm el}\rangle = n_{\rm at}\, ds\, \frac{d\sigma_{\rm el}}{d\Omega} d\Omega. \tag{6.376}$$

Multiplying this collision rate by the forward-momentum loss in Eq. (6.372) and integrating over all scattering directions gives the mean loss of forward momentum,

$$-d\langle p_{\parallel}\rangle = p\, n_{\rm at}\, ds \int (1-\cos\theta)\frac{d\sigma_{\rm el}}{d\Omega} d\Omega \tag{6.377}$$

$$= p\, n_{\rm at}\, \sigma_{\rm tr}\, ds, \tag{6.378}$$

where Eq. (6.369) has been used.

More generally, applying the same argument to the instantaneous mean forward momentum gives

$$\frac{d\langle p_\parallel\rangle}{ds} = -n_{\rm at}\sigma_{\rm tr}\langle p_\parallel\rangle. \tag{6.379}$$

For a constant transport cross section, the solution is

$$\boxed{\langle p_\parallel(s)\rangle = p_0 \exp\left(-\frac{s}{\Lambda_{\rm tr}}\right),} \tag{6.380}$$

with

$$\boxed{\Lambda_{\rm tr} = \frac{1}{n_{\rm at}\sigma_{\rm tr}}.} \tag{6.381}$$

Thus $\Lambda_{\rm tr}$ is not simply the mean distance between elastic collisions. It is the characteristic path length over which the electron loses memory of its original propagation direction.

For the strongly forward-peaked scattering relevant to high-energy electrons,

$$1 - \cos\theta = \frac{\theta^2}{2} + \mathcal{O}(\theta^4), \tag{6.382}$$

and consequently

$$\boxed{\sigma_{\rm tr} \simeq \frac{1}{2}\int \theta^2 \frac{d\sigma_{\rm el}}{d\Omega} d\Omega.} \tag{6.383}$$

The transport cross section is therefore directly related to the second moment of the angular deflection distribution.

Indeed, over a short path ds,

$$d\langle\theta^2\rangle = n_{\rm at}\, ds \int \theta^2 \frac{d\sigma_{\rm el}}{d\Omega} d\Omega, \tag{6.384}$$

so that, in the small-angle limit,

$$\boxed{\frac{d\langle\theta^2\rangle}{ds} \simeq 2n_{\rm at}\sigma_{\rm tr}.} \tag{6.385}$$

This makes explicit the connection between the microscopic elastic cross section and the gradual angular spreading of the primary-electron trajectory.

The distinction between the ordinary elastic mean free path and Λ_{tr} is therefore important. The former measures how frequently elastic collisions occur, whereas the latter measures how rapidly those collisions collectively destroy the original propagation direction. Because high-energy elastic scattering is strongly forward-peaked, many collisions may occur before the direction changes substantially.

The associated transport mean free path is

$$\boxed{\Lambda_{\mathrm{tr}} = \frac{1}{n_{\mathrm{at}}\sigma_{\mathrm{tr}}}.} \tag{6.386}$$

Whereas the ordinary elastic mean free path describes the typical distance between elastic interactions, Λ_{tr} characterizes the longer distance over which the accumulated angular deflections substantially modify the initial propagation direction.

To describe the resulting lateral displacement, write the transverse position of the primary electron as

$$\boldsymbol{\rho}_{\mathrm{p}}(s) = (x_{\mathrm{p}}(s), y_{\mathrm{p}}(s)) \,. \tag{6.387}$$

If the electron entered the detector at $\boldsymbol{\rho}_0$, its lateral displacement after travelling the path length s is

$$\boxed{\Delta\boldsymbol{\rho}_{\mathrm{p}}(s) = \boldsymbol{\rho}_{\mathrm{p}}(s) - \boldsymbol{\rho}_0.} \tag{6.388}$$

Because the sequence of elastic collisions is stochastic, $\Delta\boldsymbol{\rho}_{\mathrm{p}}(s)$ is itself a random quantity.

This immediately explains the distinction between trajectory length and penetration depth emphasized in Eq. (6.332). A tortuous trajectory may accumulate a path length much larger than its displacement along the detector normal. The CSDA range in Eq. (6.331) therefore characterizes an accumulated path length associated with energy loss, not a straight-line depth.

The distinction is particularly relevant for 300 keV electrons. Their CSDA range in silicon is hundreds of micrometres, whereas the sensitive thickness of a back-thinned direct detector may be only a few tens of micrometres. A primary electron therefore usually crosses the active sensor without stopping. Even so, elastic collisions occurring during that relatively short passage can shift the trajectory laterally and hence change the positions at which subsequent inelastic energy transfers occur.

A particularly damaging case is *backscattering*. An electron can penetrate beyond the useful sensitive region, undergo one or more large angular

deflections, and return toward the active layer at a lateral position far from its point of incidence. The resulting energy deposition produces a broad, long-range component of the detector response.

This effect explains one of the principal advantages of back-thinning. Monte Carlo calculations and direct measurements at 300 keV show that removing most of the thick supporting silicon strongly suppresses electrons that scatter deep in the substrate and subsequently return to the sensitive layer (McMullan et al., 2009b). In the measurements reported by McMullan et al. (2009b), a detector back-thinned to about 35 μm reduced the backscattered contribution to the zero-spatial-frequency signal to approximately 10%, compared with more than 40% for a detector retaining approximately 700 μm of silicon.

Thus detector thickness affects spatial resolution even before semiconductor charge diffusion is considered: reducing the available silicon restricts the distance over which the energetic primary electron can wander and suppresses the long spatial tails produced by returning electrons.

Secondary-electron cascades. Elastic scattering redistributes the trajectory of the primary electron without transferring much of its kinetic energy to the atomic electrons. The spatial distribution of the detector signal is also affected by inelastic scattering, because inelastic collisions transfer energy away from the primary track.

The hard-collision contribution was derived from relativistic electron–electron, or Møller, scattering in Sec. A.14. The corresponding detector energy-transfer cross section was given in Eq. (6.140).

Let W be the kinetic energy assigned to the less energetic of the two outgoing electrons. If n_e is the number density of target electrons, the mean number of secondary electrons generated per unit primary path length and per unit transferred energy is

$$\boxed{\frac{d^2\langle N_{\rm sec}\rangle}{ds\,dW} = n_e \frac{d\sigma_{\rm M}}{dW}.} \tag{6.389}$$

Equation (6.389) therefore gives a microscopic source term for energetic secondary electrons along the primary trajectory.

Because the two outgoing electrons are indistinguishable, the energy-transfer convention introduced in Eq. (6.145) restricts the secondary energy to

$$0 < W \leq \frac{K}{2}. \tag{6.390}$$

For a 300 keV incident electron,

$$\boxed{W_{\max} = 150 \text{ keV}.} \tag{6.391}$$

Such a secondary electron is very different from the low-energy conduction-band electron that will eventually constitute part of the detector signal. A secondary carrying tens or hundreds of kiloelectronvolts is itself an energetic projectile. It can travel an appreciable distance through silicon, undergo elastic deflections, lose energy in further inelastic collisions, and generate additional secondary electrons.

The result is a branching cascade,

$$\boxed{\begin{aligned} e^-_{\text{primary}} &\longrightarrow e^-_{\text{primary}} + e^-_{\text{secondary}} \\ &\longrightarrow \text{further elastic and inelastic interactions} \\ &\longrightarrow \text{additional energetic secondaries and excitations} \\ &\longrightarrow \text{low-energy electronic excitations} \\ &\longrightarrow \text{electron–hole pairs.} \end{aligned}} \tag{6.392}$$

The energetic electrons appearing in the first stages of Eq. (6.392) are often called *knock-on electrons*; sufficiently energetic examples are also called *delta rays*. They must not be confused with the conduction-band electrons generated at the end of the cascade. The former retain enough kinetic energy to propagate through the material as individual energetic electrons, whereas the latter are low-energy semiconductor carriers whose subsequent motion is governed by drift and diffusion.

The shape of the Møller energy-transfer spectrum has an important spatial consequence. The leading behavior of Eq. (6.140) is approximately proportional to $1/W^2$. Large energy transfers are therefore much less frequent than small transfers. When they do occur, however, they produce more energetic secondary electrons and consequently allow much longer excursions away from the primary trajectory.

A typical detector event can therefore have a relatively compact central region produced by the large number of low-energy transfers, together with much rarer long spatial tails associated with energetic secondary electrons. These tails are stochastic: two primary electrons entering at the same position and with the same kinetic energy need not produce the same secondary-electron geometry.

The primary and secondary trajectories should therefore be considered together. Elastic scattering determines how the energetic electrons change

direction, while inelastic scattering transfers energy between them and creates new branches of the cascade. The result is a three-dimensional stochastic network of energetic-electron trajectories inside the detector.

This observation also clarifies the limitation of the stopping power. The stopping power defined in Eq. (6.37) answers the question *how much energy is removed from the primary electron per unit path length.* It does not determine *where that energy is ultimately deposited.* A hard collision may remove energy from the primary at one point, while the secondary electron carrying that energy deposits it many micrometres away.

The next subsection therefore replaces the scalar deposited energy by a spatial energy-deposition density. This will allow the primary and secondary energetic-electron trajectories described here to be connected to the subsequent drift and diffusion of the low-energy semiconductor charge carriers.

6.4.2 Physical construction of a direct electron detector

The preceding discussion followed the energetic electron through the detector material and quantified the energy that it loses. In particular, Eq. (6.342) distinguished the energy deposited inside the active sensor volume from the energy lost by the primary electron. We now ask the next question: how is that deposited energy converted into an electrical signal?

For a silicon direct detector, the conversion chain introduced in Eq. (6.26) is

$$\boxed{\begin{aligned} E_{\mathrm{dep}} &\longrightarrow \text{electronic excitation in silicon} \longrightarrow \text{electron–hole pairs} \\ &\longrightarrow \text{charge transport} \longrightarrow \text{charge collection} \longrightarrow \text{electronic readout.} \end{aligned}} \tag{6.393}$$

To understand this chain, we first need a small amount of semiconductor physics. We shall introduce silicon as a semiconductor, then explain doping and the p–n junction, and only after these concepts are in place construct the pixel architecture of a monolithic active pixel sensor.

Silicon as a semiconductor. In an isolated silicon atom, electrons occupy discrete atomic energy levels. In a crystal, however, a very large number of silicon atoms are brought together in a periodic lattice. Their electronic states interact, and the discrete atomic levels broaden into ranges of allowed energies called *energy bands.*

The two bands most relevant to detector operation are the *valence band* and the *conduction band.* Between them lies an interval containing no al-

lowed electronic states, called the *band gap*. For silicon near room temperature,

$$\boxed{E_{\mathrm{g}} \simeq 1.12 \text{ eV}.} \tag{6.394}$$

In the unexcited crystal, the valence band is almost completely filled, whereas the conduction band contains relatively few mobile electrons. An electron that receives enough energy to cross the band gap can be promoted from the valence band into the conduction band:

$$\boxed{\text{valence-band electron} + \text{energy} \longrightarrow \text{conduction-band electron} + \text{vacant valence state}.} \tag{6.395}$$

The conduction-band electron can move through the crystal and carries charge $-e$. The vacant state left in the valence band is called a *hole*. A hole is not an additional elementary particle: it is the absence of an electron from an otherwise almost filled band. Nevertheless, successive motion of neighbouring valence electrons makes this vacancy behave as a mobile carrier with effective charge $+e$.

Thus the excitation in Eq. (6.395) creates an *electron–hole pair*,

$$\boxed{e^{-}_{\text{conduction}} + h^{+}_{\text{valence}}.} \tag{6.396}$$

The band-gap energy in Eq. (6.394) is the minimum energy required for an electronic excitation across the gap. It is not, however, the mean amount of energy that a fast particle must deposit in the crystal to generate one electron–hole pair. Part of the deposited energy is transferred to lattice vibrations and other excitations. The experimentally relevant quantity is therefore the mean pair-creation energy. For silicon near room temperature,

$$\boxed{w_{\mathrm{Si}} \simeq 3.6 \text{ eV},} \tag{6.397}$$

with values around 3.66 eV measured under typical conditions (Scholze et al., 2000).

Consequently, if an energy E_{dep} is deposited in the active silicon, the *mean* number of generated electron–hole pairs is

$$\boxed{\mathbb{E}\left[N_{eh} \mid E_{\mathrm{dep}}\right] = \frac{E_{\mathrm{dep}}}{w_{\mathrm{Si}}}.} \tag{6.398}$$

The conditional notation is important: even for a fixed deposited energy, the actual number of generated pairs is not deterministic.

The scale of the resulting signal can be connected directly with the stopping-power calculation. Equation (6.336) estimated a mean energy loss

of approximately 15.5 keV for a 300 keV electron crossing 35 μm of silicon. If that entire energy loss were deposited inside the active volume, Eq. (6.398) would give

$$\mathbb{E}[N_{eh}] \simeq \frac{15.5 \times 10^3 \text{ eV}}{3.6 \text{ eV}} \simeq 4.3 \times 10^3 \tag{6.399}$$

electron–hole pairs. As established in Eq. (6.343), this is an upper estimate whenever part of the energy removed from the primary electron escapes from the active volume in secondary electrons or photons.

The creation of electron–hole pairs is itself stochastic. For a fixed deposited energy, the fluctuations are conventionally described by the *Fano factor* F. For silicon, values of approximately

$$F \simeq 0.1\text{–}0.13 \tag{6.400}$$

are typical (Spieler, 2005), giving

$$\boxed{\mathrm{Var}\,(N_{eh} \mid E_{\mathrm{dep}}) = F\frac{E_{\mathrm{dep}}}{w_{\mathrm{Si}}}.} \tag{6.401}$$

If the individual pair-creation processes were statistically independent, the variance would be equal to the mean, $E_{\mathrm{dep}}/w_{\mathrm{Si}}$, as in a Poisson process. The fact that $F < 1$ means that the microscopic partition of deposited energy between electronic excitation, lattice excitation, and other channels is correlated, reducing the variance below the Poisson value.

For the numerical example above, taking

$$N_{eh} \simeq 4.3 \times 10^3 \tag{6.402}$$

and $F = 0.125$ gives an intrinsic pair-generation fluctuation

$$\sigma_{N,F} = \sqrt{F N_{eh}} \simeq 23, \tag{6.403}$$

or only about

$$\frac{\sigma_{N,F}}{N_{eh}} \simeq 0.5\%. \tag{6.404}$$

This small value must not be confused with the total event-to-event fluctuation of the signal produced by different 300 keV electrons. In a thin detector, E_{dep} is itself stochastic. Different primary electrons follow different trajectories, undergo different collision histories, generate different secondary-electron cascades, and may carry different amounts of energy out of the active volume. The fluctuation of the deposited energy must therefore be combined with the intrinsic Fano fluctuation of its conversion into electron–hole pairs.

Two contributions to the fluctuation of the generated charge.

For a given deposited energy, the conditional mean number of generated pairs is, from Eq. (6.398),

$$\mathbb{E}\left[N_{eh} \mid E_{\text{dep}}\right] = \frac{E_{\text{dep}}}{w_{\text{Si}}}, \tag{6.405}$$

while Eq. (6.401) gives the corresponding conditional variance,

$$\text{Var}\left(N_{eh} \mid E_{\text{dep}}\right) = F\frac{E_{\text{dep}}}{w_{\text{Si}}}. \tag{6.406}$$

The unconditional variance follows from the law of total variance,

$$\begin{aligned}\text{Var}(N_{eh}) = \mathbb{E}\left[\text{Var}\left(N_{eh} \mid E_{\text{dep}}\right)\right] \\ + \text{Var}\left[\mathbb{E}\left(N_{eh} \mid E_{\text{dep}}\right)\right].\end{aligned} \tag{6.407}$$

For the first term,

$$\mathbb{E}\left[\text{Var}\left(N_{eh} \mid E_{\text{dep}}\right)\right] = \mathbb{E}\left[F\frac{E_{\text{dep}}}{w_{\text{Si}}}\right] \tag{6.408}$$

$$= F\frac{\mathbb{E}[E_{\text{dep}}]}{w_{\text{Si}}}. \tag{6.409}$$

For the second term,

$$\text{Var}\left[\mathbb{E}\left(N_{eh} \mid E_{\text{dep}}\right)\right] = \text{Var}\left[\frac{E_{\text{dep}}}{w_{\text{Si}}}\right] \tag{6.410}$$

$$= \frac{\text{Var}(E_{\text{dep}})}{w_{\text{Si}}^2}. \tag{6.411}$$

Hence

$$\boxed{\text{Var}(N_{eh}) = F\frac{\mathbb{E}[E_{\text{dep}}]}{w_{\text{Si}}} + \frac{\text{Var}(E_{\text{dep}})}{w_{\text{Si}}^2}.} \tag{6.412}$$

The first term is the intrinsic Fano fluctuation associated with converting a given deposited energy into electron–hole pairs. The second arises because different primary electrons deposit different amounts of energy in the active detector volume.

For high-energy electrons traversing a thin silicon sensor, the second

term in Eq. (6.412) can be substantially larger than the first. The dominant fluctuation in the amplitude of a single-electron event may therefore originate not in the microscopic creation of electron–hole pairs, but in the stochastic sequence of interactions that determines E_{dep}.

This distinction will become important when integrating and electron-counting detectors are compared. An integrating detector retains the variable amplitude of each primary-electron event, whereas an ideal counting procedure attempts to identify each primary electron and replace that fluctuating amplitude by one standardized count.

Doping: n-type and p-type silicon. A chemically pure silicon crystal is called an *intrinsic semiconductor*. Its carrier concentration can be modified deliberately by replacing a very small fraction of the silicon atoms with impurity atoms. This process is called *doping*.

A silicon atom has four valence electrons. Consider first replacing a small fraction of silicon atoms with atoms having five valence electrons, such as phosphorus. Four of the electrons participate in the usual covalent bonding, while the fifth is relatively weakly bound and can be released into the conduction band. Such an impurity is called a *donor*.

The resulting material is called *n-type silicon*:

$$\boxed{n\text{-type silicon} \quad \Longrightarrow \quad \text{electrons are the majority mobile carriers.}} \tag{6.413}$$

Conversely, an impurity such as boron has only three valence electrons. It can accept an electron from the surrounding silicon bonds, leaving a mobile hole behind. Such an impurity is called an *acceptor*. The resulting material is called *p-type silicon*:

$$\boxed{p\text{-type silicon} \quad \Longrightarrow \quad \text{holes are the majority mobile carriers.}} \tag{6.414}$$

The letters n and p therefore identify which type of mobile carrier is more abundant. They do *not* mean that a macroscopic piece of n-type silicon has a net negative charge or that p-type silicon has a net positive charge. Away from junctions and surfaces, both materials remain approximately electrically neutral because the charge of the mobile carriers is compensated by the ionized dopants fixed in the lattice.

Detector schematics commonly use superscripts such as

$$p^-, \qquad n^+, \qquad p^{++}. \tag{6.415}$$

These symbols indicate relative *doping concentrations*, not electric charge and not electric potential. Thus, schematically,

$$\boxed{\begin{array}{lcl} p^- & : & \text{lightly doped } p\text{-type silicon,} \\ n^+ & : & \text{strongly doped } n\text{-type silicon,} \\ p^{++} & : & \text{very heavily doped } p\text{-type silicon.} \end{array}} \tag{6.416}$$

The precise dopant concentrations associated with $-$, $+$, and $++$ depend on the fabrication technology.

The p–n junction and the depletion region. We now bring a region of p-type silicon into contact with a region of n-type silicon. The resulting interface is a p–n *junction*.

Immediately after the junction is formed, the carrier concentrations are very different on the two sides. Electrons therefore diffuse from the n side toward the p side, whereas holes diffuse from the p side toward the n side.

Near the interface, electrons and holes recombine. The mobile carriers therefore disappear locally, but the dopant atoms cannot move: they are part of the crystal lattice. Ionized donor atoms of positive charge remain on the n side, while ionized acceptor atoms of negative charge remain on the p side.

The region around the interface is consequently almost devoid of mobile electrons and holes. It is called the *depletion region*:

$$\boxed{p\text{-type} \quad | \quad \underbrace{\text{very few mobile carriers}}_{\text{depletion region}} \quad | \quad n\text{-type.}} \tag{6.417}$$

The fixed positive and negative ions left inside the depletion region create an internal electric field. This field opposes further diffusion, and equilibrium is reached when the drift caused by the electric field balances the diffusion caused by the carrier-concentration gradient.

The reverse-biased p–n diode. A p–n junction equipped with electrical contacts forms a semiconductor diode. For particle detection we normally operate the junction in *reverse bias*: the n side is held at a higher electric potential than the p side.

If we choose the p side as the reference potential,

$$V_p = 0, \tag{6.418}$$

we may write the reverse bias as

$$\boxed{V_n - V_p = V_R > 0.} \tag{6.419}$$

Reverse bias increases the electrostatic barrier across the junction and widens the depletion region. The electric field points from the higher-potential n side toward the lower-potential p side,

$$\mathbf{E} = -\boldsymbol{\nabla} V. \tag{6.420}$$

An electron and a hole experience opposite forces,

$$\boxed{\mathbf{F}_e = -e\mathbf{E}, \qquad \mathbf{F}_h = +e\mathbf{E}.} \tag{6.421}$$

Consequently, electron–hole pairs created inside or sufficiently close to the depleted region are separated:

$$\boxed{\begin{array}{lcl} e^- & \longrightarrow & n\text{-type side}, \\ h^+ & \longrightarrow & p\text{-type side}. \end{array}} \tag{6.422}$$

This is the essential detector function of the diode. The diode does not create the electron–hole pairs: those pairs are created by the energy deposition described by Eqs. (6.395) and (6.398). The diode establishes an electric field that separates the generated carriers and guides them toward the appropriate electrical contacts.

Figure 6.2 summarizes the sequence from semiconductor excitation to reverse-biased charge collection.

From the diode to a monolithic active pixel sensor. We can now construct a detector pixel from the semiconductor ingredients introduced in Fig. 6.2.

A *monolithic active pixel sensor* (MAPS) integrates the sensitive silicon and the pixel readout electronics in the same semiconductor device. The essential structure is a continuous sensitive silicon volume containing a periodic array of charge-collection diodes and their associated readout circuits.

In a common MAPS architecture, the sensitive volume is a lightly doped p^- epitaxial layer. Small, more strongly doped n-type regions are formed near one surface. Each such n-type region and the surrounding p-type material constitute the p–n collection diode described above.

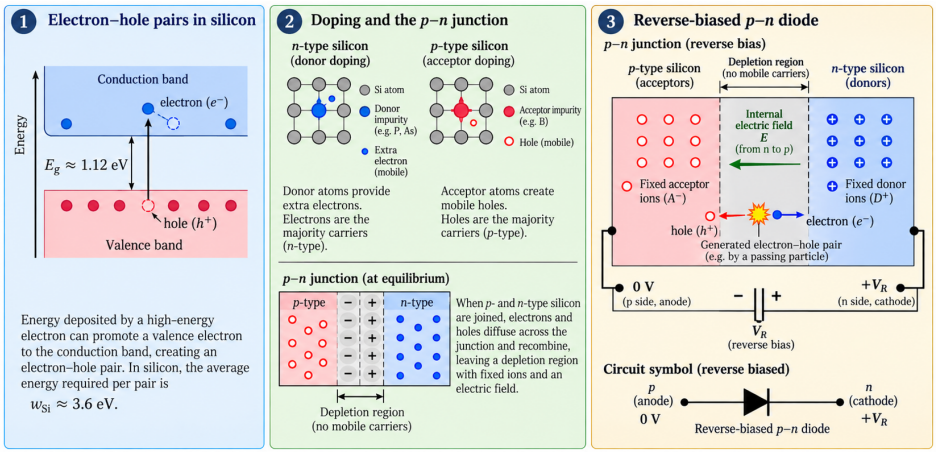


Figure 6.2: Semiconductor physics underlying charge collection in a silicon detector. (1) Energy deposited in silicon promotes an electron from the valence band to the conduction band, leaving a mobile hole and thereby creating an electron–hole pair. (2) Donor impurities produce n-type silicon, in which electrons are the majority mobile carriers, whereas acceptor impurities produce p-type silicon, in which holes are the majority carriers. When the two regions are joined, electrons and holes initially diffuse across the interface and recombine. The fixed ionized dopants that remain form a depletion region containing very few mobile carriers and generate an internal electric field. (3) Under reverse bias, the n side is held at a higher potential than the p side. The depletion region widens and electron–hole pairs created within or near it are separated by the electric field: electrons drift toward the n-type cathode and holes toward the p-type anode. The lower inset shows the corresponding reverse-biased diode circuit and the voltage convention used in the text.

The semiconductor principles summarized in Fig. 6.2 are therefore repeated across the pixel array:

$$\boxed{\begin{array}{c} p^- \text{ epitaxial silicon} + n^+ \text{ collection region} \\ \Downarrow \\ \text{reverse-biased local } p\text{–}n \text{ junction} \\ \Downarrow \\ \text{collection of signal electrons.} \end{array}} \tag{6.423}$$

Figure 6.3 shows two complementary views of this architecture.

The epitaxial layer as the sensitive interaction volume. The lightly doped epitaxial layer in Fig. 6.3 is the principal sensitive volume of the detector. It is here that the collisional stopping power derived in the preceding detector-interaction treatment produces electronic excitations and ionizations.

As the primary imaging electron traverses this volume, it transfers energy to the silicon through the processes quantified by the stopping power. The deposited part of that energy generates the electron–hole pairs described by Eq. (6.398).

The sensitive-layer thickness embodies the detector trade-off introduced in Eq. (6.36). Increasing the active thickness increases the amount of energy that can be deposited and therefore increases the generated signal. At the same time, a longer trajectory inside silicon provides more opportunity for elastic deflection, secondary-electron transport, and lateral spreading of the event.

The thin-detector calculation in Eq. (6.336) illustrates why complete stopping of the primary electron is unnecessary. Even a modest fraction of the 300 keV incident kinetic energy can generate thousands of electron–hole pairs.

The collection diode and the in-pixel circuitry. The collection diode does not itself convert the deposited energy into charge; this conversion has already occurred through the creation of electron–hole pairs. Its function is to collect a fraction of the generated mobile charge.

The n-type collection region and the surrounding p-type epitaxial silicon form the reverse-biased junction described by Eq. (6.419). Electrons that reach its collection region are accumulated at the corresponding electrical node.

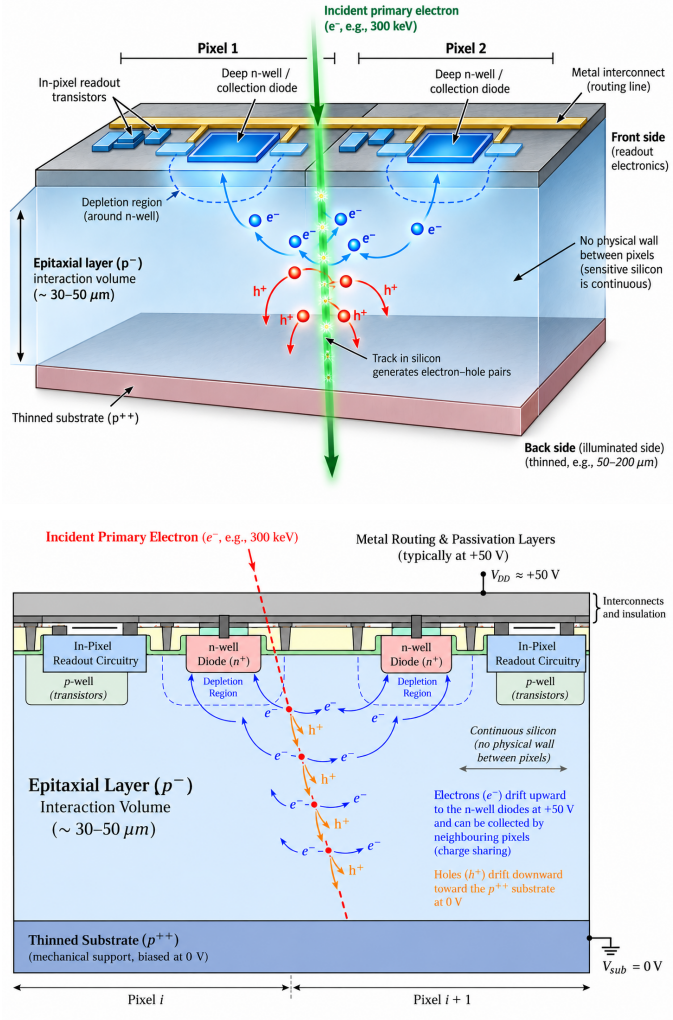


Figure 6.3: Two complementary views of a Monolithic Active Pixel Sensor (MAPS) used for direct electron detection. The first schematic shows the three-dimensional pixel arrangement, including neighbouring collection nodes, in-pixel readout circuitry, and the continuous sensitive silicon shared between adjacent pixels. The second shows a cross-section through neighbouring pixels. In a back-illuminated implementation, the high-energy imaging electron enters through the thinned back side and generates electron–hole pairs along its trajectory in the lightly doped p^- epitaxial layer. Signal electrons move by drift and diffusion toward nearby reverse-biased n-type collection diodes, whereas holes move toward the p-type side. Since the sensitive silicon contains no physical walls between pixels, carriers generated by one incident electron may reach more than one collection diode, producing charge sharing.

A defining feature of a MAPS architecture is that electronic circuitry for sensing and reading this collected charge is fabricated close to each collection node. The detector can therefore convert the local collected charge into an electrical signal without transporting the entire charge packet over macroscopic distances across the sensor, as occurs in a conventional charge-coupled device.

The detailed relation between collected charge, node voltage, frame integration, and electronic reset will be developed in Sec. 6.4.4.

Back-thinning and spatial resolution. The silicon wafer used during semiconductor fabrication is generally much thicker than the active region required for electron detection. If a large thickness of passive silicon remains in the path of a high-energy imaging electron, the primary electron can undergo additional elastic and inelastic interactions outside the useful sensitive region. In particular, electrons scattered through large angles may return to the active volume at positions far from their original point of incidence.

For detector architectures designed for back illumination, the substrate is therefore thinned substantially after fabrication. The imaging electron then enters through the thinned back side rather than first crossing the front-side metal routing and readout structures.

Back-thinning reduces the amount of passive material available for multiple scattering and backscattering. This improves spatial localization while retaining enough active silicon to generate a strong electronic signal. The quantitative importance of this trade-off is already suggested by Eq. (6.336), and its spatial consequences are developed in Sec. 6.4.1.

6.4.3 Spatial spreading from energy deposition to collected charge

The preceding subsections have described three ingredients that must now be brought together. The primary electron follows a stochastic trajectory through the silicon (Sec. 6.4.1); energetic secondary electrons can carry part of the transferred energy away from that trajectory (Sec. 6.4.1); and the deposited energy is converted into mobile electron–hole pairs that are transported toward the collection diodes (Sec. 6.4.2).

These processes act at different physical stages, but all contribute to the

spatial response of the detector. It is useful to distinguish

$$\boxed{\begin{array}{c}\text{incident electron at } \boldsymbol{\rho}_0 \\ \Downarrow \\ \text{primary and secondary electron trajectories} \\ \Downarrow \\ \text{spatial distribution of deposited energy} \\ \Downarrow \\ \text{electron–hole pairs} \\ \Downarrow \\ \text{drift and diffusion of signal carriers} \\ \Downarrow \\ \text{spatial distribution of collected charge.}\end{array}} \tag{6.424}$$

The first two stages are determined by the transport of energetic electrons, whereas the last stage is determined by semiconductor charge transport. They should not be confused. In particular, the lateral extent of the deposited energy is not the same quantity as the lateral extent of the charge finally collected by the pixel array.

From electron trajectories to deposited energy. The stopping-power treatment developed earlier quantified how much energy the primary electron loses while traversing the detector. In particular, Eq. (6.333) gave

$$\langle E_{\text{loss}} \rangle \simeq \int S\big(K(s)\big)\, ds, \tag{6.425}$$

and Eq. (6.334) reduced this result to $S(K_0)t$ for a sufficiently thin, approximately straight sensor.

We subsequently distinguished this primary-electron energy loss from the energy that actually remains inside the active detector volume. The deposited energy E_{dep} was defined in Eq. (6.342), with

$$E_{\text{dep}} \leq E_{\text{loss}} \tag{6.426}$$

as established in Eq. (6.343). The inequality arises because energetic secondary electrons or photons can carry part of the transferred energy out of the active volume.

For localization, however, even the scalar quantity E_{dep} is insufficient. We must now refine it spatially. For one microscopic event history ω, and for a primary electron entering at transverse coordinate $\boldsymbol{\rho}_0$, define

$$\varepsilon_{\text{dep}}\left(\mathbf{r} \mid \boldsymbol{\rho}_0, \omega\right) \tag{6.427}$$

as the energy deposited per unit volume at position $\mathbf{r}$.

Its integral over the active detector volume must reproduce the deposited energy introduced in Eq. (6.342):

$$\boxed{E_{\mathrm{dep}}(\omega) = \int_{V_{\mathrm{act}}} \varepsilon_{\mathrm{dep}}\left(\mathbf{r} \mid \boldsymbol{\rho}_0, \omega\right) d^3\mathbf{r}.} \tag{6.428}$$

Since detector images are two-dimensional, it is convenient to project the deposited energy along the detector depth,

$$\boxed{\varepsilon_\perp\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) = \int_{\mathrm{active}} \varepsilon_{\mathrm{dep}}\left(\boldsymbol{\rho}, z \mid \boldsymbol{\rho}_0, \omega\right) dz.} \tag{6.429}$$

By construction,

$$\int \varepsilon_\perp\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) d^2\boldsymbol{\rho} = E_{\mathrm{dep}}(\omega). \tag{6.430}$$

The function $\varepsilon_\perp$ contains two spatial properties that are useful to distinguish. The first is the position of its energy-weighted centroid,

$$\boxed{\boldsymbol{\rho}_E(\omega) = \frac{1}{E_{\mathrm{dep}}(\omega)} \int \boldsymbol{\rho}\, \varepsilon_\perp\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) d^2\boldsymbol{\rho}.} \tag{6.431}$$

The displacement

$$\boldsymbol{\rho}_E(\omega) - \boldsymbol{\rho}_0 \tag{6.432}$$

measures how far the centre of the deposited energy has moved away from the true entrance position.

The second quantity is the width of the deposited-energy distribution around its own centroid. We define its radial second moment as

$$\boxed{\sigma_E^2(\omega) = \frac{1}{E_{\mathrm{dep}}(\omega)} \int \left|\boldsymbol{\rho} - \boldsymbol{\rho}_E\right|^2 \varepsilon_\perp\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) d^2\boldsymbol{\rho}.} \tag{6.433}$$

These two effects are fundamentally different. A compact deposition event may have a centroid far from the true entrance coordinate after a large elastic deflection or a long-range secondary-electron excursion. Conversely, a broad event may remain approximately centred on the entrance coordinate.

Centroid displacement and internal event width.

Consider the energy-weighted mean squared distance from the true

entrance coordinate,

$$R_E^2 = \frac{1}{E_{\text{dep}}} \int |\boldsymbol{\rho} - \boldsymbol{\rho}_0|^2 \, \varepsilon_\perp \, d^2\boldsymbol{\rho}. \tag{6.434}$$

Write

$$\boldsymbol{\rho} - \boldsymbol{\rho}_0 = (\boldsymbol{\rho} - \boldsymbol{\rho}_E) + (\boldsymbol{\rho}_E - \boldsymbol{\rho}_0) \,. \tag{6.435}$$

Squaring,

$$\begin{aligned} |\boldsymbol{\rho} - \boldsymbol{\rho}_0|^2 = \; & |\boldsymbol{\rho} - \boldsymbol{\rho}_E|^2 + |\boldsymbol{\rho}_E - \boldsymbol{\rho}_0|^2 \\ & + 2\,(\boldsymbol{\rho} - \boldsymbol{\rho}_E) \cdot (\boldsymbol{\rho}_E - \boldsymbol{\rho}_0) \,. \end{aligned} \tag{6.436}$$

The last term vanishes after energy weighting because Eq. (6.431) implies

$$\int (\boldsymbol{\rho} - \boldsymbol{\rho}_E) \, \varepsilon_\perp \, d^2\boldsymbol{\rho} = 0. \tag{6.437}$$

Therefore,

$$\boxed{R_E^2 = |\boldsymbol{\rho}_E - \boldsymbol{\rho}_0|^2 + \sigma_E^2.} \tag{6.438}$$

Thus the mean squared lateral distance of the deposited energy from the true entrance coordinate contains two terms: displacement of the event centroid and spreading of the energy around that centroid.

It is worth emphasizing that σ_E^2 in Eq. (6.433) is a *two-dimensional radial second moment*. Later we shall introduce σ_D^2 for carrier diffusion, where σ_D^2 denotes the variance in each individual transverse coordinate. The two conventions must therefore not be compared without the corresponding factor of two.

Depth and stochasticity of energy deposition. The depth at which the energy transfers occur is also important. Equation (6.333) gives the mean energy removed from the primary trajectory,

$$\langle E_{\text{loss}} \rangle \simeq \int S(K(s)) \, ds, \tag{6.439}$$

but this continuous expression hides the microscopic structure of the event. Actual energy transfers occur at discrete and random positions along the primary trajectory.

Furthermore, the position at which energy is transferred away from the primary electron need not be the position at which it is eventually deposited.

As discussed in Sec. 6.4.1, a hard collision can produce an energetic secondary electron, which may propagate for an appreciable distance before transferring its energy to lower-energy electronic excitations.

The thin-detector estimate of Eq. (6.336) provides a useful scale. A 300 keV electron traversing 35 μm of silicon loses, in the straight-track approximation, only about

$$15.5 \text{ keV}. \tag{6.440}$$

The primary electron is therefore normally a through-going particle rather than an electron that comes to rest inside the active layer. The detector signal is generated by the comparatively small fraction of its energy transferred at stochastic positions during its passage.

The event-to-event energy fluctuations discussed in Eq. (6.340) consequently have a spatial counterpart. Even for electrons entering at the same $\boldsymbol{\rho}_0$, the functions

$$\varepsilon_{\text{dep}}\left(\mathbf{r} \mid \boldsymbol{\rho}_0, \omega\right) \tag{6.441}$$

are different. Their total deposited energies, centroids, widths, and detailed shapes all fluctuate from electron to electron.

Monte Carlo calculations for back-thinned silicon detectors show why the detector geometry matters. When long-range backscattering from a thick substrate is suppressed, the intrinsic lateral spread of deposited energy from a 300 keV electron can be only a few micrometres (McMullan et al., 2009a). This is already comparable with the physical pixel pitches used in modern direct electron detectors.

From deposited energy to mobile charge. The deposited-energy distribution is still not the quantity measured by the electronic readout. As described in Sec. 6.4.2, energy deposited in silicon creates electron–hole pairs. The mean pair-creation energy w_{Si} was defined in Eq. (6.397), and the corresponding mean number of generated pairs was obtained in Eq. (6.398).

In particular, the example based on the 15.5 keV mean energy loss of Eq. (6.336) was already shown in Sec. 6.4.2 to correspond to roughly 4.3×10^3 electron–hole pairs if that energy is fully deposited inside the active silicon. The distinction between primary-electron energy loss and energy actually deposited in the active volume is given by Eq. (6.343).

The conversion from deposited energy to charge is itself stochastic. Its intrinsic Fano fluctuations were quantified in Eq. (6.401), while the combined contribution of pair-generation statistics and fluctuations in E_{dep} was derived in Eq. (6.412).

Here we therefore do not reconsider how much charge is generated. Instead, we ask where that already-generated charge is collected. Once created, the electron–hole pairs must travel through the silicon toward the collection electrodes, and this transport introduces an additional source of spatial spreading.

Drift and diffusion of signal carriers. Once generated, the low-energy charge carriers undergo two simultaneous types of transport. The electric field inside the sensor produces a directed *drift*, while thermal motion produces random *diffusion*. These are physically different from the elastic and inelastic trajectories of the high-energy primary and secondary electrons described in Secs. 6.4.1 and 6.4.1.

A continuum description of one carrier species is provided by the drift–diffusion equation (Spieler, 2005),

$$\boxed{\frac{\partial n}{\partial t} = D\nabla^2 n - \nabla \cdot (n\mathbf{v}_{\mathrm{d}})\,,} \tag{6.442}$$

where $n(\mathbf{r}, t)$ is the carrier number density, D is the diffusion coefficient, and $\mathbf{v}_{\mathrm{d}}$ is the drift velocity.

In a sufficiently weak electric field, the mean drift velocity is proportional to the local electric field. For electrons and holes, respectively,

$$\boxed{\mathbf{v}_{\mathrm{d},e} = -\mu_e \mathbf{E}, \qquad \mathbf{v}_{\mathrm{d},h} = +\mu_h \mathbf{E},} \tag{6.443}$$

where $\mu_e > 0$ and $\mu_h > 0$ are the electron and hole mobilities. Consequently, for either carrier species,

$$\boxed{v_{\mathrm{d}} = \mu\mathcal{E},} \tag{6.444}$$

where $v_{\mathrm{d}} = |\mathbf{v}_{\mathrm{d}}|$ and $\mathcal{E} = |\mathbf{E}|$.

The opposite signs in Eq. (6.443) follow from the opposite carrier charges: electrons drift against the electric field, whereas holes drift along it, consistently with the reverse-biased p–n junction described in Sec. 6.4.2. The proportionality in Eq. (6.444) can be understood from a simple relaxation-time model.

Carrier drift velocity in the low-field regime.

A charge carrier in a semiconductor does not accelerate indefinitely under an electric field. Between scattering events it is accelerated by the field, but collisions with phonons, impurities, and other lattice excita-

tions continually randomize its momentum.

At the semiclassical level, the motion of a carrier in an approximately isotropic energy band can be described using an effective mass m^*. If the carrier has charge q, the electric force is

$$\mathbf{F} = q\mathbf{E}. \tag{6.445}$$

For an electron,

$$q = -e, \tag{6.446}$$

whereas for a hole,

$$q = +e. \tag{6.447}$$

Let $\mathbf{v}_\mathrm{d}$ denote the ensemble-averaged velocity of the carriers. In the absence of an electric field, the microscopic thermal velocities have random directions and therefore average to zero. The electric field introduces a small non-zero mean velocity.

A simple way of describing the effect of collisions is the *relaxation-time approximation*. If τ is the mean momentum relaxation time, scattering tends to return the average momentum toward its equilibrium value on the time scale τ. The equation for the mean carrier velocity is then

$$\boxed{m^* \frac{d\mathbf{v}_\mathrm{d}}{dt} = q\mathbf{E} - \frac{m^*\mathbf{v}_\mathrm{d}}{\tau}.} \tag{6.448}$$

The first term on the right-hand side is the acceleration produced by the electric field. The second represents the average loss of directed momentum caused by scattering.

After a transient of order τ, a stationary drift state is reached. Then

$$\frac{d\mathbf{v}_\mathrm{d}}{dt} = 0, \tag{6.449}$$

and Eq. (6.448) gives

$$q\mathbf{E} - \frac{m^*\mathbf{v}_\mathrm{d}}{\tau} = 0. \tag{6.450}$$

Therefore,

$$\boxed{\mathbf{v}_\mathrm{d} = \frac{q\tau}{m^*}\mathbf{E}.} \tag{6.451}$$

Mobility is defined as the positive quantity

$$\boxed{\mu = \frac{|q|\tau}{m^*}.} \tag{6.452}$$

Since $|q| = e$ for both electrons and holes, the corresponding relations are

$$\mathbf{v}_{\mathrm{d},e} = -\mu_e \mathbf{E} \tag{6.453}$$

for electrons and

$$\mathbf{v}_{\mathrm{d},h} = +\mu_h \mathbf{E} \tag{6.454}$$

for holes, proving Eq. (6.443).

Taking magnitudes gives

$$\boxed{v_{\mathrm{d}} = \mu \mathcal{E},} \tag{6.455}$$

which proves Eq. (6.444).

The linear relation Eq. (6.444) is a low-field result. Its derivation assumes that the applied field produces only a small perturbation of the equilibrium carrier distribution, so that the effective momentum-relaxation time and hence the mobility can be treated as approximately independent of the field. At sufficiently high fields, the carriers gain appreciable energy between collisions, the scattering rates change, and the drift velocity no longer grows linearly with $\mathcal{E}$; in silicon it eventually approaches a saturation velocity. For the diffusion argument below, we shall remain in the regime in which Eq. (6.444) is a suitable local approximation.

If a localized carrier packet is allowed to diffuse for a collection time t_{col}, its transverse distribution is Gaussian in the simplest homogeneous model,

$$\boxed{G_D\left(\Delta\boldsymbol{\rho}; t_{\mathrm{col}}\right) = \frac{1}{2\pi\sigma_D^2} \exp\left[-\frac{|\Delta\boldsymbol{\rho}|^2}{2\sigma_D^2}\right],} \tag{6.456}$$

where

$$\boxed{\sigma_D^2 = 2Dt_{\mathrm{col}}.} \tag{6.457}$$

Here σ_D^2 is the variance in *each* transverse coordinate.

For non-degenerate carriers sufficiently close to thermal equilibrium, the diffusion coefficient and mobility are related by the Einstein relation,

$$\boxed{D = \mu \frac{k_{\mathrm{B}}T}{e}.} \tag{6.458}$$

Suppose, for the moment, that a carrier is generated a distance ℓ from its collection electrode and travels through an approximately uniform electric field of magnitude $\mathcal{E}$. Its collection time is approximately

$$t_{\rm col} \simeq \frac{\ell}{\mu\mathcal{E}}. \tag{6.459}$$

Combining this with Eqs. (6.457) and (6.458) gives

$$\boxed{\sigma_D^2 \simeq 2\frac{k_{\rm B}T}{e}\frac{\ell}{\mathcal{E}}.} \tag{6.460}$$

The origin and interpretation of this result are sufficiently important that we derive it explicitly.

Lateral diffusion during charge collection.

Consider first a localized packet of carriers and ignore the directed motion along the electric field. We examine only the transverse coordinates

$$\boldsymbol{\rho} = (x, y). \tag{6.461}$$

For a spatially uniform diffusion coefficient D, the transverse carrier density obeys

$$\boxed{\frac{\partial n}{\partial t} = D\nabla_\perp^2 n,} \tag{6.462}$$

where

$$\nabla_\perp^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \tag{6.463}$$

Suppose that the carriers are initially localized at $\boldsymbol{\rho} = 0$. The normalized initial condition is

$$\boxed{n(\boldsymbol{\rho}, 0) = \delta^{(2)}(\boldsymbol{\rho}).} \tag{6.464}$$

Using the Fourier-transform,

$$\widehat{n}(\mathbf{R}, t) = \int n(\boldsymbol{\rho}, t)e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\rho}}d^2\boldsymbol{\rho}, \tag{6.465}$$

where

$$\mathbf{R} = (R_x, R_y), \tag{6.466}$$

the transverse derivatives transform as

$$\frac{\partial^2}{\partial x^2} \longrightarrow -(2\pi R_x)^2, \tag{6.467}$$

and

$$\frac{\partial^2}{\partial y^2} \longrightarrow -(2\pi R_y)^2. \tag{6.468}$$

Equation (6.462) therefore becomes

$$\frac{\partial \widehat{n}}{\partial t} = -4\pi^2 D \left(R_x^2 + R_y^2\right) \widehat{n}. \tag{6.469}$$

For each fixed spatial frequency $\mathbf{R}$, this is an ordinary differential equation in time. Hence

$$\widehat{n}(\mathbf{R}, t) = \widehat{n}(\mathbf{R}, 0) \exp\left[-4\pi^2 D|\mathbf{R}|^2 t\right]. \tag{6.470}$$

The Fourier transform of the point initial condition in Eq. (6.464) is unity,

$$\widehat{n}(\mathbf{R}, 0) = 1, \tag{6.471}$$

and therefore

$$\boxed{\widehat{n}(\mathbf{R}, t) = \exp\left[-4\pi^2 Dt|\mathbf{R}|^2\right].} \tag{6.472}$$

Taking the inverse Fourier transform gives

$$\boxed{n(\boldsymbol{\rho}, t) = \frac{1}{4\pi Dt} \exp\left[-\frac{|\boldsymbol{\rho}|^2}{4Dt}\right].} \tag{6.473}$$

The normalization follows directly:

$$\int n(\boldsymbol{\rho}, t)\, d^2\boldsymbol{\rho} = \frac{1}{4\pi Dt} \int_0^{2\pi} \int_0^{\infty} \exp\left[-\frac{\rho^2}{4Dt}\right] \rho\, d\rho\, d\phi \tag{6.474}$$

$$= 1. \tag{6.475}$$

Diffusion therefore redistributes the carriers spatially but, in this idealized model, does not change their total number.

Comparing Eq. (6.473) with the normalized two-dimensional Gaussian in Eq. (6.456),

$$G_D(\boldsymbol{\rho};t) = \frac{1}{2\pi\sigma_D^2}\exp\left[-\frac{|\boldsymbol{\rho}|^2}{2\sigma_D^2}\right], \tag{6.476}$$

gives

$$2\sigma_D^2 = 4Dt, \tag{6.477}$$

and hence

$$\boxed{\sigma_D^2 = 2Dt,} \tag{6.478}$$

which proves Eq. (6.457).

The meaning of σ_D must be stated carefully. It is the standard deviation in each transverse coordinate:

$$\left\langle x^2 \right\rangle = \left\langle y^2 \right\rangle = 2Dt. \tag{6.479}$$

The mean squared radial displacement is therefore

$$\left\langle |\Delta\boldsymbol{\rho}|^2 \right\rangle = \left\langle x^2 + y^2 \right\rangle \tag{6.480}$$

$$= 4Dt \tag{6.481}$$

$$= 2\sigma_D^2. \tag{6.482}$$

We now restore the directed drift toward the collection electrode. Using Eq. (6.444), the magnitude of the drift velocity in an approximately uniform field is

$$v_{\mathrm{d}} = \mu\mathcal{E}. \tag{6.483}$$

If the carrier must travel a distance ℓ, the corresponding collection time is

$$t_{\mathrm{col}} = \frac{\ell}{v_{\mathrm{d}}} \tag{6.484}$$

$$= \boxed{\frac{\ell}{\mu\mathcal{E}}.} \tag{6.485}$$

During this time, transverse diffusion continues. Substitution into Eq. (6.457) gives

$$\sigma_D^2 = 2D\frac{\ell}{\mu\mathcal{E}} \tag{6.486}$$

$$= 2\frac{D}{\mu}\frac{\ell}{\mathcal{E}}. \tag{6.487}$$

For non-degenerate carriers close to thermal equilibrium, Eq. (6.458) gives

$$\frac{D}{\mu} = \frac{k_\mathrm{B}T}{e}. \tag{6.488}$$

Thus,

$$\boxed{\sigma_D^2 = 2\frac{k_\mathrm{B}T}{e}\frac{\ell}{\mathcal{E}},} \tag{6.489}$$

which proves Eq. (6.460).

An interesting feature of this result is that the mobility μ has cancelled. Through the Einstein relation, a larger mobility implies a larger diffusion coefficient and therefore faster random spreading. The same larger mobility also causes faster drift, however, reducing the time available for diffusion. Under the assumptions of this derivation, the two effects cancel exactly.

The result therefore has the simple scaling

$$\boxed{\sigma_D^2 \propto T, \qquad \sigma_D^2 \propto \ell, \qquad \sigma_D^2 \propto \frac{1}{\mathcal{E}}.} \tag{6.490}$$

Higher temperature increases thermal spreading; charge generated farther from the collection electrode has longer to diffuse; and a stronger electric field shortens the collection time and reduces the lateral spread.

The uniform-field approximation is not essential. If the carrier follows a drift path along which the field magnitude varies with position, then

$$dt = \frac{d\ell}{\mu\mathcal{E}(\ell)}. \tag{6.491}$$

Combining this relation with Eq. (6.457) and the Einstein relation gives the more general result

$$\boxed{\sigma_D^2 = 2\frac{k_\mathrm{B}T}{e}\int_\mathrm{path}\frac{d\ell}{\mathcal{E}(\ell)}.} \tag{6.492}$$

Regions of weak electric field therefore contribute disproportionately to the accumulated diffusion.

Dependence on detector architecture. The relative importance of drift and diffusion depends strongly on the detector architecture introduced in Sec. 6.4.2.

In a fully depleted sensor, the electric field extends through most or all of the active silicon. Charge carriers are then transported toward the collection electrodes predominantly by drift. Thermal diffusion still acts during this transport, but the relatively short collection time limits its contribution to the final lateral spread.

In a conventional MAPS architecture, by contrast, much of the lightly doped epitaxial layer may be only weakly depleted or essentially undepleted. Away from the immediate depletion region surrounding the collection diode, the electric field can then be small. Carriers generated there must reach the collection region largely by thermal diffusion before the stronger junction field takes over. Their collection times are consequently longer, and diffusion can become a major contribution to the spatial extent of the event.

Detector thickness by itself is therefore insufficient to predict the charge-cloud width. The relevant quantities include the depth at which the charge is generated, the electric-field profile, the degree of depletion, the temperature, the carrier mobility, and the geometry and spacing of the collection electrodes. In particular, two detectors with similar silicon thicknesses can exhibit substantially different charge sharing if their internal electric fields are different.

From deposited energy to collected charge. We can now combine the energetic-electron stage and the semiconductor transport stage into a single mathematical expression.

Let

$$G_{\mathrm{c}}\left(\boldsymbol{\rho};\mathbf{r}\right) \tag{6.493}$$

be the normalized collection kernel for charge generated at the three-dimensional point $\mathbf{r}$. Thus

$$\int G_{\mathrm{c}}\left(\boldsymbol{\rho};\mathbf{r}\right) d^2\boldsymbol{\rho} = 1. \tag{6.494}$$

The function G_{c} describes the distribution of that charge over the readout plane after drift, diffusion, and the detailed collection geometry have acted.

Let

$$\eta_{\mathrm{c}}(\mathbf{r}) \tag{6.495}$$

denote the fraction of the generated charge at $\mathbf{r}$ that is ultimately collected. Using the mean pair-creation energy w_{Si} from Eq. (6.397), the magnitude of the collected electron-charge density produced by one microscopic event is then

$$\boxed{\begin{aligned}\mathcal{Q}\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) = \frac{e}{w_{\mathrm{Si}}} \int_{V_{\mathrm{act}}} \eta_{\mathrm{c}}(\mathbf{r})\, \varepsilon_{\mathrm{dep}}\left(\mathbf{r} \mid \boldsymbol{\rho}_0, \omega\right) \\ \times G_{\mathrm{c}}\left(\boldsymbol{\rho}; \mathbf{r}\right)\, d^3\mathbf{r}.\end{aligned}} \tag{6.496}$$

Equation (6.496) is the central result of this subsection. It separates two physically different sources of spatial spreading. The deposited-energy distribution $\varepsilon_{\mathrm{dep}}$ contains the stochastic trajectories of the primary and secondary high-energy electrons. The collection kernel G_{c} contains the subsequent transport of the low-energy semiconductor carriers.

The final charge cloud is therefore neither the primary-electron track itself nor the response obtained by applying a simple diffusion kernel to the entrance coordinate. It is the superposition of charge generated throughout a spatially distributed three-dimensional energy-deposition event, with the charge created at each point undergoing its own depth-dependent transport before collection.

Experimental scales in direct electron detectors. Published measurements provide useful scales for the spreading mechanisms described above. A single universal diffusion width cannot be assigned to modern cryo-EM cameras because it depends on the sensitive-layer thickness, doping profile, electric-field distribution, collection geometry, and operating conditions. Moreover, the detailed internal structure of commercial detectors is not always published.

Measurements on monolithic CMOS sensors developed specifically for transmission electron microscopy nevertheless show that carrier spreading can be comparable with the physical pixel pitch. For the back-thinned TEAM detector with 9.5 μm pixels, Battaglia et al. (2010) fitted the measured charge-sharing behaviour using the PixelSim transport model and obtained the diffusion parameter

$$\boxed{\sigma_{\mathrm{diff}} = 14.5^{+2.0}_{-1.0}\ \mu\mathrm{m}.} \tag{6.497}$$

The parameter σ_{diff} is the charge-spreading parameter of that particular simulation model. It is closely related to the diffusion width σ_D defined in Eq. (6.457), but the two should not be identified exactly without matching the detailed conventions of the transport models.

The same detector illustrates that several spatial widths can coexist. For 300 keV electrons, the measured line-spread scale was approximately

$$7.4 \pm 0.6\ \mu\text{m}, \tag{6.498}$$

whereas event-by-event reconstruction reduced the corresponding effective scale to approximately

$$2.4 \pm 0.2\ \mu\text{m}. \tag{6.499}$$

Thus a carrier cloud can spread over a distance comparable with or larger than the physical pixel pitch—the centre-to-centre distance between neighbouring pixels—while still containing sufficient information to determine its centre substantially more precisely.

Commercial cryo-EM cameras display the same qualitative dependence on sensor architecture. In their comparison at 300 keV, McMullan et al. (2014) reported physical pixel pitches of approximately 6.4 μm for the DE-20, 14 μm for the Falcon II, and 5.0 μm for the K2 Summit. Despite its larger pixels, the Falcon II exhibited a comparatively low MTF, which the authors associated with the larger lateral spread of charge in its sensitive layer. They also noted the important role of carrier diffusion in the spatial distribution of its single-electron events.

The thickness of the active detector is therefore part of a broader architectural trade-off. Thinning reduces long-range trajectories due to multiple elastic scattering and backscattering, while the thickness and electric-field structure of the sensitive region determine how far the generated carriers must travel before collection. For the Falcon family, Kuijper et al. (2015) reported sensor thicknesses of approximately 50 μm for Falcon 2 and 30 μm for Falcon 3.

These observations place the relevant spatial scales in perspective. For high-energy electrons in direct silicon detectors, the lateral spread produced by the energetic-electron trajectories can be of order a few micrometres in a well-thinned sensor, while subsequent semiconductor carrier transport can spread the collected charge over several to tens of micrometres. Both scales are comparable with the physical pixel pitch of current detectors.

It is therefore important to distinguish three quantities. The *pixel pitch* determines the discrete sampling grid. The *event width* describes the spatial region over which the signal from one primary electron is distributed. The *localization precision* describes how accurately the original entrance coordinate can ultimately be inferred from that distributed event. The last of these will be formulated statistically in Sec. 6.4.4.

At this point in the detection chain, the incident high-energy electron has therefore been converted into a spatially distributed cloud of collected charge. The next step is to understand how that charge is stored at the pixel collection nodes and converted into a measurable electronic value.

6.4.4 Pixel readout and localization of electron events

The preceding discussion followed the physical response of the detector up to the point at which the charge generated by one incident electron has reached the collection plane. In particular, Eq. (6.496) described the continuous charge distribution

$$\mathcal{Q}\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) \tag{6.500}$$

produced by one microscopic realization ω of an electron entering the detector at transverse coordinate $\boldsymbol{\rho}_0$.

The remaining task is performed by the pixel array. The collected charge must first be converted into an electrical signal; that signal is sampled by discrete collection nodes; and, in electron-counting operation, the resulting pattern of pixel values can be interpreted to estimate the original electron position.

It is important to keep these stages conceptually distinct. The sensitive silicon is largely continuous, whereas the electronic readout is discrete. A physical electron event can therefore extend across several readout pixels even though it originated from only one incident electron.

Charge-to-voltage conversion, digitization, integration, and reset. As illustrated in Fig. 6.3, the detector contains a periodic array of collection nodes with associated electronic circuitry. The collection node and its associated readout region define one pixel, but the boundary between neighbouring pixels is not a physical wall in the sensitive silicon.

Once signal electrons reach a collection node, their charge is converted into a voltage. Let C_{node} denote the effective capacitance of the sensing node. If the magnitude of the accumulated charge changes by Q, the corresponding magnitude of the voltage change is

$$\boxed{|\Delta V| = \frac{Q}{C_{\text{node}}}.} \tag{6.501}$$

For the n-type collection node described in Sec. 6.4.2, the collected signal carriers are electrons. If N_{col} electrons arrive at a floating sensing node, the change in charge is

$$\Delta Q = -eN_{\text{col}}, \tag{6.502}$$

and therefore

$$\boxed{\Delta V = -\frac{eN_{\mathrm{col}}}{C_{\mathrm{node}}}.} \tag{6.503}$$

In this simple circuit picture, collection of electrons lowers the voltage of the sensing node relative to its reset value. Other front-end architectures may use different electrical conventions, but the essential result is unchanged: the first electronic representation of the detector signal is an analog voltage proportional to the collected charge.

The number of collected electrons is ultimately connected to the deposited energy through the electron–hole-pair generation described in Sec. 6.4.2. If a fraction η_{c} of the generated signal electrons reaches a particular collection node, then

$$N_{\mathrm{col}} = \eta_{\mathrm{c}} N_{eh}. \tag{6.504}$$

Using Eq. (6.398), the magnitude of the corresponding voltage signal is approximately

$$\boxed{|\Delta V| \simeq \frac{e\eta_{\mathrm{c}}}{C_{\mathrm{node}} w_{\mathrm{Si}}} E_{\mathrm{dep}}.} \tag{6.505}$$

Equation (6.505) therefore completes the analog conversion from deposited energy to a voltage at one collection node.

From an analog voltage to a digital pixel value. The voltage in Eq. (6.505) is still a continuous electrical quantity. Before it can be stored or processed numerically it must be sampled and converted by an *analog-to-digital converter* (ADC).

Suppose that, after any amplification and offset correction, the ADC accepts an input range V_{FS} and represents it using b bits. The nominal voltage corresponding to one digital step, or *least-significant bit* (LSB), is

$$\boxed{\Delta V_{\mathrm{ADC}} = \frac{V_{\mathrm{FS}}}{2^b}.} \tag{6.506}$$

A measured voltage is consequently represented by an integer digital code. In a simple idealized model,

$$D_p \simeq \mathrm{round}\left(\frac{V_p - V_{\mathrm{off}}}{\Delta V_{\mathrm{ADC}}}\right), \tag{6.507}$$

where V_{off} denotes the electronic baseline and D_p is the digital value associated with pixel p.

The quantity ΔV_{ADC} is therefore one kind of "quantum" in the electronic measurement, but it should not be confused with one incident electron. A single 300 keV primary electron produces thousands of electron–hole pairs, as quantified in Sec. 6.4.2, and that charge is usually distributed over several neighbouring collection nodes. One primary electron therefore corresponds to a spatial pattern of analog voltages, and subsequently to a cluster of digital pixel values, rather than to one ADC count in one pixel.

The ADC range, amplifier gain, and bit depth are chosen so that the digital step is sufficiently smaller than the relevant electronic noise and single-electron signal variations, while the full dynamic range remains large enough to avoid saturation. A calibration then relates the digital output units to collected charge or, equivalently, to the response expected from incident electrons.

Integration, reset, and correlated double sampling. After the collected charge has been converted into a voltage, the detector must decide over what time interval that signal is accumulated and how the sensing node is prepared for the next measurement.

In a simple integrating pixel, the sensing node is first brought to a known electrical state and then allowed to accumulate charge during an integration interval T_{int}. If the kth incident-electron event contributes a charge magnitude Q_k to the node, the total accumulated signal is

$$\boxed{Q_{\mathrm{int}} = \sum_{k=1}^{N_{\mathrm{events}}} Q_k.} \tag{6.508}$$

The corresponding voltage excursion is

$$|\Delta V_{\mathrm{int}}| = \frac{Q_{\mathrm{int}}}{C_{\mathrm{node}}}, \tag{6.509}$$

using the charge-to-voltage relation of Eq. (6.501).

The accumulated charge must eventually be removed so that a new measurement can begin. In a simple CMOS active-pixel circuit, a *reset transistor* temporarily connects the sensing node to a reference potential V_{reset}. Charge then flows between the node and the external circuit until

$$V_{\mathrm{node}} \longrightarrow V_{\mathrm{reset}}. \tag{6.510}$$

When the reset transistor is opened again, the sensing node becomes approximately floating and can accumulate the charge generated during the next integration interval.

The phrase "reset after each frame" must be interpreted with some care. In a large CMOS detector, all pixels need not be reset simultaneously. Many architectures use a *rolling* sequence in which rows, groups of rows, or other portions of the array are sampled and reset at different times. What matters physically is that every sensing node undergoes a well-defined sequence of baseline establishment, charge integration, signal sampling, and restoration.

Reset itself introduces noise. Thermal fluctuations in the reset process, transistor offsets, and small differences between pixel circuits mean that the node does not return to exactly the same measured voltage on every cycle. A common way of reducing these contributions is *correlated double sampling* (CDS). The node is measured once near the beginning of the integration interval and again after charge has been accumulated. If V_{r} denotes the reference or reset measurement and V_{s} the signal measurement, one forms

$$\boxed{\Delta V_{\mathrm{CDS}} = V_{\mathrm{r}} - V_{\mathrm{s}}.} \tag{6.511}$$

For the electron-collecting convention of Eq. (6.503), this difference is positive when signal electrons have been collected.

Because several electronic offsets are common to the two samples, their subtraction suppresses part of the reset noise and fixed-pattern variation. CDS therefore improves the precision with which the relatively small signal-induced voltage change is measured, although obtaining two samples can reduce the effective maximum readout rate in architectures where the two measurements consume separate readout cycles.

Integrating and electron-counting operation. The same physical sensor can use the charge collected during these short acquisition intervals in two conceptually different ways.

In an *integrating detector*, all charge arriving during the acquisition interval contributes directly to the numerical pixel values. Individual primary electrons are not explicitly identified. If several events contribute to the same node, their signals add according to Eq. (6.508). Consequently, the event-to-event fluctuations in deposited energy and generated charge derived in Eq. (6.412) remain present in the measured amplitudes.

In an *electron-counting detector*, by contrast, the detector is sampled rapidly enough, and the incident electron fluence per sampling interval is kept sufficiently low, that the signal generated by an individual primary electron can be recognized as a distinct event. Because the charge generated by one electron generally spreads over several neighbouring pixels, one primary electron does not normally correspond to one non-zero pixel. It produces a spatial cluster of digitized pixel values.

Electron counting therefore introduces a second discretization beyond the analog-to-digital conversion discussed above. The ADC converts an analog voltage into a digital number; it does not by itself decide that one incident electron has arrived. The counting algorithm must identify a characteristic local pattern of digital values and classify that pattern as an electron event.

Let D_p denote the baseline-subtracted digitized signal in pixel p. Electronic noise alone produces small fluctuations around zero, whereas an incident electron normally produces a larger correlated cluster. A detection threshold D_{th}, together with spatial cluster-analysis rules, is therefore used schematically as

$$\boxed{\text{digitized local pixel pattern} \xrightarrow{\text{threshold and cluster analysis}} \begin{cases} \text{no electron event,} \\ \text{one electron event.} \end{cases}} \tag{6.512}$$

The threshold is not a universal number of electron–hole pairs or ADC units. It is determined from detector calibration by comparing the electronic-noise distribution with the measured single-electron signal distribution. A threshold that is too low produces false events from noise, whereas one that is too high rejects genuine low-amplitude electron events.

Once a cluster has been identified, its total amplitude need no longer be retained as the measure of the incident electron. The event can be standardized conceptually to one count. This is an important advantage of counting: the large event-to-event fluctuations in deposited energy described by Eq. (6.412) affect the analog signal amplitude, but a reliably identified primary-electron event can subsequently be represented simply by the number one.

The spatial structure of the cluster is not discarded, however. The relative values of neighbouring pixels can be used to estimate the entrance coordinate of the electron with subpixel precision, as discussed below.

Counting also introduces an important limitation. If two primary electrons arrive sufficiently close together in space during the same internal sampling interval, their charge clusters can overlap. The event-recognition algorithm may then interpret the two electrons as a single event. This produces *coincidence loss*. The probability of such overlap decreases when the internal sampling interval is shortened, which is one of the principal reasons that counting detectors operate at very high internal frame rates.

Internal frame rate and parallel readout. The *internal frame rate* of a counting detector specifies how frequently the sensor is sampled for the

purpose of separating individual electron events. It should not be confused with the much lower rate at which movie frames are normally delivered to the acquisition computer.

Typical internal sampling times are of order milliseconds or less. The K2 Summit, for example, operates internally at 400 frames s^{-1}, corresponding to

$$T_{\mathrm{int}} = \frac{1}{400\ \mathrm{s}^{-1}} = 2.5\ \mathrm{ms} \tag{6.513}$$

per internal frame (McMullan et al., 2014).

The K3 increases the internal rate to 1500 frames s^{-1}, giving

$$\boxed{T_{\mathrm{int}} = \frac{1}{1500\ \mathrm{s}^{-1}} \simeq 0.67\ \mathrm{ms}.} \tag{6.514}$$

When the K3 is operated in its correlated-double-sampling mode described above, two samples are required for each effective measurement and the effective rate is approximately 750 frames s^{-1}, corresponding to about 1.33 ms.

The Falcon 4 provides an instructive comparison. Its internal rate is approximately 250 frames s^{-1}, corresponding to about 4 ms per internal frame. A higher internal frame rate is therefore not, by itself, a complete measure of counting performance: pixel geometry, event size, detector noise, event-detection algorithms, and the maximum acceptable coincidence loss are also important.

These internal frames are normally much faster than the image frames presented to the user. The K3, for example, contains approximately 5760×4092 pixels—more than twenty million sensing elements—and performs its internal event detection at 1500 frames s^{-1}, while conventional full frames are delivered to the acquisition system at a much lower rate. Many internal sampling intervals can therefore contribute to one stored movie frame.

Such rates would be impossible if all pixels were read one after another through a single amplifier and a single ADC. A 4096×4096 detector already contains nearly seventeen million pixels. Direct-electron cameras instead exploit extensive parallelism.

All pixels collect charge concurrently during their respective integration intervals. During readout, rows or groups of rows may be addressed sequentially, while many columns are sensed simultaneously. Multiple column amplifiers, ADCs, output channels, and readout blocks operate in parallel across the detector. Thus the architecture is typically neither completely serial nor completely parallel: some coordinates of the array are scanned sequentially, while many signals are processed simultaneously.

Further parallelism is provided by dedicated digital electronics. In an electron-counting camera, thresholding, cluster identification, event localization, and temporal accumulation can be performed close to the sensor. The enormous stream of raw pixel samples therefore need not be transmitted unchanged to the acquisition computer. Depending on the architecture, the external system may receive accumulated image frames or a compact representation of individual detected events rather than every raw internal pixel measurement.

The distinction between the internal sampling rate and the delivered movie-frame rate is consequently fundamental. The former is chosen primarily to separate electron events in time and reduce coincidence loss; the latter is chosen according to the temporal resolution and data volume required for the cryo-EM experiment.

Pixel boundaries, charge sharing, and subpixel information. The pixel array discretizes the *readout*, not the underlying sensitive silicon. Consequently, the geometrical boundary between two pixels does not confine the charge generated on either side of that boundary.

The lateral spreading derived earlier means that the charge cloud from one electron can overlap several collection regions. Different parts of the same physical event may therefore be collected by different nodes. If all of the relevant charge is contained in a cluster of pixels and Q_p denotes the measured signal-charge magnitude in pixel p, charge conservation gives

$$\boxed{Q_{\text{event}} = \sum_{p \in \text{cluster}} Q_p.} \tag{6.515}$$

The division of one physical electron event among several pixels is called *charge sharing*.

Charge sharing has two different consequences depending on how the detector is operated. In an integrating detector, spreading the signal over neighbouring pixels broadens the average spatial response and therefore reduces the transfer of high spatial frequencies. For an isolated event, however, the relative pixel amplitudes contain information about where the electron arrived. An event producing nearly equal signals in two neighbouring pixels, for example, is more likely to have occurred near their common boundary than an event whose signal is almost entirely confined to one pixel.

This distinction is the reason that an event can be broader than a pixel and nevertheless be localized more accurately than the pixel size.

To quantify the simplest alternative, suppose first that all subpixel information is discarded and that the event is assigned only to the centre of the pixel in which it occurred. Let a be the *pixel pitch*, that is, the centre-to-centre separation between neighbouring pixels. For a square pixel, if the true coordinate along one direction is uniformly distributed over $[-a/2, a/2]$, the localization variance is

$$\sigma_{\text{pix}}^2 = \frac{1}{a}\int_{-a/2}^{a/2} x^2\,dx \tag{6.516}$$

$$= \frac{a^2}{12}. \tag{6.517}$$

Hence

$$\boxed{\sigma_{\text{pix}} = \frac{a}{\sqrt{12}}.} \tag{6.518}$$

Equation (6.518) is purely a quantization result. It is not a fundamental localization limit of the detector.

If the individual pixel charges within an event are retained, they can be used to estimate a continuous position. The simplest example is the charge centre of gravity,

$$\boxed{\widehat{\boldsymbol{\rho}}_{\text{cog}} = \frac{\sum\limits_{p\in\text{event}} Q_p \boldsymbol{\rho}_p}{\sum\limits_{p\in\text{event}} Q_p},} \tag{6.519}$$

where $\boldsymbol{\rho}_p$ is the position of the centre of pixel p.

The estimator in Eq. (6.519) uses the relative amplitudes of neighbouring pixels to infer the location of the event within the discrete sampling grid. Charge-centroid reconstruction has experimentally achieved localization substantially finer than the physical pixel pitch at 300 keV (Battaglia et al., 2009), and modern counting detectors use the spatial structure of individual events to obtain subpixel, or *super-resolution*, coordinates (McMullan et al., 2014).

Thus charge sharing is neither intrinsically beneficial nor intrinsically detrimental. If the distributed signal is simply integrated, it contributes to spatial blur. If the pattern of an isolated event is resolved and interpreted, the same distribution contains information about its subpixel entrance position.

Statistical description of a recorded electron event. The preceding discussion also shows that one incident electron cannot in general be represented by a unique deterministic pixel pattern. Even when the entrance coordinate and the incident kinetic energy are fixed, different electrons undergo different elastic deflections, different inelastic energy transfers, different secondary-electron cascades, and different microscopic charge-transport histories.

For one event, collect the measured pixel charges into the vector

$$\mathbf{Q} = (Q_1, Q_2, \ldots, Q_P)\,. \tag{6.520}$$

If the electron enters the detector at $\boldsymbol{\rho}_0$ with kinetic energy K, then $\mathbf{Q}$ is a random vector. Its statistics also depend on detector properties such as the active-layer thickness, electric-field distribution, collection geometry, pixel pitch, temperature, collection efficiency, and readout architecture. We denote this set of parameters collectively by Θ_{det}.

The complete single-electron response can therefore be described by the conditional probability distribution

$$\boxed{p\left(\mathbf{Q} \mid \boldsymbol{\rho}_0, K, \Theta_{\mathrm{det}}\right).} \tag{6.521}$$

This distribution contains considerably more information than a deterministic point-spread function because it describes both the mean response and the event-to-event fluctuations around that mean.

For an integrating detector, these stochastic realizations are averaged over many incident electrons. Let $\mathcal{Q}(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega)$ denote the continuous collected-charge distribution produced by one microscopic realization, as defined in Eq. (6.496). For a locally translationally invariant detector, the corresponding normalized detector point-spread function can be defined as

$$\boxed{h_{\mathrm{det}}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_0\right) = \frac{\left\langle \mathcal{Q}\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right)\right\rangle_\omega}{\left\langle Q_{\mathrm{tot}}\right\rangle_\omega}.} \tag{6.522}$$

With this normalization,

$$\int h_{\mathrm{det}}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_0\right) d^2\boldsymbol{\rho} = 1. \tag{6.523}$$

A single incident electron does not itself produce the PSF in Eq. (6.522). It produces one random realization $\mathcal{Q}$. The PSF is the ensemble average over many statistically equivalent events with the same entrance coordinate. Fluctuations that disappear from the mean response nevertheless remain

important because they contribute to detector noise and to the uncertainty with which an individual event can be localized.

An electron-counting detector uses the random pixel pattern differently. Instead of averaging it directly, an event-localization algorithm $\mathcal{A}$ maps the measured vector into an estimated entrance coordinate,

$$\mathcal{A} : \mathbf{Q} \longrightarrow \widehat{\boldsymbol{\rho}}. \tag{6.524}$$

Because $\mathbf{Q}$ is random, the reconstructed coordinate $\widehat{\boldsymbol{\rho}}$ is also random even when the true entrance coordinate $\boldsymbol{\rho}_0$ is fixed.

The probability distribution of reconstructed positions follows from Eq. (6.521). Omitting K and Θ_{det} from the notation for compactness,

$$\boxed{\begin{aligned} p_{\text{loc}}\left(\widehat{\boldsymbol{\rho}} \mid \boldsymbol{\rho}_0\right) = \int & p\left(\mathbf{Q} \mid \boldsymbol{\rho}_0\right) \\ & \times \delta^{(2)}\left(\widehat{\boldsymbol{\rho}} - \mathcal{A}(\mathbf{Q})\right) d\mathbf{Q}. \end{aligned}} \tag{6.525}$$

Equation (6.525) considers every pixel pattern that can be produced by an electron entering at $\boldsymbol{\rho}_0$, weights it by its probability, applies the localization algorithm, and accumulates the resulting reconstructed positions.

The mean localization bias is

$$\boxed{\mathbf{b}(\boldsymbol{\rho}_0) = \langle \widehat{\boldsymbol{\rho}} \rangle - \boldsymbol{\rho}_0,} \tag{6.526}$$

which measures systematic displacement of the reconstructed position from the true entrance coordinate.

After removing this bias, the localization covariance is

$$\boxed{\boldsymbol{\Sigma}_{\text{loc}} = \left\langle \left(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}_0 - \mathbf{b}\right)\left(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}_0 - \mathbf{b}\right)^T \right\rangle.} \tag{6.527}$$

If the localization errors are approximately isotropic in the detector plane, a convenient scalar measure is the root-mean-square localization uncertainty per transverse coordinate,

$$\boxed{\sigma_{\text{loc}} = \sqrt{\frac{\operatorname{Tr} \boldsymbol{\Sigma}_{\text{loc}}}{2}}.} \tag{6.528}$$

For an unbiased response, σ_{loc} measures how precisely the entrance coordinate of one primary electron can be inferred from the complete stochastic pixel event.

This makes precise the distinction introduced above between the three relevant spatial scales. The pixel pitch specifies the sampling lattice, the event width describes the physical extent of the measured charge pattern, and σ_{loc} characterizes the uncertainty with which the centre of that random event can be inferred. These quantities are not required to be equal.

6.5 Spatial response of the detector

The preceding sections described the microscopic mechanisms that spread the signal generated by one incident electron: elastic deflection of the primary trajectory, secondary-electron transport, carrier drift and diffusion, charge sharing, and, in counting operation, event localization. We can now summarize the average spatial consequence of all these processes using the language of linear systems.

This description deliberately separates two questions. The first is *how broadly the detector responds in space*; this is described by a point-spread function. The second is *how strongly spatial modulations of different frequencies survive*; this is described by the modulation transfer function.

It is equally important to distinguish the continuous physical response of the detector from the subsequent effects of finite pixel size and discrete sampling. These are related, but they are not the same operation.

6.5.1 The detector point-spread function

The detector point-spread function has already been defined in Eq. (6.522). For a linear integrating detector whose response can locally be regarded as translationally invariant,

$$h_{\mathrm{det}}\left(\boldsymbol{\rho}-\boldsymbol{\rho}_0\right) \tag{6.529}$$

is the normalized ensemble-averaged charge response to electrons entering at the same transverse position $\boldsymbol{\rho}_0$.

In particular, Eq. (6.522) gives

$$\int h_{\mathrm{det}}\left(\boldsymbol{\rho}-\boldsymbol{\rho}_0\right) d^2\boldsymbol{\rho} = 1. \tag{6.530}$$

Thus h_{det} describes the *shape* of the mean spatial response independently of the mean amount of charge produced by one electron.

The translational-invariance assumption means that the response depends only on the displacement from the entrance position, not on the absolute position on the detector. If the incident electron density in the detector

plane is $I_{\text{in}}(\boldsymbol{\rho})$, the corresponding continuous mean detector signal is then described by convolution,

$$\boxed{I_{\text{det}}(\boldsymbol{\rho}) = \int h_{\text{det}}(\boldsymbol{\rho} - \boldsymbol{\rho}')\, I_{\text{in}}(\boldsymbol{\rho}')\, d^2\boldsymbol{\rho}'.} \tag{6.531}$$

Equation (6.531) is an ensemble statement. It does not say that one electron produces the deterministic pattern h_{det}. As emphasized following Eq. (6.522), an individual electron produces one random realization of the detector response. The PSF is obtained only after averaging many statistically equivalent events.

For a counting detector the situation is slightly different. The raw charge cluster is first converted into an estimated arrival coordinate $\widehat{\boldsymbol{\rho}}$. Its probability distribution was defined in Eq. (6.525). If the localization statistics are also translationally invariant, we may write

$$p_{\text{loc}}(\widehat{\boldsymbol{\rho}} \mid \boldsymbol{\rho}_0) = h_{\text{count}}(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}_0)\,. \tag{6.532}$$

Thus the effective PSF of a counting detector describes the distribution of reconstructed event positions rather than the spatial distribution of analog charge.

The two PSFs arise at different stages of the detection chain, but once defined they can be treated with the same Fourier-space formalism.

6.5.2 The modulation transfer function

A point-spread function is intuitive in real space, but the preservation of image detail is often easier to quantify in spatial-frequency space. Let

$$\mathbf{R} = (R_x, R_y) \tag{6.533}$$

denote the two-dimensional spatial frequency, measured in cycles m^{-1}, and define the Fourier transform of the detector PSF as

$$\boxed{H_{\text{det}}(\mathbf{R}) = \widehat{h}_{\text{det}}(\mathbf{R}) = \int h_{\text{det}}(\boldsymbol{\rho}) e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\rho}} d^2\boldsymbol{\rho}.} \tag{6.534}$$

The complex quantity H_{det} is the spatial transfer function of the continuous detector response. Its magnitude determines how strongly a spatial frequency is transmitted, while its phase describes any spatial phase shift introduced at that frequency.

The *modulation transfer function* (MTF) is the normalized magnitude of this transfer function,

$$\boxed{\mathrm{MTF}_{\mathrm{det}}(\mathbf{R}) = \frac{|H_{\mathrm{det}}(\mathbf{R})|}{|H_{\mathrm{det}}(\mathbf{0})|} = \frac{\left|\widehat{h}_{\mathrm{det}}(\mathbf{R})\right|}{\left|\widehat{h}_{\mathrm{det}}(\mathbf{0})\right|}.} \tag{6.535}$$

Because h_{det} was normalized in Eq. (6.530),

$$H_{\mathrm{det}}(\mathbf{0}) = 1, \tag{6.536}$$

and therefore in the present normalization

$$\mathrm{MTF}_{\mathrm{det}}(\mathbf{R}) = \left|\widehat{h}_{\mathrm{det}}(\mathbf{R})\right|. \tag{6.537}$$

Nevertheless, the normalized form in Eq. (6.535) is useful because experimentally measured detector responses need not initially be normalized.

The physical meaning of the MTF follows directly by asking how the detector responds to a weak sinusoidal modulation.

Why the MTF is the ratio of output to input modulation.

Consider an incident electron density containing a small sinusoidal modulation,

$$I_{\mathrm{in}}(\boldsymbol{\rho}) = I_0\left[1 + m_{\mathrm{in}}\cos\left(2\pi\mathbf{R}_0\cdot\boldsymbol{\rho} + \phi\right)\right], \tag{6.538}$$

where I_0 is the mean intensity and m_{in} is the input modulation depth.

Using

$$\cos\alpha = \frac{1}{2}\left(e^{j\alpha} + e^{-j\alpha}\right), \tag{6.539}$$

the input contains only three spatial-frequency components: $\mathbf{0}$, $+\mathbf{R}_0$, and $-\mathbf{R}_0$.

For the linear shift-invariant detector of Eq. (6.531), convolution in real space becomes multiplication in Fourier space. The zero-spatial-frequency, or *DC* (direct-current), component is the spatially constant part of the image, corresponding to

$$\mathbf{R} = \mathbf{0}. \tag{6.540}$$

For the input in Eq. (6.538), this component is simply the mean intensity

I_0. The DC component is therefore multiplied by

$$H_{\mathrm{det}}(\mathbf{0}), \tag{6.541}$$

whereas the sinusoidal components at $\pm\mathbf{R}_0$ are multiplied by $H_{\mathrm{det}}(\pm\mathbf{R}_0)$.

For a real PSF,

$$H_{\mathrm{det}}(-\mathbf{R}) = H^*_{\mathrm{det}}(\mathbf{R}), \tag{6.542}$$

so the output can again be written as a real sinusoid,

$$\begin{aligned} I_{\mathrm{out}}(\boldsymbol{\rho}) &= I_0 H_{\mathrm{det}}(\mathbf{0}) \\ &\quad \times \left[1 + m_{\mathrm{in}} \frac{|H_{\mathrm{det}}(\mathbf{R}_0)|}{|H_{\mathrm{det}}(\mathbf{0})|} \cos\left(2\pi \mathbf{R}_0 \cdot \boldsymbol{\rho} + \phi + \varphi_{\mathrm{det}}\right)\right], \end{aligned} \tag{6.543}$$

where

$$\varphi_{\mathrm{det}} = \arg H_{\mathrm{det}}(\mathbf{R}_0) \tag{6.544}$$

is the phase shift introduced by the detector.

The output modulation depth is therefore

$$m_{\mathrm{out}} = m_{\mathrm{in}} \frac{|H_{\mathrm{det}}(\mathbf{R}_0)|}{|H_{\mathrm{det}}(\mathbf{0})|}. \tag{6.545}$$

Hence

$$\boxed{\mathrm{MTF}_{\mathrm{det}}(\mathbf{R}_0) = \frac{m_{\mathrm{out}}}{m_{\mathrm{in}}}.} \tag{6.546}$$

Thus the MTF measures the fraction of the input sinusoidal contrast that survives the detector response.

Equation (6.546) gives the most useful physical interpretation of the MTF:

$$\boxed{\mathrm{MTF} = \frac{\text{output modulation}}{\text{input modulation}}.} \tag{6.547}$$

At zero spatial frequency,

$$\mathrm{MTF}(\mathbf{0}) = 1. \tag{6.548}$$

As spatial frequency increases, spreading of the detector response tends to average together neighbouring bright and dark regions, reducing the modulation and therefore reducing the MTF.

A broad PSF consequently corresponds to a rapidly decaying MTF, whereas a narrow PSF corresponds to better preservation of high spatial frequencies.

Radially averaged and anisotropic MTF The detector MTF is fundamentally a two-dimensional function,

$$\mathrm{MTF} = \mathrm{MTF}(R_x, R_y). \tag{6.549}$$

It need not be isotropic.

If the detector response is rotationally symmetric, then

$$h_{\mathrm{det}}(\boldsymbol{\rho}) = h_{\mathrm{det}}(\rho), \qquad \rho = |\boldsymbol{\rho}|, \tag{6.550}$$

and its transfer function depends only on

$$R = |\mathbf{R}| = \sqrt{R_x^2 + R_y^2}. \tag{6.551}$$

In that special case,

$$\mathrm{MTF}(R_x, R_y) = \mathrm{MTF}(R). \tag{6.552}$$

Real detectors are not perfectly rotationally symmetric. The pixel array is usually square, collection-electrode geometry can be directional, readout structures may differ between the two axes, and event localization algorithms can introduce additional anisotropies. Consequently, the MTF measured along the R_x direction need not be identical to that measured along R_y or along a diagonal.

A convenient one-dimensional summary is the radial average,

$$\boxed{\mathrm{MTF}_{\mathrm{rad}}(R) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{MTF}\left(R\cos\phi, R\sin\phi\right) d\phi.} \tag{6.553}$$

This representation is useful for comparing detectors with a single curve, but it inevitably discards directional information. A detector can have an apparently smooth radial MTF while exhibiting appreciably different resolution along different directions. Whenever anisotropy is important, the full two-dimensional $\mathrm{MTF}(R_x, R_y)$ or directional cuts should therefore be examined.

6.5.3 Pixel integration

The continuous detector response described by h_{det} is not yet the final sampled image. The connection with the charge-transport picture developed in the preceding section is particularly important here.

For one microscopic electron event, the transport calculation led to the continuous collected-charge distribution

$$\mathcal{Q}\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) \tag{6.554}$$

in Eq. (6.496). This distribution describes how the charge generated by one primary electron is spread over the detector plane before we assign the signal to discrete pixel values.

A pixel does not measure the value of this distribution at a mathematical point. Rather, its collection node receives charge from a finite region of the detector. If A_p denotes the collection region associated with pixel p, the charge measured by that pixel in one event is, in the simplest idealization,

$$\boxed{Q_p(\omega) = \int_{A_p} \mathcal{Q}\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) d^2\boldsymbol{\rho}.} \tag{6.555}$$

Equation (6.555) provides the direct connection between the microscopic charge-sharing picture and the usual linear-systems description of pixel integration. The finite pixel area performs a spatial integration of the continuous charge distribution before the result is represented by one number.

We first introduce a convenient function for describing the geometrical extent of an ideal square pixel. The rectangular function is defined as

$$\boxed{\operatorname{rect}(u) = \begin{cases} 1, & |u| < \dfrac{1}{2}, \\ 0, & |u| > \dfrac{1}{2}. \end{cases}} \tag{6.556}$$

The value at $|u| = 1/2$ does not affect any of the integrals below and may conventionally be taken to be $1/2$. Consequently, $\operatorname{rect}(x/a)$ selects an interval of width a centred at $x = 0$.

Consider now an ideal square pixel of width a, centred at the origin. Its geometrical collection window can be represented by

$$\boxed{\Pi_a(\boldsymbol{\rho}) = \operatorname{rect}\left(\frac{x}{a}\right) \operatorname{rect}\left(\frac{y}{a}\right),} \tag{6.557}$$

where $\boldsymbol{\rho} = (x, y)$. Thus $\Pi_a = 1$ inside the square

$$-\frac{a}{2} < x < \frac{a}{2}, \qquad -\frac{a}{2} < y < \frac{a}{2}, \tag{6.558}$$

and vanishes outside it.

For a regular square detector array with pixel pitch a, the centre of pixel (m, n) is

$$\boxed{\boldsymbol{\rho}_{mn} = (ma, na).} \tag{6.559}$$

Here we are considering the simplest idealized geometry in which the width of the collection region is equal to the pixel pitch, so that neighbouring square regions exactly tile the detector plane. The window associated with pixel (m, n) is therefore

$$\Pi_a \left(\boldsymbol{\rho} - \boldsymbol{\rho}_{mn}\right). \tag{6.560}$$

We can now connect this geometrical description directly with the continuous collected-charge distribution introduced in Eq. (6.496). For one microscopic event ω, the charge assigned to pixel (m, n) is

$$\boxed{\begin{aligned} Q_{mn}(\omega) = \int & \mathcal{Q}\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) \\ & \times \Pi_a \left(\boldsymbol{\rho} - \boldsymbol{\rho}_{mn}\right) d^2\boldsymbol{\rho}. \end{aligned}} \tag{6.561}$$

Equation (6.561) makes explicit what is meant physically by pixel integration. The charge distribution $\mathcal{Q}$ is continuous across the detector plane, but the readout assigns to each pixel a single number obtained by integrating that distribution over the finite collection region associated with the pixel. The discrete pixel array therefore appears only at the final readout stage; it does not imply that the underlying charge cloud is itself spatially discrete.

The same operation can be written in the language of linear systems. Define the normalized pixel-aperture function

$$\boxed{p_a(\boldsymbol{\rho}) = \frac{1}{a^2}\Pi_a(\boldsymbol{\rho}) = \frac{1}{a^2}\,\mathrm{rect}\left(\frac{x}{a}\right)\mathrm{rect}\left(\frac{y}{a}\right).} \tag{6.562}$$

Its normalization is

$$\int p_a(\boldsymbol{\rho}) d^2\boldsymbol{\rho} = 1. \tag{6.563}$$

The normalization by a^2 does not mean that the physical pixel measures an average rather than a total charge. The physical charge is given by Eq. (6.561). The normalized function p_a is introduced because it removes the trivial factor associated with pixel area and gives a transfer function equal to unity at zero spatial frequency.

With this definition, Eq. (6.561) can be written as

$$\boxed{Q_{mn}(\omega) = a^2 \left[\mathcal{Q}(\cdot \mid \boldsymbol{\rho}_0, \omega) * p_a\right](\boldsymbol{\rho}_{mn}).} \tag{6.564}$$

Averaging Eq. (6.564) over many statistically equivalent electron events gives the same operation for the mean detector response. Since h_{det} was defined in Eq. (6.522) from the ensemble average of $\mathcal{Q}$, the effective continuous response immediately before sampling is

$$\boxed{h_{\mathrm{pre}} = h_{\mathrm{det}} * p_a.} \tag{6.565}$$

Thus the pixel-aperture convolution is not a new physical model unrelated to the charge-transport description. It is the ensemble-averaged representation of the finite-area charge integration performed by each pixel.

Taking Fourier transforms of Eq. (6.565) gives

$$\boxed{H_{\mathrm{pre}}(R_x, R_y) = H_{\mathrm{det}}(R_x, R_y) P_a(R_x, R_y),} \tag{6.566}$$

where P_a is the Fourier transform of the normalized pixel aperture.

The square-window model is an idealization. In a real semiconductor detector, the effective collection region of a node need not have a perfectly sharp square boundary: the electric-field geometry and carrier diffusion can make the assignment of charge to neighbouring nodes position dependent. Such effects can be represented more generally by replacing the rectangular window in Eq. (6.561) by a pixel collection weighting function. The rectangular aperture is nevertheless the appropriate model for isolating the purely geometrical effect of finite pixel integration.

The MTF introduced in Eq. (6.535) describes the continuous physical response of the detector associated with h_{det}. It summarizes the spatial spreading produced by the microscopic detector physics discussed previously: energetic-electron transport, secondary-electron cascades, charge transport, and, depending on the operating mode, event localization.

The finite area of a pixel introduces an additional and conceptually different spatial filtering operation. Even if h_{det} were infinitely narrow, a pixel would still integrate the signal over a finite region before assigning one numerical value to that pixel. We therefore introduce a separate *pixel-aperture transfer function* and its corresponding *pixel-aperture MTF*.

Using the normalized pixel-aperture function introduced in Eq. (6.562), the frequency response associated purely with this finite-area integration is obtained from its Fourier transform.

MTF associated with the finite area of a square pixel.

For a square pixel of width a, the normalized pixel-aperture function

is

$$p_a(x,y) = \frac{1}{a^2}\,\mathrm{rect}\left(\frac{x}{a}\right)\mathrm{rect}\left(\frac{y}{a}\right), \tag{6.567}$$

as defined in Eq. (6.562).

Its Fourier transform is

$$P_a(R_x,R_y) = \int_{-a/2}^{a/2}\int_{-a/2}^{a/2}\frac{1}{a^2}e^{-j2\pi(R_x x+R_y y)}\,dx\,dy \tag{6.568}$$

$$= \left[\frac{1}{a}\int_{-a/2}^{a/2}e^{-j2\pi R_x x}\,dx\right]\left[\frac{1}{a}\int_{-a/2}^{a/2}e^{-j2\pi R_y y}\,dy\right]. \tag{6.569}$$

For the x factor,

$$\frac{1}{a}\int_{-a/2}^{a/2}e^{-j2\pi R_x x}\,dx = \frac{\sin(\pi a R_x)}{\pi a R_x} \tag{6.570}$$

$$= \mathrm{sinc}(aR_x), \tag{6.571}$$

where

$$\boxed{\mathrm{sinc}(u) \equiv \frac{\sin(\pi u)}{\pi u}.} \tag{6.572}$$

The same calculation in y gives

$$\boxed{P_a(R_x,R_y) = \mathrm{sinc}(aR_x)\,\mathrm{sinc}(aR_y).} \tag{6.573}$$

Because p_a is normalized to unit integral,

$$P_a(\mathbf{0}) = 1. \tag{6.574}$$

The modulation transfer function associated solely with finite pixel integration is therefore

$$\boxed{\mathrm{MTF}_{\mathrm{pix}}(R_x,R_y) = |P_a(R_x,R_y)| = |\mathrm{sinc}(aR_x)\,\mathrm{sinc}(aR_y)|\,.} \tag{6.575}$$

Equation (6.575) should not be confused with the intrinsic detector MTF defined in Eq. (6.535). The two describe different stages of the imaging process:

$$\boxed{\begin{array}{ll} \mathrm{MTF}_{\mathrm{det}}: & \text{physical spreading of the detector response,} \\ \mathrm{MTF}_{\mathrm{pix}}: & \text{finite-area integration by one pixel.} \end{array}} \tag{6.576}$$

The finite pixel area attenuates spatial modulation even before the resulting pixel values are discretely sampled. Along one of the pixel axes, for example, setting $R_y = 0$ gives

$$\mathrm{MTF}_{\mathrm{pix}}(R, 0) = |\operatorname{sinc}(aR)|. \tag{6.577}$$

At the axial Nyquist frequency,

$$R = \frac{1}{2a}, \tag{6.578}$$

the pixel-aperture contribution is

$$\mathrm{MTF}_{\mathrm{pix}}\left(\frac{1}{2a}, 0\right) = \operatorname{sinc}\left(\frac{1}{2}\right) \tag{6.579}$$

$$= \boxed{\frac{2}{\pi} \simeq 0.637.} \tag{6.580}$$

Thus even a hypothetical detector with a perfect intrinsic response,

$$\mathrm{MTF}_{\mathrm{det}}(\mathbf{R}) = 1, \tag{6.581}$$

would transmit only about 64% of the axial modulation at Nyquist because of finite pixel integration alone.

In a real detector, the physical detector response acts first and the finite pixel aperture acts afterwards. Since both operations are convolutions in the continuous detector plane, their transfer functions multiply:

$$\boxed{H_{\mathrm{pre}}(R_x, R_y) = H_{\mathrm{det}}(R_x, R_y) P_a(R_x, R_y).} \tag{6.582}$$

The corresponding *presampling MTF* is therefore

$$\boxed{\mathrm{MTF}_{\mathrm{pre}}(R_x, R_y) = \mathrm{MTF}_{\mathrm{det}}(R_x, R_y)\, \mathrm{MTF}_{\mathrm{pix}}(R_x, R_y).} \tag{6.583}$$

The three quantities should therefore be distinguished carefully. $\mathrm{MTF}_{\mathrm{det}}$ describes the continuous physical response represented by h_{det}; $\mathrm{MTF}_{\mathrm{pix}}$ describes the additional spatial averaging associated with assigning the continuous detector signal to finite pixels; and $\mathrm{MTF}_{\mathrm{pre}}$ describes their combined continuous response immediately before discrete sampling.

There are, however, two important qualifications to this simple factorization.

First, Eq. (6.566) assumes that the detector is sufficiently spatially homogeneous that the response of one pixel is just a translated copy of the

response of any other pixel. In other words, all collection diodes and their surrounding electric fields are assumed to behave in the same way, apart from translation by an integer number of pixel pitches. Only under this shift-invariance assumption can the pixel-collection process be represented by a convolution with a single function and therefore by multiplication by a well-defined Fourier-space transfer function.

If the response varies significantly from pixel to pixel, the more general description is not

$$h(\boldsymbol{\rho} - \boldsymbol{\rho}_0), \tag{6.584}$$

but rather a position-dependent response

$$h\left(\boldsymbol{\rho}, \boldsymbol{\rho}_0\right), \tag{6.585}$$

and a single global MTF is no longer sufficient to characterize the detector. One must then use a position-dependent response, local transfer functions, or an equivalent discrete response matrix.

Second, the square aperture of width a used in Eq. (6.562) is an idealization. The geometrical pixel pitch a specifies the separation between neighbouring collection nodes, but it does not imply that charge generated directly above a given pixel is collected exclusively by that pixel. As discussed in the preceding charge-transport description, carriers generated inside the silicon bulk can diffuse or drift laterally before collection. A carrier generated closer to one diode may therefore be collected by a neighbouring diode instead.

A more general description replaces the sharp square aperture by an effective collection weighting function

$$w_p(\boldsymbol{\rho}), \tag{6.586}$$

such that the charge assigned to pixel p is

$$\boxed{Q_p(\omega) = \int w_p(\boldsymbol{\rho}) \mathcal{Q}\left(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0, \omega\right) d^2\boldsymbol{\rho}.} \tag{6.587}$$

For a translationally invariant detector,

$$w_p(\boldsymbol{\rho}) = w\left(\boldsymbol{\rho} - \boldsymbol{\rho}_p\right), \tag{6.588}$$

so that a single weighting function w again describes every pixel.

Unlike the ideal rectangular window, the effective collection weighting functions of neighbouring pixels need not have disjoint support. They may

overlap substantially: charge generated in one region of the silicon may have a non-zero probability of being collected by several nearby nodes. In an ideal charge-conserving detector, these weighting functions satisfy approximately

$$\sum_p w_p(\boldsymbol{\rho}) \simeq 1 \tag{6.589}$$

throughout the active region, although the contribution associated with any individual pixel may extend beyond one geometrical pixel pitch.

This overlap is precisely the pixel-level manifestation of charge sharing. Signals in neighbouring pixels are then statistically correlated, because the charge generated by one primary electron can be distributed among several collection nodes. In frequency space, a broader effective collection weighting function generally produces stronger attenuation of high spatial frequencies than the ideal square-aperture model of Eq. (6.575).

The simple result

$$\mathrm{MTF}_{\mathrm{pix}}(R_x, R_y) = |\mathrm{sinc}(aR_x)\,\mathrm{sinc}(aR_y)| \tag{6.590}$$

should therefore be understood as the MTF of an ideal geometrical square-pixel integration model, not as a universal description of the complete charge-collection process of a real direct electron detector.

Discrete sampling is a separate operation again. It does not contribute another multiplicative MTF; instead, it replicates the continuous spectrum periodically in spatial-frequency space. The frequency at which this sampling becomes ambiguous, the Nyquist frequency, and the resulting phenomenon of aliasing are introduced in next section.

6.5.4 Nyquist frequency and aliasing

The final operation is spatial sampling. Let the square detector array have pixel pitch a, defined previously as the centre-to-centre separation between neighbouring pixels. The continuous pre-sampled signal is measured only at positions

$$\boldsymbol{\rho}_{mn} = (ma, na), \qquad m, n \in \mathbb{Z}. \tag{6.591}$$

Along either Cartesian direction, the sampling frequency is

$$R_{\mathrm{s}} = \frac{1}{a}, \tag{6.592}$$

and the corresponding Nyquist frequency is

$$\boxed{R_{\mathrm{Nyq}} = \frac{1}{2a}.} \tag{6.593}$$

The meaning of Eq. (6.593) deserves particular care. The Nyquist frequency is determined entirely by the spacing of the sampling grid and therefore by the pixel pitch. It does *not* state that the detector preserves high contrast up to that frequency. That question is governed by the presampling detector response, as quantified by the MTF. Nyquist instead addresses a different issue: whether a continuous spatial oscillation can be represented unambiguously by its sampled values. Figure 6.4 illustrates this distinction for a set of one-dimensional sinusoidal signals sampled on the same grid.

Indeed, Eq. (6.580) already showed that finite pixel integration alone reduces the axial modulation at Nyquist to approximately 0.637, even before any additional spreading from electron trajectories or charge transport is included. A real detector may therefore have a much smaller total MTF at Nyquist.

Consequently,

$$\boxed{\text{Nyquist frequency} \neq \text{detector spatial resolution.}} \tag{6.594}$$

The effect of sampling can be seen explicitly in Fourier space. A two-dimensional sampling lattice can be represented by

$$s_a(\boldsymbol{\rho}) = \sum_{m,n\in\mathbb{Z}} \delta(x - ma)\delta(y - na). \tag{6.595}$$

Sampling the continuous pre-sampled signal gives

$$I_{\text{samp}} = I_{\text{pre}} s_a. \tag{6.596}$$

Multiplication by the sampling lattice in real space has an important and perhaps initially surprising consequence in Fourier space: it creates periodically repeated copies of the continuous spectrum. We derive this result explicitly.

Why spatial sampling creates periodic spectral replicas.

Consider a square sampling lattice with pitch a. Its two-dimensional Dirac comb is

$$\boxed{s_a(x, y) = \sum_{p,q\in\mathbb{Z}} \delta(x - pa)\,\delta(y - qa).} \tag{6.597}$$

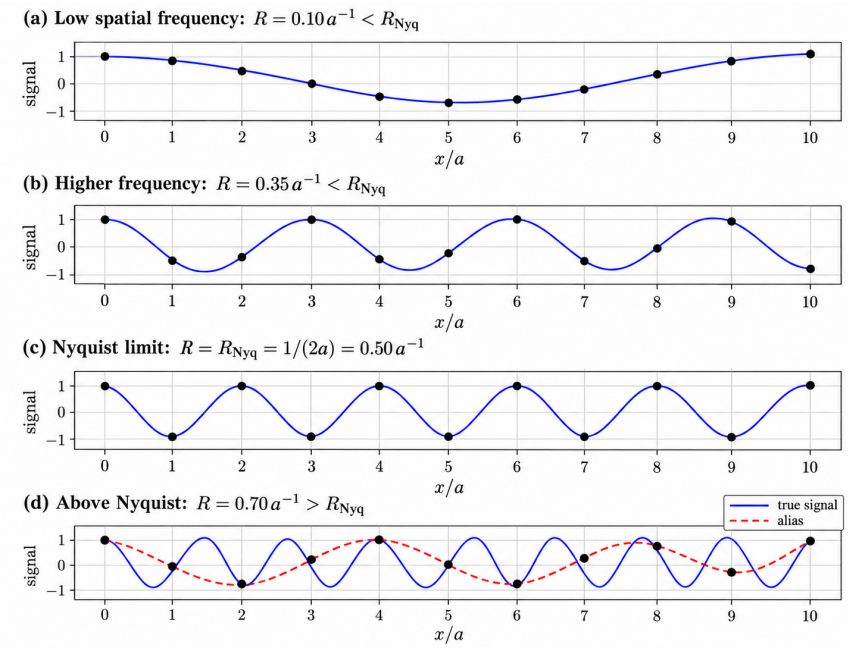


Figure 6.4: Sampling of one-dimensional sinusoidal signals on a fixed grid with spacing a. The black dots are the sampled values; the blue curve is the underlying continuous signal. In (a) and (b), the spatial frequency R lies below the Nyquist frequency $R_{\mathrm{Nyq}} = 1/(2a)$, so the samples represent the oscillation unambiguously. In (c), the signal is sampled exactly at the Nyquist limit, so only two samples per period are obtained. In (d), the true signal frequency lies above Nyquist. The same sampled values are then also consistent with a lower-frequency sinusoid (red dashed), which is the aliased interpretation. This figure emphasizes that the Nyquist frequency is determined solely by the sampling pitch a. It does not describe how much physical contrast is preserved at that frequency, which is instead governed by the presampling detector response through the MTF in Eqs. (6.535) and (6.583).

Multiplying the continuous presampling image by this comb gives

$$I_{\text{samp}}(x, y) = I_{\text{pre}}(x, y)\, s_a(x, y). \tag{6.598}$$

Indeed, using the sifting property of the Dirac delta,

$$I_{\text{samp}}(x, y) = \sum_{p,q\in\mathbb{Z}} I_{\text{pre}}(pa, qa)\, \delta(x - pa)\, \delta(y - qa), \tag{6.599}$$

so this continuous representation contains exactly the values of I_{pre} at the sampling positions.

We first determine the Fourier transform of the sampling comb. From Eq. (6.597),

$$\begin{aligned}
\widehat{s}_a(R_x, R_y) &= \sum_{p,q\in\mathbb{Z}} \iint \delta(x - pa)\delta(y - qa) e^{-j2\pi(R_x x + R_y y)}\, dx\, dy && (6.600)\\
&= \sum_{p,q\in\mathbb{Z}} e^{-j2\pi a(pR_x + qR_y)} && (6.601)\\
&= \left[\sum_{p\in\mathbb{Z}} e^{-j2\pi apR_x}\right] \left[\sum_{q\in\mathbb{Z}} e^{-j2\pi aqR_y}\right]. && (6.602)
\end{aligned}$$

The Poisson summation identity gives

$$\boxed{\sum_{p\in\mathbb{Z}} e^{-j2\pi apR} = \frac{1}{a} \sum_{m\in\mathbb{Z}} \delta\left(R - \frac{m}{a}\right).} \tag{6.603}$$

Applying this identity independently in the two spatial-frequency directions gives

$$\boxed{\widehat{s}_a(R_x, R_y) = \frac{1}{a^2} \sum_{m,n\in\mathbb{Z}} \delta\left(R_x - \frac{m}{a}\right) \delta\left(R_y - \frac{n}{a}\right).} \tag{6.604}$$

The important result is that a comb of spacing a in real space becomes another comb in Fourier space, now with spacing $1/a$.

We next use the Fourier-transform property that multiplication in real space becomes convolution in Fourier space. From Eq. (6.598),

$$\widehat{I}_{\text{samp}} = \widehat{I}_{\text{pre}} * \widehat{s}_a. \tag{6.605}$$

Substituting Eq. (6.604),

$$\widehat{I}_{\text{samp}}(R_x, R_y) = \frac{1}{a^2}\widehat{I}_{\text{pre}} * \sum_{m,n\in\mathbb{Z}} \delta\left(R_x - \frac{m}{a}\right)\delta\left(R_y - \frac{n}{a}\right) \tag{6.606}$$

$$= \boxed{\frac{1}{a^2}\sum_{m,n\in\mathbb{Z}} \widehat{I}_{\text{pre}}\left(R_x - \frac{m}{a}, R_y - \frac{n}{a}\right).} \tag{6.607}$$

Thus the spectrum of the sampled image consists of translated copies of the original presampling spectrum, centred at

$$\left(\frac{m}{a}, \frac{n}{a}\right), \qquad m, n \in \mathbb{Z}. \tag{6.608}$$

If the pre-sampled image contains frequencies extending beyond the Nyquist region, these replicated spectra overlap. Components lying above Nyquist are then folded into lower frequencies. This phenomenon is called *aliasing* (see Fig. 6.5).

In one dimension, for example, a sinusoid with frequency

$$R = R_{\text{s}} - R_0 \tag{6.609}$$

produces exactly the same samples as a sinusoid at frequency R_0, apart from a possible phase convention. Frequencies outside the Nyquist interval can therefore masquerade as lower frequencies in the sampled image.

For a square two-dimensional lattice, the unaliased fundamental region is naturally the square

$$\boxed{|R_x| \le \frac{1}{2a}, \qquad |R_y| \le \frac{1}{2a}.} \tag{6.610}$$

The commonly quoted scalar Nyquist frequency $1/(2a)$ corresponds to the boundary along either pixel axis. The corner of the two-dimensional Nyquist square lies at the larger radial frequency

$$R_{\text{corner}} = \frac{\sqrt{2}}{2a}. \tag{6.611}$$

This geometrical fact is another reason why a genuinely two-dimensional MTF contains more information than a single radial curve.

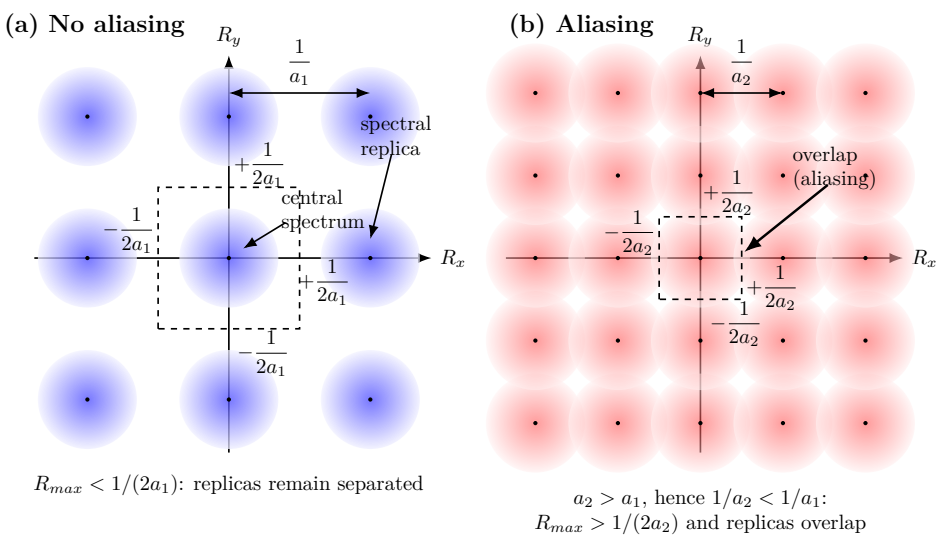


Figure 6.5: Aliasing viewed in spatial-frequency space. Sampling with pixel pitch a replicates the presampling spectrum at reciprocal-lattice positions $(m/a, n/a)$, with $m, n \in \mathbb{Z}$, as derived in Eq. (6.607). The spectral support is identical in the two panels. In (a), the sampling pitch a_1 is sufficiently small that the reciprocal-lattice spacing $1/a_1$ is large and neighbouring replicas remain separated. In (b), the sampling pitch is increased to $a_2 > a_1$; consequently $1/a_2 < 1/a_1$, the replicas move closer together, and they overlap. The dashed square is the first Nyquist region, bounded by $|R_x| \leq 1/(2a)$ and $|R_y| \leq 1/(2a)$. Overlap of neighbouring spectral replicas is the Fourier-space origin of aliasing.

The complete sequence is therefore

$$\boxed{\begin{array}{c} I_{\text{in}} \\ \Downarrow \\ h_{\text{det}} \quad \text{intrinsic detector spreading} \\ \Downarrow \\ p_a \quad \text{finite pixel integration} \\ \Downarrow \\ \text{sampling on pitch } a \\ \Downarrow \\ \text{spectral replicas and possible aliasing.} \end{array}} \tag{6.612}$$

The first two operations attenuate spatial frequencies and are naturally described by transfer functions and MTFs. Sampling is fundamentally different: it determines which spatial frequencies can be represented without ambiguity and, if sufficiently high frequencies remain in the pre-sampled signal, allows them to alias into the recorded frequency band.

Aliasing in an atomic-resolution image. The strict sampling theorem assumes that the continuous signal is *band limited*: its Fourier transform must vanish identically outside some finite region of spatial-frequency space. It is worth asking whether this assumption is actually satisfied by the signals encountered in electron microscopy.

The answer is, strictly speaking, no. In Sec. 5.3.3 we showed that the electron atomic scattering factor is proportional to the Fourier transform of the atomic electrostatic potential. Furthermore, the Fourier projection theorem led to Eq. (5.263),

$$\widehat{V}_a^{\text{proj}}(R) = \frac{2\pi\hbar^2}{m_0 e} f_e(R), \tag{6.613}$$

so the electron scattering factor directly determines the spatial frequency content of the projected atomic potential.

Most importantly, the high-spatial-frequency behaviour derived in Eq. (5.117) is

$$\boxed{f_e(R) \sim \frac{Z}{2\pi^2 a_0 R^2}, \qquad R \longrightarrow \infty.} \tag{6.614}$$

The atomic spectrum therefore decreases strongly with spatial frequency, but it is not band limited: there is no finite frequency $R_{\max}$ above which the atomic scattering factor becomes identically zero.

For a specimen containing many atoms, the corresponding electron structure factor contains coherent sums of these atomic contributions, as described by Eq. (5.307). Interference between different atoms can produce zeros at particular spatial frequencies, but it does not in general produce a finite frequency beyond which the complete spectrum vanishes.

Consequently, the ideal situation illustrated in Fig. 6.5, in which neighbouring spectral replicas have strictly disjoint support, is not realized exactly for an atomic specimen. For any finite sampling pitch a, some spectral content exists beyond the Nyquist region and can therefore be folded back into the recorded frequency band.

This does not make the Nyquist criterion useless. The important distinction is between *exactly zero aliasing* and *negligible aliasing*. Before sampling, the high-frequency specimen signal is attenuated both by its intrinsic decrease with spatial frequency and by the continuous image-formation and detector responses. In particular, the detector contributes the presampling MTF of Eq. (6.583). Thus the spectrum reaching the sampling operation can become very small outside the Nyquist region, even though it is not mathematically zero.

To make this statement quantitative, let

$$\boxed{\mathcal{B}_a = \left\{ (R_x, R_y) : |R_x| \leq \frac{1}{2a}, \quad |R_y| \leq \frac{1}{2a} \right\}} \tag{6.615}$$

denote the first Nyquist region. From Eq. (6.607), the contribution folded into this region from all neighbouring replicas is

$$\boxed{\widehat{I}_{\text{alias}}(\mathbf{R}) = \sum_{(m,n)\neq(0,0)} \widehat{I}_{\text{pre}}\left(R_x - \frac{m}{a}, R_y - \frac{n}{a}\right), \qquad \mathbf{R} \in \mathcal{B}_a,} \tag{6.616}$$

apart from the common factor $1/a^2$ appearing in Eq. (6.607).

In practice, the sampling pitch is therefore chosen so that this aliased contribution is sufficiently small for the accuracy required by the experiment. If ϵ denotes an acceptable relative aliasing level, the practical requirement is not

$$\widehat{I}_{\text{alias}} = 0, \tag{6.617}$$

but rather schematically

$$\boxed{\text{aliased signal} \ < \ \epsilon \times \text{relevant in-band signal.}} \tag{6.618}$$

Thus the Nyquist frequency should not be interpreted as a frequency at which the specimen suddenly ceases to contain information. Atomic specimens possess spatial-frequency content beyond any finite Nyquist frequency.

The role of sufficiently fine sampling, together with the presampling attenuation produced by the microscope and detector, is to make the folded contribution of that high-frequency tail acceptably small.

6.6 Physical sources of noise in a direct electron detector

The preceding sections described several stages at which a primary electron is transformed into a recorded detector signal. Randomness can enter at essentially every one of these stages. It is useful, however, to begin with a source of fluctuations that is present even before the electron reaches the detector.

Even an ideal detector cannot produce a noiseless image from a finite number of imaging electrons. The electrons themselves arrive as individual discrete particles, and the number arriving in any finite region during a finite exposure fluctuates from one realization to another. This fundamental fluctuation is called *electron shot noise*.

It must be distinguished from noise introduced subsequently by the detector. The distinction that will be used throughout this section is therefore

incident-electron shot noise	:	randomness in how many imaging electrons arrive,
detector-added fluctuations	:	randomness introduced while those electrons are detected.

(6.619)

The first contribution would remain even for a perfect detector. The second includes the stochastic energy deposition, charge generation, charge transport, electronic readout, and other detector processes considered below.

6.6.1 Incident-electron shot noise

Let

$$\Phi(\boldsymbol{\rho}) \tag{6.620}$$

denote the mean electron fluence at the detector entrance plane during one exposure, measured in electrons per unit area. The function $\Phi(\boldsymbol{\rho})$ contains the spatial variation produced by the specimen and by the electron-optical image-formation process. It is an *expected* electron density: the actual image is formed by a finite set of individual electron arrivals.

Consider an arbitrary region A of the detector entrance plane. Its expected number of incident electrons is

$$\boxed{\Lambda_A = \int_A \Phi(\boldsymbol{\rho})\, d^2\boldsymbol{\rho}.} \tag{6.621}$$

If the individual electron arrivals are statistically independent, the number $N(A)$ arriving inside this region is described by a Poisson distribution,

$$\boxed{N(A) \sim \text{Poisson}(\Lambda_A),} \tag{6.622}$$

or explicitly

$$P[N(A) = n] = e^{-\Lambda_A} \frac{\Lambda_A^n}{n!}, \qquad n = 0, 1, 2, \ldots. \tag{6.623}$$

For approximately uniform illumination,

$$\Phi(\boldsymbol{\rho}) \simeq \Phi_0, \tag{6.624}$$

Eq. (6.621) reduces to

$$\Lambda_A = \Phi_0 |A|, \tag{6.625}$$

where $|A|$ is the area of the region. Thus doubling the exposure, electron fluence, or observation area doubles the expected number of incident electrons.

Mean and variance of Poisson electron arrivals.

Let

$$N \sim \text{Poisson}(\mu), \tag{6.626}$$

with

$$P(N = n) = e^{-\mu} \frac{\mu^n}{n!}. \tag{6.627}$$

Its expectation is

$$\mathbb{E}[N] = \sum_{n=0}^{\infty} n e^{-\mu} \frac{\mu^n}{n!} \tag{6.628}$$

$$= \mu e^{-\mu} \sum_{n=1}^{\infty} \frac{\mu^{n-1}}{(n-1)!} \tag{6.629}$$

$$= \mu. \tag{6.630}$$

To obtain the variance, it is convenient first to calculate the second

factorial moment,

$$\mathbb{E}[N(N-1)] = \sum_{n=0}^{\infty} n(n-1)e^{-\mu}\frac{\mu^n}{n!} \tag{6.631}$$

$$= \mu^2 e^{-\mu} \sum_{n=2}^{\infty} \frac{\mu^{n-2}}{(n-2)!} \tag{6.632}$$

$$= \mu^2. \tag{6.633}$$

Since

$$N^2 = N(N-1) + N, \tag{6.634}$$

we obtain

$$\mathbb{E}[N^2] = \mu^2 + \mu. \tag{6.635}$$

Therefore,

$$\mathrm{Var}(N) = \mathbb{E}[N^2] - \mathbb{E}[N]^2 \tag{6.636}$$

$$= \boxed{\mu.} \tag{6.637}$$

Thus, for Poisson electron arrivals,

$$\boxed{\mathbb{E}[N] = \mathrm{Var}(N) = \mu.} \tag{6.638}$$

Equation (6.638) is the defining statistical property of ideal electron shot noise. For an expected number μ of incident electrons, the mean signal is

$$\mathbb{E}[N] = \mu, \tag{6.639}$$

whereas the noise power, identified with the variance of the fluctuation about this mean, is

$$\sigma_N^2 = \mathrm{Var}(N) = \mu. \tag{6.640}$$

If the signal power is represented by the squared mean,

$$P_{\text{signal}} = \mathbb{E}[N]^2 = \mu^2, \tag{6.641}$$

and the noise power by the variance,

$$P_{\text{noise}} = \sigma_N^2 = \mu, \tag{6.642}$$

the signal-to-noise ratio is

$$\boxed{\mathrm{SNR} = \frac{P_{\text{signal}}}{P_{\text{noise}}} = \frac{\mu^2}{\mu} = \mu.} \tag{6.643}$$

Thus, for an ideal Poisson-limited measurement, the SNR is directly proportional to the number of incident electrons. Equivalently, the ratio of signal amplitude to the root-mean-square noise amplitude is

$$\frac{\mathbb{E}[N]}{\sigma_N} = \sqrt{\mu}, \tag{6.644}$$

whose square is precisely the power SNR of Eq. (6.643).

Collecting four times as many electrons therefore increases the power SNR by a factor of four, while increasing the signal-amplitude to noise-amplitude ratio by only a factor of two. This square-root behavior is the familiar manifestation of Poisson statistics in the measured amplitudes. In cryo-EM, improving the SNR simply by increasing the electron fluence is severely constrained by radiation damage to the specimen.

Shot noise and the detectability of image contrast. The importance of the result above becomes particularly clear when we consider the distinction between two regions of an image. Suppose that two regions have expected electron counts

$$\mu_0 \qquad \text{and} \qquad \mu_1, \tag{6.645}$$

and let

$$N_0 \sim \text{Poisson}(\mu_0), \qquad N_1 \sim \text{Poisson}(\mu_1) \tag{6.646}$$

be independent measurements of those regions.

The image signal that distinguishes the two regions is their difference,

$$\Delta\mu = \mu_1 - \mu_0. \tag{6.647}$$

If we estimate this difference from the measured counts,

$$\Delta N = N_1 - N_0, \tag{6.648}$$

then independence gives

$$\text{Var}(\Delta N) = \text{Var}(N_1) + \text{Var}(N_0) \tag{6.649}$$

$$= \mu_1 + \mu_0. \tag{6.650}$$

Using the power definition of SNR introduced above, the SNR with which the contrast between the two regions can be distinguished is therefore

$$\boxed{\text{SNR}_{\text{contrast}} = \frac{(\mu_1 - \mu_0)^2}{\mu_1 + \mu_0}.} \tag{6.651}$$

Equivalently, the separation between the two expected signals measured in units of the root-mean-square fluctuation of their difference is

$$\boxed{\frac{|\mu_1 - \mu_0|}{\sqrt{\mu_1 + \mu_0}} = \sqrt{\mathrm{SNR}_{\mathrm{contrast}}}.} \tag{6.652}$$

Thus two regions become progressively easier to distinguish as this quantity increases.

Suppose now that the two regions receive mean electron rates λ_0 and λ_1, measured in electrons per unit time. After an acquisition time T,

$$\mu_0 = \lambda_0 T, \qquad \mu_1 = \lambda_1 T. \tag{6.653}$$

Substitution into Eq. (6.651) gives

$$\boxed{\mathrm{SNR}_{\mathrm{contrast}} = T\frac{(\lambda_1 - \lambda_0)^2}{\lambda_1 + \lambda_0}.} \tag{6.654}$$

Consequently, the power SNR increases linearly with acquisition time, whereas the separation expressed in root-mean-square noise units increases only as

$$\sqrt{\mathrm{SNR}_{\mathrm{contrast}}} \propto \sqrt{T}. \tag{6.655}$$

It is important to distinguish this increased *detectability* from an increase in the physical contrast itself. For example, define the relative contrast between the two regions as

$$\boxed{C = \frac{\mu_1 - \mu_0}{\mu_1 + \mu_0}.} \tag{6.656}$$

If the acquisition time is increased, both μ_0 and μ_1 grow by the same factor, so C remains unchanged. What improves is our ability to measure that fixed contrast in the presence of shot noise.

Indeed, Eq. (6.651) can be written as

$$\boxed{\mathrm{SNR}_{\mathrm{contrast}} = C^2\,(\mu_0 + \mu_1) = C^2(\lambda_0 + \lambda_1)T.} \tag{6.657}$$

Thus weak contrast can in principle be made detectable by collecting more electrons: the contrast C remains fixed, while the number of electrons available to measure it increases.

Distribution of the measured contrast. The distinction between the two regions can be described more explicitly by considering the random difference between their measured electron counts,

$$\Delta N = N_1 - N_0, \tag{6.658}$$

where

$$N_0 \sim \text{Poisson}(\mu_0), \qquad N_1 \sim \text{Poisson}(\mu_1), \tag{6.659}$$

and the two counts are assumed independent.

The difference of two independent Poisson random variables follows a *Skellam distribution*. Its probability mass function is

$$\boxed{\begin{aligned} P(\Delta N = k) = {} & \exp[-(\mu_0 + \mu_1)] \left(\frac{\mu_1}{\mu_0}\right)^{k/2} \\ & \times I_{|k|}\left(2\sqrt{\mu_0 \mu_1}\right), \qquad k \in \mathbb{Z}, \end{aligned}} \tag{6.660}$$

where I_ν is the modified Bessel function of the first kind.

Its mean and variance are particularly simple:

$$\boxed{\mathbb{E}[\Delta N] = \mu_1 - \mu_0, \qquad \text{Var}(\Delta N) = \mu_1 + \mu_0.} \tag{6.661}$$

Suppose now that the two regions produce mean count rates λ_0 and λ_1. After an acquisition time T,

$$\mu_0 = \lambda_0 T, \qquad \mu_1 = \lambda_1 T. \tag{6.662}$$

Equation (6.661) then becomes

$$\boxed{\mathbb{E}[\Delta N] = (\lambda_1 - \lambda_0)T,} \tag{6.663}$$

and

$$\boxed{\sigma_{\Delta N} = \sqrt{(\lambda_0 + \lambda_1)T}.} \tag{6.664}$$

Thus increasing the acquisition time has two effects. The separation between the two expected signals grows *linearly* with T, whereas the width of the statistical fluctuations grows only as $\sqrt{T}$. Consequently,

$$\frac{\mathbb{E}[\Delta N]}{\sigma_{\Delta N}} = \sqrt{T}\,\frac{\lambda_1 - \lambda_0}{\sqrt{\lambda_0 + \lambda_1}}. \tag{6.665}$$

Figure 6.6 illustrates this effect. Although the Skellam distribution becomes broader as more electrons are collected, its mean moves away from zero faster than its width grows. The probability of confusing the two regions therefore progressively decreases.

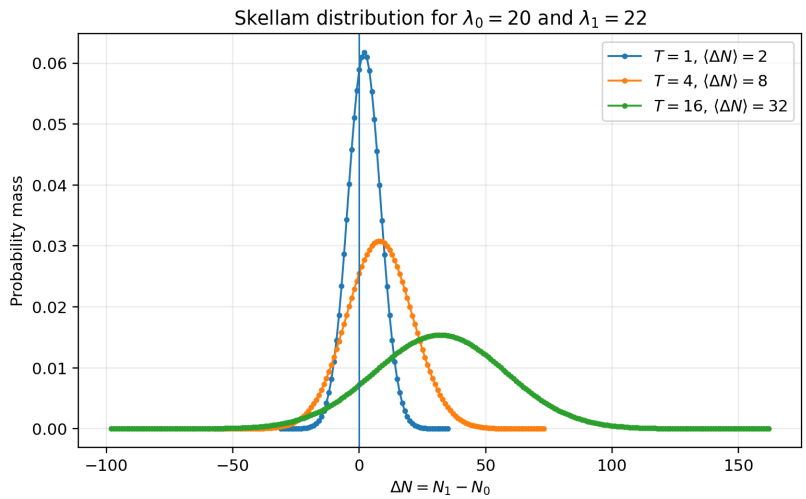


Figure 6.6: Distribution of the measured count difference $\Delta N = N_1 - N_0$ for two independent Poisson regions with mean rates $\lambda_0 = 20$ and $\lambda_1 = 22$ electrons per unit time. The curves are the corresponding Skellam distributions for increasing acquisition times T. The mean difference grows as $(\lambda_1 - \lambda_0)T$, whereas the standard deviation grows only as $\sqrt{(\lambda_0 + \lambda_1)T}$. Consequently, the separation in units of the noise standard deviation increases as $\sqrt{T}$, or equivalently the power SNR increases linearly with T, as shown in Eq. (6.657).

Contrast in CryoEM. This is one of the central difficulties of cryo-EM. The image must distinguish a biological macromolecule from the surrounding vitreous ice, yet the two materials are composed predominantly of light elements. Proteins consist mainly of hydrogen, carbon, nitrogen, oxygen, and smaller amounts of sulfur, with atomic numbers

$$Z_{\mathrm{H}} = 1, \qquad Z_{\mathrm{C}} = 6, \qquad Z_{\mathrm{N}} = 7, \qquad Z_{\mathrm{O}} = 8, \tag{6.666}$$

whereas the surrounding ice is composed primarily of water, H_2O, and is therefore also dominated by hydrogen and oxygen.

As discussed in the specimen-interaction chapter, the electron image contrast ultimately arises from differences in the projected electrostatic potential and therefore from differences in the electron scattering factors of the atoms composing the specimen. Because both protein and vitreous ice are made largely of light elements and have comparable electron densities, the difference between their average projected scattering strengths is small. The relative contrast C in Eq. (6.656) is consequently often weak.

Equation (6.657) shows the statistical consequence directly:

$$\mathrm{SNR}_{\mathrm{contrast}} = C^2 \left(\mu_0 + \mu_1\right). \tag{6.667}$$

When C is small, a large number of detected electrons is required to obtain an appreciable contrast SNR. In principle, one could simply increase the acquisition time or electron fluence so that $\mu_0 + \mu_1$ becomes large enough to separate the protein and ice distributions reliably.

Cryo-EM cannot exploit this strategy without limit, because the same electrons that carry structural information also damage the biological specimen through inelastic interactions. Radiation damage therefore places a severe upper bound on the usable electron fluence. Cryo-EM is consequently a problem of detecting very small differences in electron scattering between protein and surrounding ice using only a limited number of electrons.

Spatial independence before detection. The Poisson model also describes the spatial independence of the incident arrivals. To make this explicit, divide the detector entrance plane mathematically into non-overlapping regions A_p and define

$$N_p = N(A_p), \qquad \mu_p = \int_{A_p} \Phi(\boldsymbol{\rho})\, d^2\boldsymbol{\rho}. \tag{6.668}$$

For independent arrivals,

$$N_p \sim \mathrm{Poisson}(\mu_p), \tag{6.669}$$

and the counts in two disjoint regions are independent.

Define the zero-mean fluctuation

$$\delta N_p = N_p - \mu_p. \tag{6.670}$$

Then

$$\boxed{\langle \delta N_p \delta N_q \rangle = \mu_p \delta_{pq}.} \tag{6.671}$$

For uniform illumination, $\mu_p = \mu$, and Eq. (6.671) becomes

$$\langle \delta N_p \delta N_q \rangle = \mu \delta_{pq}. \tag{6.672}$$

The regions A_p in this argument are only a mathematical partition of the *incident* electron distribution. They should not yet be identified with independent detector pixels. As established by the charge-sharing discussion, the silicon is continuous and one incident electron can produce signals in several neighbouring readout pixels. The detector can therefore transform initially independent arrival statistics into spatially correlated recorded signals.

Shot noise follows the local image intensity. Shot noise is not generally a spatially uniform additive background. If the expected electron image varies with position, then $\Phi(\boldsymbol{\rho})$ varies as well, and so does the local Poisson variance. From Eq. (6.638),

$$\boxed{\mathrm{Var}(N_p) = \mu_p.} \tag{6.673}$$

Bright regions therefore have a larger *absolute* shot-noise variance than dark regions, although their relative fluctuation is smaller.

This is sometimes called *signal-dependent* or *heteroscedastic* noise. Only under approximately uniform illumination can the incident shot noise be regarded as having the same variance everywhere.

Shot noise is not detector-added noise. It is important to locate this fluctuation correctly in the physical detection chain. The randomness described by Eq. (6.622) exists before any interaction with the detector material:

$$\boxed{\begin{array}{c} \text{mean incident electron image} \\ \Downarrow \\ \text{random realization of discrete electron arrivals} \\ \Downarrow \\ \text{detector interaction and readout.} \end{array}} \tag{6.674}$$

The subsequent detector response introduces additional randomness. For one primary electron, the complete stochastic detector response was described by Eq. (6.521). Likewise, Eq. (6.412) showed explicitly that the charge generated by one electron fluctuates because both the deposited energy and the microscopic pair-creation process fluctuate.

These mechanisms must not be confused with Poisson arrival noise. Electron shot noise answers the question

$$\text{“How many primary electrons arrived?”} \tag{6.675}$$

whereas the subsequent detector statistics answer questions such as

$$\text{“What signal did each arriving electron produce?”} \tag{6.676}$$

This distinction also clarifies the difference between integrating and electron-counting operation. In an integrating detector, Eq. (6.508) shows that the signals from all arriving electrons are added. Consequently, fluctuations in the number of incident electrons and fluctuations in the response of each electron both affect the final signal. In an ideal counting detector, the procedure summarized in Eq. (6.512) attempts to replace each successfully identified primary electron by one standardized count. This can suppress fluctuations in event amplitude, but it cannot remove the fundamental Poisson fluctuation in the number of electrons that actually arrived.

At sufficiently high event density, coincidence loss can modify even these counting statistics because nearby primary electrons may no longer be recorded independently. That effect is not an additional source of Poisson shot noise: it is a nonlinear detector process acting on the incident Poisson events and will therefore be treated separately.

The spatial-frequency consequence of Eq. (6.671) will be developed when the noise power spectrum is introduced. Under uniform illumination, independent Poisson arrivals lead to a flat, or *white*, input noise spectrum. The detector need not preserve that spectral shape, because spatial spreading, charge sharing, event localization, and nonlinear counting can introduce correlations between the recorded pixels.

6.6.2 Fluctuations of the single-electron response

The Poisson statistics discussed in Sec. 6.6.1 describe the randomness in *how many* primary electrons arrive at the detector. They do not yet describe a second source of randomness: two primary electrons entering the detector under otherwise identical conditions need not produce identical measured signals.

This second form of stochasticity appeared naturally in the microscopic detector description. For one incident electron, the measured charges in the detector pixels were collected into the vector

$$\mathbf{Q} = (Q_1, Q_2, \ldots, Q_P), \tag{6.677}$$

whose complete conditional probability distribution was introduced in Eq. (6.521),

$$p\left(\mathbf{Q} \mid \boldsymbol{\rho}_0, K, \Theta_{\text{det}}\right). \tag{6.678}$$

Even when the entrance position $\boldsymbol{\rho}_0$, incident kinetic energy K, and detector configuration Θ_{det} are fixed, $\mathbf{Q}$ remains random.

A first characterization of this distribution is given by its mean response,

$$\boxed{\overline{\mathbf{Q}} = \mathbb{E}[\mathbf{Q}],} \tag{6.679}$$

and its covariance matrix,

$$\boxed{\boldsymbol{\Sigma}_Q = \mathbb{E}\left[\left(\mathbf{Q} - \overline{\mathbf{Q}}\right)\left(\mathbf{Q} - \overline{\mathbf{Q}}\right)^T\right].} \tag{6.680}$$

The diagonal elements of $\boldsymbol{\Sigma}_Q$ describe the event-to-event variance measured in individual pixels, whereas the off-diagonal elements describe correlations between different pixels.

Fluctuations of event amplitude and spatial shape. The response of one electron can fluctuate in two conceptually different ways. First, the *total amount of collected charge* can change from one event to another. Define

$$\boxed{Q_{\text{tot}} = \sum_p Q_p.} \tag{6.681}$$

Its mean and variance will be denoted by

$$\boxed{\overline{Q} = \mathbb{E}[Q_{\text{tot}}], \qquad \sigma_Q^2 = \text{Var}(Q_{\text{tot}}).} \tag{6.682}$$

Second, even events producing the same total charge can distribute that charge differently among neighbouring pixels. For $Q_{\text{tot}} > 0$, define the normalized charge fractions

$$\pi_p = \frac{Q_p}{Q_{\text{tot}}}, \qquad \sum_p \pi_p = 1. \tag{6.683}$$

Then the complete pixel response can be written as

$$\boxed{\mathbf{Q} = Q_{\text{tot}}\boldsymbol{\pi},} \tag{6.684}$$

where

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_P) \tag{6.685}$$

describes the spatial shape of the event.

Thus Q_{tot} describes the event amplitude, whereas $\boldsymbol{\pi}$ describes how that amplitude is distributed across the detector. Both quantities are random, and they need not be statistically independent. An energetic secondary electron, for example, can carry energy far from the primary trajectory while at the same time changing the amount of energy that remains deposited inside the active detector volume.

The physical origin of these fluctuations has already been developed in the preceding detector model. Equation (6.412) showed that fluctuations in the number of generated electron–hole pairs contain both the Fano fluctuation at fixed deposited energy and the event-to-event fluctuation of E_{dep}. Spatial fluctuations arise from the stochastic primary and secondary trajectories followed by carrier drift, diffusion, and charge sharing. Event amplitude and event shape are therefore two manifestations of the same underlying microscopic transport process.

Combining Poisson arrivals with fluctuating event amplitudes. We can first study the effect of the amplitude fluctuations alone, without yet considering their spatial distribution. This is particularly transparent for an integrating detector.

Suppose that a detector region receives

$$N \sim \text{Poisson}(\mu) \tag{6.686}$$

primary electrons during one acquisition interval. Let

$$Q_1, Q_2, \ldots, Q_N \tag{6.687}$$

be the total collected charges produced by those electrons. We assume, for the moment, that the Q_i are independent and identically distributed, with the mean and variance defined in Eq. (6.682), and that they are independent of N.

The total integrated detector signal is

$$\boxed{S = \sum_{i=1}^{N} Q_i.} \tag{6.688}$$

The random variable S is a *compound Poisson* variable: both the number of contributions and the magnitude of each contribution are random.

Mean, variance, and SNR of a fluctuating single-electron response.

Conditioned on the arrival of N electrons,

$$\mathbb{E}[S \mid N] = N\overline{Q} \tag{6.689}$$

and

$$\mathrm{Var}(S \mid N) = N\sigma_Q^2. \tag{6.690}$$

Taking the expectation over the Poisson distribution of N,

$$\mathbb{E}[S] = \mathbb{E}\left[\mathbb{E}[S \mid N]\right] \tag{6.691}$$

$$= \overline{Q}\,\mathbb{E}[N] \tag{6.692}$$

$$= \boxed{\mu\overline{Q}.} \tag{6.693}$$

We now calculate the variance of the integrated signal S. There are two independent reasons why S can vary from one acquisition to another:

1. even if exactly the same number N of primary electrons arrived, the individual responses Q_i would fluctuate;
2. the number N of primary electrons itself fluctuates according to the Poisson arrival statistics.

It is useful to treat these two sources of randomness separately.

Suppose first that the number of incident electrons is known to be exactly

$$N = n. \tag{6.694}$$

Then

$$S = Q_1 + Q_2 + \cdots + Q_n. \tag{6.695}$$

Since the individual single-electron responses are assumed independent and have

$$\mathbb{E}[Q_i] = \overline{Q}, \qquad \mathrm{Var}(Q_i) = \sigma_Q^2, \tag{6.696}$$

the conditional mean of the integrated signal is

$$\boxed{\mathbb{E}[S \mid N = n] = n\overline{Q},} \tag{6.697}$$

and its conditional variance is

$$\boxed{\mathrm{Var}(S \mid N = n) = n\sigma_Q^2.} \tag{6.698}$$

Equation (6.698) answers a very specific question: *if exactly n electrons arrived, how much would the detector signal still fluctuate because different electrons produce different amounts of charge?*

We can now allow N itself to be random. Since Eq. (6.698) depends on n, when N is random we may write

$$\mathrm{Var}(S \mid N) = N\sigma_Q^2. \tag{6.699}$$

This quantity is itself random because its value depends on the random number N. Averaging it over the Poisson distribution of N gives

$$\mathbb{E}\left[\mathrm{Var}(S \mid N)\right] = \mathbb{E}\left[N\sigma_Q^2\right] \tag{6.700}$$

$$= \sigma_Q^2 \mathbb{E}[N] \tag{6.701}$$

$$= \boxed{\mu\sigma_Q^2.} \tag{6.702}$$

This first contribution is therefore the variance produced by *single-electron response fluctuations*, averaged over the random number of electrons in the exposure.

There is, however, a second source of variability. Even if every electron produced exactly its mean charge $\overline{Q}$, the total signal would still fluctuate because different acquisitions contain different numbers of primary electrons.

For a given value of N, the conditional mean signal is

$$\mathbb{E}[S \mid N] = N\overline{Q}. \tag{6.703}$$

Since N fluctuates, this conditional mean also fluctuates from one acquisition to another. Its variance is

$$\mathrm{Var}\left(\mathbb{E}[S \mid N]\right) = \mathrm{Var}\left(N\overline{Q}\right) \tag{6.704}$$

$$= \overline{Q}^{\,2}\,\mathrm{Var}(N). \tag{6.705}$$

For Poisson arrivals,

$$\mathrm{Var}(N) = \mu, \tag{6.706}$$

and therefore

$$\boxed{\mathrm{Var}\left(\mathbb{E}[S \mid N]\right) = \mu\overline{Q}^{\,2}.} \tag{6.707}$$

This second term is the variance produced purely by fluctuations in the number of incident electrons: it is the contribution of electron shot noise.

The total variance is obtained by adding these two contributions. This is the content of the *law of total variance*,

$$\boxed{\mathrm{Var}(S) = \underbrace{\mathbb{E}\left[\mathrm{Var}(S \mid N)\right]}_{\substack{\text{average variance remaining}\\ \text{when } N \text{ is fixed}}} + \underbrace{\mathrm{Var}\left(\mathbb{E}[S \mid N]\right)}_{\substack{\text{variance caused by}\\ \text{the fluctuation of } N}}.} \tag{6.708}$$

Substituting Eqs. (6.702) and (6.707) gives

$$\mathrm{Var}(S) = \mu\sigma_Q^2 + \mu\overline{Q}^{\,2} \tag{6.709}$$

$$= \boxed{\mu\left(\overline{Q}^{\,2} + \sigma_Q^2\right).} \tag{6.710}$$

Since

$$\mathbb{E}[Q_{\mathrm{tot}}^2] = \mathrm{Var}(Q_{\mathrm{tot}}) + \mathbb{E}[Q_{\mathrm{tot}}]^2 = \sigma_Q^2 + \overline{Q}^{\,2}, \tag{6.711}$$

the same result can also be written as

$$\boxed{\mathrm{Var}(S) = \mu\mathbb{E}\left[Q_{\mathrm{tot}}^2\right].} \tag{6.712}$$

The decomposition

$$\boxed{\mathrm{Var}(S) = \underbrace{\mu\overline{Q}^{\,2}}_{\text{incident-electron shot noise}} + \underbrace{\mu\sigma_Q^2}_{\text{single-electron response fluctuations}}} \tag{6.713}$$

is important because the two terms have different physical origins.

The first term would remain even for a perfectly reproducible detector: if every primary electron produced exactly the same charge $\overline{Q}$, the random number of incident electrons would still make the total signal fluctuate.

The second term would remain even if the number of incident electrons were fixed exactly: different primary electrons would still deposit different energies and produce different collected charges.

A real integrating detector is subject to both effects simultaneously.

Using the power definition of SNR adopted in Sec. 6.6.1,

$$\mathrm{SNR}_{\mathrm{int}} = \frac{\mathbb{E}[S]^2}{\mathrm{Var}(S)} \tag{6.714}$$

$$= \frac{\mu^2 \overline{Q}^2}{\mu\left(\overline{Q}^2 + \sigma_Q^2\right)}, \tag{6.715}$$

and therefore

$$\boxed{\mathrm{SNR}_{\mathrm{int}} = \frac{\mu}{1 + \dfrac{\sigma_Q^2}{\overline{Q}^2}}.} \tag{6.716}$$

Equation (6.716) shows directly that event-to-event fluctuations in the response of one electron reduce the SNR of an integrating detector below the ideal Poisson value.

It is convenient to define the single-electron-response excess-noise factor

$$\boxed{\mathcal{F}_{\mathrm{SER}} = \frac{\mathbb{E}[Q_{\mathrm{tot}}^2]}{\mathbb{E}[Q_{\mathrm{tot}}]^2} = 1 + \frac{\sigma_Q^2}{\overline{Q}^2}.} \tag{6.717}$$

This quantity should not be confused with the semiconductor Fano factor F introduced in the discussion of electron–hole-pair generation.

Equation (6.716) can then be written as

$$\boxed{\mathrm{SNR}_{\mathrm{int}} = \frac{\mu}{\mathcal{F}_{\mathrm{SER}}}.} \tag{6.718}$$

If every incident electron produced exactly the same total collected charge, then

$$\sigma_Q^2 = 0, \qquad \mathcal{F}_{\mathrm{SER}} = 1, \tag{6.719}$$

and Eq. (6.718) reduces to the ideal Poisson result

$$\mathrm{SNR} = \mu \tag{6.720}$$

derived in Sec. 6.6.1. Any event-to-event amplitude fluctuation instead gives

$$\mathcal{F}_{\mathrm{SER}} > 1 \tag{6.721}$$

and therefore lowers the SNR even if the subsequent electronic readout were completely noiseless.

Why electron counting suppresses amplitude fluctuations. Equation (6.718) also explains one of the principal advantages of electron counting. An integrating detector retains the variable value Q_i contributed by every primary electron. Fluctuations of deposited energy and generated charge therefore appear directly in the measured signal.

An ideal counting detector instead attempts to recognize the spatial pattern produced by each primary electron and replace its variable analog amplitude by one standardized count,

$$Q_i \longrightarrow 1. \tag{6.722}$$

For perfectly detected, isolated events, the event-amplitude variance is therefore removed:

$$\sigma_Q^2 \longrightarrow 0. \tag{6.723}$$

Counting does not, however, make the complete single-electron response deterministic. The reconstructed event position remains random. Its probability distribution was defined in Eq. (6.525), and its covariance quantifies the uncertainty in where the incident electron is finally recorded. Counting can therefore suppress fluctuations in *event amplitude* while retaining fluctuations in *event position*.

Real counting detectors introduce still other effects, including missed events, false detections, and coincidence loss. These will be treated separately.

From the random single-electron response to the detector PSF. We now restore the spatial structure of the single-electron response. Consider primary electrons entering at the same transverse position $\boldsymbol{\rho}_0$, and define

$$\boldsymbol{u} = \boldsymbol{\rho} - \boldsymbol{\rho}_0. \tag{6.724}$$

For one microscopic realization ω, let

$$\boxed{q_\omega(\boldsymbol{u}) = \mathcal{Q}\left(\boldsymbol{\rho}_0 + \boldsymbol{u} \mid \boldsymbol{\rho}_0, \omega\right)} \tag{6.725}$$

be the continuous collected-charge distribution produced by that electron.

Different values of ω correspond to different primary trajectories, secondary-electron cascades, deposited energies, and carrier-transport histories. Consequently, $q_\omega(\boldsymbol{u})$ is itself a random spatial function.

Its ensemble mean is

$$\boxed{\overline{q}(\boldsymbol{u}) = \mathbb{E}_\omega\left[q_\omega(\boldsymbol{u})\right].} \tag{6.726}$$

Because

$$Q_{\text{tot}}(\omega) = \int q_\omega(\boldsymbol{u})\, d^2\boldsymbol{u}, \tag{6.727}$$

the mean total charge introduced in Eq. (6.682) is also

$$\overline{Q} = \int \overline{q}(\boldsymbol{u})\, d^2\boldsymbol{u}. \tag{6.728}$$

The detector PSF defined in Eq. (6.522) can therefore be written as

$$\boxed{h_{\text{det}}(\boldsymbol{u}) = \frac{\overline{q}(\boldsymbol{u})}{\overline{Q}}.} \tag{6.729}$$

Equation (6.729) makes the meaning of the PSF precise. The PSF is not the spatial pattern produced by one primary electron. It is the normalized *mean* spatial pattern obtained after averaging many statistically equivalent events.

Individual events can be narrower or broader than this mean pattern, can deposit more or less total charge, and can distribute that charge differently among neighbouring positions. All of these differences are lost when only the ensemble mean is retained.

The fluctuations discarded by the PSF. We can make this loss of information explicit by decomposing each event into its mean response and a zero-mean fluctuation,

$$\boxed{q_\omega(\boldsymbol{u}) = \overline{q}(\boldsymbol{u}) + \delta q_\omega(\boldsymbol{u}),} \tag{6.730}$$

where, by construction,

$$\mathbb{E}_\omega\left[\delta q_\omega(\boldsymbol{u})\right] = 0. \tag{6.731}$$

The PSF contains only $\overline{q}(\boldsymbol{u})$. The fluctuations $\delta q_\omega(\boldsymbol{u})$ do not contribute to the mean PSF, but they remain physically present in every individual detector event.

Their spatial statistics are described by the single-electron covariance

$$\boxed{K_q\left(\boldsymbol{u},\boldsymbol{v}\right) = \mathbb{E}_\omega\left[\delta q_\omega(\boldsymbol{u})\,\delta q_\omega(\boldsymbol{v})\right].} \tag{6.732}$$

The diagonal

$$K_q(\boldsymbol{u},\boldsymbol{u}) \tag{6.733}$$

measures the event-to-event variance of the response at one position, whereas the off-diagonal terms measure correlations between different positions.

Thus two detectors can, in principle, have the same mean PSF and hence the same MTF while having different noise properties. One detector may produce almost the same spatial charge cloud for every primary electron, whereas another may alternate between weak, strong, narrow, broad, or displaced events. If the averages of those events are the same, the two detectors have the same PSF, but their second-order statistics are different.

How the random single-electron response enters image noise. We can now combine the spatial response above with the Poisson arrival process introduced in Sec. 6.6.1.

For simplicity, consider uniform illumination with mean incident electron fluence λ per unit area, and let the entrance positions $\boldsymbol{\rho}_i$ form a homogeneous Poisson point process. The continuous detector signal can then be written as

$$\boxed{I(\boldsymbol{\rho}) = \sum_i q_i\left(\boldsymbol{\rho}-\boldsymbol{\rho}_i\right),} \tag{6.734}$$

where each q_i is an independent realization of the random single-electron spatial response.

The mean image is determined by $\overline{q}$. The image fluctuations depend additionally on the second-order statistics of the individual events.

From the single-electron response to the image covariance.

For a homogeneous Poisson distribution of incident electrons, the covariance between detector signals at two positions separated by $\Delta\boldsymbol{\rho}$ is

$$\boxed{C_I(\Delta\boldsymbol{\rho}) = \lambda\int \mathbb{E}_\omega\left[q_\omega(\boldsymbol{u})q_\omega\left(\boldsymbol{u}+\Delta\boldsymbol{\rho}\right)\right]d^2\boldsymbol{u}.} \tag{6.735}$$

Using Eq. (6.730),

$$\mathbb{E}_\omega\left[q_\omega(\boldsymbol{u})q_\omega(\boldsymbol{v})\right] = \overline{q}(\boldsymbol{u})\overline{q}(\boldsymbol{v}) \tag{6.736}$$

$$+ K_q(\boldsymbol{u},\boldsymbol{v}), \tag{6.737}$$

because the terms linear in δq_ω have zero expectation.

Therefore,

$$C_I(\Delta\boldsymbol{\rho}) = \lambda\int \overline{q}(\boldsymbol{u})\overline{q}\left(\boldsymbol{u}+\Delta\boldsymbol{\rho}\right)d^2\boldsymbol{u} \tag{6.738}$$

$$+\lambda\int K_q\left(\boldsymbol{u},\boldsymbol{u}+\Delta\boldsymbol{\rho}\right)d^2\boldsymbol{u}. \tag{6.739}$$

Hence

$$\boxed{C_I = C_{\text{shot}} + C_{\text{SER}},} \tag{6.740}$$

where the first term is the covariance produced by Poisson fluctuations in the number and positions of electrons, propagated through the *mean* single-electron response, and the second term arises from event-to-event fluctuations of the response itself.

Equation (6.740) is the spatial analogue of the scalar decomposition in Eq. (6.713). In the scalar case, the two contributions were

$$\mu\overline{Q}^2 \qquad \text{and} \qquad \mu\sigma_Q^2. \tag{6.741}$$

Here the same distinction is retained, but the fluctuations can now be correlated across different spatial positions.

Connection with the MTF and the later noise power spectrum. The distinction becomes particularly compact in Fourier space. Let

$$\widehat{q}_\omega(\mathbf{R}) \tag{6.742}$$

be the Fourier transform of one random single-electron response. Equation (6.730) gives

$$\widehat{q}_\omega(\mathbf{R}) = \overline{\widehat{q}}(\mathbf{R}) + \delta\widehat{q}_\omega(\mathbf{R}). \tag{6.743}$$

Taking the squared magnitude and averaging,

$$\boxed{\mathbb{E}_\omega\left[\left|\widehat{q}_\omega(\mathbf{R})\right|^2\right] = \left|\overline{\widehat{q}}(\mathbf{R})\right|^2 + \mathbb{E}_\omega\left[\left|\delta\widehat{q}_\omega(\mathbf{R})\right|^2\right].} \tag{6.744}$$

From Eq. (6.729),

$$\overline{\widehat{q}}(\mathbf{R}) = \overline{Q}\, H_{\mathrm{det}}(\mathbf{R}), \tag{6.745}$$

where H_{det} is the detector transfer function introduced in Sec. 6.5.2. Consequently,

$$\boxed{\mathbb{E}_\omega\left[|\widehat{q}_\omega(\mathbf{R})|^2\right] = \overline{Q}^2\,|H_{\mathrm{det}}(\mathbf{R})|^2 + \mathbb{E}_\omega\left[|\delta\widehat{q}_\omega(\mathbf{R})|^2\right].} \tag{6.746}$$

Equation (6.746) is the frequency-dependent counterpart of Eq. (6.713). The first term is what would remain if every electron produced exactly the same spatial response. In that idealized case, the spatial distribution of the shot noise would be shaped by

$$|H_{\mathrm{det}}(\mathbf{R})|^2\,. \tag{6.747}$$

The second term exists because the response to one electron is itself random. It contains fluctuations of event amplitude, event width, charge-sharing pattern, and localization. Consequently, the noise properties of a real detector cannot in general be predicted simply by squaring its MTF.

The fundamental distinction is therefore

$$\boxed{\begin{array}{lcl} \text{mean single-electron response} & \Longrightarrow & h_{\mathrm{det}} \Longrightarrow \text{MTF}, \\ \text{second-order single-electron statistics} & \Longrightarrow & \text{spatial noise covariance} \Longrightarrow \text{noise power spectrum}. \end{array}} \tag{6.748}$$

Equivalently,

$$\boxed{\text{MTF depends on } |\mathbb{E}[\widehat{q}_\omega]|\,, \qquad \text{noise power depends on } \mathbb{E}\left[|\widehat{q}_\omega|^2\right].} \tag{6.749}$$

In general,

$$\mathbb{E}\left[|\widehat{q}_\omega|^2\right] \neq |\mathbb{E}[\widehat{q}_\omega]|^2\,. \tag{6.750}$$

The difference between these two quantities is the Fourier-space signature of the event-to-event fluctuations. The formal definition of the noise power spectrum and its relation to spatial covariance will be introduced later. At this stage, the important conclusion is that the PSF and MTF characterize only the *average* single-electron response. A complete description of detector performance also requires the second-order statistics of that random response.

6.6.3 Energy-deposition and charge-generation fluctuations

The preceding subsection treated the response of one incident electron as a random quantity and introduced the single-electron-response variance σ_Q^2. We now identify the first microscopic mechanisms that contribute to this variance.

The relevant sequence is

$$\boxed{\begin{array}{c}\text{primary electron}\\ \Downarrow\\ \text{stochastic inelastic collisions}\\ \Downarrow\\ E_{\text{dep}}\\ \Downarrow\\ N_{eh}\\ \Downarrow\\ \text{collected electrical signal.}\end{array}} \tag{6.751}$$

Randomness enters at two distinct stages. First, different primary electrons deposit different amounts of energy in the active silicon. Second, even for a fixed deposited energy, the number of electron–hole pairs generated from that energy fluctuates.

Fluctuations of the deposited energy. The stopping-power calculation developed earlier describes an average energy loss. For example, Eq. (6.336) gives an average loss of about 15.5 keV for a 300 keV electron traversing 35 μm of silicon in the straight-track approximation. Individual electrons, however, do not all lose this same amount of energy.

The reason is that energy loss occurs through a discrete sequence of inelastic interactions. Most collisions transfer relatively modest amounts of energy, whereas occasional hard collisions transfer much larger energies. The number, positions, and transferred energies of these interactions differ from one primary-electron trajectory to another. This event-to-event variability is the *energy straggling* already discussed in connection with Eq. (6.340).

For a sufficiently thin absorber, the distribution of energy loss is therefore strongly asymmetric. It has a relatively pronounced maximum and a long tail toward large energy transfers and is commonly described approximately by Landau or Landau–Vavilov statistics. In particular, the most probable loss is generally smaller than the mean, as expressed in Eq. (6.339).

This behaviour has been observed directly in silicon electron detectors. Simulations of 300 keV electrons traversing thin silicon produce strongly

asymmetric single-electron signal distributions, and experimental single-electron responses of direct detectors show similar Landau-like shapes (Dourki et al., 2017; McMullan et al., 2014).

It is important, however, to distinguish once again between the energy lost by the primary electron and the energy actually deposited in the active detector. The former was defined in Eq. (6.341), whereas the latter was defined in Eq. (6.342). They obey

$$E_{\text{dep}} \leq E_{\text{loss}}, \tag{6.752}$$

as established in Eq. (6.343).

An energetic secondary electron produced by a hard collision may escape from the active layer before depositing all of its kinetic energy. Similarly, a backscattered or transmitted primary electron may leave the sensor carrying a substantial fraction of its original energy. Consequently, the deposited energy

$$E_{\text{dep}} \tag{6.753}$$

is itself a random variable whose distribution reflects not only the statistics of the inelastic collisions, but also the geometry and thickness of the active detector volume.

The quantity that matters for the electrical signal is E_{dep}, not E_{loss}, because only energy deposited in the active silicon can generate charge available for collection.

From deposited energy to electron–hole pairs. Once an amount of energy E_{dep} has been deposited in the active silicon, it is converted into electronic excitations and ultimately into mobile electron–hole pairs. The mean energy required to create one pair was introduced in Eq. (6.397),

$$w_{\text{Si}} \simeq 3.6 \text{ eV}, \tag{6.754}$$

more precisely about 3.66 eV under the conditions measured by Scholze et al. (2000).

Therefore, for a specified deposited energy, Eq. (6.398) gives the conditional mean

$$\boxed{\mathbb{E}\left[N_{eh} \mid E_{\text{dep}}\right] = \frac{E_{\text{dep}}}{w_{\text{Si}}}.} \tag{6.755}$$

The word *mean* is essential. Even if two events deposited exactly the same energy, they would not necessarily generate exactly the same number of electron–hole pairs.

Fano fluctuations at fixed deposited energy. The fluctuations in pair creation at fixed deposited energy are described by the semiconductor Fano factor F. As established in Eq. (6.401),

$$\boxed{\mathrm{Var}\,(N_{eh} \mid E_{\mathrm{dep}}) = F\frac{E_{\mathrm{dep}}}{w_{\mathrm{Si}}},} \tag{6.756}$$

with

$$F \simeq 0.1\text{–}0.13 \tag{6.757}$$

for silicon.

If the production of the individual electron–hole pairs were an independent Poisson process, the variance would instead be equal to the mean. The fact that

$$F < 1 \tag{6.758}$$

means that pair creation is *sub-Poissonian*. The deposited energy is shared among electronic excitation, lattice excitation, and other microscopic channels in a correlated manner, so the final number of pairs fluctuates less than it would if every pair-creation event were independent.

This Fano fluctuation must not be confused with the incident-electron shot noise discussed in Sec. 6.6.1. They occur at completely different levels:

incident shot noise	:	how many primary electrons arrive,
Fano fluctuation	:	how many electron–hole pairs are generated by one specified deposited energy.

(6.759)

Two successive contributions to the generated-charge variance. The actual detector event contains both sources of randomness simultaneously. The deposited energy changes from electron to electron, and the number of pairs generated from a specified deposited energy also fluctuates.

This decomposition was derived previously in Eq. (6.412):

$$\boxed{\mathrm{Var}(N_{eh}) = F\frac{\mathbb{E}[E_{\mathrm{dep}}]}{w_{\mathrm{Si}}} + \frac{\mathrm{Var}(E_{\mathrm{dep}})}{w_{\mathrm{Si}}^2}.} \tag{6.760}$$

The two terms have a particularly clear interpretation:

$$\boxed{\underbrace{F\frac{\mathbb{E}[E_{\mathrm{dep}}]}{w_{\mathrm{Si}}}}_{\text{pair-generation fluctuations}} + \underbrace{\frac{\mathrm{Var}(E_{\mathrm{dep}})}{w_{\mathrm{Si}}^2}}_{\text{energy-deposition fluctuations}}.} \tag{6.761}$$

The first contribution would remain even if every primary electron deposited exactly the same energy. The second would remain even if the conversion of deposited energy into electron–hole pairs were perfectly deterministic.

Relative importance of the two contributions. The relative size of the two mechanisms becomes especially transparent if Eq. (6.760) is divided by the square of the mean number of generated pairs.

From Eq. (6.755),

$$\mathbb{E}[N_{eh}] = \frac{\mathbb{E}[E_{\mathrm{dep}}]}{w_{\mathrm{Si}}}. \tag{6.762}$$

Therefore,

$$\boxed{\frac{\mathrm{Var}(N_{eh})}{\mathbb{E}[N_{eh}]^2} = \underbrace{F\frac{w_{\mathrm{Si}}}{\mathbb{E}[E_{\mathrm{dep}}]}}_{\text{relative Fano contribution}} + \underbrace{\frac{\mathrm{Var}(E_{\mathrm{dep}})}{\mathbb{E}[E_{\mathrm{dep}}]^2}}_{\text{relative energy-straggling contribution}}.} \tag{6.763}$$

The second term is simply the squared coefficient of variation of the deposited energy,

$$\frac{\mathrm{Var}(E_{\mathrm{dep}})}{\mathbb{E}[E_{\mathrm{dep}}]^2} = \left[\frac{\sigma_{E_{\mathrm{dep}}}}{\mathbb{E}[E_{\mathrm{dep}}]}\right]^2. \tag{6.764}$$

Equation (6.763) exposes an important difference between the two mechanisms. The relative Fano contribution decreases inversely with the mean deposited energy, whereas the energy-straggling contribution is governed by the relative width of the deposited-energy distribution itself.

For the scale considered in Eq. (6.336), if approximately 15.5 keV were deposited in the active silicon, the mean number of pairs would be of order

$$N_{eh} \simeq 4.3 \times 10^3. \tag{6.765}$$

Taking

$$F = 0.125, \tag{6.766}$$

the standard deviation associated purely with Fano pair creation would be

$$\sigma_{N,F} \simeq \sqrt{F N_{eh}} \simeq 23, \tag{6.767}$$

corresponding to a relative fluctuation of only about 0.5%.

This number should not be interpreted as the total fluctuation of the single-electron signal. The distribution of E_{dep} in a thin silicon sensor is

much broader and strongly asymmetric because of energy straggling and the possible escape of energetic secondaries. For energetic electrons, the second term in Eq. (6.763) can therefore dominate the intrinsic Fano contribution.

Connection with the single-electron excess-noise factor. We can now connect these microscopic fluctuations with the single-electron-response description of Sec. 6.6.2.

Suppose temporarily that every generated charge carrier is collected with the same deterministic efficiency, so that the total collected charge is proportional to the number of generated pairs,

$$Q_{\text{tot}} = g\, N_{eh}, \tag{6.768}$$

where g is a constant conversion factor. In the simplest case of collecting one carrier species perfectly,

$$g = e. \tag{6.769}$$

Because multiplication by a constant does not change the relative variance,

$$\frac{\text{Var}(Q_{\text{tot}})}{\mathbb{E}[Q_{\text{tot}}]^2} = \frac{\text{Var}(N_{eh})}{\mathbb{E}[N_{eh}]^2}. \tag{6.770}$$

Using Eq. (6.763),

$$\boxed{\frac{\sigma_Q^2}{\overline{Q}^2} = F\frac{w_{\text{Si}}}{\mathbb{E}[E_{\text{dep}}]} + \frac{\text{Var}(E_{\text{dep}})}{\mathbb{E}[E_{\text{dep}}]^2}.} \tag{6.771}$$

The corresponding contribution to the single-electron excess-noise factor introduced in Eq. (6.717) is therefore

$$\boxed{\mathcal{F}_{\text{SER}} = 1 + F\frac{w_{\text{Si}}}{\mathbb{E}[E_{\text{dep}}]} + \frac{\text{Var}(E_{\text{dep}})}{\mathbb{E}[E_{\text{dep}}]^2},} \tag{6.772}$$

under the idealized assumption of deterministic charge collection.

Equation (6.772) provides the direct connection between the microscopic interaction physics and the SNR degradation derived in Eq. (6.718). Even before readout noise, dark current, gain variation, or charge-collection fluctuations are considered, an integrating detector can have

$$\mathcal{F}_{\text{SER}} > 1 \tag{6.773}$$

simply because one primary electron does not always deposit the same energy and a fixed deposited energy does not always generate exactly the same number of charge carriers.

The assumption leading to Eq. (6.772) is deliberately restrictive. Real charge collection introduces additional fluctuations and redistributes the generated carriers among neighbouring pixels. Those effects do not belong to energy deposition or pair generation themselves and will be considered separately in the following subsections.

Integrating versus counting operation. These fluctuations affect integrating and electron-counting detectors differently. An integrating detector retains the analog amplitude of each event. Consequently, the fluctuations quantified by Eq. (6.771) contribute directly to the single-electron variance σ_Q^2 and therefore reduce the SNR according to Eq. (6.718).

An ideal counting detector instead attempts to recognize that a primary electron has arrived and replace its variable analog amplitude by one standardized count. Once an isolated event has been detected reliably, the fact that it generated, for example, 3500 rather than 5000 electron–hole pairs need not appear in the final counted image.

The amplitude fluctuations have not disappeared physically; they are used internally to detect and localize the event. They cease to contribute directly to the final image amplitude only after successful event classification. If the deposited energy is unusually small, however, the event may fall below the detection threshold, whereas a very broad or unusually energetic event can affect several pixels. Thus energy-deposition statistics remain relevant even for counting detectors, although their effect is no longer simply an analog amplitude fluctuation.

6.6.4 Charge-sharing and localization fluctuations

The preceding subsection considered fluctuations in the *amount* of charge generated by one primary electron. A second type of single-electron variability concerns *where* that charge is collected.

As discussed in Sec. 6.4.4, the geometrical boundaries between pixels are boundaries of the readout architecture, not physical walls inside the sensitive silicon. The charge generated by one primary electron may therefore be divided among several neighbouring collection nodes. For an event contained within a cluster of pixels, Eq. (6.515) gives

$$Q_{\text{event}} = \sum_{p \in \text{cluster}} Q_p. \tag{6.774}$$

The important point for the present discussion is that this division is not identical from one primary electron to another. Even for electrons entering at the same position and with the same kinetic energy, the primary trajectory, secondary-electron cascade, depth of energy deposition, and subsequent carrier transport fluctuate. Consequently, the fractions of the total event charge reaching the different pixels also fluctuate.

In the notation introduced in Eq. (6.684),

$$\mathbf{Q} = Q_{\text{tot}}\boldsymbol{\pi}, \tag{6.775}$$

where

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_P), \qquad \sum_p \pi_p = 1. \tag{6.776}$$

The random variable Q_{tot} describes the event amplitude, whereas the random vector $\boldsymbol{\pi}$ describes the spatial sharing of that amplitude among the pixels.

Thus a single-electron event can fluctuate even when its total collected charge remains unchanged: the same amount of charge may simply be redistributed differently over the pixel array.

Charge sharing as a redistribution fluctuation. To isolate fluctuations of the *spatial distribution* of an event from fluctuations of its total amplitude, suppose temporarily that every primary electron produces exactly the same total collected charge,

$$Q_{\text{tot}} = Q. \tag{6.777}$$

The only randomness then lies in how this fixed amount of charge is distributed among the pixels.

Let

$$\pi_p \tag{6.778}$$

denote the fraction of the total event charge collected by pixel p. Then

$$\boxed{Q_p = Q\pi_p,} \tag{6.779}$$

and conservation of the total charge requires

$$\boxed{\sum_p \pi_p = 1.} \tag{6.780}$$

Because the microscopic trajectory and charge-transport history vary from event to event, the fractions π_p are random variables. We write each of them as its mean value plus a fluctuation,

$$\pi_p = \overline{\pi}_p + \delta\pi_p, \tag{6.781}$$

where

$$\overline{\pi}_p = \mathbb{E}[\pi_p], \qquad \mathbb{E}[\delta\pi_p] = 0. \tag{6.782}$$

Taking the expectation of Eq. (6.780) gives

$$\sum_p \overline{\pi}_p = 1. \tag{6.783}$$

Subtracting this mean relation from the event-by-event normalization in Eq. (6.780) yields

$$\boxed{\sum_p \delta\pi_p = 0.} \tag{6.784}$$

Equation (6.784) expresses the essential physics of pure charge redistribution. If one pixel receives a larger fraction of the charge than usual, some other part of the cluster must receive less, because the total available charge has been held fixed.

Using Eq. (6.779), the fluctuation of the charge measured in pixel p is

$$\delta Q_p = Q\,\delta\pi_p. \tag{6.785}$$

Summing over all pixels and using Eq. (6.784),

$$\boxed{\sum_p \delta Q_p = 0.} \tag{6.786}$$

Thus pure charge sharing changes the distribution of charge among the pixels but does not change the total event charge.

We can now examine the correlations that this redistribution produces. For fixed total charge Q, the covariance between the signals measured in pixels p and q is

$$\boxed{\mathrm{Cov}(Q_p, Q_q \mid Q) = \mathbb{E}\left[\delta Q_p \delta Q_q \mid Q\right].} \tag{6.787}$$

Using Eq. (6.785), this can also be written as

$$\boxed{\mathrm{Cov}(Q_p, Q_q \mid Q) = Q^2\,\mathrm{Cov}(\pi_p, \pi_q).} \tag{6.788}$$

Now fix one pixel p and sum its covariance with all pixels q:

$$\sum_q \mathrm{Cov}(Q_p, Q_q \mid Q) = \sum_q \mathbb{E}\left[\delta Q_p \delta Q_q \mid Q\right] \tag{6.789}$$

$$= \mathbb{E}\left[\delta Q_p \sum_q \delta Q_q \mid Q\right]. \tag{6.790}$$

From Eq. (6.786), the sum inside the expectation is identically zero. Therefore,

$$\boxed{\sum_q \mathrm{Cov}(Q_p, Q_q \mid Q) = 0.} \tag{6.791}$$

The term with $q = p$ is simply

$$\mathrm{Cov}(Q_p, Q_p \mid Q) = \mathrm{Var}(Q_p \mid Q). \tag{6.792}$$

Equation (6.791) can therefore be rewritten as

$$\boxed{\sum_{q \neq p} \mathrm{Cov}(Q_p, Q_q \mid Q) = -\,\mathrm{Var}(Q_p \mid Q).} \tag{6.793}$$

This result gives a precise statistical expression of charge redistribution. The variance of the charge measured in one pixel is necessarily balanced by negative covariance with the other pixels in the event cluster. If pixel p receives more charge than usual in a particular event, the remaining pixels must, collectively, receive less.

Charge sharing therefore does not merely broaden the spatial response of an electron. Its event-to-event fluctuations also generate correlations between neighbouring pixel signals. These correlations will later contribute to the spatial covariance and hence to the noise power spectrum of an integrating detector.

A two-pixel example. Consider the simplest possible shared event, in which the fixed total charge Q is divided between two neighbouring pixels. Let the random fraction collected by the first pixel be f. Then

$$Q_1 = fQ, \qquad Q_2 = (1 - f)Q. \tag{6.794}$$

If

$$\mathbb{E}[f] = \overline{f}, \qquad \mathrm{Var}(f) = \sigma_f^2, \tag{6.795}$$

then

$$\mathbb{E}[Q_1] = Q\overline{f}, \qquad \mathbb{E}[Q_2] = Q(1 - \overline{f}). \tag{6.796}$$

The two variances are

$$\boxed{\mathrm{Var}(Q_1) = \mathrm{Var}(Q_2) = Q^2\sigma_f^2,} \tag{6.797}$$

whereas their covariance is

$$\mathrm{Cov}(Q_1, Q_2) = \mathrm{Cov}\,(fQ, (1-f)Q) \tag{6.798}$$

$$= Q^2\,\mathrm{Cov}(f, 1-f) \tag{6.799}$$

$$= \boxed{-Q^2\sigma_f^2.} \tag{6.800}$$

Thus, in this simple example, the charge-sharing fluctuation is perfectly anticorrelated: whenever pixel 1 receives more charge than usual, pixel 2 necessarily receives less.

Nevertheless,

$$Q_1 + Q_2 = Q \tag{6.801}$$

for every event. Therefore,

$$\mathrm{Var}(Q_1 + Q_2) = 0. \tag{6.802}$$

The fluctuation affects the *spatial distribution* of the signal, but not its total amplitude.

Consequences for an integrating detector. This distinction is important for an integrating detector. The *mean* charge-sharing pattern contributes to the average single-electron response and therefore to the detector PSF. Charge sharing that is reproducible from event to event is therefore primarily a source of spatial blur, not noise.

The event-to-event deviations from that mean pattern are different. They contribute to the covariance matrix introduced in Eq. (6.680) and, in the continuous description, to the single-electron spatial covariance K_q introduced in Eq. (6.732).

It is therefore useful to distinguish

$$\boxed{\begin{array}{ccl} \text{mean charge sharing} & \Longrightarrow & \text{PSF broadening and MTF attenuation,} \\ \text{fluctuations of charge sharing} & \Longrightarrow & \text{spatially correlated detector fluctuations.} \end{array}} \tag{6.803}$$

Pure charge redistribution has a further useful property. If the total charge of an event does not fluctuate, then

$$\int \delta q_\omega(\boldsymbol{u})\, d^2\boldsymbol{u} = 0, \tag{6.804}$$

because every excess in one part of the charge cloud is compensated by a deficit elsewhere.

Equivalently, at zero spatial frequency,

$$\boxed{\delta\widehat{q}_\omega(\mathbf{0}) = 0} \tag{6.805}$$

for a pure redistribution fluctuation. Thus charge-sharing fluctuations can modify the spatial structure of the noise without necessarily changing the total single-electron signal.

In a real detector the separation is not exact, because event amplitude and event shape need not be independent. An energetic secondary electron, for example, can simultaneously alter both the total deposited energy and the spatial extent of the charge cloud. Nevertheless, the distinction between amplitude fluctuations and redistribution fluctuations remains useful for understanding their different effects.

From charge sharing to localization fluctuations. For an electron-counting detector, the same spatially distributed charge cluster is used differently. Rather than retaining the complete analog vector $\mathbf{Q}$ in the final image, the detector applies an event-localization algorithm

$$\mathcal{A} : \mathbf{Q} \longrightarrow \widehat{\boldsymbol{\rho}}, \tag{6.806}$$

as introduced in Eq. (6.525).

The estimated coordinate $\widehat{\boldsymbol{\rho}}$ is random because the input cluster $\mathbf{Q}$ is random. Define the localization error of one primary electron as

$$\boxed{\boldsymbol{\epsilon} = \widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}_0.} \tag{6.807}$$

Its mean is the localization bias already defined in Eq. (6.526),

$$\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{b}(\boldsymbol{\rho}_0), \tag{6.808}$$

and the fluctuations around that mean are described by $\boldsymbol{\Sigma}_{\mathrm{loc}}$ in Eq. (6.527).

For an approximately isotropic, unbiased detector, Eq. (6.528) summarized the localization uncertainty by

$$\sigma_{\mathrm{loc}} = \sqrt{\frac{\mathrm{Tr}\,\boldsymbol{\Sigma}_{\mathrm{loc}}}{2}}. \tag{6.809}$$

The origin of $\boldsymbol{\epsilon}$ is therefore not a new physical interaction occurring after charge collection. It is the statistical consequence of trying to infer the unknown entrance position $\boldsymbol{\rho}_0$ from a stochastic charge-sharing pattern.

Localization fluctuations and the counting-detector PSF. Consider an ideal counting detector that recognizes every primary electron exactly once and assigns it to the reconstructed position

$$\widehat{\boldsymbol{\rho}} = \boldsymbol{\rho}_0 + \boldsymbol{\epsilon}. \tag{6.810}$$

The recorded spatial event corresponding to one primary electron can then be represented as

$$\boxed{c_\omega(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0) = \delta^{(2)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_0 - \boldsymbol{\epsilon}_\omega\right).} \tag{6.811}$$

Averaging over many identical primary electrons gives

$$\overline{c}(\boldsymbol{\rho} \mid \boldsymbol{\rho}_0) = \mathbb{E}_\omega\left[\delta^{(2)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_0 - \boldsymbol{\epsilon}_\omega\right)\right] \tag{6.812}$$

$$= p_\epsilon\left(\boldsymbol{\rho} - \boldsymbol{\rho}_0\right), \tag{6.813}$$

where p_ϵ is the probability density of the localization error. Hence

$$\boxed{h_{\mathrm{loc}}(\boldsymbol{u}) = p_\epsilon(\boldsymbol{u})} \tag{6.814}$$

for a translationally invariant and unbiased localization process.

Thus the distribution of localization errors is itself the single-electron PSF of an ideal counting detector. A narrow error distribution gives a narrow counting PSF; a broad localization distribution gives a broader one.

Transfer function produced by localization uncertainty.

The Fourier transform of the localized event is

$$\widehat{c}_\omega(\mathbf{R}) = \exp\left[-j2\pi\mathbf{R}\cdot\left(\boldsymbol{\rho}_0 + \boldsymbol{\epsilon}_\omega\right)\right]. \tag{6.815}$$

Because the true entrance coordinate $\boldsymbol{\rho}_0$ is fixed whereas the localization error $\boldsymbol{\epsilon}_\omega$ is random, this can be factored as

$$\widehat{c}_\omega(\mathbf{R}) = e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\rho}_0} e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\epsilon}_\omega}. \tag{6.816}$$

Taking the ensemble average over many statistically equivalent localization events gives

$$\mathbb{E}_\omega\left[\widehat{c}_\omega(\mathbf{R})\right] = e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\rho}_0}\mathbb{E}_\omega\left[e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\epsilon}_\omega}\right]. \tag{6.817}$$

The first factor is the Fourier phase of an ideal point event located at $\boldsymbol{\rho}_0$. It describes where the electron actually entered the detector. The

second factor describes only the loss of transfer caused by uncertainty in the reconstructed position.

We therefore define the localization transfer function as

$$\boxed{H_{\mathrm{loc}}(\mathbf{R}) = \mathbb{E}\left[e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\epsilon}}\right].} \tag{6.818}$$

Hence the full averaged response is

$$\boxed{\mathbb{E}_\omega\left[\widehat{c}_\omega(\mathbf{R})\right] = e^{-j2\pi\mathbf{R}\cdot\boldsymbol{\rho}_0} H_{\mathrm{loc}}(\mathbf{R}).} \tag{6.819}$$

Thus H_{loc} is the characteristic function of the localization-error distribution. It multiplies the ideal Fourier response of a point located at $\boldsymbol{\rho}_0$.

If the localization error is a zero-mean Gaussian random vector,

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{loc}}\right), \tag{6.820}$$

its characteristic function gives

$$\boxed{H_{\mathrm{loc}}(\mathbf{R}) = \exp\left(-2\pi^2\mathbf{R}^T\boldsymbol{\Sigma}_{\mathrm{loc}}\mathbf{R}\right).} \tag{6.821}$$

For an isotropic detector,

$$\boldsymbol{\Sigma}_{\mathrm{loc}} = \sigma_{\mathrm{loc}}^2\mathbf{I}, \tag{6.822}$$

so

$$\boxed{\mathrm{MTF}_{\mathrm{loc}}(R) = \exp\left(-2\pi^2\sigma_{\mathrm{loc}}^2R^2\right).} \tag{6.823}$$

Localization uncertainty is not simply additive noise. It would be tempting to interpret $\boldsymbol{\epsilon}$ as an additional additive detector-noise term. For an ideal counting detector, however, this interpretation is not generally correct.

Every successfully detected primary electron in Eq. (6.811) still contributes exactly one count. Its position changes, but its amplitude does not. Indeed,

$$\left|\widehat{c}_\omega(\mathbf{R})\right|^2 = 1 \tag{6.824}$$

for every event and every spatial frequency, because a spatial displacement changes only the Fourier phase.

Consequently,

$$\boxed{\mathbb{E}\left[|\widehat{c}_\omega(\mathbf{R})|^2\right] = 1.} \tag{6.825}$$

This result has an important interpretation. If the incident electrons form a homogeneous Poisson process, every electron is detected exactly once, and the localization errors of different electrons are independent, random localization displacements do not by themselves create an additional count variance. They primarily reduce the spatial transfer of the signal through Eq. (6.818).

Equivalently, an independently displaced homogeneous Poisson point process is still a homogeneous Poisson point process. Thus ideal localization uncertainty can reduce the MTF while leaving the shot-noise statistics of a uniformly illuminated counting image Poissonian.

This distinction will become important when the noise power spectrum and DQE are introduced: loss of spatial information need not always appear as an increase in noise power. It may instead appear through a reduction of the transferred signal.

Real localization algorithms. The ideal result above assumes that every primary electron produces exactly one recorded event. Real counting detectors need not satisfy this assumption.

The stochastic charge-sharing pattern can interact with the event threshold and localization algorithm. A weak or unusually diffuse event may fail to cross the detection threshold. A cluster may be split incorrectly into more than one event, or two nearby primary-electron clusters may be merged. The latter mechanism leads to coincidence loss and will be treated separately.

Thus charge-sharing fluctuations can affect a counting detector in two different ways:

$$\boxed{\begin{array}{lcl} \text{successful event localization} & \Longrightarrow & \text{random position error and MTF attenuation,} \\ \text{detection or classification failure} & \Longrightarrow & \text{modified count statistics.} \end{array}} \tag{6.826}$$

The distinction is important. Charge sharing itself is neither simply "blur" nor simply "noise." Its mean spatial pattern contributes to the PSF, its event-to-event variation produces spatial covariance in an integrating detector, and in a counting detector the same spatial information can be used to estimate the incident position with subpixel precision.

The physical pixel pitch, the width of the charge cloud, and the localization uncertainty are therefore different quantities. As already shown by

Eqs. (6.518), (6.519), and (6.528), an event may extend over several pixels while its entrance coordinate can nevertheless be estimated with substantially better precision than one pixel.

6.6.5 Coincidence loss in electron counting

The ideal counting picture developed above assumes that every primary electron can be recognized as a separate event. As summarized in Eq. (6.512), the detector examines a local pattern of digitized pixel values and decides whether that pattern corresponds to one incident electron.

This assumption eventually fails as the incident dose rate increases. One primary electron normally generates a cluster extending over several neighbouring pixels. If another electron arrives sufficiently close in space during the same internal sampling interval, the two clusters may overlap. The counting algorithm may then be unable to determine that two primary electrons were present and may reconstruct them as only one event.

This loss of separately identifiable events is called *coincidence loss*. It has been characterized experimentally in electron-counting direct detectors and places an important upper limit on the useful incident dose rate (Li et al., 2013; McMullan et al., 2014).

Coincidence loss should be distinguished from the stochastic noise sources considered so far. It is not an independent fluctuation added to the measured signal. It is a *nonlinear counting process*: the probability that an electron is recorded correctly depends on how many other electrons arrive nearby during the same internal frame.

An effective event-recognition area. To obtain a simple quantitative model, it is useful first to recall the orders of magnitude obtained in Sec. 2.4.6. For the representative cryo-EM acquisition of Eq. (2.122),

$$D_{\text{tot}} = 50\ e^{-}/\text{Å}^{2}, \qquad T_{\text{exp}} = 2\ \text{s}, \tag{6.827}$$

over a 4096×4096 image sampled at 0.8 Å/pixel, the complete recorded field receives approximately

$$2.68 \times 10^{8} \quad \text{electrons per second}, \tag{6.828}$$

as calculated in Eq. (2.144). The mean temporal separation between successive electrons arriving *somewhere in the complete field of view* is therefore only

$$\Delta t_e \simeq 3.73\ \text{ns}, \tag{6.829}$$

as obtained in Eq. (2.146).

At first sight, such a short interval might suggest that overlapping electron events should be unavoidable. However, coincidence is not determined by the arrival rate over the complete detector. Two electrons can arrive almost simultaneously at widely separated positions and be recognized independently. What matters is the arrival rate within the small spatial neighbourhood over which two electron events can no longer be reliably separated.

The same representative acquisition illustrates the difference. One image pixel corresponds at the specimen to an area

$$A_{\text{pix}} = (0.8\ \text{Å})^2 = 0.64\ \text{Å}^2. \tag{6.830}$$

The mean dose rate is

$$\frac{D_{\text{tot}}}{T_{\text{exp}}} = 25\ e^-\ \text{Å}^{-2}\,\text{s}^{-1}, \tag{6.831}$$

so, for this sampling, the mean rate associated with one image pixel is

$$\dot{D} = (25)(0.64) \simeq 16\ e^-\ \text{pixel}^{-1}\,\text{s}^{-1}. \tag{6.832}$$

The corresponding mean interval between electrons associated with one such pixel is therefore

$$\boxed{\Delta t_{\text{pix}} = \frac{1}{\dot{D}} \simeq 62.5\ \text{ms}.} \tag{6.833}$$

Thus the same exposure has two very different characteristic times:

$$\boxed{\begin{array}{rcl} \text{complete field of view} & : & \Delta t_e \simeq 3.73\ \text{ns}, \\ \text{one image-pixel area} & : & \Delta t_{\text{pix}} \simeq 62.5\ \text{ms}. \end{array}} \tag{6.834}$$

The enormous difference arises simply because the incident electrons are distributed over millions of spatial positions.

For the coincidence calculation we therefore consider a spatially uniform incident rate

$$\dot{D}, \tag{6.835}$$

measured in primary electrons per physical pixel per second, and let

$$F \tag{6.836}$$

be the internal detector frame rate, measured in s^{-1}. The duration of one internal sampling interval is

$$\boxed{T_{\text{int}} = \frac{1}{F}.} \tag{6.837}$$

Coincidence is controlled not by comparing T_{int} with the global 3.73 ns interval, but by comparing it with the local arrival time within the complete event-recognition neighbourhood.

The event generated by one primary electron does not occupy only one pixel. Let

$$s_{\text{eff}} \tag{6.838}$$

denote an *effective coincidence area*, expressed as the number of physical-pixel areas over which another electron would interfere with the recognition of the first event.

The quantity s_{eff} should not be interpreted simply as the geometrical area of the charge cloud. It depends on the spatial extent of the event, the detection threshold, the clustering rules, and the localization algorithm. It is therefore an effective property of the complete counting system.

Probability of coincidence during one internal frame.

During one internal sampling interval, the expected number of additional primary electrons entering an effective area of s_{eff} pixel areas is

$$\boxed{\lambda_{\text{c}} = D\, s_{\text{eff}}\, T_{\text{int}} = \frac{D\, s_{\text{eff}}}{F}.} \tag{6.839}$$

The incident electrons obey the Poisson arrival statistics derived in Sec. 6.6.1. Therefore, using Eq. (6.623), the probability that exactly k additional electrons enter this neighbourhood during the same sampling interval is

$$P(K = k) = \frac{\lambda_{\text{c}}^{k} e^{-\lambda_{\text{c}}}}{k!}. \tag{6.840}$$

The reference electron is isolated if no additional electron enters its effective neighbourhood. Hence

$$\boxed{P_{\text{iso}} = P(K = 0) = e^{-\lambda_{\text{c}}} = \exp\left(-\frac{D\, s_{\text{eff}}}{F}\right).} \tag{6.841}$$

The probability that at least one additional electron enters the same

effective neighbourhood is therefore

$$P_{\text{coin}} = 1 - P_{\text{iso}} \tag{6.842}$$

$$= \boxed{1 - \exp\left(-\frac{D\, s_{\text{eff}}}{F}\right).} \tag{6.843}$$

When the occupancy is small,

$$\lambda_{\text{c}} \ll 1, \tag{6.844}$$

the exponential can be expanded as

$$e^{-\lambda_{\text{c}}} = 1 - \lambda_{\text{c}} + \frac{\lambda_{\text{c}}^2}{2} - \cdots, \tag{6.845}$$

giving

$$\boxed{P_{\text{coin}} \simeq \lambda_{\text{c}} = \frac{D\, s_{\text{eff}}}{F}.} \tag{6.846}$$

Equation (6.846) explains the essential design dependence of coincidence loss. It increases with the incident dose rate D and with the effective spatial extent s_{eff} of an electron event, whereas increasing the internal frame rate F reduces the probability that two nearby events occur during the same sampling interval.

Thus fast readout and spatially compact, accurately localized single-electron responses are both advantageous for electron counting.

Coincidence probability and count loss are not identical. Some care is required when interpreting Eq. (6.843). It gives the probability that a given primary electron has at least one additional electron inside its effective coincidence neighbourhood during the same internal frame. It does not imply that every such electron simply disappears from the output.

For example, two overlapping primary electrons may be reconstructed as one event. Two electrons entered the detector, but one event was recorded. More complicated overlaps may be split, merged, or rejected depending on the counting algorithm.

Consequently, the exact relation between P_{coin} and the number of lost counts is algorithm-dependent. The effective area s_{eff} provides a useful phenomenological way of summarizing this complicated spatial decision process (Li et al., 2013).

A simple dose-dependent counting efficiency. Let

$$\eta_0 \tag{6.847}$$

denote the probability that an isolated primary electron is successfully detected and counted. From Eq. (6.841), the probability that the electron has no additional primary electron inside its effective coincidence neighbourhood during the same internal frame is

$$P_{\text{iso}} = \exp\left(-\frac{D\, s_{\text{eff}}}{F}\right). \tag{6.848}$$

If, as a simple approximation, we assume that an electron contributes an independent count only when it is both successfully detected and spatially isolated, then its effective counting probability is

$$\boxed{\eta_{\text{eff}}(D) = \eta_0 \exp\left(-\frac{D\, s_{\text{eff}}}{F}\right).} \tag{6.849}$$

Multiplying this probability by the incident electron rate gives the rate of independently resolved events and the corresponding recorded count rate

$$\boxed{D_{\text{out}} = D_{\text{iso}} \simeq \eta_0 D \exp\left(-\frac{D\, s_{\text{eff}}}{F}\right).} \tag{6.850}$$

Equation (6.850) should be regarded as an effective low-to-moderate-occupancy model rather than a universal high-dose counting law. Its important message is the departure from linearity:

$$D_{\text{out}} \not\propto D \tag{6.851}$$

once coincidence becomes appreciable.

At very low dose rate,

$$\frac{D\, s_{\text{eff}}}{F} \ll 1, \tag{6.852}$$

Eq. (6.850) approaches the linear response

$$D_{\text{out}} \simeq \eta_0 D. \tag{6.853}$$

As the dose rate increases, a progressively larger fraction of the electron events cannot be separated from nearby events.

Numerical scale for the K2 Summit. The K2 Summit provides a useful experimental example. Li et al. (2013) modeled coincidence by assuming that a primary electron is affected when another event occurs within a fixed area s during the same internal accumulation period. Fitting this model to the experimentally measured dose-rate-dependent counting efficiency gave

$$s = 5.1459 \text{ pixel}^2, \tag{6.854}$$

corresponding to a circular region with diameter approximately 2.57 physical pixels. The corresponding effective diameter is about 2.6 physical pixels. The internal frame rate discussed in Sec. 6.4.4 is

$$F = 400 \text{ s}^{-1}. \tag{6.855}$$

For an incident dose rate

$$D = 8 \text{ e}^- \text{ pixel}^{-1} \text{ s}^{-1}, \tag{6.856}$$

Eq. (6.839) gives

$$\lambda_c = \frac{8 \times 5.15}{400} \simeq 0.103. \tag{6.857}$$

Therefore,

$$\boxed{P_{\text{coin}} \simeq 1 - e^{-0.103} \simeq 0.098.} \tag{6.858}$$

Thus, in this effective-area description, an incident electron has approximately a 10% probability of having another primary electron inside its coincidence neighbourhood during the same internal frame.

Coincidence is therefore already appreciable even though the mean number of electrons incident on any *single* physical pixel during one internal frame is much smaller than one. The relevant quantity is not the occupancy of one pixel, but the occupancy of the complete spatial neighbourhood over which two events become difficult to separate.

Why increasing the internal frame rate helps. Equation (6.839) also makes clear why modern counting detectors use extremely high internal frame rates. If the dose rate and effective event area were held fixed, increasing F reduces the coincidence occupancy in direct proportion to $1/F$.

For example, increasing the internal frame rate from 400 s^{-1} to 1500 s^{-1}, the latter being the K3 internal rate associated with Eq. (6.514), reduces λ_c by a factor

$$\frac{1500}{400} = 3.75. \tag{6.859}$$

For the same incident dose rate

$$D = 8\ \mathrm{e}^{-}\ \mathrm{pixel}^{-1}\,\mathrm{s}^{-1} \tag{6.860}$$

and the same effective coincidence area

$$s_{\mathrm{eff}} = 5.15\ \mathrm{pixel}^{2}, \tag{6.861}$$

the coincidence occupancy becomes

$$\lambda_{\mathrm{c}} = \frac{8 \times 5.15}{1500} \simeq 0.0275. \tag{6.862}$$

Therefore,

$$\boxed{P_{\mathrm{coin}} = 1 - e^{-0.0275} \simeq 0.027,} \tag{6.863}$$

corresponding to an overlap probability of only about 2.7%.

For comparison, the same calculation at 400 s^{-1} gave approximately 9.8%. Thus, within this simple effective-area model, increasing the internal frame rate from 400 to 1500 s^{-1} reduces the coincidence probability by roughly a factor of 3.6 at this dose rate.

This comparison illustrates only the frame-rate scaling. It should not be interpreted as a quantitative prediction of K3 coincidence loss, because the effective event area and counting algorithm are not necessarily the same for the two detectors.

Coincidence loss changes both signal and noise statistics. Coincidence loss is particularly important for the statistical description of a counting detector because it changes more than the mean number of recorded electrons.

The incident electrons obey Poisson statistics, but the counted events are no longer an independent Poisson sample when neighbouring primary electrons can be merged. Once one event has been accepted, the probability of accepting another event very close to it during the same internal frame is reduced. The counting process therefore introduces short-range exclusion and correlations between recorded events.

Experimentally and theoretically, this leads to a characteristic modification of the noise power spectrum of an electron-counting detector. In particular, coincidence loss can suppress the measured noise power at low spatial frequencies (Li et al., 2013; McMullan et al., 2014). The reduction of noise power should not be interpreted as an improvement by itself: the useful image signal is also reduced and the detector response has become

nonlinear. The consequences for the NPS and DQE will be examined after those quantities have been formally introduced.

Coincidence loss therefore differs fundamentally from an additive electronic-noise source:

$$\boxed{\begin{array}{lcl} \text{readout noise} & \Longrightarrow & \text{random signal added by the electronics,} \\ \text{coincidence loss} & \Longrightarrow & \text{nonlinear merging of otherwise valid electron events.} \end{array}} \tag{6.864}$$

The practical objective in counting operation is consequently to remain in the regime

$$\boxed{\lambda_{\mathrm{c}} = \frac{D\, s_{\mathrm{eff}}}{F} \ll 1,} \tag{6.865}$$

so that individual-electron events remain approximately separable and the counted image remains close to a linear representation of the incident electron distribution.

6.6.6 Electronic readout and reset noise

The fluctuations considered in the preceding subsections originate before the detector electronics measures the signal. The deposited energy fluctuates, the number of generated electron–hole pairs fluctuates, and the spatial distribution of the collected charge fluctuates. Even if all of these processes were perfectly reproducible, however, the electronic measurement of the collected charge would still not be noiseless.

Once the signal electrons have reached a collection node, their charge is converted into a voltage through the effective node capacitance C_{node}. As established in Eq. (6.501),

$$|\Delta V| = \frac{Q}{C_{\mathrm{node}}}. \tag{6.866}$$

For the electron-collecting convention used in Eq. (6.503),

$$\Delta V = -\frac{eN_{\mathrm{col}}}{C_{\mathrm{node}}}. \tag{6.867}$$

The electronic readout problem is therefore ultimately one of measuring a small voltage change. Thermal fluctuations in the reset operation, noise in the in-pixel transistor, column and output amplifiers, and other electronic stages perturb this voltage measurement.

Thermal noise introduced during reset. Before a new integration interval, the sensing node must be restored to its reset potential V_{reset}, as described in Sec. 6.4.4. In a simple CMOS pixel, the reset transistor temporarily connects the node capacitance to the reset-voltage circuit.

If the reset were perfectly deterministic, disconnecting the reset transistor would leave the node at exactly

$$V_{\text{reset}}. \tag{6.868}$$

In reality, the reset path is at a finite temperature T, and the electrical carriers in that path undergo thermal fluctuations. Consequently, when the reset transistor is opened, the voltage left on the node is more properly written as

$$\boxed{V_{\text{r},true} = V_{\text{reset}} + \delta V_{\text{reset}},} \tag{6.869}$$

where δV_{reset} changes randomly from one reset cycle to another.

Thermal reset noise of a capacitive sensing node.

Consider a capacitance C_{node} connected through the reset circuit to a thermal reservoir at temperature T. Let δV_{reset} denote the fluctuation of the capacitor voltage about its mean equilibrium value.

The electrostatic energy associated with this fluctuation is

$$U = \frac{1}{2} C_{\text{node}} \delta V_{\text{reset}}^2. \tag{6.870}$$

In classical thermal equilibrium, the equipartition theorem assigns an average energy

$$\mathbb{E}[U] = \frac{1}{2} k_{\text{B}} T \tag{6.871}$$

to this quadratic degree of freedom. Therefore,

$$\frac{1}{2} C_{\text{node}} \mathbb{E}\left[\delta V_{\text{reset}}^2\right] = \frac{1}{2} k_{\text{B}} T. \tag{6.872}$$

Hence

$$\boxed{\text{Var}\left(\delta V_{\text{reset}}\right) = \frac{k_{\text{B}} T}{C_{\text{node}}}.} \tag{6.873}$$

The root-mean-square reset-voltage fluctuation is therefore

$$\boxed{\sigma_{V,\mathrm{reset}} = \sqrt{\frac{k_\mathrm{B}T}{C_\mathrm{node}}}.} \tag{6.874}$$

This result is commonly called *kTC noise*.

The corresponding fluctuation expressed as charge on the sensing node is obtained from

$$\delta Q_\mathrm{reset} = C_\mathrm{node}\delta V_\mathrm{reset}. \tag{6.875}$$

Therefore,

$$\boxed{\sigma_{Q,\mathrm{reset}} = \sqrt{k_\mathrm{B}TC_\mathrm{node}}.} \tag{6.876}$$

If this charge fluctuation is expressed as an equivalent number of electrons,

$$\boxed{\sigma_{N,\mathrm{reset}} = \frac{\sqrt{k_\mathrm{B}TC_\mathrm{node}}}{e}.} \tag{6.877}$$

Equations (6.874) and (6.877) illustrate an apparently opposite dependence on the node capacitance. Reducing C_node increases the reset fluctuation measured in volts, but decreases the same fluctuation when expressed as an equivalent number of signal electrons. This is because the voltage signal produced by one collected electron also increases as $1/C_\mathrm{node}$, according to Eq. (6.503). A small sensing capacitance therefore gives a large charge-to-voltage conversion gain.

To put these scales into perspective, consider room temperature,

$$T = 300\ \mathrm{K}. \tag{6.878}$$

Using

$$k_\mathrm{B} = 1.380649 \times 10^{-23}\ \mathrm{J\,K^{-1}}, \tag{6.879}$$

Eq. (6.874) gives

$$\sigma_{V,\mathrm{reset}} = \sqrt{\frac{k_\mathrm{B}T}{C_\mathrm{node}}}. \tag{6.880}$$

Sensing-node capacitances in CMOS image sensors are commonly in the femtofarad range, with specially optimized high-conversion-gain pixels reaching values below 1 fF (Kusuhara et al., 2016). For an illustrative calculation, let us take

$$C_\mathrm{node} = 5\ \mathrm{fF}, \tag{6.881}$$

then

$$\boxed{\sigma_{V,\mathrm{reset}} \simeq 0.91 \text{ mV}.} \tag{6.882}$$

The voltage change produced by one collected electron is, from Eq. (6.503),

$$|\Delta V_{1e^-}| = \frac{e}{C_{\mathrm{node}}} \simeq 32\ \mu\mathrm{V}. \tag{6.883}$$

Thus the reset fluctuation corresponds to approximately

$$\boxed{\sigma_{N,\mathrm{reset}} = \frac{0.91 \text{ mV}}{32\ \mu\mathrm{V}} \simeq 28\ e^-,} \tag{6.884}$$

in agreement with Eq. (6.877).

C_{node}	$\sigma_{V,\mathrm{reset}}$	e/C_{node}	$\sigma_{N,\mathrm{reset}}$
1 fF	2.04 mV	160 μV/e^-	12.7 e^-
5 fF	0.91 mV	32.0 μV/e^-	28.4 e^-
10 fF	0.64 mV	16.0 μV/e^-	40.2 e^-
20 fF	0.46 mV	8.0 μV/e^-	56.8 e^-

Table 6.1: Illustrative thermal reset-noise scales at $T = 300$ K for different sensing-node capacitances. Decreasing C_{node} increases the voltage fluctuation $\sigma_{V,\mathrm{reset}} \propto C_{\mathrm{node}}^{-1/2}$, but the voltage produced by one collected electron increases faster, as C_{node}^{-1}. Consequently, the equivalent reset noise expressed in electrons decreases as $\sigma_{N,\mathrm{reset}} \propto \sqrt{C_{\mathrm{node}}}$.

The comparison becomes even more relevant for a fast primary electron. As discussed previously, a deposited energy of approximately 15.5 keV would generate of order

$$N_{eh} \simeq 4.3 \times 10^3 \tag{6.885}$$

electron–hole pairs in silicon. If all of the corresponding signal charge were collected at a node with $C_{\mathrm{node}} = 5$ fF, the voltage excursion would be of order

$$\boxed{|\Delta V_{\mathrm{event}}| \simeq 4.3 \times 10^3 \times 32\ \mu\mathrm{V} \simeq 0.14 \text{ V}.} \tag{6.886}$$

The raw kTC fluctuation of about 28 equivalent electrons is therefore small compared with the several thousand charge carriers generated by such an event. This comparison should not be interpreted as the final detector SNR, because the deposited energy itself fluctuates strongly, the charge is generally distributed among several collection nodes, and CDS is specifically designed to suppress the reset contribution. Nevertheless, it gives the correct order of magnitude of the electronic scales involved.

Noise introduced while reading the node voltage. Reset noise is not the only electronic fluctuation. Even if the voltage present on the sensing node were known exactly, measuring that voltage through the pixel transistor, column electronics, amplifiers, and output circuitry would introduce additional noise.

We can describe one voltage measurement schematically as

$$\boxed{V_{\mathrm{meas}} = V_{\mathrm{node}} + n_{\mathrm{ro}},} \tag{6.887}$$

where n_{ro} denotes the electronic readout noise referred to the sensing-node voltage.

The detailed origin of n_{ro} depends on the detector architecture. Thermal fluctuations in the in-pixel source follower and other transistor or resistive elements contribute a broadband component, whereas MOS transistors also exhibit low-frequency flicker ($1/f$) noise and, in some devices, random-telegraph noise. Subsequent stages of the readout chain, such as column amplifiers and sample-and-hold circuitry, introduce further fluctuations. These contributions can be referred back to the detector input and expressed as an equivalent fluctuation of the sensing-node voltage (Boukhayma et al., 2016; Konnik and Welsh, 2014).

It is useful at this stage not to assume that successive measurements of n_{ro} are independent. Slowly varying electronic fluctuations can be strongly correlated between two measurements made close together in time. This observation is central to correlated double sampling.

Why correlated double sampling suppresses reset noise. The reset fluctuation in Eq. (6.869) might appear to produce a large uncertainty in the signal. However, the important point is that the same reset offset is present both before and after the charge integration interval.

Suppose that immediately after reset the true node voltage is

$$V_{\mathrm{r},true} = V_{\mathrm{reset}} + \delta V_{\mathrm{reset}}. \tag{6.888}$$

Let $Q_{\mathrm{col}} > 0$ denote the magnitude of the electron charge collected during the subsequent integration interval. Using the electron-collecting sign convention of Eq. (6.503), the signal-level voltage immediately before readout is

$$V_{\mathrm{s},true} = V_{\mathrm{reset}} + \delta V_{\mathrm{reset}} - \frac{Q_{\mathrm{col}}}{C_{\mathrm{node}}}. \tag{6.889}$$

The electronics does not measure these two voltages perfectly. Let n_{r} and n_{s} denote the readout noise associated with measuring the reset and

signal levels, respectively. The measured values are therefore

$$V_{\mathrm{r}} = V_{\mathrm{reset}} + \delta V_{\mathrm{reset}} + n_{\mathrm{r}} \tag{6.890}$$

and

$$V_{\mathrm{s}} = V_{\mathrm{reset}} + \delta V_{\mathrm{reset}} - \frac{Q_{\mathrm{col}}}{C_{\mathrm{node}}} + n_{\mathrm{s}}. \tag{6.891}$$

The correlated-double-sampling measurement introduced in Eq. (6.511) is

$$\Delta V_{\mathrm{CDS}} = V_{\mathrm{r}} - V_{\mathrm{s}} \tag{6.892}$$

$$= (V_{\mathrm{reset}} + \delta V_{\mathrm{reset}} + n_{\mathrm{r}}) - \left(V_{\mathrm{reset}} + \delta V_{\mathrm{reset}} - \frac{Q_{\mathrm{col}}}{C_{\mathrm{node}}} + n_{\mathrm{s}}\right). \tag{6.893}$$

The reset voltage and, importantly, the particular reset fluctuation $\delta V_{\mathrm{reset}}$ cancel. Thus

$$\boxed{\Delta V_{\mathrm{CDS}} = \frac{Q_{\mathrm{col}}}{C_{\mathrm{node}}} + n_{\mathrm{r}} - n_{\mathrm{s}}.} \tag{6.894}$$

This calculation explains why measuring the reset level is so useful. The detector does not need the reset voltage to be identical from one cycle to the next. It needs to measure the reset level associated with the *same integration interval* and subtract it from the subsequent signal measurement.

In this idealized model, the kTC reset fluctuation is therefore removed completely by CDS. Real circuits can leave a residual contribution because of leakage, imperfect sampling, finite amplifier bandwidth, and other non-idealities, but the cancellation mechanism is the one shown in Eq. (6.894).

What happens to the readout noise in CDS? The readout-noise terms do not necessarily cancel. From Eq. (6.894), the electronic noise remaining in the CDS signal is

$$n_{\mathrm{CDS}} = n_{\mathrm{r}} - n_{\mathrm{s}}. \tag{6.895}$$

Its variance is

$$\boxed{\begin{aligned}\sigma^2_{\mathrm{CDS}} = {} & \mathrm{Var}(n_{\mathrm{r}}) + \mathrm{Var}(n_{\mathrm{s}}) \\ & - 2\,\mathrm{Cov}\,(n_{\mathrm{r}}, n_{\mathrm{s}})\,.\end{aligned}} \tag{6.896}$$

Equation (6.896) shows why the word *correlated* is important. Noise components that are nearly the same in the two measurements have a large positive covariance and are therefore strongly suppressed by subtraction.

For example, suppose that both measurements have the same readout-noise variance,

$$\mathrm{Var}(n_\mathrm{r}) = \mathrm{Var}(n_\mathrm{s}) = \sigma_\mathrm{ro}^2, \tag{6.897}$$

and define their correlation coefficient ρ_ro by

$$\mathrm{Cov}\,(n_\mathrm{r}, n_\mathrm{s}) = \rho_\mathrm{ro}\sigma_\mathrm{ro}^2. \tag{6.898}$$

Then

$$\boxed{\sigma_\mathrm{CDS}^2 = 2\sigma_\mathrm{ro}^2\,(1-\rho_\mathrm{ro})\,.} \tag{6.899}$$

Two limiting cases are instructive. If a noise component varies very slowly compared with the interval between the two measurements, then

$$\rho_\mathrm{ro} \simeq 1, \tag{6.900}$$

and it is strongly suppressed by CDS. Conversely, if the two readout noise samples are statistically independent,

$$\rho_\mathrm{ro} = 0, \tag{6.901}$$

then

$$\sigma_\mathrm{CDS} = \sqrt{2}\,\sigma_\mathrm{ro}. \tag{6.902}$$

Thus CDS does not reduce every type of electronic noise. It suppresses components that are correlated between the reset and signal measurements, while two independent noise measurements add in quadrature.

Equivalent noise charge. Electronic noise is naturally generated and measured in volts, whereas the detector signal originates as collected charge. It is therefore useful to express the electronic noise as the amount of input charge that would produce the same voltage fluctuation.

If the root-mean-square voltage noise referred to the sensing node is σ_V, Eq. (6.501) gives the corresponding equivalent noise charge

$$\boxed{\mathrm{ENC} = C_\mathrm{node}\sigma_V.} \tag{6.903}$$

The same quantity expressed as an equivalent number of electrons is

$$\boxed{\mathrm{ENC}_{e^-} = \frac{C_\mathrm{node}\sigma_V}{e}.} \tag{6.904}$$

The equivalent-noise-charge representation is particularly useful because it allows the electronic noise to be compared directly with the number of

signal electrons collected from one primary imaging electron. For a given voltage-noise level, decreasing C_{node} decreases ENC_{e^-} because the voltage change produced by each collected carrier becomes larger.

For example, consider again a sensing-node capacitance

$$C_{\text{node}} = 5 \text{ fF}. \tag{6.905}$$

The voltage change produced by one collected electron is

$$\frac{e}{C_{\text{node}}} \simeq 32 \ \mu\text{V}/e^-, \tag{6.906}$$

as obtained previously.

Suppose, as an illustrative example, that after the relevant readout and correlated-double-sampling operations the root-mean-square electronic noise referred to the sensing node is

$$\sigma_V = 100 \ \mu\text{V}. \tag{6.907}$$

Equation (6.904) then gives

$$\text{ENC}_{e^-} = \frac{C_{\text{node}}\sigma_V}{e} \tag{6.908}$$

$$= \frac{(5 \times 10^{-15}\,\text{F})(100 \times 10^{-6}\,\text{V})}{1.602 \times 10^{-19}\,\text{C}} \tag{6.909}$$

$$\simeq \boxed{3.1 \ e^-_{\text{rms}}.} \tag{6.910}$$

Thus an electronic fluctuation of 100 μV at a 5 fF sensing node is equivalent to the charge carried by only about three electrons.

The subscript "rms" (root-mean-square) in Eq. (6.910) emphasizes that this quantity is the standard deviation of a statistical fluctuation, not a discrete count of physical particles. Because the underlying electronic voltage fluctuation is a continuous random variable, the equivalent charge inferred from it is also continuous.

A value of 3.1 e^-_{rms} therefore means that if the same pixel is read out repeatedly in the absence of incident electrons, the distribution of the measured equivalent charge will have a standard deviation of 3.1 elementary charges. Assuming the readout noise is approximately Gaussian, roughly 68% of these empty-pixel measurements will fall within a band of ± 3.1 equivalent electrons around the electronic baseline.

This root-mean-square width defines the intrinsic electronic noise floor of the sensing node. It establishes the fundamental requirement for electron

counting: to be reliably recognized as a true arrival event rather than a random electronic fluctuation, the localized charge cloud generated by an incident primary electron must produce a signal that exceeds this rms noise floor by a statistically significant margin.

This scale should be compared with the several thousand electron–hole pairs generated by one energetic primary electron. For example, the 15.5 keV deposited-energy scale considered previously corresponds to approximately

$$N_{eh} \simeq 4.3 \times 10^3 \tag{6.911}$$

electron–hole pairs.

If all of this charge were collected by a single node, the ratio of the single-event signal to the equivalent electronic noise in the example above would be of order

$$\frac{4.3 \times 10^3}{3.1} \simeq 1.4 \times 10^3. \tag{6.912}$$

The corresponding amplitude SNR would therefore be extremely large.

In practice, however, the charge generated by one primary electron is usually distributed among several neighbouring pixels, and the deposited energy itself fluctuates strongly from event to event. Consequently, the signal relevant for thresholding or localization is the charge collected by the individual pixels of the event cluster, rather than the total charge generated by the primary electron.

C_{node}	σ_V	ENC_{e^-}
1 fF	100 μV	0.62 e^-
5 fF	100 μV	3.1 e^-
10 fF	100 μV	6.2 e^-
5 fF	300 μV	9.4 e^-
5 fF	500 μV	15.6 e^-

Table 6.2: Illustrative equivalent-noise-charge values obtained from Eq. (6.904). The values are not intended to represent a particular commercial cryo-EM detector; they show how the input-referred electronic noise scales with node capacitance and voltage noise.

The equivalent-noise-charge representation is particularly useful because it allows the electronic noise to be compared directly with the charge generated by one primary imaging electron. For example, with $C_{\text{node}} = 5$ fF, one collected electron produces approximately 32 μV. An input-referred voltage noise of 100 μV_{rms} therefore corresponds to only about 3.1 electrons rms. Increasing the node capacitance to 10 fF while keeping the same voltage-noise

level doubles the ENC to about 6.2 electrons. Thus, for a fixed voltage-noise level, decreasing C_{node} improves charge sensitivity because the voltage signal produced by each collected carrier increases as $1/C_{\text{node}}$.

Contribution to the SNR of an integrating detector. We can now connect the electronic fluctuations to the SNR decomposition derived in Sec. 6.6.2.

Let

$$S = \sum_{i=1}^{N} Q_i \tag{6.913}$$

be the integrated charge generated by the incident primary electrons, as in Eq. (6.688). Suppose that the electronics adds an independent equivalent input-charge fluctuation n_{el} with

$$\mathbb{E}[n_{\text{el}}] = 0, \qquad \text{Var}(n_{\text{el}}) = \sigma_{\text{el}}^2. \tag{6.914}$$

The measured signal is then

$$\boxed{S_{\text{meas}} = S + n_{\text{el}}.} \tag{6.915}$$

Because the electronic noise is assumed independent of the electron-induced signal,

$$\text{Var}(S_{\text{meas}}) = \text{Var}(S) + \sigma_{\text{el}}^2 \tag{6.916}$$

$$= \boxed{\mu\left(\overline{Q}^2 + \sigma_Q^2\right) + \sigma_{\text{el}}^2.} \tag{6.917}$$

Here the first two contributions are those derived in Eq. (6.710): incident-electron shot noise and single-electron-response fluctuations. The final term is introduced by the electronic measurement itself.

Since

$$\mathbb{E}[S_{\text{meas}}] = \mu\overline{Q}, \tag{6.918}$$

the corresponding power SNR is

$$\boxed{\text{SNR}_{\text{int}} = \frac{\mu^2\overline{Q}^2}{\mu\left(\overline{Q}^2 + \sigma_Q^2\right) + \sigma_{\text{el}}^2}.} \tag{6.919}$$

Equation (6.919) shows an important difference between electronic readout noise and the two previous contributions. The shot-noise and single-electron-response terms grow with the number μ of incident electrons. By contrast, the electronic readout contribution associated with one frame does not in this simple model grow in proportion to μ. It can therefore be especially important at low electron fluence.

Consequences for electron counting. Electron counting changes the role of electronic readout noise, but does not make it irrelevant. The counting procedure introduced in Eq. (6.512) must decide whether a digitized local pixel pattern corresponds to a genuine primary-electron event.

Electronic fluctuations perturb the pixel values used in this decision. If the threshold is set too low, a sufficiently large electronic fluctuation can be mistaken for an electron event. If it is set too high, a genuine but weak single-electron event can be rejected. Readout noise can also perturb the relative pixel amplitudes within a detected cluster and thereby degrade the localization estimate.

Once an isolated event has nevertheless been identified correctly, its variable analog amplitude can be replaced by one standardized count. Electronic readout noise is then no longer transferred directly to the final image as an additive analog amplitude fluctuation. Its principal effect is instead through the probability of detection, false-event generation, and localization accuracy.

Thus the same electronic noise has different consequences in the two operating modes:

$$\boxed{\begin{array}{lcl} \text{integrating detector} & : & \text{electronic noise adds directly to the measured signal,} \\ \text{counting detector} & : & \text{electronic noise perturbs event detection and localization.} \end{array}} \tag{6.920}$$

From temporal electronic noise to spatial image noise. Most electronic noise mechanisms originate in circuits that evolve in time. The noise power spectrum considered later in this chapter, however, will describe fluctuations as a function of *spatial* frequency. The two should not be confused.

If every pixel had completely independent and statistically identical readout electronics, electronic fluctuations could produce an approximately spatially uncorrelated contribution to the image. Real arrays share column amplifiers, row-selection circuitry, power supplies, reference voltages, and output channels. A fluctuation common to one column, row, or readout block can therefore appear as correlated spatial structure in the recorded frame.

Consequently, a temporal low-frequency electronic fluctuation does not automatically imply a low-spatial-frequency NPS. Its spatial signature depends on how the temporal fluctuation is mapped onto the pixel array by the detector readout architecture. The spatial covariance and NPS formalism introduced later will provide the appropriate language for describing these

correlations.

6.6.7 Dark-current and thermal fluctuations

The electronic noise considered in the preceding subsection is introduced while the charge stored at the sensing node is reset, amplified, and measured. There is another source of detector signal that arises earlier, inside the semiconductor itself: even when no imaging electron enters the detector, electron–hole pairs can be generated thermally.

As discussed in Sec. 6.4.2, silicon is a semiconductor with a band gap

$$E_{\mathrm{g}} \simeq 1.12 \text{ eV} \tag{6.921}$$

near room temperature, as given in Eq. (6.394). At finite temperature, thermal fluctuations can promote carriers into electrically active states. Generation can occur in the semiconductor bulk, in the depletion region, and through defect or interface states that provide intermediate energy levels inside the band gap.

Once such an electron–hole pair is generated inside a region from which charge can be collected, the detector electric field acts on it in essentially the same way as on a pair generated by an imaging electron: the carriers are separated and transported toward the collection electrodes. The resulting current exists even in the absence of an incident imaging signal and is therefore called *dark current* (Murari et al., 2009; Konnik and Welsh, 2014).

Mean dark signal. Let

$$r_{\mathrm{d}} \tag{6.922}$$

denote the mean rate at which thermally generated signal electrons are collected by one pixel, measured in

$$\text{electrons}\,\mathrm{s}^{-1}\,\text{pixel}^{-1}. \tag{6.923}$$

During an integration time T_{int}, the expected number of dark electrons is

$$\boxed{\mu_{\mathrm{d}} = r_{\mathrm{d}} T_{\mathrm{int}}.} \tag{6.924}$$

The corresponding mean collected dark charge is

$$\boxed{\overline{Q}_{\mathrm{d}} = e\, r_{\mathrm{d}} T_{\mathrm{int}}.} \tag{6.925}$$

Thus, unlike the readout-noise contribution considered in the preceding subsection, the mean dark signal accumulates continuously with integration time.

If the temperature and detector operating conditions remain stable, this mean contribution can in principle be measured in the absence of imaging electrons and subtracted as a dark offset. Subtracting the mean, however, does not remove the statistical fluctuations associated with the discrete generation of the dark carriers.

Dark-current shot noise. Suppose that the individual thermal-generation events are independent. The number of dark electrons collected during one integration interval is then naturally described by a Poisson distribution (Konnik and Welsh, 2014),

$$\boxed{N_{\mathrm{d}} \sim \mathrm{Poisson}\left(r_{\mathrm{d}} T_{\mathrm{int}}\right).} \tag{6.926}$$

Using the Poisson property derived in Sec. 6.6.1,

$$\mathbb{E}[N_{\mathrm{d}}] = \mathrm{Var}(N_{\mathrm{d}}) = r_{\mathrm{d}} T_{\mathrm{int}}. \tag{6.927}$$

Therefore,

$$\boxed{\sigma_{N,\mathrm{d}} = \sqrt{r_{\mathrm{d}} T_{\mathrm{int}}}.} \tag{6.928}$$

Expressed as charge,

$$\boxed{\sigma_{Q,\mathrm{d}} = e\sqrt{r_{\mathrm{d}} T_{\mathrm{int}}}.} \tag{6.929}$$

Thus the mean dark signal grows linearly with integration time,

$$\mathbb{E}[N_{\mathrm{d}}] \propto T_{\mathrm{int}}, \tag{6.930}$$

whereas its root-mean-square fluctuation grows only as

$$\sigma_{N,\mathrm{d}} \propto \sqrt{T_{\mathrm{int}}}. \tag{6.931}$$

This is exactly the same square-root dependence encountered for incident-electron shot noise, because both arise from the counting statistics of independent discrete carriers.

Why dark-frame subtraction cannot remove dark-current noise. The distinction between the mean dark signal and its fluctuation is important. Suppose that the mean dark contribution μ_{d} is known perfectly. A dark-corrected measurement would contain

$$N_{\mathrm{d}} - \mu_{\mathrm{d}}. \tag{6.932}$$

Its expectation is indeed zero,

$$\mathbb{E}\left[N_{\mathrm{d}} - \mu_{\mathrm{d}}\right] = 0, \tag{6.933}$$

but subtracting a deterministic mean does not change a variance:

$$\begin{aligned} \mathrm{Var}\left(N_{\mathrm{d}} - \mu_{\mathrm{d}}\right) &= \mathrm{Var}(N_{\mathrm{d}}) && (6.934) \\ &= \boxed{\mu_{\mathrm{d}}.} && (6.935) \end{aligned}$$

A calibration can therefore remove the *mean* dark signal, but it cannot predict which thermally generated carriers will appear in a particular exposure.

This is the same distinction encountered earlier for incident-electron shot noise: knowing the expected number of electrons does not determine the random realization that will actually be observed.

Noise introduced by a measured dark reference. There is a further practical point. The dark signal is not generally known exactly; it must itself be estimated from measurements.

Suppose that M independent dark frames, each with the same integration time as the image, are averaged to estimate the mean dark level:

$$\widehat{\mu}_{\mathrm{d}} = \frac{1}{M} \sum_{m=1}^{M} N_{\mathrm{d}}^{(m)}. \tag{6.936}$$

Since every dark frame has variance μ_{d},

$$\boxed{\mathrm{Var}\left(\widehat{\mu}_{\mathrm{d}}\right) = \frac{\mu_{\mathrm{d}}}{M}.} \tag{6.937}$$

If a science frame contains an independent dark realization N_{d} and we subtract the estimated dark reference, the residual dark contribution is

$$N_{\mathrm{d}} - \widehat{\mu}_{\mathrm{d}}. \tag{6.938}$$

Because these quantities are independent,

$$\begin{aligned} \mathrm{Var}\left(N_{\mathrm{d}} - \widehat{\mu}_{\mathrm{d}}\right) &= \mathrm{Var}(N_{\mathrm{d}}) + \mathrm{Var}\left(\widehat{\mu}_{\mathrm{d}}\right) && (6.939) \\ &= \boxed{\mu_{\mathrm{d}}\left(1 + \frac{1}{M}\right).} && (6.940) \end{aligned}$$

Thus subtracting one single dark frame,

$$M = 1, \tag{6.941}$$

would actually give a dark-noise variance

$$2\mu_{\mathrm{d}}, \tag{6.942}$$

because both the science frame and the dark reference contain independent dark-current shot noise. Averaging many dark frames makes the calibration term small,

$$\frac{\mu_{\mathrm{d}}}{M} \longrightarrow 0, \tag{6.943}$$

but the intrinsic dark-current shot noise μ_{d} of the science exposure remains.

Why dark current depends strongly on temperature. Dark current is strongly temperature dependent because the generation of carriers is thermally activated. Its precise temperature dependence depends on the microscopic generation mechanism, the detector architecture, and the density and energies of defect states (Murari et al., 2009; Konnik and Welsh, 2014).

Over a restricted temperature range it is often useful to describe the dark-generation rate by an Arrhenius-type expression,

$$\boxed{r_{\mathrm{d}}(T) \simeq A T^{\alpha} \exp\left(-\frac{E_{\mathrm{a}}}{k_{\mathrm{B}} T}\right),} \tag{6.944}$$

where A is a device-dependent prefactor, α describes a possible additional algebraic temperature dependence, and E_{a} is an effective activation energy.

The parameter E_{a} should not automatically be identified with the full silicon band gap. Different physical mechanisms can produce different effective activation energies. In particular, generation through defect states inside the band gap can require substantially less thermal activation than a direct transition across the complete band gap.

The strong exponential factor in Eq. (6.944) explains why cooling the sensor can reduce dark current by a large factor. For two temperatures,

$$\boxed{\frac{r_{\mathrm{d}}(T_2)}{r_{\mathrm{d}}(T_1)} \simeq \left(\frac{T_2}{T_1}\right)^{\alpha} \exp\left[-\frac{E_{\mathrm{a}}}{k_{\mathrm{B}}}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)\right].} \tag{6.945}$$

Because the dark-current shot noise is proportional to the square root of the dark current, Eq. (6.928), cooling reduces both the mean dark signal and its statistical fluctuation.

An illustrative temperature scale. The exponential dependence can be substantial even for moderate changes in temperature. As an illustrative example, suppose that the dominant generation process has an effective activation energy

$$E_{\mathrm{a}} \simeq \frac{E_{\mathrm{g}}}{2} \simeq 0.56 \text{ eV}, \tag{6.946}$$

and neglect the algebraic prefactor for this estimate, $\alpha = 0$.

Cooling from

$$T_1 = 300 \text{ K} \tag{6.947}$$

to

$$T_2 = 273 \text{ K} \tag{6.948}$$

then gives, from Eq. (6.945),

$$\frac{r_{\mathrm{d}}(273 \text{ K})}{r_{\mathrm{d}}(300 \text{ K})} \simeq 0.12. \tag{6.949}$$

Thus, in this illustrative model, cooling by only 27 K reduces the dark-generation rate by approximately one order of magnitude.

This numerical value is not universal: the actual temperature dependence must be measured for a particular sensor because E_{a} and the prefactor depend on its materials, defects, junction structure, and operating conditions.

Dark current is not the same as electronic thermal noise. The word *thermal* appears in both this subsection and the preceding discussion of electronic readout noise, but the underlying processes must be distinguished.

The kTC reset fluctuation derived previously is an equilibrium voltage fluctuation of a capacitance connected temporarily to a resistive thermal reservoir. Johnson–Nyquist noise and transistor thermal noise similarly originate from thermal fluctuations in the electronic readout circuitry.

Dark current has a different origin. Thermal activation creates *additional charge carriers inside the sensitive semiconductor*. Once generated, these carriers can be transported and collected just as signal carriers are. Their discreteness then gives rise to dark-current *shot noise*.

The distinction is therefore

electronic thermal noise	:	fluctuation of voltages and currents in the readout circuitry,
dark current	:	thermally generated charge carriers in the sensor,
dark-current shot noise	:	statistical fluctuation in their number.

(6.950)

Pixel-to-pixel variation and hot pixels. The mean rate r_{d} need not be the same for all pixels. Local variations in semiconductor defects, interfaces, junction geometry, and fabrication conditions produce a distribution of dark rates across the detector (Konnik and Welsh, 2014). It is therefore more generally appropriate to write

$$r_{\mathrm{d},p} \tag{6.951}$$

for pixel p.

A pixel with an unusually large value of $r_{\mathrm{d},p}$ produces an unusually large dark signal and may appear as a *hot pixel*. This pixel-to-pixel variation is conceptually different from dark-current shot noise. At fixed temperature and operating conditions, the mean pattern

$$r_{\mathrm{d},p} \tag{6.952}$$

can be approximately reproducible and therefore belongs to dark-signal non-uniformity or fixed-pattern response. Around that mean, however, the number of thermally generated carriers still fluctuates randomly.

Thus,

$$\boxed{\begin{array}{ccl} r_{\mathrm{d},p} & \Longrightarrow & \text{mean spatial dark pattern,} \\ N_{\mathrm{d},p} - r_{\mathrm{d},p}T_{\mathrm{int}} & \Longrightarrow & \text{temporal dark-current shot noise.} \end{array}} \tag{6.953}$$

The systematic spatial non-uniformities of the detector will be considered separately. The important point here is that calibration of the mean dark pattern does not remove its shot-noise fluctuations.

Consequences for integrating and counting detectors. For an integrating detector, thermally generated carriers simply add to the collected signal. If the dark process is independent of the electron-induced signal, its variance contributes additively to the total variance. The dark-current contribution expressed in equivalent electrons is

$$\boxed{\sigma_{\mathrm{d}}^2 = r_{\mathrm{d}}T_{\mathrm{int}}.} \tag{6.954}$$

Dark current therefore becomes progressively more important for long integration times and weak signals.

In an electron-counting detector, the situation is somewhat different. One primary imaging electron typically produces a cluster containing many generated electron–hole pairs, whereas an isolated thermal generation event

corresponds to only a very small amount of charge. Appropriate event thresholds can therefore reject many dark-generated carriers.

Dark current nevertheless remains relevant. Accumulated thermally generated charge can perturb the local baseline, unusually active pixels can produce spurious structures, and sufficiently large fluctuations can affect event detection. The short internal integration times used for electron counting are advantageous because Eq. (6.924) shows that the mean number of dark carriers accumulated in each internal frame decreases linearly with the frame duration.

Dark current is therefore fundamentally different from the single-electron response fluctuations considered earlier: it generates signal even when no imaging electron is present. Its mean component can be calibrated, but its Poisson fluctuation represents a genuine additional source of detector noise.

6.6.8 Gain variations and fixed-pattern noise

The noise sources considered so far were stochastic from one acquisition to another. The number of incident electrons fluctuates, the response of each primary electron fluctuates, thermally generated carriers fluctuate, and the electronic readout introduces random voltage fluctuations.

A detector can also exhibit spatial non-uniformities that are much more reproducible in time. Two pixels exposed to exactly the same incident electron fluence need not produce exactly the same mean output. Their conversion gain, collection efficiency, dark offset, threshold, or readout response may differ slightly because of fabrication variations and local defects.

Such reproducible spatial variations are commonly referred to as *fixed-pattern noise*. The terminology should be interpreted with care: a perfectly stable fixed pattern is not a stochastic fluctuation from one frame to the next. It is a spatially varying detector response. It behaves as noise in the recorded image because it is superimposed on the unknown specimen signal and can be removed only to the accuracy with which the detector response has been calibrated.

Amplification of the pixel signal. The voltage change produced at the sensing node is generally too small and too high-impedance to be sent directly through the detector array. It is therefore buffered and amplified before digitization. A simplified CMOS readout chain is shown schematically in Fig. 6.7.

For pixel p, the collected charge produces the sensing-node voltage change

derived in Eq. (6.503),

$$\Delta V_{\mathrm{node},p} = -\frac{Q_p}{C_{\mathrm{node},p}}. \tag{6.955}$$

The node voltage is commonly buffered by an in-pixel transistor, often implemented as a source follower, and subsequently processed by column-level and output amplifiers. To first order, these stages can be represented by linear voltage gains. If $A_{\mathrm{pix},p}$ denotes the gain of the in-pixel buffer and $A_{\mathrm{col},c}$ the gain of the column circuitry associated with column c, the signal reaching the output can be written schematically as

$$\boxed{\Delta V_{\mathrm{out},p} = A_{\mathrm{col},c(p)} A_{\mathrm{pix},p} \Delta V_{\mathrm{node},p}.} \tag{6.956}$$

Substitution of Eq. (6.955) gives

$$\boxed{\Delta V_{\mathrm{out},p} = -\frac{A_{\mathrm{col},c(p)} A_{\mathrm{pix},p}}{C_{\mathrm{node},p}} Q_p.} \tag{6.957}$$

Equation (6.957) already shows why two pixels need not have exactly the same conversion gain. Small fabrication differences can change the effective sensing-node capacitance $C_{\mathrm{node},p}$ or the characteristics of the in-pixel transistor. In addition, different columns or readout blocks can have slightly different amplifier gains. The complete conversion from collected charge to digitized output is therefore not perfectly identical across the array.

The same circuitry can also introduce additive differences. Variations in transistor threshold voltage, reset level, amplifier offset, and ADC (Analog-to-Digital Converter) offset can shift the baseline of one pixel or readout channel relative to another. Thus the detector naturally acquires both a *multiplicative* non-uniformity, associated with gain, and an *additive* non-uniformity, associated with offset.

A simple pixel-response model. The detailed electronic chain of Fig. 6.7 can be summarized, for many purposes, by an effective gain and offset for each pixel. Let X_p denote the signal that an ideal detector would assign to pixel p. A simple model for the actual measured value is

$$\boxed{Y_p = g_p X_p + b_p + n_p,} \tag{6.958}$$

where

$$g_p \tag{6.959}$$

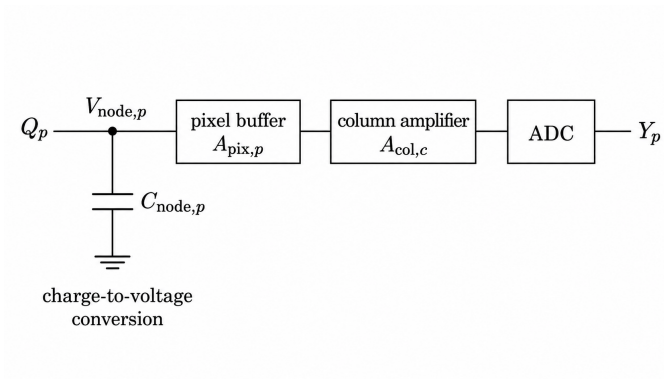


Figure 6.7: Simplified readout chain for one pixel of a CMOS direct electron detector. The collected charge Q_p changes the voltage of the sensing node through $C_{\text{node},p}$. The node voltage is buffered by in-pixel electronics, amplified by column or output circuitry, and finally digitized. Pixel-to-pixel and column-to-column variations in capacitance, transistor characteristics, amplifier gain, and electronic offsets give rise to gain and fixed-pattern non-uniformities.

is the gain of pixel p,

$$b_p \tag{6.960}$$

is its additive offset, and

$$n_p \tag{6.961}$$

contains stochastic contributions such as readout noise and dark-current fluctuations.

Equation (6.958) separates three different effects:

$$\boxed{\begin{array}{lcl} b_p & : & \text{additive spatial non-uniformity,} \\ g_p & : & \text{multiplicative spatial non-uniformity,} \\ n_p & : & \text{random frame-to-frame fluctuation.} \end{array}} \tag{6.962}$$

The distinction between offset and gain variation is important. An offset remains present even in the absence of signal, whereas a gain difference becomes visible only when signal is present.

Offset fixed pattern. Suppose first that the detector is measured in the absence of imaging electrons. The mean response of pixel p is then

$$\mathbb{E}[Y_p] = b_p. \tag{6.963}$$

Differences in b_p across the array produce a fixed spatial pattern.

Dark-current non-uniformity discussed in Sec. 6.6.7 is one possible contribution. For example, the mean dark rate $r_{\mathrm{d},p}$ can differ from pixel to pixel because of local defects or junction variations, as summarized in Eq. (6.953).

The mean offset pattern can in principle be estimated from many dark frames and subtracted. Such subtraction removes the reproducible mean pattern but, as shown in Eq. (6.940), it cannot remove the stochastic dark-current fluctuations occurring during the science exposure itself.

Gain fixed pattern. Now consider a spatially uniform illumination for which an ideal detector would have the same mean signal

$$\mu \tag{6.964}$$

in every pixel. After subtracting the offset, the mean output of pixel p is

$$\mathbb{E}[Y_p - b_p] = g_p \mu. \tag{6.965}$$

It is useful to write the gain as

$$g_p = \overline{g}\,(1 + \varepsilon_p)\,, \tag{6.966}$$

where $\overline{g}$ is the mean detector gain and

$$\mathbb{E}_p[\varepsilon_p] = 0 \tag{6.967}$$

when the expectation is taken over pixels.

The mean signal then becomes

$$\mathbb{E}[Y_p - b_p] = \overline{g}\mu + \overline{g}\mu\varepsilon_p. \tag{6.968}$$

Thus the gain non-uniformity produces a spatial deviation

$$\boxed{\delta Y_{p,\mathrm{gain}} = \overline{g}\mu\varepsilon_p.} \tag{6.969}$$

Unlike shot noise, whose standard deviation grows as $\sqrt{\mu}$, the amplitude of this multiplicative fixed pattern grows directly as μ.

Why an uncorrected gain pattern can dominate at high dose. The different scaling with signal level has an important consequence.

Suppose that, after normalizing the mean gain to unity, the incident electron counts are Poisson distributed with mean μ, and suppose that the relative pixel gains have variance

$$\mathrm{Var}_p(\varepsilon_p) = \sigma_g^2. \tag{6.970}$$

For a uniform illumination, the variance associated with incident electron shot noise is

$$\sigma_{\text{shot}}^2 = \mu, \tag{6.971}$$

whereas the spatial variance of the fixed gain pattern is

$$\boxed{\sigma_{\text{gain}}^2 = \mu^2 \sigma_g^2.} \tag{6.972}$$

The ratio between the two contributions is therefore

$$\boxed{\frac{\sigma_{\text{gain}}^2}{\sigma_{\text{shot}}^2} = \mu \sigma_g^2.} \tag{6.973}$$

At low dose,

$$\mu \sigma_g^2 \ll 1, \tag{6.974}$$

and Poisson fluctuations dominate. At sufficiently high dose, however, even a small uncorrected gain variation can dominate because its variance grows as μ^2, whereas shot-noise variance grows only as μ.

SNR limitation produced by uncorrected gain variation.

For a uniform signal with mean μ, let the useful signal power be

$$P_{\text{signal}} = \mu^2. \tag{6.975}$$

If the only two fluctuations are Poisson shot noise and a fixed gain variation with relative variance σ_g^2, the total apparent noise power is

$$P_{\text{noise}} = \mu + \mu^2 \sigma_g^2. \tag{6.976}$$

Using the power definition of SNR adopted in Sec. 6.6.1,

$$\boxed{\text{SNR} = \frac{\mu^2}{\mu + \mu^2 \sigma_g^2} = \frac{\mu}{1 + \mu \sigma_g^2}.} \tag{6.977}$$

For small μ,

$$\text{SNR} \simeq \mu, \tag{6.978}$$

and the detector is approximately shot-noise limited.

For very large μ,

$$\boxed{\mathrm{SNR} \longrightarrow \frac{1}{\sigma_g^2}.} \tag{6.979}$$

Thus increasing the number of electrons eventually ceases to improve the apparent SNR if the gain non-uniformity is not corrected.

For example, consider a relatively small root-mean-square pixel-to-pixel gain variation of

$$\sigma_g = 0.01, \tag{6.980}$$

corresponding to only 1% rms gain non-uniformity. From Eq. (6.973), the gain-pattern variance becomes equal to the Poisson variance when

$$\mu\sigma_g^2 = 1. \tag{6.981}$$

Therefore,

$$\boxed{\mu = \frac{1}{(0.01)^2} = 10^4} \tag{6.982}$$

electrons per pixel.

Thus a gain non-uniformity that appears very small can become important when many frames or many electrons are averaged. This is why accurate gain normalization is essential in high-precision detector measurements.

Flat-field or gain correction. The gain map can be estimated experimentally by illuminating the detector as uniformly as possible. If the mean response measured in pixel p is

$$\overline{Y}_p, \tag{6.983}$$

a relative gain estimate can be constructed by comparing it with the array-average response.

Once estimates

$$\widehat{b}_p \qquad \text{and} \qquad \widehat{g}_p \tag{6.984}$$

are available, a corrected pixel value can be written schematically as

$$\boxed{X_{p,\mathrm{corr}} = \frac{Y_p - \widehat{b}_p}{\widehat{g}_p}.} \tag{6.985}$$

If

$$\widehat{b}_p = b_p, \qquad \widehat{g}_p = g_p, \tag{6.986}$$

and the detector response is stable, the deterministic offset and gain patterns are removed exactly by Eq. (6.985).

Real calibration maps are estimated from a finite number of electrons and therefore contain statistical uncertainty. Detector gains may also drift with time, temperature, accumulated radiation damage, or operating conditions. Consequently, the corrected image generally contains a residual fixed-pattern component.

Residual gain error after calibration. Suppose, for simplicity, that the offset has been removed perfectly and that the estimated gain differs slightly from the true gain:

$$\widehat{g}_p = g_p \left(1 + \eta_p\right), \tag{6.987}$$

where

$$|\eta_p| \ll 1. \tag{6.988}$$

Using Eq. (6.985),

$$X_{p,\mathrm{corr}} = \frac{g_p X_p}{g_p(1+\eta_p)} \tag{6.989}$$

$$= \frac{X_p}{1+\eta_p}. \tag{6.990}$$

For small calibration errors,

$$\frac{1}{1+\eta_p} \simeq 1 - \eta_p, \tag{6.991}$$

so

$$\boxed{X_{p,\mathrm{corr}} \simeq X_p \left(1 - \eta_p\right).} \tag{6.992}$$

Thus a small uncertainty in the gain map reappears as a multiplicative error in the corrected image.

For an approximately uniform signal $X_p \simeq \mu$, the residual pattern is

$$\delta X_{p,\mathrm{res}} \simeq -\mu \eta_p, \tag{6.993}$$

and therefore its spatial variance is approximately

$$\boxed{\mathrm{Var}_p\left(\delta X_{p,\mathrm{res}}\right) \simeq \mu^2 \, \mathrm{Var}_p(\eta_p).} \tag{6.994}$$

The same quadratic dependence on signal level therefore applies to residual gain-calibration errors.

Fixed pattern versus stochastic noise. It is useful to be precise about terminology. If the gain map $\{g_p\}$ and offset map $\{b_p\}$ are perfectly stable, then repeatedly recording the same uniform illumination produces the same mean spatial pattern. The pattern is therefore not random in time.

In statistical terms, it belongs to the spatially varying mean response, not to the frame-to-frame covariance around that mean.

Nevertheless, if one analyzes a single detector image without knowing the pattern, its spatial variations are indistinguishable from structure in the specimen. For this reason the term *fixed-pattern noise* is widely used.

After gain and offset correction, the distinction becomes

$$\boxed{\begin{array}{rcl} \text{reproducible gain/offset map} & \Longrightarrow & \text{systematic detector non-uniformity,} \\ \text{uncertainty or drift of that map} & \Longrightarrow & \text{residual detector fluctuations.} \end{array}} \tag{6.995}$$

When the NPS is introduced later, this distinction will be important. A deterministic pattern has a Fourier spectrum, but it is not by itself a noise power spectrum unless an ensemble is defined in which the pattern, detector position, or calibration residual is treated as random.

Hot, cold, and defective pixels. The spatial response need not vary only by small amounts. Some pixels can have unusually large dark current, unusually low or high gain, unstable offsets, or essentially no usable response.

A pixel with an anomalously large dark signal is commonly called a *hot pixel*, whereas a pixel with little or no useful response may be described as a *dead* or *cold pixel*. These are extreme forms of detector non-uniformity rather than a separate microscopic noise mechanism.

Because their positions are fixed, defective pixels can often be identified from calibration data and masked or replaced by interpolation from neighbouring pixels. Such replacement prevents the defect from dominating the image but does not recover the information that a functioning pixel at that location would have measured.

Gain non-uniformity in electron-counting operation. Electron counting suppresses the direct effect of analog gain variation once an event has been identified successfully, because the variable analog charge cluster is replaced by one standardized count. Nevertheless, pixel gain and offset variations remain important *before* that decision is made.

The event-recognition algorithm operates on the measured pixel amplitudes. A pixel with lower analog gain, a larger electronic offset, or a higher

effective detection threshold may therefore contribute less reliably to the recognition of a weak event. Conversely, a pixel with larger gain or unusually large fluctuations may contribute more readily to false detections.

To describe this effect, let $\mathcal{D}$ denote the event that an isolated primary electron is successfully detected and counted, and let $\mathcal{A}_p$ denote the event that its detector response is associated with pixel p. We define the *effective counting efficiency* of pixel p as

$$\boxed{\eta_p = P\left(\mathcal{D} \mid \mathcal{A}_p\right).} \tag{6.996}$$

By definition,

$$0 \leq \eta_p \leq 1. \tag{6.997}$$

The quantity η_p therefore represents the probability that an isolated primary electron associated with pixel p is successfully recognized as one electron event. It includes the combined influence of the local analog gain, electronic noise, detection threshold, charge-sharing pattern, and event-recognition algorithm.

If the mean number of incident primary electrons associated with pixel p is μ_p, and if coincidence loss, false detections, and migration of events between pixels are neglected for the moment, the expected recorded count is

$$\boxed{\mathbb{E}\left[N_{p,\mathrm{out}}\right] = \eta_p \mu_p.} \tag{6.998}$$

Under spatially uniform illumination,

$$\mu_p = \mu, \tag{6.999}$$

so that

$$\mathbb{E}\left[N_{p,\mathrm{out}}\right] = \eta_p \mu. \tag{6.1000}$$

Therefore, pixel-to-pixel variations of η_p generate a multiplicative fixed pattern in the counted image even though the final pixel values are integer-valued electron counts.

It is useful to separate the mean counting efficiency from its spatial variation by writing

$$\eta_p = \overline{\eta}\left(1 + \varepsilon_{\eta,p}\right), \tag{6.1001}$$

with

$$\mathbb{E}_p\left[\varepsilon_{\eta,p}\right] = 0. \tag{6.1002}$$

For uniform illumination,

$$\mathbb{E}\left[N_{p,\mathrm{out}}\right] = \overline{\eta}\mu\left(1 + \varepsilon_{\eta,p}\right), \tag{6.1003}$$

which has exactly the same multiplicative structure as the analog gain variation discussed above.

Accurate flat-field calibration is therefore required in counting mode as well as in integrating mode. The quantity being calibrated is no longer merely the analog charge-to-voltage gain; it is the complete effective probability that an incident primary electron produces a recorded count.

The description in Eq. (6.996) is a local approximation. In a real counting detector, an electron entering near a pixel boundary may generate a cluster spanning several pixels and may ultimately be assigned to a neighbouring output pixel. A more complete description would therefore specify the probability that an electron entering at $\boldsymbol{\rho}_0$ is detected and localized at $\widehat{\boldsymbol{\rho}}$. The scalar efficiency η_p is the simpler quantity obtained when only the probability of successful detection is of interest.

6.6.9 Analog-to-digital quantization

The detector signal is ultimately stored and processed numerically. Before this can happen, the continuous voltage produced by the pixel and readout electronics must be converted into a finite set of digital values by an *analog-to-digital converter* (ADC).

This conversion introduces another form of discretization. It should not be confused with either the discreteness of the incident imaging electrons or the discreteness of the electron–hole pairs generated in silicon:

primary-electron discreteness	:	how many imaging electrons arrive,
charge-carrier discreteness	:	how many electron–hole pairs are generated,
ADC discreteness	:	which numerical code represents an analog voltage.

(6.1004)

The first two are physical properties of the detection process. The third is introduced by the electronic representation of the measured signal.

From a continuous voltage to a digital code. Consider an ideal b-bit ADC whose usable input-voltage interval extends from $V_{\min}$ to $V_{\max}$. The total full-scale voltage range is

$$V_{\mathrm{FS}} = V_{\max} - V_{\min}. \tag{6.1005}$$

A b-bit converter has

$$2^b \tag{6.1006}$$

digital codes, so the width of one quantization interval is

$$\boxed{\Delta V_{\mathrm{ADC}} = \frac{V_{\mathrm{FS}}}{2^b}.} \tag{6.1007}$$

This voltage interval is commonly called one *least significant bit* (LSB).

Thus the ADC does not retain the exact continuous input voltage. All voltages falling inside one interval of width ΔV_{ADC} are represented by the same digital code.

Ignoring saturation at the two ends of the ADC range, the conversion can be represented schematically as

$$\boxed{D = \mathrm{round}\left(\frac{V - V_{\mathrm{off}}}{\Delta V_{\mathrm{ADC}}}\right),} \tag{6.1008}$$

where D is an integer digital code and V_{off} represents the voltage corresponding to the chosen numerical origin.

Equation (6.1008) is an idealized local model. A real ADC also has finite input range, offset, gain error, and differential and integral nonlinearities. The essential point here is simply that the continuous voltage axis is replaced by discrete levels separated by ΔV_{ADC}.

Quantization error. Let V_Q denote the voltage represented by the digital code. The *quantization error* is

$$\boxed{\epsilon_Q = V_Q - V.} \tag{6.1009}$$

For an ideal converter that rounds to the nearest quantization level, the error lies within half one LSB:

$$\boxed{-\frac{\Delta V_{\mathrm{ADC}}}{2} \le \epsilon_Q < \frac{\Delta V_{\mathrm{ADC}}}{2}.} \tag{6.1010}$$

Assuming a uniformly distributed noise, we would have

$$\mathrm{Var}(\epsilon_Q) = \mathbb{E}[\epsilon_Q^2] \tag{6.1011}$$

$$= \frac{1}{\Delta V_{\mathrm{ADC}}} \int_{-\Delta V_{\mathrm{ADC}}/2}^{\Delta V_{\mathrm{ADC}}/2} \epsilon_Q^2 \, d\epsilon_Q \tag{6.1012}$$

$$= \boxed{\frac{\Delta V_{\mathrm{ADC}}^2}{12}.} \tag{6.1013}$$

The important distinction from the noise mechanisms considered earlier is that ϵ_Q is not intrinsically random. For a given exact input voltage and a deterministic ADC, the quantization error is completely determined. It becomes useful to describe it statistically only when the input voltage varies from measurement to measurement, or when other analog fluctuations effectively randomize its position within a quantization interval.

Input-referred quantization in electrons. The voltage step of the ADC, ΔV_{ADC}, describes the digitization resolution at the *input of the converter*. By itself, however, this number is not particularly informative about detector performance. What matters physically is how much collected charge at the sensing node is required to change the digital output by one code.

Between the sensing node and the ADC, the signal normally passes through one or more analog amplification stages. Let

$$A_{\mathrm{el}} \tag{6.1014}$$

denote their overall voltage gain. The ADC quantization interval can then be referred backwards through the electronic chain and expressed as an equivalent charge, or more conveniently as an equivalent number of collected electrons.

This provides a common scale on which ADC quantization can be compared directly with the electronic noise expressed as an equivalent noise charge and with the charge generated by an incident imaging electron. The conversion is derived below.

Input-referred ADC quantization in equivalent electrons.

For a b-bit ADC with full-scale voltage range V_{FS}, Eq. (6.1007) gives

$$\Delta V_{\mathrm{ADC}} = \frac{V_{\mathrm{FS}}}{2^b}. \tag{6.1015}$$

If the analog electronics between the sensing node and the ADC has voltage gain A_{el}, this interval corresponds at the sensing node to

$$\Delta V_{\mathrm{node,LSB}} = \frac{\Delta V_{\mathrm{ADC}}}{A_{\mathrm{el}}}. \tag{6.1016}$$

Using the charge-to-voltage relation of Eq. (6.501), the corresponding

charge interval is

$$\Delta Q_{\mathrm{LSB}} = C_{\mathrm{node}} \Delta V_{\mathrm{node,LSB}} = \frac{C_{\mathrm{node}} \Delta V_{\mathrm{ADC}}}{A_{\mathrm{el}}}. \tag{6.1017}$$

Dividing by the elementary charge gives the number of collected electrons represented by one ADC step,

$$\boxed{N_{\mathrm{LSB}} = \frac{C_{\mathrm{node}} \Delta V_{\mathrm{ADC}}}{e A_{\mathrm{el}}}.} \tag{6.1018}$$

Under the uniform quantization-error approximation derived in Eq. (6.1010), the rms error is one quantization interval divided by $\sqrt{12}$. Therefore,

$$\boxed{\sigma_{N,Q} = \frac{C_{\mathrm{node}} \Delta V_{\mathrm{ADC}}}{\sqrt{12}\, e A_{\mathrm{el}}}.} \tag{6.1019}$$

Finally, substituting $\Delta V_{\mathrm{ADC}} = V_{\mathrm{FS}}/2^b$ gives

$$\boxed{\sigma_{N,Q} = \frac{C_{\mathrm{node}} V_{\mathrm{FS}}}{\sqrt{12}\, e A_{\mathrm{el}}\, 2^b}.} \tag{6.1020}$$

Equation (6.1020) shows directly how the electronic design controls the significance of digitization. Adding one ADC bit halves the input-referred rms quantization error. Increasing the analog gain also makes each ADC code correspond to a smaller amount of charge at the sensing node, while decreasing C_{node} increases the voltage produced by each collected electron. Conversely, increasing the full-scale ADC range without increasing the bit depth makes the quantization coarser.

The relevant design criterion is therefore not simply to use a large number of ADC bits, but to make the input-referred quantization scale sufficiently smaller than the other fluctuations already present in the detector. Once this condition is satisfied, increasing the digital resolution further provides little improvement because the measurement is already limited by the analog detector and readout processes rather than by digitization.

An illustrative numerical example. Consider, for illustration, a 12-bit ADC with a 1 V full-scale input range. From Eq. (6.1007),

$$\Delta V_{\mathrm{ADC}} = \frac{1\ \mathrm{V}}{4096} \simeq 244\ \mu\mathrm{V}. \tag{6.1021}$$

Suppose also that

$$C_{\text{node}} = 5 \text{ fF} \tag{6.1022}$$

and that the analog electronics provides a voltage gain

$$A_{\text{el}} = 4. \tag{6.1023}$$

One ADC step then corresponds at the sensing node to

$$\Delta V_{\text{node,LSB}} \simeq 61 \ \mu\text{V}. \tag{6.1024}$$

Since one collected electron produces approximately

$$\frac{e}{C_{\text{node}}} \simeq 32 \ \mu\text{V}, \tag{6.1025}$$

one LSB corresponds to approximately

$$\boxed{N_{\text{LSB}} \simeq 1.9 \ e^{-}.} \tag{6.1026}$$

The rms quantization uncertainty is therefore

$$\boxed{\sigma_{N,Q} \simeq \frac{1.9}{\sqrt{12}} \simeq 0.55 \ e^{-}.} \tag{6.1027}$$

These values are illustrative rather than representative of a particular commercial cryo-EM detector. They show, however, that suitable combination of analog conversion gain and ADC resolution can make the quantization uncertainty substantially smaller than the analog electronic noise and the charge generated by one primary imaging electron.

Quantization is not necessarily white noise. The standard deviation in Eq. (6.1020) is often referred to as *quantization noise*. This terminology can be misleading if the assumptions behind the uniform-error model are forgotten.

An ideal ADC is deterministic. If the same noiseless analog voltage is presented repeatedly, it produces the same digital code and hence the same quantization error every time. The error is then correlated with the signal rather than an independent random fluctuation.

The approximation

$$\epsilon_Q \sim \mathcal{U}\left(-\frac{\Delta V_{\text{ADC}}}{2}, \frac{\Delta V_{\text{ADC}}}{2}\right) \tag{6.1028}$$

becomes appropriate when the analog signal and its existing noise vary enough that the position of the input within one ADC interval is effectively randomized. Under additional assumptions, quantization error may then behave approximately as an uncorrelated noise contribution.

It should therefore not be assumed automatically that quantization gives a spatially white NPS. If the analog signal explores only a few ADC levels, or if structured signals repeatedly fall at similar positions within the quantization intervals, the quantization error can remain signal-dependent and can produce structured spectral components.

The NPS treatment developed later should therefore regard Eq. (6.1020) as an approximation valid under the usual high-resolution or effectively dithered quantization conditions, not as a universal property of digitization.

Quantization and saturation are different effects. The bounded error in Eq. (6.1010) applies only while the analog voltage remains inside the ADC input range. If the signal exceeds $V_{\max}$ or falls below $V_{\min}$, the output code cannot represent the true voltage and is clipped to the highest or lowest available code.

This is *saturation*, not ordinary quantization noise. Whereas quantization introduces a small local error of at most half one LSB in the ideal rounding model, saturation can produce an arbitrarily large error and destroys information about the true signal amplitude.

The ADC range must therefore be chosen to balance two requirements:

$$\boxed{\begin{array}{lcl} \text{small enough voltage range} & \Longrightarrow & \text{fine quantization,} \\ \text{large enough voltage range} & \Longrightarrow & \text{avoid saturation.} \end{array}} \tag{6.1029}$$

For fixed bit depth b, increasing the full-scale range increases ΔV_{ADC} through Eq. (6.1007). Thus digital resolution and dynamic range cannot be chosen independently.

ADC quantization is not electron counting. In an electron-counting detector, the existence of integer-valued output images can obscure an important distinction. The ADC does not normally decide whether zero, one, or two primary imaging electrons entered a pixel.

The ADC first converts the analog voltage of every relevant pixel into a digital numerical value. One primary electron generally generates thousands of electron–hole pairs and produces a spatial cluster spanning several pixels. The digitized amplitudes of that cluster are then examined by the event-recognition algorithm.

As summarized in Eq. (6.512), the counting operation is therefore schematically

$$\boxed{\begin{array}{c}\text{analog pixel voltages}\\ \Downarrow\\ \text{ADC codes}\\ \Downarrow\\ \text{threshold and cluster analysis}\\ \Downarrow\\ \text{identified primary-electron event.}\end{array}} \tag{6.1030}$$

ADC quantization can nevertheless influence counting indirectly. If the quantization step is too coarse, it perturbs the measured amplitudes within the event cluster and can affect threshold crossing, cluster classification, and subpixel localization. When the quantization step is small compared with the analog electronic fluctuations and with the single-electron signal scale, these effects become correspondingly small.

Thus analog-to-digital quantization is best regarded as the final discretization of the electronic measurement, not as the physical counting of the incident electrons.

6.7 Noise power spectrum

The preceding section identified the principal physical mechanisms that make the detector output fluctuate. We considered incident-electron shot noise, fluctuations of the single-electron response, charge sharing and localization uncertainty, coincidence loss, electronic readout noise, dark current, gain non-uniformity, and analog-to-digital quantization.

Those mechanisms answer the question

$$\textit{why does the measured signal fluctuate?} \tag{6.1031}$$

We now ask a different question:

$$\boxed{\textit{how are those fluctuations distributed in space?}} \tag{6.1032}$$

This distinction is essential for imaging. Two detectors can produce noise with the same variance while distributing that noise very differently among neighbouring pixels and, consequently, among spatial frequencies. One detector may produce a fine, nearly pixel-to-pixel grain, whereas another may produce broad correlated fluctuations over several pixels. A single variance cannot distinguish these cases.

The appropriate quantity is the *noise power spectrum* (NPS). The NPS describes how the power of the random detector fluctuations is distributed as a function of spatial frequency. It is therefore the noise counterpart of the frequency-domain quantities introduced earlier for deterministic image transfer.

The need for such a description has already appeared in Sec. 6.6.2. There we introduced the spatial covariance of the random single-electron response in Eq. (6.732), and showed that the image covariance contains distinct shot-noise and single-electron-response contributions in Eq. (6.740). The relation between first- and second-order detector statistics was summarized in Eq. (6.748) as

$$\text{mean response} \longrightarrow \text{PSF and MTF}, \qquad \text{covariance} \longrightarrow \text{NPS}. \tag{6.1033}$$

We now develop the second of these chains explicitly.

6.7.1 Why a single variance is not enough

Consider one recorded detector image and let

$$I_p \tag{6.1034}$$

denote the measured value in pixel p. Repeating the same experiment many times under statistically identical conditions would in general produce different values of I_p. We separate the mean response from the random fluctuation by writing

$$\boxed{I_p = \overline{I}_p + n_p,} \tag{6.1035}$$

where

$$\overline{I}_p = \mathbb{E}[I_p] \tag{6.1036}$$

is the ensemble-mean pixel value and

$$\boxed{n_p = I_p - \overline{I}_p} \tag{6.1037}$$

is the random noise fluctuation. By construction,

$$\mathbb{E}[n_p] = 0. \tag{6.1038}$$

This separation is also useful for interpreting the fixed-pattern effects discussed previously. A perfectly reproducible pixel-dependent offset or gain belongs to the mean response $\overline{I}_p$. It becomes part of the stochastic noise

n_p only to the extent that the pattern is imperfectly calibrated, drifts with time, or otherwise varies between realizations.

The most immediate scalar measure of the fluctuations in pixel p is their variance,

$$\boxed{\sigma_p^2 = \mathrm{Var}(I_p) = \mathbb{E}[n_p^2].} \tag{6.1039}$$

For a spatially uniform detector under uniform illumination, the variance may be approximately the same in every pixel,

$$\sigma_p^2 \simeq \sigma^2. \tag{6.1040}$$

This number tells us the typical squared amplitude of the noise at one pixel. It does *not*, however, tell us whether the fluctuation in one pixel is related to the fluctuations in neighbouring pixels.

The missing information is spatial correlation. To see what the variance fails to describe, consider two pixels p and q. Their joint fluctuation is characterized by the covariance

$$\boxed{C_{pq} = \mathbb{E}\left[n_p n_q\right].} \tag{6.1041}$$

For $p = q$,

$$C_{pp} = \mathbb{E}[n_p^2] = \sigma_p^2, \tag{6.1042}$$

so the ordinary variance is simply the diagonal of the covariance matrix.

For $p \neq q$, however,

$$C_{pq} \tag{6.1043}$$

contains additional information:

$$\boxed{\begin{array}{lll} C_{pq} > 0 & \Longrightarrow & \text{the two pixels tend to fluctuate in the same direction,} \\ C_{pq} < 0 & \Longrightarrow & \text{the two pixels tend to fluctuate in opposite directions,} \\ C_{pq} = 0 & \Longrightarrow & \text{no linear correlation between their fluctuations.} \end{array}} \tag{6.1044}$$

Two noise patterns with the same variance.

Consider a one-dimensional sequence of independent random variables

$$u_p, \tag{6.1045}$$

with

$$\mathbb{E}[u_p] = 0, \qquad \mathrm{Var}(u_p) = \sigma^2, \tag{6.1046}$$

and

$$\mathbb{E}[u_p u_q] = 0, \qquad p \neq q. \tag{6.1047}$$

Define a first noise image by

$$n_p^{(1)} = u_p. \tag{6.1048}$$

Its variance is

$$\mathrm{Var}\left(n_p^{(1)}\right) = \sigma^2, \tag{6.1049}$$

and neighbouring pixels are uncorrelated:

$$\mathrm{Cov}\left(n_p^{(1)}, n_{p+1}^{(1)}\right) = 0. \tag{6.1050}$$

Now define a second noise image by averaging neighbouring independent fluctuations,

$$\boxed{n_p^{(2)} = \frac{u_p + u_{p-1}}{\sqrt{2}}.} \tag{6.1051}$$

Its variance is

$$\mathrm{Var}\left(n_p^{(2)}\right) = \frac{1}{2}\,\mathrm{Var}\left(u_p + u_{p-1}\right) \tag{6.1052}$$

$$= \frac{1}{2}\left(\mathrm{Var}(u_p) + \mathrm{Var}(u_{p-1})\right) \tag{6.1053}$$

$$= \boxed{\sigma^2.} \tag{6.1054}$$

Thus the two noise images have exactly the same pixel variance.

Their spatial statistics are nevertheless different. Two neighbouring samples of the second process are

$$n_p^{(2)} = \frac{u_p + u_{p-1}}{\sqrt{2}}, \qquad n_{p+1}^{(2)} = \frac{u_{p+1} + u_p}{\sqrt{2}}. \tag{6.1055}$$

They share the random variable u_p, and therefore

$$\mathrm{Cov}\left(n_p^{(2)}, n_{p+1}^{(2)}\right) = \frac{1}{2}\mathbb{E}\left[(u_p + u_{p-1})(u_{p+1} + u_p)\right] \tag{6.1056}$$

$$= \frac{1}{2}\mathbb{E}[u_p^2] \tag{6.1057}$$

$$= \boxed{\frac{\sigma^2}{2}.} \tag{6.1058}$$

Thus

$$\boxed{\begin{array}{lcll} \text{noise 1} & : & \mathrm{Var} = \sigma^2, & C(1) = 0, \\ \text{noise 2} & : & \mathrm{Var} = \sigma^2, & C(1) = \sigma^2/2. \end{array}} \tag{6.1059}$$

The two processes therefore have the same variance but different spatial structure.

The second noise image in Eq. (6.1051) would appear spatially smoother because neighbouring pixels tend to rise and fall together. The first would appear more finely grained because adjacent pixels fluctuate independently. Yet a calculation of the pixel variance alone would assign exactly the same value σ^2 to both.

Detector physics naturally produces such correlations. The distinction is not merely mathematical. Several physical mechanisms discussed earlier generate spatial correlations.

For an ideal incident Poisson process, counts in disjoint spatial regions are independent, as expressed by Eq. (6.671). At the detector output, however, one primary electron can distribute signal among several neighbouring pixels. The corresponding event-to-event fluctuations were described by the covariance K_q in Eq. (6.732).

Charge redistribution provides a particularly clear example. The two-pixel model derived earlier showed that fluctuations in the two shared charges can be anticorrelated: if one pixel receives more charge, the neighbouring pixel must receive less. Thus two detectors could have similar per-pixel variances while having very different correlations between pixels.

Other mechanisms can produce positive correlations. A broad single-electron charge cloud, shared electronic circuitry, or a slowly varying baseline can make neighbouring pixels fluctuate together. Coincidence processing can introduce still different correlations by making the probability of recording

one event depend on the presence of another nearby event.

Therefore,

$$\boxed{\text{variance} \quad \text{measures the amount of fluctuation,}} \tag{6.1060}$$

but

$$\boxed{\text{covariance} \quad \text{also measures its spatial organization.}} \tag{6.1061}$$

From covariance to spatial frequency. For imaging, the spatial covariance is already a much richer description than a single variance. It is nevertheless often more useful to express the same information in terms of spatial frequency.

A slowly varying fluctuation contributes primarily at low spatial frequencies. Pixel-to-pixel grain contains stronger high-frequency components. Periodic row or column structures produce concentrated spectral components in particular directions. The detector noise is therefore naturally described by asking how much fluctuation power is associated with each spatial frequency

$$\mathbf{R} = (R_x, R_y). \tag{6.1062}$$

For imaging, the spatial covariance is already a much richer description than a single variance. It is nevertheless often more useful to express exactly the same second-order information in the spatial- frequency domain.

Consider a zero-mean noise field

$$n(\boldsymbol{\rho}), \qquad \mathbb{E}\left[n(\boldsymbol{\rho})\right] = 0. \tag{6.1063}$$

If the noise is statistically homogeneous, so that its second-order statistics depend only on the spatial separation $\Delta\boldsymbol{\rho}$ and not on the absolute position, its autocovariance function can be written as

$$\boxed{C_n(\Delta\boldsymbol{\rho}) = \mathbb{E}\left[n(\boldsymbol{\rho})n(\boldsymbol{\rho} + \Delta\boldsymbol{\rho})\right].} \tag{6.1064}$$

Because the noise has zero mean, this autocovariance is also its autocorrelation function.

The *power spectral density* (PSD) of the noise is the Fourier transform of this correlation function. In detector imaging it is conventionally called the *noise power spectrum* (NPS):

$$\boxed{\mathrm{NPS}(\mathbf{R}) = \int_{\mathbb{R}^2} C_n(\Delta\boldsymbol{\rho}) \exp\left(-j2\pi\mathbf{R}\cdot\Delta\boldsymbol{\rho}\right) d^2\Delta\boldsymbol{\rho}.} \tag{6.1065}$$

Equation (6.1065) is the spatial form of the Wiener–Khinchin theorem. It shows that the covariance and the NPS are not two independent descriptions of detector noise: they contain the same second-order statistical information, represented respectively in the spatial and spatial-frequency domains.

The inverse relation is

$$\boxed{C_n(\Delta\boldsymbol{\rho}) = \int_{\mathbb{R}^2} \mathrm{NPS}(\mathbf{R}) \exp\left(j2\pi\mathbf{R}\cdot\Delta\boldsymbol{\rho}\right) d^2\mathbf{R}.} \tag{6.1066}$$

In particular, evaluating the covariance at zero separation gives

$$C_n(\mathbf{0}) = \mathbb{E}\left[n^2(\boldsymbol{\rho})\right] = \sigma_n^2. \tag{6.1067}$$

Equation (6.1066) therefore gives the important relation

$$\boxed{\sigma_n^2 = \int_{\mathbb{R}^2} \mathrm{NPS}(\mathbf{R})\, d^2\mathbf{R}.} \tag{6.1068}$$

Thus the ordinary noise variance measures the *total* noise power, whereas the NPS reveals how that same power is distributed among spatial frequencies.

The conceptual progression is

$$\boxed{\begin{array}{ccccc} \text{random noise field} & \longrightarrow & \text{autocovariance} & \xleftrightarrow{\mathcal{F}} & \text{power spectral density} \\ n(\boldsymbol{\rho}) & & C_n(\Delta\boldsymbol{\rho}) & & \mathrm{NPS}(\mathbf{R}). \end{array}} \tag{6.1069}$$

Stationarity of the detector noise. The random-field description required for the NPS has already appeared in Sec. 6.6.2, where the random detector image was written as a superposition of stochastic single-electron responses. To define a spatial power spectral density, however, one additional assumption must be made explicit.

In general, the covariance of a noise field depends on the two spatial positions separately,

$$C_n\left(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2\right) = \mathbb{E}\left[n(\boldsymbol{\rho}_1)n(\boldsymbol{\rho}_2)\right]. \tag{6.1070}$$

For a statistically homogeneous detector under uniform illumination, we assume that the second-order statistics are invariant under translation. The covariance then depends only on the displacement between the two positions,

$$\boxed{C_n\left(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2\right) = C_n\left(\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1\right).} \tag{6.1071}$$

Together with a position-independent mean,

$$\mathbb{E}\left[n(\boldsymbol{\rho})\right] = 0, \tag{6.1072}$$

Eq. (6.1071) is the condition of *wide-sense stationarity* for the spatial noise field.

This assumption is what makes it possible to describe the detector noise by a single spatial-frequency function $\mathrm{NPS}(\mathbf{R})$. If the noise statistics varied strongly across the detector, for example because different regions had different gain, dark current, or readout circuitry, a single global NPS would no longer provide a complete description.

Ensemble averages and spatial averages. The definitions above use ensemble expectations such as

$$\mathbb{E}\left[n(\boldsymbol{\rho})n(\boldsymbol{\rho}+\Delta\boldsymbol{\rho})\right], \tag{6.1073}$$

which conceptually require many statistically equivalent realizations of the detector image.

In practice, the NPS is often estimated from a finite number of flat-field images and from many spatial regions within those images. This procedure implicitly assumes that sufficiently large spatial averages reproduce the ensemble averages of interest. Such a process is called *ergodic* with respect to the corresponding statistics.

Stationarity and ergodicity are therefore distinct ideas:

$$\boxed{\begin{array}{lcl} \text{stationarity} & : & \text{the statistics do not depend on absolute position,} \\ \text{ergodicity} & : & \text{spatial averages can estimate ensemble statistics.} \end{array}} \tag{6.1074}$$

6.7.2 Definition and estimation of the NPS

Equation (6.1065) defined the noise power spectrum for a continuous, stationary random field as the Fourier transform of its autocovariance. Detector data, however, are not continuous fields. They consist of a finite rectangular array of sampled pixel values.

We therefore need to translate the continuous definition into the quantity that is actually calculated from detector images.

The NPS of a sampled detector image. Let the pixel spacings along the two detector coordinates be

$$a_x \qquad \text{and} \qquad a_y, \tag{6.1075}$$

and let

$$n_{mn} \tag{6.1076}$$

denote the zero-mean noise value measured in pixel (m, n).

Under the wide-sense stationarity assumption introduced previously, the covariance between two pixels depends only on their displacement. We therefore define the discrete covariance

$$\boxed{C_{rs} = \mathbb{E}\left[n_{mn} n_{m+r,n+s}\right],} \tag{6.1077}$$

where r and s are integer pixel displacements.

The sampled spatial power spectral density is the discrete-space counterpart of Eq. (6.1065):

$$\boxed{\begin{aligned}\mathrm{NPS}_{\mathrm{samp}}(R_x, R_y) = a_x a_y \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} C_{rs} \\ \times \exp\left[-j2\pi\left(R_x r a_x + R_y s a_y\right)\right].\end{aligned}} \tag{6.1078}$$

The factor $a_x a_y$ is important. It ensures that $\mathrm{NPS}_{\mathrm{samp}}$ has the units of a spatial power spectral density and that integration over spatial frequency returns the noise variance.

Because the detector has been sampled on a lattice, Eq. (6.1078) is periodic in spatial frequency:

$$\mathrm{NPS}_{\mathrm{samp}}\left(R_x + \frac{1}{a_x}, R_y\right) = \mathrm{NPS}_{\mathrm{samp}}(R_x, R_y), \tag{6.1079}$$

and similarly along R_y. It is therefore sufficient to consider one Nyquist region,

$$\boxed{\mathcal{B} = \left[-\frac{1}{2a_x}, \frac{1}{2a_x}\right] \times \left[-\frac{1}{2a_y}, \frac{1}{2a_y}\right].} \tag{6.1080}$$

Within this region, the inverse relation is

$$\boxed{\begin{aligned}C_{rs} = \int_{\mathcal{B}} \mathrm{NPS}_{\mathrm{samp}}(R_x, R_y) \\ \times \exp\left[j2\pi\left(R_x r a_x + R_y s a_y\right)\right] dR_x\, dR_y.\end{aligned}} \tag{6.1081}$$

For zero displacement,

$$C_{00} = \mathrm{Var}(n_{mn}), \tag{6.1082}$$

so

$$\boxed{\sigma_n^2 = \int_{\mathcal{B}} \mathrm{NPS}_{\mathrm{samp}}(R_x, R_y)\, dR_x\, dR_y.} \tag{6.1083}$$

Equation (6.1083) is the sampled-detector version of Eq. (6.1068). The variance is the total area under the NPS within one Nyquist region.

Estimating the NPS from a finite image. In an experiment we do not know the ensemble covariance C_{rs} exactly. Instead, we observe one or more finite noise images.

Consider an $N_x \times N_y$ region containing noise samples n_{mn}, with

$$m = 0, \ldots, N_x - 1, \qquad n = 0, \ldots, N_y - 1. \tag{6.1084}$$

Its discrete Fourier transform is

$$\boxed{\widetilde{n}_{kl} = \sum_{m=0}^{N_x-1} \sum_{n=0}^{N_y-1} n_{mn} \exp\left[-j2\pi\left(\frac{km}{N_x} + \frac{ln}{N_y}\right)\right].} \tag{6.1085}$$

The corresponding spatial frequencies are

$$\boxed{R_{x,k} = \frac{k}{N_x a_x}, \qquad R_{y,l} = \frac{l}{N_y a_y},} \tag{6.1086}$$

with negative indices used for the negative-frequency half of the Fourier grid.

The simplest estimator of the NPS is the two-dimensional *periodogram*,

$$\boxed{\widehat{\mathrm{NPS}}_{kl} = \frac{a_x a_y}{N_x N_y} \left|\widetilde{n}_{kl}\right|^2.} \tag{6.1087}$$

The squared Fourier amplitude therefore becomes a physical power spectral density only after the normalization in Eq. (6.1087) is applied.

Normalization of the discrete NPS estimator.

Let

$$N = N_x N_y. \tag{6.1088}$$

For the Fourier-transform convention of Eq. (6.1085), Parseval's identity

gives

$$\sum_{k,l} |\widetilde{n}_{kl}|^2 = N \sum_{m,n} |n_{mn}|^2. \tag{6.1089}$$

The frequency spacing of the discrete Fourier grid is

$$\Delta R_x = \frac{1}{N_x a_x}, \qquad \Delta R_y = \frac{1}{N_y a_y}. \tag{6.1090}$$

Hence the area represented by one Fourier sample is

$$\Delta R_x \Delta R_y = \frac{1}{N a_x a_y}. \tag{6.1091}$$

Integrating the estimated NPS numerically over the complete Nyquist region therefore gives

$$\sum_{k,l} \widehat{\mathrm{NPS}}_{kl} \Delta R_x \Delta R_y = \sum_{k,l} \frac{a_x a_y}{N} |\widetilde{n}_{kl}|^2 \frac{1}{N a_x a_y} \tag{6.1092}$$

$$= \frac{1}{N^2} \sum_{k,l} |\widetilde{n}_{kl}|^2 \tag{6.1093}$$

$$= \frac{1}{N} \sum_{m,n} |n_{mn}|^2. \tag{6.1094}$$

For a zero-mean noise image, the right-hand side is the measured mean-square fluctuation. Thus

$$\boxed{\sum_{k,l} \widehat{\mathrm{NPS}}_{kl} \Delta R_x \Delta R_y = \widehat{\sigma}_n^2.} \tag{6.1095}$$

The normalization of Eq. (6.1087) therefore guarantees that the area under the estimated NPS equals the measured noise variance.

This normalization is more than a mathematical convention. It provides a useful consistency check in practical calculations: the variance calculated directly from the pixels and the integrated NPS should agree, apart from finite-sample and preprocessing effects.

Why one periodogram is not enough. Although Eq. (6.1087) is conceptually simple, the periodogram of one finite noise realization fluctuates strongly from one acquisition to another. A smooth underlying NPS can

therefore produce a very irregular-looking individual periodogram.

For this reason, the NPS is normally estimated by averaging several statistically equivalent realizations. If

$$\widehat{\mathrm{NPS}}_{kl}^{(j)} \tag{6.1096}$$

is the periodogram obtained from realization j, then

$$\boxed{\widehat{\mathrm{NPS}}_{kl}^{\mathrm{avg}} = \frac{1}{M}\sum_{j=1}^{M}\widehat{\mathrm{NPS}}_{kl}^{(j)}\,.} \tag{6.1097}$$

The realizations may be separate flat-field exposures, separate approximately homogeneous regions from larger images, or both. The ergodicity assumption discussed previously justifies using spatially separated regions to help estimate the same ensemble statistics.

Difference images. Detector NPS measurements are often obtained from pairs of flat-field images recorded under the same experimental conditions. This is useful because a reproducible mean illumination pattern and stable fixed-pattern response can then be removed before calculating the spectrum.

Let two independent flat-field images be

$$I_{mn}^{(1)} = \overline{I}_{mn} + n_{mn}^{(1)}, \tag{6.1098}$$

and

$$I_{mn}^{(2)} = \overline{I}_{mn} + n_{mn}^{(2)}. \tag{6.1099}$$

Their difference is

$$d_{mn} = I_{mn}^{(1)} - I_{mn}^{(2)} = n_{mn}^{(1)} - n_{mn}^{(2)}. \tag{6.1100}$$

If the two noise realizations are independent and have the same statistics,

$$\mathrm{Var}(d_{mn}) = 2\sigma_n^2, \tag{6.1101}$$

and the same factor of two applies at every spatial frequency:

$$\boxed{\mathrm{NPS}_d(\mathbf{R}) = 2\,\mathrm{NPS}(\mathbf{R}).} \tag{6.1102}$$

The NPS of one detector image is therefore estimated from the difference image as

$$\boxed{\widehat{\mathrm{NPS}}(\mathbf{R}) = \frac{1}{2}\widehat{\mathrm{NPS}}_d(\mathbf{R}).} \tag{6.1103}$$

This procedure removes spatial structures that are identical in the two frames while preserving their independent stochastic fluctuations. Consequently, it estimates the stochastic NPS rather than the power spectrum of a stable fixed-pattern response.

Removing the mean and slow trends. Before calculating an NPS, the deterministic mean signal must be removed. For an ideally uniform flat field this may be done by subtracting the mean value of the region,

$$n_{mn} = I_{mn} - \overline{I}. \tag{6.1104}$$

If the illumination contains a slow gradient, subtracting only one constant leaves part of that deterministic structure in the image. Its Fourier transform then appears as excess low-frequency power and may be mistaken for detector noise. Slowly varying illumination backgrounds are therefore often estimated and removed before calculating the NPS.

There is, however, a consequence. Subtracting the sample mean forces the zero-frequency Fourier coefficient to vanish:

$$\widetilde{n}_{00} = 0. \tag{6.1105}$$

The zero-frequency bin of such a periodogram is therefore not an experimental estimate of the physical value $\mathrm{NPS}(\mathbf{0})$. More aggressive detrending can also modify the lowest non-zero spatial frequencies and must consequently be used with care.

Finite regions and windowing. Extracting a finite region from an image is equivalent to multiplying the infinite random field by a finite spatial window. The resulting Fourier transform is therefore affected by the Fourier transform of that window.

The simplest choice is a rectangular region, which is implicitly used in Eq. (6.1087). If strong gradients or structures occur near its boundaries, spectral leakage can contaminate neighbouring frequencies. A smoothly tapered window w_{mn} can reduce this leakage.

When a window is applied, however, its loss of mean-square amplitude must be included in the normalization. A corresponding periodogram is

$$\boxed{\widehat{\mathrm{NPS}}_{kl}^{(w)} = \frac{a_x a_y}{\sum_{m,n} w_{mn}^2} \left| \sum_{m,n} w_{mn} n_{mn} e^{-j2\pi(km/N_x + ln/N_y)} \right|^2 .} \tag{6.1106}$$

Large regions and averaging over many realizations are generally preferable whenever sufficient statistically homogeneous data are available.

Two-dimensional and radial NPS. The fundamental detector quantity is the two-dimensional spectrum

$$\mathrm{NPS}(R_x, R_y), \tag{6.1107}$$

because it retains directional information. Row noise, column noise, anisotropic charge spreading, or readout structures can all produce an NPS that depends on direction.

If the detector noise is approximately rotationally symmetric, the two-dimensional NPS can be summarized by averaging over spatial frequency direction. Writing

$$R = \sqrt{R_x^2 + R_y^2}, \tag{6.1108}$$

the radial NPS is conceptually

$$\boxed{\mathrm{NPS}_{\mathrm{rad}}(R) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{NPS}\left(R\cos\theta, R\sin\theta\right) d\theta.} \tag{6.1109}$$

For a discrete Fourier grid this integral is replaced by an average over frequency samples lying within an annulus around R.

Radial averaging considerably reduces the statistical fluctuations of the estimated spectrum, but it also destroys directional information. The two-dimensional NPS should therefore be inspected before assuming that radial averaging is appropriate.

Units of the NPS. The units follow directly from Eq. (6.1065). If the detector noise $n(\boldsymbol{\rho})$ is measured in some signal unit U, then the covariance has units

$$[C_n] = U^2. \tag{6.1110}$$

Because the Fourier transform integrates over an area,

$$\boxed{[\mathrm{NPS}] = U^2 \times \mathrm{length}^2.} \tag{6.1111}$$

Thus, if the image has been calibrated in equivalent electrons and spatial frequency is expressed in cycles $\mu\mathrm{m}^{-1}$, the NPS may be expressed in

$$e^{-2}\,\mu\mathrm{m}^2. \tag{6.1112}$$

If detector coordinates are instead measured in pixels, the corresponding units are signal-unit squared times pixel squared.

Different detector publications sometimes adopt different numerical normalizations. The convention used here is chosen so that the integrated NPS

always reproduces the variance through Eq. (6.1083). Stating the normalization and units is therefore essential whenever NPS values from different detectors are compared.

The acquisition conditions belong to the NPS measurement. Finally, the NPS is not necessarily an immutable property of the detector. Several noise mechanisms derived earlier depend on the operating conditions.

Incident shot noise changes with electron fluence. Dark-current noise depends on integration time and temperature. Electronic readout noise depends on the readout mode and bandwidth. In counting detectors, coincidence loss makes the statistics depend explicitly on the incident dose rate.

Consequently, an experimentally measured NPS should always be associated with the conditions under which it was obtained:

$$\boxed{\begin{array}{c}\text{dose and dose rate,}\\ \text{integration or internal-frame time,}\\ \text{detector operating mode,}\\ \text{temperature and readout configuration,}\\ \text{gain and fixed-pattern corrections.}\end{array}} \tag{6.1113}$$

This dependence will become particularly important when the NPS is combined with the MTF to define the detective quantum efficiency.

Normalized noise power spectrum The absolute NPS depends on the scale in which the detector output is expressed. If the electronic gain is doubled, for example, the measured pixel values double and the NPS increases by a factor of four, even though the underlying detector statistics have not changed.

For detector characterization it is therefore useful to divide the NPS by the square of the mean detector signal. If

$$\overline{I} \tag{6.1114}$$

is the mean signal under uniform illumination, the *normalized noise power spectrum* (NNPS) is defined as

$$\boxed{\mathrm{NNPS}(\mathbf{R}) = \frac{\mathrm{NPS}(\mathbf{R})}{\overline{I}^{\,2}}.} \tag{6.1115}$$

This normalization removes any purely multiplicative change in the detector output. Indeed, if the measured signal is multiplied by a constant gain g,

$$I \longrightarrow gI, \tag{6.1116}$$

then

$$\overline{I} \longrightarrow g\overline{I}, \tag{6.1117}$$

while

$$\mathrm{NPS} \longrightarrow g^2\,\mathrm{NPS}\,. \tag{6.1118}$$

Therefore,

$$\mathrm{NNPS} \longrightarrow \frac{g^2\,\mathrm{NPS}}{g^2\overline{I}^{\,2}} = \mathrm{NNPS}\,. \tag{6.1119}$$

The NNPS therefore describes noise relative to the mean signal rather than in arbitrary detector output units.

Because Eq. (6.1083) gives

$$\sigma_I^2 = \int_{\mathcal{B}} \mathrm{NPS}(\mathbf{R})\,d^2\mathbf{R}, \tag{6.1120}$$

division by $\overline{I}^{\,2}$ gives

$$\boxed{\frac{\sigma_I^2}{\overline{I}^{\,2}} = \int_{\mathcal{B}} \mathrm{NNPS}(\mathbf{R})\,d^2\mathbf{R}.} \tag{6.1121}$$

Thus the area under the NNPS is the relative noise variance, or equivalently the square of the coefficient of variation of the detector signal under uniform illumination.

The NNPS will be particularly useful when the noise spectrum is combined with the modulation transfer function to define the frequency-dependent detective quantum efficiency.

6.8 Spectral signatures of the physical noise mechanisms

Having defined and discussed how to estimate the NPS, we can now return to the physical noise mechanisms introduced earlier and ask what spectral signature each one produces.

Some care is required in this decomposition. The mechanisms discussed below do not all correspond to statistically independent additive noise terms. In particular, the energy-deposition, charge-generation, and charge-sharing

fluctuations are different microscopic contributions to the random single-electron response introduced in Sec. 6.6.2. Coincidence loss is different again: it modifies the statistics of the recorded electron events themselves and cannot in general be represented as an additive noise source. Finally, a perfectly reproducible fixed pattern belongs to the mean detector response rather than to the stochastic NPS.

The purpose of the following subsections is therefore not simply to construct a list of terms to be added, but to understand how each physical mechanism changes the spatial covariance and, consequently, the NPS.

6.8.1 Incident-electron shot noise

Consider first an ideal detector that could record the entrance coordinates of the incident primary electrons without any additional detector spreading or noise. Let

$$\Phi_0 \tag{6.1122}$$

denote the mean incident electron fluence per unit detector area during one exposure.

The incident electrons form the spatial Poisson process described in Sec. 6.6.1. In the continuous limit of Eq. (6.671), the covariance of the zero-mean incident-electron density is

$$\boxed{C_{\text{shot,in}}(\Delta\boldsymbol{\rho}) = \Phi_0 \delta^{(2)}(\Delta\boldsymbol{\rho}).} \tag{6.1123}$$

The Dirac delta expresses the spatial independence of the incident Poisson process: before detection, the fluctuation at one position is not correlated with the fluctuation at another position.

Taking the Fourier transform according to Eq. (6.1065) immediately gives

$$\boxed{\text{NPS}_{\text{shot,in}}(\mathbf{R}) = \Phi_0.} \tag{6.1124}$$

Thus ideal incident-electron shot noise is *white*: every spatial frequency contains the same noise power density.

Propagation of shot noise through a deterministic detector response. White input shot noise does not imply that the output noise of an integrating detector is white. Suppose that every primary electron produces exactly the same spatial response,

$$q(\boldsymbol{\rho}) = \overline{Q}\, h_{\text{det}}(\boldsymbol{\rho}), \tag{6.1125}$$

where h_{det} is the normalized detector PSF defined in Eq. (6.729) and $\overline{Q}$ is the charge generated by each electron.

For a spatially uniform Poisson input, the covariance derived in Eq. (6.735) becomes the autocorrelation of this deterministic single-electron response. Consequently,

$$\boxed{\mathrm{NPS}_{\mathrm{shot,out}}(\mathbf{R}) = \Phi_0 \overline{Q}^{\,2} \left|H_{\mathrm{det}}(\mathbf{R})\right|^2 .} \tag{6.1126}$$

The detector therefore correlates the output noise spatially for exactly the same reason that it blurs the mean signal: one incident electron contributes signal to several neighbouring positions.

If finite pixel integration is included, the relevant presampling transfer function is H_{pre} from Eq. (6.566), giving

$$\boxed{\mathrm{NPS}_{\mathrm{shot,pre}}(\mathbf{R}) = \Phi_0 \overline{Q}^{\,2} \left|H_{\mathrm{pre}}(\mathbf{R})\right|^2 .} \tag{6.1127}$$

Thus it is useful to distinguish two statements:

$$\boxed{\begin{array}{ccc} \text{incident Poisson shot noise} & \Longrightarrow & \text{white NPS,} \\ \text{shot noise after deterministic spreading} & \Longrightarrow & \text{NPS shaped by } |H|^2. \end{array}} \tag{6.1128}$$

For the deterministic model above, the mean output signal under uniform illumination is

$$\overline{I} = \Phi_0 \overline{Q}. \tag{6.1129}$$

Hence

$$\boxed{\mathrm{NNPS}_{\mathrm{shot}}(\mathbf{R}) = \frac{|H_{\mathrm{det}}(\mathbf{R})|^2}{\Phi_0}.} \tag{6.1130}$$

For an ideal detector with $H_{\mathrm{det}} = 1$, this reduces to the particularly simple result

$$\mathrm{NNPS}_{\mathrm{shot}} = \frac{1}{\Phi_0}. \tag{6.1131}$$

6.8.2 Fluctuations of the single-electron response

A real direct detector does not produce exactly the same response for every primary electron. The random event response $q_\omega(\boldsymbol{\rho})$ was introduced in Eq. (6.725), and its mean and fluctuation were separated in Eq. (6.730) as

$$q_\omega = \overline{q} + \delta q_\omega. \tag{6.1132}$$

The NPS produced by a Poisson stream of such random events can now be obtained directly from the covariance derived earlier.

NPS of a Poisson stream with a random single-electron response.

Equation (6.735) gives

$$C_I(\Delta\boldsymbol{\rho}) = \Phi_0 \int \mathbb{E}\left[q_\omega(\mathbf{u})q_\omega(\mathbf{u}+\Delta\boldsymbol{\rho})\right] d^2\mathbf{u}. \tag{6.1133}$$

Taking the Fourier transform with respect to $\Delta\boldsymbol{\rho}$ and using the autocorrelation theorem gives

$$\boxed{\mathrm{NPS}_I(\mathbf{R}) = \Phi_0 \mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right].} \tag{6.1134}$$

Using the decomposition already obtained in Eq. (6.744),

$$\mathbb{E}\left[|\widehat{q}_\omega|^2\right] = \left|\overline{\widehat{q}}\right|^2 + \mathbb{E}\left[|\delta\widehat{q}_\omega|^2\right], \tag{6.1135}$$

we obtain

$$\boxed{\begin{aligned}\mathrm{NPS}_I(\mathbf{R}) = & \underbrace{\Phi_0 \left|\overline{\widehat{q}}(\mathbf{R})\right|^2}_{\text{propagated incident shot noise}} \\ & + \underbrace{\Phi_0 \mathbb{E}\left[|\delta\widehat{q}_\omega(\mathbf{R})|^2\right]}_{\text{single-electron-response fluctuations}}.\end{aligned}} \tag{6.1136}$$

Since Eq. (6.746) gives

$$\overline{\widehat{q}}(\mathbf{R}) = \overline{Q} H_{\mathrm{det}}(\mathbf{R}), \tag{6.1137}$$

the first term is

$$\Phi_0 \overline{Q}^2 |H_{\mathrm{det}}(\mathbf{R})|^2, \tag{6.1138}$$

in agreement with Eq. (6.1126).

Equation (6.1134) is one of the central results of detector-noise theory. The NPS depends on

$$\mathbb{E}\left[|\widehat{q}_\omega|^2\right], \tag{6.1139}$$

whereas the MTF depends on the magnitude of the mean response,

$$|\mathbb{E}[\widehat{q}_\omega]|. \tag{6.1140}$$

This is precisely the distinction anticipated in Eq. (6.749).

At zero spatial frequency,

$$\widehat{q}_\omega(\mathbf{0}) = Q_{\text{tot}}, \tag{6.1141}$$

so

$$\boxed{\text{NPS}_I(\mathbf{0}) = \Phi_0 \mathbb{E}\left[Q_{\text{tot}}^2\right].} \tag{6.1142}$$

This is the spatial-frequency version of the compound-Poisson variance derived in Eq. (6.712).

6.8.3 Energy-deposition and charge-generation fluctuations

The simplest single-electron fluctuation is a change in event amplitude without a change in spatial shape. Suppose, as an idealized model, that

$$q_\omega(\boldsymbol{\rho}) = Q_\omega h_{\text{det}}(\boldsymbol{\rho}), \tag{6.1143}$$

where Q_ω fluctuates from event to event but h_{det} is fixed.

Then

$$\widehat{q}_\omega(\mathbf{R}) = Q_\omega H_{\text{det}}(\mathbf{R}), \tag{6.1144}$$

and Eq. (6.1134) becomes

$$\boxed{\text{NPS}_I(\mathbf{R}) = \Phi_0 \mathbb{E}[Q_\omega^2] |H_{\text{det}}(\mathbf{R})|^2.} \tag{6.1145}$$

Since

$$\mathbb{E}[Q_\omega^2] = \overline{Q}^2 + \sigma_Q^2, \tag{6.1146}$$

we can separate

$$\boxed{\begin{aligned}\text{NPS}_I(\mathbf{R}) = \Phi_0 \overline{Q}^2 |H_{\text{det}}(\mathbf{R})|^2 \\ + \Phi_0 \sigma_Q^2 |H_{\text{det}}(\mathbf{R})|^2.\end{aligned}} \tag{6.1147}$$

The first term is propagated incident shot noise. The second is the additional NPS caused by event-to-event amplitude fluctuations:

$$\boxed{\text{NPS}_{\text{amp}}(\mathbf{R}) = \Phi_0 \sigma_Q^2 |H_{\text{det}}(\mathbf{R})|^2.} \tag{6.1148}$$

Thus, when amplitude and spatial shape are separable, energy-deposition fluctuations do not introduce a new spectral shape. They multiply the same $|H_{\text{det}}|^2$ dependence by an excess-noise factor.

Indeed, using Eq. (6.717),

$$\boxed{\mathrm{NPS}_I(\mathbf{R}) = \Phi_0 \overline{Q}^2 \mathcal{F}_{\mathrm{SER}} |H_{\mathrm{det}}(\mathbf{R})|^2.} \tag{6.1149}$$

Under the deterministic charge-collection approximation used in Sec. 6.6.3, Eq. (6.771) gives

$$\frac{\sigma_Q^2}{\overline{Q}^2} = F \frac{w_{\mathrm{Si}}}{\mathbb{E}[E_{\mathrm{dep}}]} + \frac{\mathrm{Var}(E_{\mathrm{dep}})}{\mathbb{E}[E_{\mathrm{dep}}]^2}. \tag{6.1150}$$

Consequently,

$$\boxed{\begin{aligned}\mathrm{NPS}_{\mathrm{amp}}(\mathbf{R}) = \Phi_0 \overline{Q}^2 \Bigg[& F \frac{w_{\mathrm{Si}}}{\mathbb{E}[E_{\mathrm{dep}}]} \\ & + \frac{\mathrm{Var}(E_{\mathrm{dep}})}{\mathbb{E}[E_{\mathrm{dep}}]^2} \Bigg] |H_{\mathrm{det}}(\mathbf{R})|^2.\end{aligned}} \tag{6.1151}$$

This expression shows directly how semiconductor Fano fluctuations and the much larger possible energy-straggling fluctuations enter the NPS.

The separable model is not exact in general. A large energy-transfer event may also produce a spatially broader charge cloud, so event amplitude and event shape can be correlated. In that case the more general second moment in Eq. (6.1134) must be retained.

6.8.4 Charge-sharing and localization fluctuations

Charge-sharing fluctuations have a different spectral character from pure amplitude fluctuations. If the total charge is fixed, charge sharing only redistributes signal among neighbouring pixels.

As shown in Eq. (6.805), conservation of the total event charge implies

$$\delta \widehat{q}_\omega(\mathbf{0}) = 0. \tag{6.1152}$$

Therefore,

$$\boxed{\mathrm{NPS}_{\mathrm{share}}(\mathbf{0}) = 0} \tag{6.1153}$$

for pure redistribution at fixed total charge.

This does not mean that charge-sharing fluctuations produce no noise. Rather, they generate fluctuations at non-zero spatial frequencies.

Spectral signature of the two-pixel charge-sharing fluctuation.

Consider again the two-pixel charge-sharing model introduced in Eq. (6.794),

$$Q_1 = fQ, \qquad Q_2 = (1-f)Q, \tag{6.1154}$$

where the two collection positions are separated by one pixel pitch a. The variance of the sharing fraction is

$$\mathrm{Var}(f) = \sigma_f^2, \tag{6.1155}$$

and the resulting pixel fluctuations were shown in Eqs. (6.797) and (6.800) to be perfectly anticorrelated.

Representing the two collected charges at positions $x = 0$ and $x = a$, the Fourier transform of one event is

$$\widehat{q}_f(R) = Q\left[f + (1-f)e^{-j2\pi Ra}\right]. \tag{6.1156}$$

Writing

$$f = \overline{f} + \delta f, \tag{6.1157}$$

the fluctuating part of the Fourier response is

$$\boxed{\delta\widehat{q}_f(R) = Q\delta f\left(1 - e^{-j2\pi Ra}\right).} \tag{6.1158}$$

Its mean squared magnitude is therefore

$$\mathbb{E}\left[|\delta\widehat{q}_f(R)|^2\right] = Q^2\sigma_f^2\left|1 - e^{-j2\pi Ra}\right|^2 \tag{6.1159}$$

$$= 4Q^2\sigma_f^2\sin^2(\pi Ra). \tag{6.1160}$$

Hence, for a Poisson incident fluence Φ_0, Eq. (6.1136) gives the charge-sharing contribution

$$\boxed{\mathrm{NPS}_{\mathrm{share}}(R) = 4\Phi_0 Q^2\sigma_f^2\sin^2(\pi Ra).} \tag{6.1161}$$

Equation (6.1161) is the frequency-domain counterpart of the negative covariance derived in Eq. (6.800). At zero frequency,

$$\mathrm{NPS}_{\mathrm{share}}(0) = 0, \tag{6.1162}$$

in agreement with Eq. (6.805): pure charge redistribution cannot fluctuate

the total collected charge.

At non-zero spatial frequency, however, redistribution between the two pixels generates noise power. In this simple model the spectral dependence is

$$\mathrm{NPS}_{\mathrm{share}}(R) \propto \sin^2(\pi R a), \tag{6.1163}$$

showing explicitly how anticorrelated neighbouring-pixel fluctuations suppress low-frequency noise while transferring fluctuation power toward higher spatial frequencies.

Localization noise in an ideal counting detector. Electron counting leads to a particularly instructive limiting case. For a successfully detected electron with localization error $\boldsymbol{\epsilon}_\omega$, the Fourier transform of the localized point event differs from the ideal point only by a phase factor, as derived in Sec. 6.6.4. Its magnitude is therefore

$$|\widehat{c}_\omega(\mathbf{R})|^2 = 1 \tag{6.1164}$$

for every event and every spatial frequency.

Equation (6.1134) consequently gives

$$\boxed{\mathrm{NPS}_{\mathrm{count}}(\mathbf{R}) = \Phi_0.} \tag{6.1165}$$

Thus independent localization errors do *not* by themselves make the count NPS non-white. Randomly displacing each point of a homogeneous Poisson process produces another homogeneous Poisson process.

Nevertheless, the useful signal transfer is reduced by $H_{\mathrm{loc}}(\mathbf{R})$ from Eq. (6.818). Using the single-electron decomposition,

$$\boxed{\begin{aligned}\Phi_0 = \Phi_0 |H_{\mathrm{loc}}(\mathbf{R})|^2 \\ + \Phi_0 \left[1 - |H_{\mathrm{loc}}(\mathbf{R})|^2\right].\end{aligned}} \tag{6.1166}$$

The first term is associated with the squared mean response, whereas the second is the positional fluctuation around that mean. Their sum remains white.

This result is important: localization uncertainty can reduce the MTF without increasing the total Poisson NPS. The resulting loss of signal relative to noise will appear naturally in the DQE.

6.8.5 Coincidence loss

Coincidence loss produces a qualitatively different NPS from the mechanisms considered above. It does not add an independent fluctuation to each electron event. Instead, the counting algorithm makes the probability of recording one event depend on whether another event occurs nearby during the same internal sampling interval.

At sufficiently low dose rate, independently localized electron counts form an approximately Poisson point process and therefore have the white NPS of Eq. (6.1165). Coincidence loss introduces a short-range exclusion: after one event has been recorded, the probability of recording another very nearby event is reduced.

This can be described using the pair-correlation function

$$g_2(\Delta\boldsymbol{\rho}). \tag{6.1167}$$

For an ideal Poisson process,

$$g_2(\Delta\boldsymbol{\rho}) = 1 \tag{6.1168}$$

away from zero separation. Coincidence loss produces

$$g_2(\Delta\boldsymbol{\rho}) < 1 \tag{6.1169}$$

over the spatial range in which two events interfere with one another.

NPS of a correlated counting process.

Let the recorded electron positions be

$$\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots \tag{6.1170}$$

and represent the recorded electron density as a sum of point events,

$$\boxed{\xi(\boldsymbol{\rho}) = \sum_i \delta^{(2)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_i\right).} \tag{6.1171}$$

Thus each recorded electron contributes one Dirac delta at its measured position.

Let

$$\Phi_{\text{out}} \tag{6.1172}$$

be the mean density of recorded electron events. For a stationary and

spatially uniform counting process,

$$\boxed{\mathbb{E}\left[\xi(\boldsymbol{\rho})\right] = \Phi_{\text{out}},} \tag{6.1173}$$

independent of $\boldsymbol{\rho}$.

The zero-mean fluctuation of the counted-electron density is therefore

$$\delta\xi(\boldsymbol{\rho}) = \xi(\boldsymbol{\rho}) - \Phi_{\text{out}}. \tag{6.1174}$$

Its spatial covariance is

$$\begin{aligned} C_{\text{count}}(\boldsymbol{\rho}, \boldsymbol{\rho}') &= \mathbb{E}\left[\delta\xi(\boldsymbol{\rho})\delta\xi(\boldsymbol{\rho}')\right] \\ &= \mathbb{E}\left[\xi(\boldsymbol{\rho})\xi(\boldsymbol{\rho}')\right] - \Phi_{\text{out}}^2. \end{aligned} \tag{6.1175}$$

We must therefore calculate the second moment

$$\mathbb{E}\left[\xi(\boldsymbol{\rho})\xi(\boldsymbol{\rho}')\right]. \tag{6.1176}$$

Using Eq. (6.1171),

$$\xi(\boldsymbol{\rho})\xi(\boldsymbol{\rho}') = \sum_i \sum_j \delta^{(2)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_i\right) \delta^{(2)}\left(\boldsymbol{\rho}' - \boldsymbol{\rho}_j\right). \tag{6.1177}$$

The double sum contains two physically different kinds of terms:

$$\boxed{\sum_i \sum_j = \sum_{i=j} + \sum_{i \neq j}.} \tag{6.1178}$$

The first sum corresponds to observing the *same recorded electron* at both positions. The second corresponds to pairs of *different recorded electrons*. We consider them separately.

Contribution of the same electron. For $i = j$, the same point event must contribute simultaneously at $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$. This is possible only when the two positions coincide. Consequently,

$$\boxed{\mathbb{E}\left[\sum_i \delta^{(2)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_i\right) \delta^{(2)}\left(\boldsymbol{\rho}' - \boldsymbol{\rho}_i\right)\right] = \Phi_{\text{out}} \delta^{(2)}\left(\boldsymbol{\rho}' - \boldsymbol{\rho}\right).} \tag{6.1179}$$

This is the familiar shot-noise term. The Dirac delta expresses the fact that one point electron is perfectly correlated with itself only at zero separation.

Contribution of two different electrons. We now need to describe how likely it is to observe two *distinct* recorded electrons separated by

$$\Delta\boldsymbol{\rho} = \boldsymbol{\rho}' - \boldsymbol{\rho}. \tag{6.1180}$$

For this purpose we introduce the pair-correlation function

$$g_2(\Delta\boldsymbol{\rho}). \tag{6.1181}$$

It is defined so that the probability of finding one recorded electron inside a small area dA_1 around $\boldsymbol{\rho}$ and a *different* recorded electron inside a small area dA_2 around $\boldsymbol{\rho} + \Delta\boldsymbol{\rho}$ is

$$\boxed{dP_{12} = \Phi_{\text{out}}^2 g_2(\Delta\boldsymbol{\rho})\, dA_1\, dA_2.} \tag{6.1182}$$

Thus

$$\boxed{\begin{aligned} g_2(\Delta\boldsymbol{\rho}) = 1 \quad &\Longrightarrow \quad \text{independent Poisson-like positions,} \\ g_2(\Delta\boldsymbol{\rho}) < 1 \quad &\Longrightarrow \quad \text{fewer pairs than expected from independence,} \\ g_2(\Delta\boldsymbol{\rho}) > 1 \quad &\Longrightarrow \quad \text{more pairs than expected from independence.} \end{aligned}} \tag{6.1183}$$

Equation (6.1182) implies that the contribution of distinct electron pairs to the second moment is

$$\boxed{\mathbb{E}\left[\sum_{i\neq j} \delta^{(2)}\left(\boldsymbol{\rho} - \boldsymbol{\rho}_i\right) \delta^{(2)}\left(\boldsymbol{\rho}' - \boldsymbol{\rho}_j\right)\right] = \Phi_{\text{out}}^2 g_2\left(\boldsymbol{\rho}' - \boldsymbol{\rho}\right).} \tag{6.1184}$$

Combining the two contributions. Adding the same-electron contribution in Eq. (6.1179) and the distinct-electron contribution in Eq. (6.1184) gives

$$\boxed{\begin{aligned} \mathbb{E}\left[\xi(\boldsymbol{\rho})\xi(\boldsymbol{\rho}')\right] = \Phi_{\text{out}}\delta^{(2)}\left(\boldsymbol{\rho}' - \boldsymbol{\rho}\right) \\ + \Phi_{\text{out}}^2 g_2\left(\boldsymbol{\rho}' - \boldsymbol{\rho}\right). \end{aligned}} \tag{6.1185}$$

The covariance is obtained by subtracting the product of the mean densities, as required by Eq. (6.1175). Therefore,

$$C_{\text{count}}(\boldsymbol{\rho}, \boldsymbol{\rho}') = \Phi_{\text{out}} \delta^{(2)} (\boldsymbol{\rho}' - \boldsymbol{\rho}) \tag{6.1186}$$

$$+ \Phi_{\text{out}}^2 g_2 (\boldsymbol{\rho}' - \boldsymbol{\rho}) - \Phi_{\text{out}}^2. \tag{6.1187}$$

Because the process is stationary, the covariance depends only on

$$\Delta\boldsymbol{\rho} = \boldsymbol{\rho}' - \boldsymbol{\rho}. \tag{6.1188}$$

Hence

$$\boxed{\begin{aligned} C_{\text{count}}(\Delta\boldsymbol{\rho}) = \Phi_{\text{out}} \delta^{(2)}(\Delta\boldsymbol{\rho}) \\ + \Phi_{\text{out}}^2 \left[g_2(\Delta\boldsymbol{\rho}) - 1 \right]. \end{aligned}} \tag{6.1189}$$

The two terms have different physical meanings. The first is the ordinary point-process shot noise: every electron is necessarily correlated with itself at zero separation. The second describes correlations *between different recorded electrons*. For an ideal Poisson process,

$$g_2(\Delta\boldsymbol{\rho}) = 1, \tag{6.1190}$$

so this second term vanishes and only the usual white shot-noise term remains.

For a counting detector affected by coincidence loss,

$$g_2(\Delta\boldsymbol{\rho}) < 1 \tag{6.1191}$$

at small separations because two nearby incident electrons are less likely to appear as two distinct recorded events. The second term in Eq. (6.1189) is then negative at short distances.

Taking the Fourier transform according to Eq. (6.1065) finally gives

$$\boxed{\begin{aligned} \text{NPS}_{\text{count}}(\mathbf{R}) = \Phi_{\text{out}} \\ + \Phi_{\text{out}}^2 \mathcal{F} \left\{ g_2 - 1 \right\} (\mathbf{R}). \end{aligned}} \tag{6.1192}$$

Coincidence loss makes $g_2 - 1$ negative at small separations. At zero spatial frequency,

$$\text{NPS}_{\text{count}}(\mathbf{0}) = \Phi_{\text{out}} + \Phi_{\text{out}}^2 \int \left[g_2(\boldsymbol{\rho}) - 1 \right] d^2\boldsymbol{\rho}. \tag{6.1193}$$

The integral is negative when the counted events exhibit short-range exclu-

sion, and therefore

$$\boxed{\mathrm{NPS}_{\mathrm{count}}(\mathbf{0}) < \Phi_{\mathrm{out}}.} \tag{6.1194}$$

At spatial frequencies much higher than the inverse of the coincidence distance, the Fourier transform of this finite-range correlation tends toward zero and the NPS approaches the Poisson level again.

This explains why coincidence loss produces a characteristic *depression of the low-spatial-frequency NPS* rather than a simple uniform change in noise level.

Experimental K2 spectrum. This behaviour was measured directly by Li et al. (2013) for the K2 Summit. The experimentally measured NPS at different dose rates was described empirically by

$$\boxed{N_{\mathrm{norm}}(k) = a - b\,\mathrm{sinc}^2(kd),} \tag{6.1195}$$

where k is spatial frequency in inverse pixels and the normalized sinc function is defined in Eq. (6.572).

Across the measured dose rates, the fitted high-frequency level was approximately

$$a = 1.005 \pm 0.005, \tag{6.1196}$$

while the characteristic coincidence diameter was

$$\boxed{d = 2.584 \pm 0.044 \quad \text{pixels.}} \tag{6.1197}$$

The coefficient b increased strongly with dose rate, reflecting the increasing coincidence loss (Li et al., 2013).

For example, at

$$D = 8.1623\ e^-\ \mathrm{pixel}^{-1}\,\mathrm{s}^{-1}, \tag{6.1198}$$

the fitted values were approximately

$$a = 1.0001, \qquad b = 0.2223, \qquad d = 2.5900 \text{ pixels.} \tag{6.1199}$$

Since $\mathrm{sinc}(0) = 1$, the fitted zero-frequency level is

$$\boxed{N_{\mathrm{norm}}(0) = a - b \simeq 0.778.} \tag{6.1200}$$

Thus the low-frequency noise power was suppressed by more than 20% relative to the nearly Poisson high-frequency level at this dose rate. This is consistent with the spatial exclusion interpretation of Eq. (6.1192). (Li et al., 2013)

The decrease of the NPS must not be interpreted as an improvement in detector performance. Coincidence loss also removes useful electron events, introduces nonlinearity, and changes the signal transfer. A reduced NPS caused by rejected counts is therefore not equivalent to a detector preserving the same signal with less noise. These effects must be considered together in the DQE.

6.8.6 Electronic readout and reset noise

Electronic noise is additive to the detector signal to a good approximation, but its spatial NPS depends on the readout architecture.

Consider first the simplest case in which every pixel receives an independent zero-mean electronic fluctuation with variance

$$\sigma_{\mathrm{el}}^2. \tag{6.1201}$$

The discrete spatial covariance is then

$$C_{rs}^{\mathrm{el}} = \sigma_{\mathrm{el}}^2 \delta_{r0} \delta_{s0}. \tag{6.1202}$$

Substitution into Eq. (6.1078) gives

$$\boxed{\mathrm{NPS}_{\mathrm{el}}(R_x, R_y) = a_x a_y \sigma_{\mathrm{el}}^2.} \tag{6.1203}$$

Independent pixel readout fluctuations provide the simplest electronic example of a white spatial NPS. If the electronic noise in different pixels is uncorrelated and has voltage variance σ_V^2, its covariance is

$$C_{rs}^{\mathrm{el}} = \sigma_V^2 \delta_{r0} \delta_{s0}, \tag{6.1204}$$

and Eq. (6.1078) gives

$$\boxed{\mathrm{NPS}_{\mathrm{el},V} = a_x a_y \, \sigma_V^2.} \tag{6.1205}$$

The actual value of σ_V^2 depends on which stage of the readout chain is being considered. In particular, reset noise and the noise introduced while measuring the reset and signal levels must be distinguished.

Reset noise before correlated double sampling. Immediately after resetting the sensing node, thermal equilibrium with the reset transistor leaves a random voltage offset. As derived in Eq. (6.873), its variance is

$$\boxed{\sigma_{V,\mathrm{reset}}^2 = \frac{k_{\mathrm{B}} T}{C_{\mathrm{node}}}.} \tag{6.1206}$$

If the reset fluctuations of different pixels are independent, the corresponding spatial NPS of the *uncorrected reset level* is therefore

$$\boxed{\mathrm{NPS}_{kTC,V} = a_x a_y \frac{k_\mathrm{B} T}{C_\mathrm{node}}.} \tag{6.1207}$$

Using the equivalent-electron representation derived in Eq. (6.877),

$$\sigma^2_{N,\mathrm{reset}} = \frac{k_\mathrm{B} T C_\mathrm{node}}{e^2}, \tag{6.1208}$$

the same spectrum expressed at the detector input is

$$\boxed{\mathrm{NPS}_{kTC,e^-} = a_x a_y \frac{k_\mathrm{B} T C_\mathrm{node}}{e^2}.} \tag{6.1209}$$

These equations describe the reset fluctuation *before* its cancellation by correlated double sampling.

Residual noise after correlated double sampling. As discussed in Sec. 6.6.6, correlated double sampling measures both the reset level and the subsequent signal level and subtracts them. If the same reset offset is present in both measurements, the kTC contribution cancels.

The two voltage measurements nevertheless contain their own electronic readout-noise terms. Let

$$n_\mathrm{r} \qquad \text{and} \qquad n_\mathrm{s} \tag{6.1210}$$

denote the noise in the reset and signal measurements, respectively, and assume for simplicity that

$$\mathrm{Var}(n_\mathrm{r}) = \mathrm{Var}(n_\mathrm{s}) = \sigma^2_\mathrm{ro}. \tag{6.1211}$$

Here σ_ro is the rms voltage noise of one individual readout measurement.

Let their correlation coefficient be

$$\boxed{\rho_\mathrm{ro} = \frac{\mathrm{Cov}(n_\mathrm{r}, n_\mathrm{s})}{\sigma^2_\mathrm{ro}}.} \tag{6.1212}$$

The residual CDS noise is

$$n_\mathrm{CDS} = n_\mathrm{r} - n_\mathrm{s}, \tag{6.1213}$$

and therefore, using Eq. (6.899),

$$\boxed{\sigma_{\mathrm{CDS}}^2 = 2\sigma_{\mathrm{ro}}^2 \left(1 - \rho_{\mathrm{ro}}\right).} \tag{6.1214}$$

The meaning of this result is important. Slowly varying electronic fluctuations that are nearly identical in the two measurements have $\rho_{\mathrm{ro}} \simeq 1$ and are strongly suppressed by the subtraction. Independent fluctuations have $\rho_{\mathrm{ro}} = 0$, in which case

$$\sigma_{\mathrm{CDS}} = \sqrt{2}\,\sigma_{\mathrm{ro}}, \tag{6.1215}$$

so CDS actually combines the noise of the two independent measurements.

If the residual CDS fluctuations are also independent from pixel to pixel, their spatial NPS is white:

$$\boxed{\mathrm{NPS}_{\mathrm{CDS},V} = a_x a_y \sigma_{\mathrm{CDS}}^2 = 2a_x a_y \sigma_{\mathrm{ro}}^2 \left(1 - \rho_{\mathrm{ro}}\right).} \tag{6.1216}$$

Expressed in equivalent collected electrons, using the charge-to-voltage conversion at the sensing node, this becomes

$$\boxed{\mathrm{NPS}_{\mathrm{CDS},e^-} = a_x a_y \left(\frac{C_{\mathrm{node}}\sigma_{\mathrm{CDS}}}{e}\right)^2.} \tag{6.1217}$$

The white spectra above apply only when the residual electronic fluctuations are independent among pixels. Real detector electronics can introduce spatial correlations because pixels may share column amplifiers, row-selection circuitry, reference voltages, ADCs, or output channels. In that case the electronic NPS is no longer spatially white.

Real readout noise need not be independent among pixels. Pixels may share column amplifiers, row-selection circuitry, reference voltages, ADCs, or output channels. Fluctuations common to one column produce vertical stripe patterns and spectral power concentrated along the corresponding Fourier axis; row-correlated fluctuations generate the orthogonal structure. Shared global fluctuations can concentrate power very close to zero spatial frequency.

This also explains why a temporal electronic spectrum cannot be identified directly with the spatial NPS. A transistor may exhibit a temporal $1/f$ spectrum, but the spatial pattern seen in the recorded image depends on the sequence and architecture by which that temporal fluctuation is mapped onto rows, columns, and frames (Boukhayma et al., 2016; Konnik and Welsh, 2014).

6.8.7 Dark-current fluctuations

The mean number of thermally generated electrons accumulated by one pixel during an integration interval is, from Eq. (6.924),

$$\mu_{\mathrm{d}} = r_{\mathrm{d}} T_{\mathrm{int}}. \tag{6.1218}$$

Under the independent Poisson model of Sec. 6.6.7,

$$\mathrm{Var}(N_{\mathrm{d}}) = \mu_{\mathrm{d}}. \tag{6.1219}$$

If different pixels generate independent dark carriers, the spatial covariance is

$$C^{\mathrm{d}}_{rs} = \mu_{\mathrm{d}} \delta_{r0} \delta_{s0}, \tag{6.1220}$$

and therefore

$$\boxed{\mathrm{NPS}_{\mathrm{d}} = a_x a_y \mu_{\mathrm{d}} = a_x a_y r_{\mathrm{d}} T_{\mathrm{int}}.} \tag{6.1221}$$

Thus ideal dark-current shot noise is spatially white and its NPS grows linearly with integration time.

If the image is corrected using the average of M independent dark frames, Eq. (6.940) showed that the residual variance is multiplied by

$$1 + \frac{1}{M}. \tag{6.1222}$$

The same factor applies frequency by frequency when the dark realizations have the same NPS:

$$\boxed{\mathrm{NPS}_{\mathrm{d,corr}} = \left(1 + \frac{1}{M}\right) \mathrm{NPS}_{\mathrm{d}}\,.} \tag{6.1223}$$

Pixel-to-pixel differences in the *mean* dark rate $r_{\mathrm{d},p}$ are not part of this white stochastic term. If stable, they constitute a dark fixed pattern; only their residual calibration error or temporal drift contributes to a stochastic NPS.

6.8.8 Gain variations and fixed-pattern structure

A perfectly stable gain or offset map requires special treatment. As discussed in Sec. 6.6.8, a reproducible spatial pattern belongs to the detector mean response rather than to the frame-to-frame stochastic covariance. Its Fourier transform has a power spectrum, but that spectrum should not automatically be called an NPS.

A stochastic NPS arises when the gain correction is uncertain, drifts with time, or when different detector regions are treated as members of an ensemble.

For a uniform mean signal $\overline{I}$, let ε_p denote a small residual relative gain error. To first order, Eq. (6.992) gives a residual fluctuation of the form

$$n_{p,\text{gain}} \simeq \overline{I}\,\varepsilon_p, \tag{6.1224}$$

apart from an irrelevant sign convention.

If

$$C_\varepsilon(r,s) = \mathbb{E}[\varepsilon_{mn}\varepsilon_{m+r,n+s}] \tag{6.1225}$$

is the spatial covariance of the residual gain map, define its spatial power spectral density by

$$S_\varepsilon(\mathbf{R}) = a_x a_y \sum_{r,s} C_\varepsilon(r,s) e^{-j2\pi(R_x r a_x + R_y s a_y)}. \tag{6.1226}$$

Then

$$\boxed{\text{NPS}_{\text{gain}}(\mathbf{R}) = \overline{I}^{\,2} S_\varepsilon(\mathbf{R}).} \tag{6.1227}$$

This result is the frequency-domain counterpart of Eq. (6.994). In particular, residual gain noise scales as the *square* of the signal level.

If the residual gain errors are independent among pixels with variance σ_g^2,

$$\boxed{\text{NPS}_{\text{gain}} = a_x a_y \overline{I}^{\,2} \sigma_g^2,} \tag{6.1228}$$

which is white. Real gain maps, however, often contain spatial organization associated with columns, readout blocks, or fabrication patterns, in which case their residual spectrum is not white.

The different dose dependence is useful experimentally:

$$\boxed{\begin{array}{lcl} \text{Poisson shot-noise variance} & \propto & \overline{I}, \\ \text{residual gain-pattern variance} & \propto & \overline{I}^{\,2}. \end{array}} \tag{6.1229}$$

6.8.9 Analog-to-digital quantization

Under the uniform quantization-error approximation, the ADC error has variance

$$\text{Var}(\epsilon_Q) = \frac{\Delta V_{\text{ADC}}^2}{12}, \tag{6.1230}$$

as derived in Eq. (6.1013).

If the quantization errors of different pixels are effectively independent, their spatial covariance is diagonal and the NPS is white:

$$\boxed{\mathrm{NPS}_{Q,V} = a_x a_y \frac{\Delta V_{\mathrm{ADC}}^2}{12}.} \tag{6.1231}$$

When referred to the sensing-node input and expressed in equivalent electrons, the rms quantization error is $\sigma_{N,Q}$, so

$$\boxed{\mathrm{NPS}_{Q,e^-} = a_x a_y \sigma_{N,Q}^2.} \tag{6.1232}$$

Using the input-referred result derived previously,

$$\boxed{\mathrm{NPS}_{Q,e^-} = a_x a_y \frac{C_{\mathrm{node}}^2 V_{\mathrm{FS}}^2}{12 e^2 A_{\mathrm{el}}^2 2^{2b}}.} \tag{6.1233}$$

The white spectrum in Eq. (6.1231) is not universal. It relies on the same assumptions used to obtain the uniform quantization-error model: the analog input must explore the quantization intervals sufficiently that the error behaves approximately as an independent random variable. If the signal occupies only a few digital levels, the quantization error can remain correlated with the image and can produce structured spectral components rather than white noise.

6.8.10 Combination of noise sources and the origin of the background PSD

The preceding subsections considered the different physical sources of detector noise separately. We can now ask how they combine in a measured image and how this description connects with the empirical background of the image power spectrum introduced in Sec. 4.16.

This requires some care. Not every physical fluctuation discussed above is an independent additive noise source. Some mechanisms are different parts of the response to the *same* incident electron, some modify the statistics of the electron point process itself, and some are not stochastic at all.

The mathematical distinction is important because only statistically uncorrelated additive noise sources have NPS values that can simply be added frequency by frequency.

The general rule: auto-spectra and cross-spectra. Suppose first that the total zero-mean noise can be written as a sum of M random fields,

$$\boxed{n(\boldsymbol{\rho}) = \sum_{\alpha=1}^{M} n_\alpha(\boldsymbol{\rho}).} \tag{6.1234}$$

The covariance is then

$$C_n(\Delta\boldsymbol{\rho}) = \mathbb{E}\left[n(\boldsymbol{\rho})n(\boldsymbol{\rho}+\Delta\boldsymbol{\rho})\right] \tag{6.1235}$$

$$= \sum_\alpha C_{\alpha\alpha}(\Delta\boldsymbol{\rho}) + \sum_{\alpha\neq\beta} C_{\alpha\beta}(\Delta\boldsymbol{\rho}), \tag{6.1236}$$

where

$$\boxed{C_{\alpha\beta}(\Delta\boldsymbol{\rho}) = \mathbb{E}\left[n_\alpha(\boldsymbol{\rho})n_\beta(\boldsymbol{\rho}+\Delta\boldsymbol{\rho})\right]} \tag{6.1237}$$

is the cross-covariance between noise sources α and β.

Taking the Fourier transform according to Eq. (6.1065), define the corresponding cross-spectral density

$$\boxed{S_{\alpha\beta}(\mathbf{R}) = \mathcal{F}\left\{C_{\alpha\beta}\right\}(\mathbf{R}).} \tag{6.1238}$$

The total NPS is therefore

$$\boxed{\begin{aligned}\mathrm{NPS}_{\mathrm{tot}}(\mathbf{R}) = &\sum_\alpha \mathrm{NPS}_\alpha(\mathbf{R}) \\ &+ 2\,\mathrm{Re}\sum_{\alpha<\beta} S_{\alpha\beta}(\mathbf{R}).\end{aligned}} \tag{6.1239}$$

Equation (6.1239) is the general addition rule. Only when

$$C_{\alpha\beta}(\Delta\boldsymbol{\rho}) = 0, \qquad \alpha \neq \beta, \tag{6.1240}$$

for every spatial displacement, do the cross-spectra vanish and the familiar additive result follow:

$$\boxed{\mathrm{NPS}_{\mathrm{tot}}(\mathbf{R}) = \sum_\alpha \mathrm{NPS}_\alpha(\mathbf{R}).} \tag{6.1241}$$

Statistical independence is sufficient for Eq. (6.1241), although the weaker condition of vanishing cross-covariance is enough.

The electron-induced term must first be treated as one compound process. The fluctuations generated by the incident imaging electrons should not initially be decomposed into separate additive terms. For an integrating detector, one incident electron produces the random spatial response $q_\omega(\boldsymbol{\rho})$ introduced in Sec. 6.6.2. For a homogeneous Poisson fluence Φ_0, we derived

$$\boxed{\mathrm{NPS}_{\mathrm{electron}}(\mathbf{R}) = \Phi_0 \mathbb{E}\left[\left|\widehat{q}_\omega(\mathbf{R})\right|^2\right].} \tag{6.1242}$$

This single expression already contains several physical mechanisms:

$$\begin{gathered}\text{incident-electron shot noise,}\\ \text{energy-deposition fluctuations,}\\ \text{electron–hole-pair fluctuations,}\\ \text{primary- and secondary-electron track variations,}\\ \text{carrier diffusion,}\\ \text{charge-sharing fluctuations,}\\ \text{event-width and event-shape fluctuations.}\end{gathered} \tag{6.1243}$$

They are all properties of the response to the *same primary electron*. Adding their separately calculated NPS values to Eq. (6.1242) would therefore in general double count the same stochastic event.

The decomposition into the mean single-electron response and its fluctuation, Eq. (6.730), does provide one exact separation:

$$\widehat{q}_\omega = \overline{\widehat{q}} + \delta\widehat{q}_\omega. \tag{6.1244}$$

Because

$$\mathbb{E}[\delta\widehat{q}_\omega] = 0, \tag{6.1245}$$

we obtained

$$\boxed{\begin{aligned}\mathrm{NPS}_{\mathrm{electron}}(\mathbf{R}) = \quad & \underbrace{\Phi_0 \left|\overline{\widehat{q}}(\mathbf{R})\right|^2}_{\text{propagated incident shot noise}} \\ & + \underbrace{\Phi_0 \mathbb{E}\left[\left|\delta\widehat{q}_\omega(\mathbf{R})\right|^2\right]}_{\text{single-electron-response fluctuations}}.\end{aligned}} \tag{6.1246}$$

This decomposition is exact because the cross term between the mean and the zero-mean fluctuation vanishes.

Further decomposition of the single-electron fluctuation. It may sometimes be useful to attribute the fluctuation δq_ω to several microscopic mechanisms. Schematically, one might attempt to write

$$\delta\widehat{q}_\omega = \delta\widehat{q}_E + \delta\widehat{q}_{\mathrm{Fano}} + \delta\widehat{q}_{\mathrm{shape}} + \delta\widehat{q}_{\mathrm{share}} + \cdots . \tag{6.1247}$$

The squared magnitude is then

$$\mathbb{E}\left[|\delta\widehat{q}_\omega|^2\right] = \sum_\alpha \mathbb{E}\left[|\delta\widehat{q}_\alpha|^2\right] \tag{6.1248}$$

$$+ 2\,\mathrm{Re}\sum_{\alpha<\beta} \mathbb{E}\left[\delta\widehat{q}_\alpha \delta\widehat{q}_\beta^*\right]. \tag{6.1249}$$

Thus even within one primary-electron event the individual contributions are additive only if their cross terms vanish.

There is no reason for this to hold in general. For example, an event that deposits an unusually large amount of energy may also generate a different secondary-electron cascade and consequently a different spatial width. Event amplitude and event shape can therefore be correlated. Similarly, the depth at which energy is deposited affects the amount of carrier diffusion before collection and hence the charge-sharing pattern.

For this reason, Eq. (6.1242) is the safest general description. Separate energy-deposition, Fano, diffusion, and charge-sharing spectra are useful when a model provides the statistical independences required to justify such a separation.

Which detector contributions can be added independently? After the electron-induced signal has been generated, several additional fluctuations can, to a good approximation, be treated as additive and independent of it.

For example, under the assumptions developed in the corresponding subsections, independent post-CDS electronic noise, dark-current shot noise, and approximately independent ADC quantization noise may be written schematically as

$$n_{\mathrm{post}} = n_{\mathrm{el}} + n_{\mathrm{d}} + n_Q. \tag{6.1250}$$

If these processes are mutually uncorrelated and are also independent of the electron-induced signal,

$$\boxed{\begin{aligned}\mathrm{NPS}_{\mathrm{linear}}(\mathbf{R}) = {} & \mathrm{NPS}_{\mathrm{electron}}(\mathbf{R}) \\ & + \mathrm{NPS}_{\mathrm{el}}(\mathbf{R}) + \mathrm{NPS}_{\mathrm{d}}(\mathbf{R}) + \mathrm{NPS}_Q(\mathbf{R}).\end{aligned}} \tag{6.1251}$$

The assumptions behind Eq. (6.1251) should remain explicit.

Ideal dark-current shot noise is independent among pixels and therefore white, as shown in Eq. (6.1221). Independent post-CDS readout noise is also white according to Eq. (6.1216). ADC quantization can be represented as an additive white contribution only under the uniform-error approximation; for coarse or signal-correlated quantization this approximation fails.

Gain errors are fundamentally multiplicative. Gain non-uniformity does not naturally belong in Eq. (6.1251), because it multiplies the signal rather than being added to it.

If the residual gain error after flat-field correction is $\varepsilon_g(\boldsymbol{\rho})$, then to first order

$$I_{\mathrm{corr}}(\boldsymbol{\rho}) \simeq I(\boldsymbol{\rho})\left[1+\varepsilon_g(\boldsymbol{\rho})\right]. \tag{6.1252}$$

For a uniform flat field,

$$I(\boldsymbol{\rho}) \simeq \overline{I}, \tag{6.1253}$$

and the gain error behaves approximately like an additive random field,

$$n_g(\boldsymbol{\rho}) \simeq \overline{I}\varepsilon_g(\boldsymbol{\rho}). \tag{6.1254}$$

Under these restricted conditions, Eq. (6.1227) gives

$$\mathrm{NPS}_g(\mathbf{R}) = \overline{I}^{\,2} S_{\varepsilon_g}(\mathbf{R}). \tag{6.1255}$$

For a non-uniform specimen image, however, multiplication in real space becomes convolution in Fourier space. Gain errors then mix spatial frequencies of the specimen and of the calibration map, and cannot in general be represented as a simple independent additive NPS.

A perfectly stable gain or offset pattern is even more distinct: it is not a stochastic noise term at all. It has a Fourier power spectrum, but it belongs formally to the deterministic detector response. Only calibration uncertainty, temporal drift, or an ensemble treatment of the fixed pattern gives rise to an NPS.

Coincidence loss changes the point process itself. Coincidence loss is not an additive noise source either. It changes which electron events are recorded and introduces correlations between their positions.

For an ideal counting detector at sufficiently low dose rate, the recorded events form approximately a Poisson point process and their NPS is white. With coincidence loss, nearby events are less likely to be recorded as two

independent counts. The correct framework is therefore the pair-correlation description derived in Sec. 6.8.5,

$$\boxed{\begin{aligned}\mathrm{NPS}_{\mathrm{count}}(\mathbf{R}) = \Phi_{\mathrm{out}} \\ + \Phi_{\mathrm{out}}^2 \mathcal{F}\left\{g_2 - 1\right\}(\mathbf{R}).\end{aligned}} \tag{6.1256}$$

There is no physically independent quantity "$\mathrm{NPS}_{\mathrm{coincidence}}$" that should simply be added to a Poisson NPS. Coincidence changes the statistics of the recorded electron population itself.

The same reasoning applies more generally to missed events, merged events, split events, and threshold-dependent detection probabilities: when a processing step changes which point events exist in the output, the resulting point process must be modeled directly rather than by adding an independent noise field.

Spatial correlation determines the spectral shape. The discussion above also explains why experimentally measured detector backgrounds are generally not flat.

White noise is obtained only when the covariance is localized entirely at zero separation,

$$C(\Delta\boldsymbol{\rho}) \propto \delta^{(2)}(\Delta\boldsymbol{\rho}). \tag{6.1257}$$

Whenever one fluctuation affects several spatial positions, the covariance extends away from the origin and the NPS becomes frequency-dependent.

The sign and spatial scale of that covariance determine the shape of the spectrum.

A particularly important case is a *positive* short-range correlation. Suppose, for illustration, that

$$\boxed{C_c(\Delta\boldsymbol{\rho}) = \sigma_c^2 \exp\left[-\frac{\|\Delta\boldsymbol{\rho}\|^2}{2\ell_c^2}\right],} \tag{6.1258}$$

where ℓ_c is the correlation length. Its two-dimensional Fourier transform is

$$\boxed{\mathrm{NPS}_c(\mathbf{R}) = 2\pi\sigma_c^2\ell_c^2 \exp\left(-2\pi^2\ell_c^2 R^2\right).} \tag{6.1259}$$

The NPS is largest at low spatial frequency and decreases as R increases. A larger real-space correlation length produces a narrower spectrum concentrated still more strongly at low frequency.

This provides the fundamental connection

$$\boxed{\begin{array}{c}\text{positive correlation over neighbouring pixels}\\ \Downarrow\\ \text{smooth fluctuations in real space}\\ \Downarrow\\ \text{power concentrated at low spatial frequency}\\ \Downarrow\\ \text{a decaying NPS.}\end{array}} \tag{6.1260}$$

The Gaussian model in Eqs. (6.1258) and (6.1259) is only an illustration. An exponential, a mixture of correlation lengths, or an anisotropic covariance produces a different mathematical decay. The general statement is not that the NPS must be Gaussian, but that spatially extended positive correlations preferentially place power at low spatial frequencies.

Which of the mechanisms discussed above are spatially correlated? Several of the physical mechanisms developed in this chapter produce such correlations.

The first and most fundamental is the spatial response of the detector to one primary electron. Primary- and secondary-electron transport, carrier diffusion, and finite pixel integration spread one event over a neighbourhood rather than confining it to one mathematical point. Consequently, even an initially white Poisson electron process becomes spatially correlated after deterministic detector spreading. As derived in Sec. 6.8.1,

$$\mathrm{NPS}_{\mathrm{shot,out}}(\mathbf{R}) = \Phi_0 \overline{Q}^2 |H_{\mathrm{det}}(\mathbf{R})|^2. \tag{6.1261}$$

Since the detector MTF normally decreases toward high spatial frequency, this contribution naturally produces a decreasing NPS.

Event-to-event fluctuations of the single-electron spatial response also correlate neighbouring pixels. One primary electron generates an entire charge cloud, so the random values measured in nearby pixels do not fluctuate independently. These correlations are contained in

$$\mathbb{E}\left[|\delta\widehat{q}_\omega(\mathbf{R})|^2\right] \tag{6.1262}$$

of Eq. (6.1246).

Electronic readout can generate correlations over still larger spatial scales. Pixels sharing column amplifiers, row circuitry, reference voltages, ADCs,

or readout blocks may fluctuate together. Such correlations can generate enhanced low-frequency power, stripes, and anisotropic structures in the two-dimensional NPS. This is one reason why the detector background need not be radially symmetric.

Residual gain and offset calibration errors can likewise be spatially correlated because neighbouring pixels may share fabrication, electronic, or calibration properties. If the residual map varies smoothly over the detector, its PSD is concentrated preferentially at low spatial frequency.

Not every correlation produces a decaying spectrum. It is important not to replace the statement above by the incorrect rule

$$\text{correlated noise} \Longrightarrow \text{decaying NPS}. \tag{6.1263}$$

The *sign* of the correlation matters.

Pure charge redistribution provides the clearest counterexample. When the total event charge is fixed, an excess of charge collected by one pixel must be balanced by a deficit elsewhere. The resulting neighbouring-pixel fluctuations can therefore be negatively correlated. For the two-pixel model derived previously,

$$\mathrm{NPS}_{\mathrm{share}}(R) \propto \sin^2(\pi R a), \tag{6.1264}$$

so

$$\mathrm{NPS}_{\mathrm{share}}(0) = 0. \tag{6.1265}$$

This contribution is suppressed at low frequency and grows toward higher frequency rather than decaying from the origin.

Coincidence loss provides a second important example. It produces short-range exclusion between recorded electron events,

$$g_2(\Delta\boldsymbol{\rho}) < 1 \tag{6.1266}$$

at small separations. This is a negative correlation relative to a Poisson process and therefore depresses the low-frequency NPS, as shown in Sec. 6.8.5.

Ideal localization errors in counting mode provide yet another useful contrast. They reduce the MTF, but independently displacing the points of a homogeneous Poisson process leaves the count process Poisson. Consequently, the NPS can remain white even though the signal transfer decays strongly with frequency.

The spectral signatures can therefore be summarized conceptually as

$$\boxed{\begin{aligned} \text{independent fluctuations} &\Longrightarrow \text{white NPS}, \\ \text{positive short-range correlations} &\Longrightarrow \text{enhanced low-frequency power}, \\ \text{negative short-range correlations} &\Longrightarrow \text{suppressed low-frequency power}, \\ \text{anisotropic correlations} &\Longrightarrow \text{anisotropic NPS}. \end{aligned}} \tag{6.1267}$$

Connection with the background of the cryo-EM image power spectrum. We can now return to the background $B(\mathbf{R})$ introduced in Sec. 4.16. There, the observed image power spectrum was written in the practical form

$$S_{\text{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})|\,\text{CTF}(\mathbf{R})|^2, \tag{6.1268}$$

as given in Eq. (4.2162). The background was deliberately defined operationally in Eq. (4.2165) as the power not explained by the chosen CTF-modulated structural model, rather than as the PSD of one specific microscopic process.

The analysis developed in the present chapter provides a microscopic interpretation of two important contributions to this background: the fluctuations associated with the finite number of imaging electrons and the additional fluctuations introduced by the detector. For bookkeeping purposes, we may therefore write schematically

$$\boxed{B(\mathbf{R}) \simeq B_{\text{shot}}(\mathbf{R}) + B_{\text{det}}(\mathbf{R}) + B_{\text{other}}(\mathbf{R}).} \tag{6.1269}$$

This decomposition is not intended to imply that these terms can be uniquely separated from an experimental power spectrum. Rather, it identifies the contributions for which the present detector model provides an explicit statistical description.

The term B_{shot} originates from the finite number of incident electrons. As shown in Sec. 6.8.1, the incident Poisson process is spatially white, but its contribution at the detector output need not be white because a primary electron generally produces a spatially extended response. The term B_{det} includes additional detector-related fluctuations, such as variations of the single-electron response, electronic readout noise, dark-current noise, residual gain and offset errors, and quantization effects. Coincidence loss also modifies the recorded background, but, as discussed in Sec. 6.8.5, it must not be regarded as an independent additive NPS because it changes the

statistics of the recorded point process itself. The remaining term B_{other} represents whatever power is not described by these detector contributions or by the chosen CTF-modulated structural model.

This statistical description also explains why $B(\mathbf{R})$ need not be flat. Spatially independent fluctuations produce approximately white spectra, whereas spatial correlations generate frequency-dependent spectra. It is therefore useful to write, at a different level of description,

$$\boxed{B(\mathbf{R}) = B_{\text{white}} + B_{\text{colored}}(\mathbf{R}),} \tag{6.1270}$$

where B_{white} collects approximately uncorrelated contributions and B_{colored} represents the effect of spatial correlations and other frequency-dependent processes.

Positive correlations extending over finite spatial distances concentrate fluctuation power toward low spatial frequencies and can therefore contribute to the smooth decay commonly observed beneath the Thon rings. This provides a physical justification for allowing the empirical background introduced in Sec. 4.16 to vary smoothly with spatial frequency rather than treating it as a constant.

The precise spectral shape, however, is not universal. Negative short-range correlations can instead suppress low-frequency power: charge redistribution provides one example, while coincidence loss, Sec. 6.8.5, provides another. Thus the measured background $B(\mathbf{R})$ should be understood as the combined spectral effect of several physical processes, rather than as the signature of a single noise mechanism.

The role of sampling and aliasing. All of the spectra described above are ultimately observed on a sampled pixel array. Noise generated before pixel integration is modified by the pixel aperture, and spectral power outside the Nyquist region is folded back by sampling.

The same replica structure derived for the image spectrum in Eq. (6.607) therefore applies to noise power. Schematically, if $\text{NPS}_{\text{pre}}(\mathbf{R})$ denotes the continuous presampling NPS, the sampled spectrum contains contributions of the form

$$\boxed{\text{NPS}_{\text{samp}}(\mathbf{R}) \propto \sum_{\mathbf{m}} \text{NPS}_{\text{pre}}\left(\mathbf{R} - \mathbf{R}_{\mathbf{m}}\right),} \tag{6.1271}$$

with the appropriate pixel-aperture factors and normalization.

Aliasing can therefore modify the apparent high-frequency shape of the measured background. In particular, a strongly decreasing presampling NPS

need not continue to decrease in the same way close to Nyquist, because higher-frequency noise power can be folded back into the measured band.

Final hierarchy. The complete statistical picture can now be organized into three different mathematical levels:

$$\boxed{\begin{aligned}
\text{independent additive sources} &\Longrightarrow \text{add their NPS values,}\\
\text{coupled fluctuations within one electron event} &\Longrightarrow \Phi_0 \mathbb{E}\left[|\widehat{q}_\omega|^2\right],\\
\text{nonlinear detection and coincidence} &\Longrightarrow \text{model the point process through } g_2,\\
\text{multiplicative calibration errors} &\Longrightarrow \text{model signal–gain coupling,}\\
\text{stable fixed patterns} &\Longrightarrow \text{deterministic power, not NPS.}
\end{aligned}} \tag{6.1272}$$

This hierarchy prevents the same physical fluctuation from being counted twice and clarifies the physical origin of the smooth background used in the interpretation and fitting of experimental cryo-EM power spectra.

The preceding analysis has established an important distinction that will be central to the definition of detector quantum efficiency. If every incident electron produced the same deterministic spatial response, white Poisson shot noise would be shaped by the squared detector transfer function,

$$\mathrm{NPS}_{\mathrm{electron}}(\mathbf{R}) = \Phi_0 \overline{Q}^2 |H_{\mathrm{det}}(\mathbf{R})|^2. \tag{6.1273}$$

For a real detector, however, the single-electron response fluctuates, and Eq. (6.746) shows that the NPS depends on the second moment

$$\mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right], \tag{6.1274}$$

whereas the MTF depends on

$$|\mathbb{E}[\widehat{q}_\omega(\mathbf{R})]|\,. \tag{6.1275}$$

This distinction between coherent signal transfer and total noise power is the key statistical ingredient of the detector quantum efficiency derived next.

6.9 Detective quantum efficiency

The preceding analysis has established an important distinction that is central to the quantitative characterization of a detector. If every incident electron produced exactly the same deterministic spatial response, the white

Poisson fluctuations of the incident electron field would be shaped by the same detector transfer function as the signal. For a real detector, however, the response to an individual electron fluctuates, and Eq. (6.746) showed that its noise depends on

$$\mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right], \tag{6.1276}$$

whereas the mean signal transfer depends on

$$|\mathbb{E}\left[\widehat{q}_\omega(\mathbf{R})\right]|^2. \tag{6.1277}$$

This distinction was summarized explicitly in Eq. (6.749).

We are therefore ready to introduce a quantity that measures not merely how much signal modulation survives the detector, but how much of the *input signal-to-noise ratio* survives. This quantity is the *detective quantum efficiency*, or DQE.

Why quantum efficiency alone is insufficient A first measure of detector performance is its *quantum efficiency*. For an electron-counting detector operating at sufficiently low occupancy, let

$$\eta = P\left(\text{an incident isolated electron is successfully recorded}\right). \tag{6.1278}$$

Thus

$$0 \le \eta \le 1. \tag{6.1279}$$

Quantum efficiency answers an important question:

$$\boxed{\eta \quad \Longrightarrow \quad \text{what fraction of the incident electrons are recorded?}} \tag{6.1280}$$

It does not, however, describe what happens to the information carried by the recorded electrons.

For example, two detectors could have the same quantum efficiency while having very different localization precision. Both might record nearly every incident electron, but one might assign their positions much more accurately than the other. Their high-frequency MTF values, and hence their ability to preserve fine spatial information, would then be very different.

Similarly, an integrating detector may collect signal from essentially every incident electron while retaining the event-to-event fluctuations of deposited energy and generated charge derived in Sec. 6.6.2. Those fluctuations increase the output noise without necessarily changing the fraction of electrons that produce a measurable signal.

Electronic readout noise, dark-current noise, residual gain errors, and other detector fluctuations can also decrease image quality without changing the fraction of primary electrons that interact with the sensor.

Thus quantum efficiency alone does not answer the question that matters most in a dose-limited experiment:

$$\boxed{\text{how much of the signal-to-noise ratio carried by the incident electron field survives the complete detection process?}} \tag{6.1281}$$

This is the question answered by the detective quantum efficiency.

Definition from input and output SNR The most fundamental definition of DQE is expressed directly in terms of signal-to-noise ratio.

A notation issue must first be made explicit. In much of the detector literature, SNR denotes the ratio of signal amplitude to rms noise, and the DQE is consequently written as the ratio of the *squares* of the input and output SNR values (McMullan et al., 2009a). In this book, however, we have consistently defined SNR as

$$\mathrm{SNR} = \frac{\text{signal power}}{\text{noise power}}, \tag{6.1282}$$

as used, for example, in the derivation leading to Eq. (6.716).

With this convention, the detective quantum efficiency is

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{\mathrm{SNR}_{\mathrm{out}}(\mathbf{R})}{\mathrm{SNR}_{\mathrm{in}}(\mathbf{R})}.} \tag{6.1283}$$

Equation (6.1283) is equivalent to the more common detector-literature expression

$$\mathrm{DQE} = \frac{\mathrm{SNR}^2_{\mathrm{out,amp}}}{\mathrm{SNR}^2_{\mathrm{in,amp}}}, \tag{6.1284}$$

because

$$\mathrm{SNR}_{\mathrm{power}} = \mathrm{SNR}^2_{\mathrm{amp}}\,. \tag{6.1285}$$

The interpretation of Eq. (6.1283) is immediate:

$$\boxed{\mathrm{SNR}_{\mathrm{out}}(\mathbf{R}) = \mathrm{DQE}(\mathbf{R})\,\mathrm{SNR}_{\mathrm{in}}(\mathbf{R}).} \tag{6.1286}$$

The same concept appears in communications engineering in the inverse form. There, the degradation of signal-to-noise ratio produced by a device is characterized by its *noise factor*,

$$\boxed{F(\mathbf{R}) = \frac{\mathrm{SNR}_{\mathrm{in}}(\mathbf{R})}{\mathrm{SNR}_{\mathrm{out}}(\mathbf{R})}.} \tag{6.1287}$$

Comparison with Eq. (6.1283) shows that

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{1}{F(\mathbf{R})}.} \tag{6.1288}$$

Thus the noise factor measures the *degradation* of SNR, whereas the DQE measures the fraction of SNR that is *preserved*. The corresponding noise figure, commonly used in communications engineering, is

$$\mathrm{NF}(\mathbf{R}) = 10 \log_{10} F(\mathbf{R}) = -10 \log_{10} \mathrm{DQE}(\mathbf{R}) \quad \mathrm{dB}. \tag{6.1289}$$

An ideal detector has

$$\mathrm{DQE}(\mathbf{R}) = 1, \tag{6.1290}$$

meaning that it preserves all of the input SNR at that spatial frequency.

For a Poisson-limited experiment, input SNR is proportional to electron fluence. Consequently, a detector with

$$\mathrm{DQE}(\mathbf{R}) = 0.5 \tag{6.1291}$$

produces, at that frequency, the same SNR that an ideal detector would produce from half as many incident electrons.

This interpretation is particularly important in cryo-EM, where the electron fluence cannot simply be increased indefinitely because of radiation damage. DQE therefore measures the efficiency with which a limited electron dose is converted into useful image information.

Frequency-dependent DQE and its relation to MTF and NPS To connect the definition in Eq. (6.1283) with the MTF and NPS, consider a weak deterministic input modulation at spatial frequency $\mathbf{R}$. Let its Fourier amplitude be

$$S_{\mathrm{in}}(\mathbf{R}), \tag{6.1292}$$

and let

$$\mathrm{NPS}_{\mathrm{in}}(\mathbf{R}) \tag{6.1293}$$

be the input noise power spectrum.

For a linear shift-invariant detector, let

$$G(\mathbf{R}) \tag{6.1294}$$

denote the mean detector transfer function, including its dimensional gain. The output signal amplitude is then

$$S_{\text{out}}(\mathbf{R}) = G(\mathbf{R})S_{\text{in}}(\mathbf{R}). \tag{6.1295}$$

DQE in terms of transfer and noise power.

Using the power definition of SNR,

$$\text{SNR}_{\text{in}}(\mathbf{R}) = \frac{|S_{\text{in}}(\mathbf{R})|^2}{\text{NPS}_{\text{in}}(\mathbf{R})}. \tag{6.1296}$$

At the detector output,

$$\text{SNR}_{\text{out}}(\mathbf{R}) = \frac{|S_{\text{out}}(\mathbf{R})|^2}{\text{NPS}_{\text{out}}(\mathbf{R})} \tag{6.1297}$$

$$= \frac{|G(\mathbf{R})|^2|S_{\text{in}}(\mathbf{R})|^2}{\text{NPS}_{\text{out}}(\mathbf{R})}. \tag{6.1298}$$

Taking their ratio according to Eq. (6.1283),

$$\text{DQE}(\mathbf{R}) = \frac{|G(\mathbf{R})|^2|S_{\text{in}}(\mathbf{R})|^2}{\text{NPS}_{\text{out}}(\mathbf{R})}\frac{\text{NPS}_{\text{in}}(\mathbf{R})}{|S_{\text{in}}(\mathbf{R})|^2} \tag{6.1299}$$

$$= \boxed{\frac{|G(\mathbf{R})|^2\,\text{NPS}_{\text{in}}(\mathbf{R})}{\text{NPS}_{\text{out}}(\mathbf{R})}.} \tag{6.1300}$$

The arbitrary test-signal amplitude has disappeared. DQE is therefore a property of the detector and of its operating conditions, not of the strength of the particular weak modulation used to measure it.

For a homogeneous incident Poisson electron field of mean fluence Φ_0, Sec. 6.8.1 showed that

$$\text{NPS}_{\text{in}}(\mathbf{R}) = \Phi_0. \tag{6.1301}$$

For an integrating detector whose mean response to one electron is $\overline{q}$, Eq. (6.746) gives

$$G(\mathbf{R}) = \overline{\widehat{q}}(\mathbf{R}) = \overline{Q}H_{\text{det}}(\mathbf{R}). \tag{6.1302}$$

With the usual normalization

$$H_{\mathrm{det}}(\mathbf{0}) = 1, \tag{6.1303}$$

the MTF is

$$\mathrm{MTF}(\mathbf{R}) = |H_{\mathrm{det}}(\mathbf{R})|. \tag{6.1304}$$

Equation (6.1300) therefore becomes

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{\Phi_0 \overline{Q}^{\,2}\,\mathrm{MTF}^2(\mathbf{R})}{\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}.} \tag{6.1305}$$

Unlike the schematic proportionality often used when introducing DQE, Eq. (6.1305) includes the normalization required to make DQE dimensionless.

If the mean flat-field output is

$$\overline{I} = \Phi_0 \overline{Q}, \tag{6.1306}$$

define the normalized NPS, as in the preceding NPS discussion, by

$$\mathrm{NNPS}(\mathbf{R}) = \frac{\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}{\overline{I}^{\,2}}. \tag{6.1307}$$

Substituting

$$\mathrm{NPS}_{\mathrm{out}} = \Phi_0^2 \overline{Q}^{\,2}\,\mathrm{NNPS} \tag{6.1308}$$

into Eq. (6.1305) gives

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{\mathrm{MTF}^2(\mathbf{R})}{\Phi_0\,\mathrm{NNPS}(\mathbf{R})}.} \tag{6.1309}$$

The units are consistent. If Φ_0 is expressed in electrons per unit area, NNPS has units of area, and their product is dimensionless.

For a discrete detector it is common instead to work with the mean number of incident electrons per pixel,

$$\mu_0 = \Phi_0 a_x a_y. \tag{6.1310}$$

With the corresponding pixel-normalized definition of the NPS, Eq. (6.1309) takes the equivalent form

$$\mathrm{DQE}(\mathbf{R}) = \frac{\mathrm{MTF}^2(\mathbf{R})}{\mu_0\,\mathrm{NNPS}_{\mathrm{pix}}(\mathbf{R})}. \tag{6.1311}$$

This is why apparently different DQE formulas are encountered in the literature: the physical definition is the same, but the normalization of fluence and NPS may differ.

Another normalization, particularly common when displaying detector measurements, is to divide the NPS by its zero-frequency value. To avoid confusing this quantity with the NNPS of Eq. (6.1307), define

$$\boxed{\mathcal{N}(\mathbf{R}) = \frac{\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}{\mathrm{NPS}_{\mathrm{out}}(\mathbf{0})}.} \tag{6.1312}$$

Since

$$\mathrm{MTF}(\mathbf{0}) = 1, \tag{6.1313}$$

division of Eq. (6.1305) by its value at zero frequency gives

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \mathrm{DQE}(\mathbf{0})\frac{\mathrm{MTF}^2(\mathbf{R})}{\mathcal{N}(\mathbf{R})}.} \tag{6.1314}$$

Equation (6.1314) is a common experimental form of the DQE relation for electron detectors (McMullan et al., 2009a). It is important to remember that $\mathcal{N}$ in this equation is normalized by the zero-frequency NPS, whereas the NNPS in Eq. (6.1309) is normalized by the square of the mean flat-field signal.

6.9.1 DQE in the single-electron-response framework

The statistical description developed earlier allows us to express DQE more fundamentally than through separately measured MTF and NPS curves.

For a linear integrating detector, one primary electron produces the random Fourier response

$$\widehat{q}_\omega(\mathbf{R}). \tag{6.1315}$$

Its mean controls signal transfer,

$$G(\mathbf{R}) = \mathbb{E}\left[\widehat{q}_\omega(\mathbf{R})\right], \tag{6.1316}$$

whereas Sec. 6.6.2 showed that, for Poisson fluence Φ_0, the electron-induced output NPS is

$$\mathrm{NPS}_{\mathrm{electron}}(\mathbf{R}) = \Phi_0 \mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right]. \tag{6.1317}$$

DQE from the statistics of one electron.

Assume first that there are no additional electronic, dark-current, or other additive noise sources. Substituting

$$\mathrm{NPS}_{\mathrm{in}} = \Phi_0, \tag{6.1318}$$

$$G(\mathbf{R}) = \mathbb{E}[\widehat{q}_\omega(\mathbf{R})], \tag{6.1319}$$

and

$$\mathrm{NPS}_{\mathrm{out}} = \Phi_0 \mathbb{E}[|\widehat{q}_\omega(\mathbf{R})|^2] \tag{6.1320}$$

into Eq. (6.1300) gives

$$\mathrm{DQE}_{\mathrm{electron}}(\mathbf{R}) = \frac{\Phi_0 \left|\mathbb{E}[\widehat{q}_\omega(\mathbf{R})]\right|^2}{\Phi_0 \mathbb{E}[|\widehat{q}_\omega(\mathbf{R})|^2]} \tag{6.1321}$$

$$= \boxed{\frac{\left|\mathbb{E}[\widehat{q}_\omega(\mathbf{R})]\right|^2}{\mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right]}.} \tag{6.1322}$$

Equation (6.1322) is one of the central results of the detector analysis. It connects DQE directly with the distinction established in Eq. (6.749):

$$\boxed{\begin{array}{lcl} \left|\mathbb{E}[\widehat{q}_\omega]\right|^2 & \Longrightarrow & \text{coherent mean signal transfer,} \\ \mathbb{E}\left[|\widehat{q}_\omega|^2\right] & \Longrightarrow & \text{total electron-induced noise power.} \end{array}} \tag{6.1323}$$

Using the exact decomposition Eq. (6.746),

$$\mathbb{E}\left[|\widehat{q}_\omega|^2\right] = \overline{Q}^{\,2} |H_{\mathrm{det}}|^2 + \mathbb{E}\left[|\delta\widehat{q}_\omega|^2\right], \tag{6.1324}$$

we obtain

$$\boxed{\mathrm{DQE}_{\mathrm{electron}}(\mathbf{R}) = \frac{\overline{Q}^{\,2} |H_{\mathrm{det}}(\mathbf{R})|^2}{\overline{Q}^{\,2} |H_{\mathrm{det}}(\mathbf{R})|^2 + \mathbb{E}\left[|\delta\widehat{q}_\omega(\mathbf{R})|^2\right]}.} \tag{6.1325}$$

Thus event-to-event fluctuations reduce DQE because they contribute noise power that is not part of the coherent mean response.

By the inequality

$$\left|\mathbb{E}[\widehat{q}_\omega]\right|^2 \leq \mathbb{E}\left[|\widehat{q}_\omega|^2\right], \tag{6.1326}$$

Eq. (6.1322) gives, within this linear single-electron model,

$$\boxed{0 \leq \mathrm{DQE}_{\mathrm{electron}}(\mathbf{R}) \leq 1.} \tag{6.1327}$$

This result also clarifies the relation between DQE and MTF. In general,

$$\boxed{\mathrm{DQE}(\mathbf{R}) \text{ is not determined by } \mathrm{MTF}^2(\mathbf{R}).} \tag{6.1328}$$

The two quantities answer different questions:

$$\boxed{\begin{array}{lcl} \mathrm{MTF}(\mathbf{R}) & : & \text{how much mean modulation survives?} \\ \mathrm{DQE}(\mathbf{R}) & : & \text{how much input SNR survives?} \end{array}} \tag{6.1329}$$

A deterministic blur illustrates the difference especially clearly. Suppose every incident electron produces exactly

$$q_\omega(\boldsymbol{\rho}) = Q\, h_{\mathrm{det}}(\boldsymbol{\rho}), \tag{6.1330}$$

with no event-to-event fluctuation. Then

$$\mathbb{E}[|\delta \widehat{q}_\omega|^2] = 0, \tag{6.1331}$$

and Eq. (6.1325) gives

$$\mathrm{DQE}_{\mathrm{electron}}(\mathbf{R}) = 1 \tag{6.1332}$$

at frequencies where the transfer function is non-zero, even though

$$\mathrm{MTF}(\mathbf{R}) < 1. \tag{6.1333}$$

The deterministic blur attenuates the signal, but it attenuates the Poisson noise by exactly the same factor, so the presampling SNR is not lost. Information is lost only when stochastic response fluctuations, additional noise, sampling, aliasing, or other non-ideal processes prevent this attenuation from being perfectly reversible.

Conversely, two detectors can have the same MTF but different DQE values if their single-electron response fluctuations are different.

If independent additive detector noise is also present, let

$$\mathrm{NPS}_{\mathrm{add}}(\mathbf{R}) \tag{6.1334}$$

denote the sum of those contributions that can legitimately be treated as independent additive noise according to the framework developed in the preceding section. Then

$$\mathrm{NPS}_{\mathrm{out}} = \Phi_0 \mathbb{E}\left[|\widehat{q}_\omega|^2\right] + \mathrm{NPS}_{\mathrm{add}}, \tag{6.1335}$$

and

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{\Phi_0 \left|\mathbb{E}[\widehat{q}_\omega(\mathbf{R})]\right|^2}{\Phi_0 \mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right] + \mathrm{NPS}_{\mathrm{add}}(\mathbf{R})}.} \tag{6.1336}$$

Independent electronic, dark-current, and quantization noise therefore reduce DQE further because they increase the denominator without increasing the transferred electron signal.

6.9.2 DQE at zero spatial frequency

At zero spatial frequency,

$$H_{\mathrm{det}}(\mathbf{0}) = 1, \tag{6.1337}$$

and the Fourier transform of one complete single-electron event is its total collected charge,

$$\widehat{q}_\omega(\mathbf{0}) = Q. \tag{6.1338}$$

Equation (6.1322) therefore becomes

$$\boxed{\mathrm{DQE}(\mathbf{0}) = \frac{\mathbb{E}[Q]^2}{\mathbb{E}[Q^2]},} \tag{6.1339}$$

when no additional noise is present.

Using the single-electron-response excess-noise factor defined in Eq. (6.717),

$$\mathcal{F}_{\mathrm{SER}} = \frac{\mathbb{E}[Q^2]}{\mathbb{E}[Q]^2}, \tag{6.1340}$$

we immediately obtain

$$\boxed{\mathrm{DQE}(\mathbf{0}) = \frac{1}{\mathcal{F}_{\mathrm{SER}}}.} \tag{6.1341}$$

This result is the frequency-domain counterpart of Eq. (6.718). Event-to-event amplitude fluctuations reduce the zero-frequency DQE even when every primary electron is detected and the subsequent electronic readout is perfect.

Quantum efficiency appears as an important special case. Suppose an ideal counting detector either records one standardized event or misses the incident electron. Let

$$D = \begin{cases} 1, & \text{electron detected}, \\ 0, & \text{electron missed}, \end{cases} \tag{6.1342}$$

with

$$P(D = 1) = \eta. \tag{6.1343}$$

For the zero-frequency response,

$$Q = D, \tag{6.1344}$$

so

$$\mathbb{E}[Q] = \eta, \qquad \mathbb{E}[Q^2] = \eta. \tag{6.1345}$$

Therefore,

$$\boxed{\mathrm{DQE}(\mathbf{0}) = \eta.} \tag{6.1346}$$

Thus quantum efficiency and zero-frequency DQE are equal only for this idealized counting situation: every successfully detected electron contributes exactly one standardized count and no additional noise is present.

For an integrating detector with fluctuating event amplitude, or for a detector with additional readout noise,

$$\mathrm{DQE}(\mathbf{0}) \tag{6.1347}$$

can be smaller than the fraction of incident electrons that produce a detectable signal.

For example, suppose an electron is detected with probability η and, when detected, produces a random total amplitude A independent of the detection decision:

$$Q = DA. \tag{6.1348}$$

Then

$$\mathbb{E}[Q] = \eta\mathbb{E}[A], \tag{6.1349}$$

whereas

$$\mathbb{E}[Q^2] = \eta\mathbb{E}[A^2]. \tag{6.1350}$$

Hence

$$\boxed{\mathrm{DQE}(\mathbf{0}) = \eta\frac{\mathbb{E}[A]^2}{\mathbb{E}[A^2]}.} \tag{6.1351}$$

The first factor measures whether the primary electron is retained; the second measures the additional loss caused by fluctuations in the amplitude assigned to the retained event. Ideal electron counting makes the second factor equal to unity.

If independent additive detector noise is present, the more general zero-frequency expression is

$$\boxed{\mathrm{DQE}(\mathbf{0}) = \frac{\Phi_0 \overline{Q}^2}{\Phi_0 \mathbb{E}[Q^2] + \mathrm{NPS}_{\mathrm{add}}(\mathbf{0})}.} \tag{6.1352}$$

At sufficiently high fluence the electron-induced term may dominate the additive electronic noise, whereas at very low fluence the latter can substantially reduce DQE. This is one reason why electron counting, which can strongly suppress readout-noise contributions once events are reliably detected, is advantageous for low-dose cryo-EM.

A useful counting example also shows why quantum efficiency alone is not enough away from zero frequency. Suppose an isolated electron is detected with probability η and, when detected, is assigned a position with localization error $\boldsymbol{\epsilon}$. The Fourier response of one incident electron is then

$$\widehat{q}_\omega(\mathbf{R}) = D e^{-j2\pi \mathbf{R}\cdot\boldsymbol{\epsilon}}. \tag{6.1353}$$

Assuming the detection decision and localization error are independent,

$$\mathbb{E}[\widehat{q}_\omega(\mathbf{R})] = \eta H_{\mathrm{loc}}(\mathbf{R}), \tag{6.1354}$$

where H_{loc} is the localization transfer function derived previously. Since

$$\left| e^{-j2\pi \mathbf{R}\cdot\boldsymbol{\epsilon}} \right|^2 = 1, \tag{6.1355}$$

we also have

$$\mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2 \right] = \eta. \tag{6.1356}$$

Equation (6.1322) therefore gives

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \eta |H_{\mathrm{loc}}(\mathbf{R})|^2.} \tag{6.1357}$$

At zero frequency,

$$\mathrm{DQE}(\mathbf{0}) = \eta, \tag{6.1358}$$

but at higher frequency localization uncertainty reduces the DQE even though the fraction of detected electrons is unchanged.

For perfect detection, $\eta = 1$, this special model gives

$$\mathrm{DQE}(\mathbf{R}) = \mathrm{MTF}^2(\mathbf{R}). \tag{6.1359}$$

This equality is therefore possible, but it is a special consequence of the ideal counting-localization model and must not be interpreted as a general identity between DQE and MTF.

6.9.3 DQE near the sampling limits and beyond the physical Nyquist frequency

The frequency dependence of DQE becomes especially important as a detector approaches the sampling limit of its *output representation*. The relevant sampling limit, however, is not always the Nyquist frequency associated with the physical detector pixels.

The basic mechanism is the same in all cases. From Eq. (6.1305),

$$\mathrm{DQE}(\mathbf{R}) = \frac{\Phi_0 \overline{Q}^2 \, \mathrm{MTF}^2(\mathbf{R})}{\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}, \tag{6.1360}$$

so the high-frequency behavior of the DQE is determined by a competition between

$$\boxed{\begin{array}{c}\text{the loss of coherent signal transfer}\\ \text{and}\\ \text{the noise power that remains after detection and sampling.}\end{array}} \tag{6.1361}$$

The MTF generally decreases toward high spatial frequency because of primary-electron spreading, carrier diffusion, charge sharing, finite pixel integration, and, in counting detectors, localization uncertainty. The numerator of the DQE therefore normally decreases with increasing spatial frequency.

The NPS need not decrease in the same way. As shown in the preceding sections, single-electron-response fluctuations, electronic noise, charge-sharing correlations, and counting correlations have their own spectral dependences. Sampling can also fold presampling noise from frequencies outside the fundamental sampling region back into the measured NPS.

A falling MTF therefore does not by itself determine how rapidly DQE falls. If signal and noise were attenuated by exactly the same factor, their ratio could remain unchanged. DQE decreases when the useful modulation is attenuated more strongly than the total noise power against which that modulation must be detected.

Conventional and integrating detectors. Let the physical detector pixels have pitch a. If the detector output contains one value per physical pixel, the Nyquist frequency along a detector axis is

$$\boxed{R_{\mathrm{N,phys}} = \frac{1}{2a}.} \tag{6.1362}$$

This is the relevant sampling limit for an integrating detector, and also for a counting detector whose final output contains one count value per physical pixel.

As $R_{\mathrm{N,phys}}$ is approached, the useful modulation is already attenuated by the presampling detector response and by the finite pixel aperture. At the same time, the measured NPS contains aliased contributions from presampling frequencies outside the physical Nyquist region.

The important asymmetry is that this aliased noise contributes to the denominator of the measured DQE without providing useful signal for the particular baseband modulation being measured. Thus aliasing can decrease high-frequency DQE even when the corresponding presampling detector response would retain more information.

For a conventional pixel output, frequencies above $R_{\mathrm{N,phys}}$ cannot be represented uniquely: sampling folds them back into the fundamental frequency region.

Electron counting and super-resolution. The situation changes for an electron-counting detector capable of estimating the position of an individual electron with subpixel precision.

A primary electron generally produces a multi-pixel event. The relative signals within that event contain information about the impact position inside the physical pixel. If this information is analyzed *before* the event is reduced to one count per physical pixel, the estimated electron position can be placed on a finer output lattice.

Suppose each physical pixel is divided into $u \times u$ output subpixels. The corresponding output pitch is

$$a_{\mathrm{sr}} = \frac{a}{u}, \tag{6.1363}$$

and its Nyquist frequency is

$$\boxed{R_{\mathrm{N,sr}} = \frac{1}{2a_{\mathrm{sr}}} = uR_{\mathrm{N,phys}}.} \tag{6.1364}$$

For the common 2×2 subdivision,

$$R_{\mathrm{N,sr}} = 2R_{\mathrm{N,phys}}. \tag{6.1365}$$

Consequently, the interval

$$R_{\mathrm{N,phys}} < R < R_{\mathrm{N,sr}} \tag{6.1366}$$

is beyond the Nyquist frequency of the *physical-pixel output*, but not beyond the Nyquist frequency of the recorded super-resolution image.

The physical-pixel Nyquist frequency is therefore not necessarily a sharp information boundary for an event-localizing detector.

Why DQE can remain non-zero beyond the physical Nyquist. The localization model developed previously makes this particularly clear. For an electron detected with probability η and assigned a position with localization error $\boldsymbol{\epsilon}$, the single-electron Fourier response is

$$\widehat{q}_\omega(\mathbf{R}) = De^{-j2\pi\mathbf{R}\cdot\boldsymbol{\epsilon}}, \tag{6.1367}$$

where $D = 1$ for a detected event and $D = 0$ otherwise.

As derived in Eq. (6.1357),

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \eta\,|H_{\mathrm{loc}}(\mathbf{R})|^2\,,} \tag{6.1368}$$

when localization uncertainty is the only frequency-dependent imperfection.

Nothing in Eq. (6.1368) is singular at

$$R = R_{\mathrm{N,phys}}. \tag{6.1369}$$

The DQE decreases according to the precision with which the electron position is inferred, not because the physical-pixel Nyquist frequency has been crossed.

For an isotropic Gaussian localization error with standard deviation σ_{loc},

$$H_{\mathrm{loc}}(\mathbf{R}) = \exp\left(-2\pi^2\sigma_{\mathrm{loc}}^2 R^2\right), \tag{6.1370}$$

and hence

$$\boxed{\mathrm{DQE}(R) = \eta\exp\left(-4\pi^2\sigma_{\mathrm{loc}}^2 R^2\right).} \tag{6.1371}$$

If $\sigma_{\mathrm{loc}} \ll a$, appreciable DQE can therefore survive for

$$R > R_{\mathrm{N,phys}}. \tag{6.1372}$$

This is genuine super-resolution and should not be confused with interpolation. Interpolating an ordinary pixel image onto a finer grid cannot recover information already discarded during pixel sampling. In event-based super-resolution, the subpixel position is estimated from the multi-pixel electron event *before* that information is discarded.

The same competition reappears near the super-resolution Nyquist frequency. Super-resolution changes the sampling boundary, but it does not remove the high-frequency competition expressed in Eq. (6.1361).

The DQE in the super-resolution representation is still

$$\mathrm{DQE}_{\mathrm{sr}}(\mathbf{R}) = \frac{\mathrm{SNR}_{\mathrm{out,sr}}(\mathbf{R})}{\mathrm{SNR}_{\mathrm{in}}(\mathbf{R})}, \tag{6.1373}$$

and, under the linear small-signal assumptions of Sec. 6.9,

$$\boxed{\mathrm{DQE}_{\mathrm{sr}}(\mathbf{R}) = \frac{\Phi_0 \overline{Q}^{\,2}\,\mathrm{MTF}_{\mathrm{sr}}^2(\mathbf{R})}{\mathrm{NPS}_{\mathrm{sr}}(\mathbf{R})}.} \tag{6.1374}$$

Beyond the physical Nyquist frequency, the coherent signal continues to decrease because localization is imperfect and because the original electron event has a finite spatial extent. Meanwhile, the noise retains its own spectral structure. Coincidence effects, event-to-event response fluctuations, localization errors, and any additional noise introduced by the event-processing algorithm may all contribute.

Super-resolution also pushes the sampling replicas farther apart and therefore reduces aliasing within the frequency range around and below $R_{\mathrm{N,phys}}$. This is an important advantage: information that would have been prematurely folded by a one-value-per-physical-pixel representation can instead be represented explicitly on the finer lattice.

However, as

$$R \to R_{\mathrm{N,sr}}, \tag{6.1375}$$

the same sampling problem arises again. Frequencies outside the super-resolution Nyquist region can now alias into its fundamental region, while the useful modulation has already been strongly attenuated by the localization transfer function and the other high-frequency components of the detector response.

Thus the super-resolution Nyquist frequency does not remove the competition of Eq. (6.1361); it merely moves the sampling boundary to a higher spatial frequency.

Three distinct frequency scales. It is therefore useful to distinguish three different spatial-frequency scales:

$$\boxed{\begin{aligned} R_{\mathrm{N,phys}} &= \text{Nyquist frequency of the physical-pixel lattice,} \\ R_{\mathrm{N,sr}} &= \text{Nyquist frequency of the super-resolution output lattice,} \\ R_{\mathrm{info}} &= \text{highest frequency carrying useful detector information.} \end{aligned}} \tag{6.1376}$$

For an integrating detector or a conventional pixel output, $R_{\mathrm{N,phys}}$ is the relevant sampling limit. For an event-localizing counting detector,

$$R_{\mathrm{info}} > R_{\mathrm{N,phys}} \tag{6.1377}$$

is possible, provided that the output grid is fine enough to represent that information.

Neither Nyquist frequency alone determines R_{info}. The useful frequency range is determined by the combined behavior of signal transfer, noise, localization accuracy, and sampling—that is, ultimately, by the DQE.

Two-dimensional sampling. Finally, the Nyquist boundary is intrinsically two-dimensional. For a rectangular detector grid it is a region in (R_x, R_y) rather than one unique radial frequency. Detector response, localization accuracy, and readout noise may also be anisotropic. The fundamental quantity is therefore

$$\mathrm{DQE}(R_x, R_y), \tag{6.1378}$$

even when a radially averaged $\mathrm{DQE}(R)$ is reported for convenience.

Representative DQE values of current direct detectors. It is useful to put the preceding discussion into numerical perspective. Table 6.3 lists several recently reported or specified DQE values at the *physical-pixel Nyquist frequency.* The numbers should not be interpreted as a strict ranking of detectors because DQE depends on accelerating voltage, dose rate, detector operating mode, event-processing algorithm, and the experimental procedure used to estimate MTF and NPS.

For the Apollo, a direct experimental characterization reported $\mathrm{DQE}(0) \simeq 0.94$, $\mathrm{DQE}(R_{\mathrm{N}}/2) \simeq 0.66$, and $\mathrm{DQE}(R_{\mathrm{N}}) \simeq 0.23$ (Peng et al., 2023). More recently, the Falcon C, optimized for 100-keV electrons, was reported to reach approximately 0.96, 0.56, and 0.22 at zero, half-Nyquist, and physical Nyquist frequency, respectively (Karia et al., 2025). Current manufacturer

Table 6.3: Representative DQE values at the physical-pixel Nyquist frequency for modern direct electron detectors.

Detector	Voltage	Mode	DQE at physical Nyquist
Falcon 4i	300 kV	counting, EER	0.50
K3	300 kV	counting, super-resolution	> 0.40
Apollo	300 kV	counting	$\simeq 0.23$
Falcon C	100 kV	counting, EER	$\simeq 0.22$

specifications quote $\mathrm{DQE}(R_{\mathrm{N}}) = 0.50$ at 300 kV for the Falcon 4i in EER mode and a value above 0.40 at 300 kV for the K3 in super-resolution counting mode.

These values illustrate how demanding the Nyquist region remains. Modern detectors can preserve more than 90% of the input SNR at low spatial frequency, while typically retaining only a few tens of percent at the physical Nyquist frequency. The loss is precisely the result of the competition emphasized in Eq. (6.1361): coherent signal transfer becomes progressively weaker while noise power, including aliased noise, remains in the measurement.

The comparatively large values quoted at the physical Nyquist frequency for modern event-localizing detectors should not be confused with a fundamental sampling limit. In EER or super-resolution operation, subpixel event information has been retained before formation of the final image, and the DQE can remain non-zero beyond $R_{\mathrm{N,phys}}$ until the higher-frequency competition between localization accuracy, signal transfer, noise, and the super-resolution sampling limit becomes dominant.

6.10 Noise-equivalent quanta

The detective quantum efficiency tells us what *fraction* of the input signal-to-noise ratio survives the detector. It is dimensionless and therefore provides a convenient way of comparing detectors, but its numerical value can still feel somewhat abstract.

The same result can be expressed in a more physically intuitive form. Instead of asking what fraction of the input SNR survives, we may ask:

$$\boxed{\text{How many electrons would an ideal detector require to produce the same SNR as the real detector?}} \tag{6.1379}$$

This leads to the *noise-equivalent quanta*, abbreviated NEQ.

In detector literature the incident fluence is often denoted by q. We retain the notation Φ_0 used throughout this chapter in order to avoid confusion with the single-electron response $q_\omega(\boldsymbol{\rho})$.

Definition of NEQ Let Φ_0 be the mean incident electron fluence, expressed, for example, in electrons per unit area. The noise-equivalent quanta at spatial frequency $\mathbf{R}$ are defined by

$$\boxed{\mathrm{NEQ}(\mathbf{R}) = \Phi_0 \, \mathrm{DQE}(\mathbf{R}).} \tag{6.1380}$$

Since the DQE is dimensionless, NEQ has the same units as the incident fluence. Thus, if

$$[\Phi_0] = \text{electrons}\,\mathrm{m}^{-2}, \tag{6.1381}$$

then

$$[\mathrm{NEQ}] = \text{electrons}\,\mathrm{m}^{-2}. \tag{6.1382}$$

If instead the incident exposure is expressed as the mean number of electrons per pixel,

$$\mu_0 = \Phi_0 a_x a_y, \tag{6.1383}$$

the corresponding pixel-based quantity is

$$\boxed{\mathrm{NEQ}_{\mathrm{pix}}(\mathbf{R}) = \mu_0 \, \mathrm{DQE}(\mathbf{R}),} \tag{6.1384}$$

and is naturally interpreted as an equivalent number of ideal electrons per pixel.

The meaning of Eq. (6.1380) follows immediately from Eq. (6.1286). Suppose, for example, that

$$\Phi_0 = 100 \tag{6.1385}$$

electrons in some chosen unit of area contribute to the measurement and that, at a particular spatial frequency,

$$\mathrm{DQE}(\mathbf{R}) = 0.7. \tag{6.1386}$$

Then

$$\mathrm{NEQ}(\mathbf{R}) = 70. \tag{6.1387}$$

At that frequency, the real detector therefore produces the same SNR as an *ideal* detector receiving only 70 electrons in the same unit of area.

This does not mean that only 70 of the 100 electrons were physically detected. All 100 electrons might have produced detector events. Localization uncertainty, fluctuations of the single-electron response, electronic noise, sampling, or other imperfections can nevertheless make their information content equivalent to that of only 70 perfectly measured electrons.

Thus

$$\boxed{\begin{array}{c}\text{DQE} = \text{fraction of ideal SNR retained,}\\ \text{NEQ} = \text{equivalent ideal-electron fluence retained.}\end{array}} \tag{6.1388}$$

Relationship to DQE, MTF, and NPS Equation (6.1380) may be combined directly with the frequency-dependent DQE derived in Sec. 6.9. From Eq. (6.1305),

$$\text{DQE}(\mathbf{R}) = \frac{\Phi_0 \overline{Q}^2 \,\text{MTF}^2(\mathbf{R})}{\text{NPS}_{\text{out}}(\mathbf{R})}. \tag{6.1389}$$

Multiplication by Φ_0 gives the NEQ.

NEQ in terms of MTF and NPS.

Starting from Eq. (6.1380),

$$\text{NEQ}(\mathbf{R}) = \Phi_0 \,\text{DQE}(\mathbf{R}) \tag{6.1390}$$

$$= \Phi_0 \frac{\Phi_0 \overline{Q}^2 \,\text{MTF}^2(\mathbf{R})}{\text{NPS}_{\text{out}}(\mathbf{R})} \tag{6.1391}$$

$$= \boxed{\frac{\Phi_0^2 \overline{Q}^2 \,\text{MTF}^2(\mathbf{R})}{\text{NPS}_{\text{out}}(\mathbf{R})}}. \tag{6.1392}$$

The mean flat-field detector output is

$$\overline{I} = \Phi_0 \overline{Q}. \tag{6.1393}$$

Hence

$$\boxed{\text{NEQ}(\mathbf{R}) = \frac{\overline{I}^2 \,\text{MTF}^2(\mathbf{R})}{\text{NPS}_{\text{out}}(\mathbf{R})}.} \tag{6.1394}$$

Equation (6.1394) is particularly useful experimentally. It shows directly

that NEQ increases with the square of the transferred mean signal and decreases with the noise power at the same spatial frequency.

Using the normalized NPS introduced previously,

$$\mathrm{NNPS}(\mathbf{R}) = \frac{\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}{\overline{I}^{\,2}}, \tag{6.1395}$$

Eq. (6.1394) becomes

$$\boxed{\mathrm{NEQ}(\mathbf{R}) = \frac{\mathrm{MTF}^2(\mathbf{R})}{\mathrm{NNPS}(\mathbf{R})}.} \tag{6.1396}$$

The parallel among the three detector quantities is therefore

$$\boxed{\begin{array}{lcl} \mathrm{MTF}(\mathbf{R}) & \Longrightarrow & \text{mean signal transfer,} \\ \mathrm{DQE}(\mathbf{R}) & \Longrightarrow & \text{fraction of input SNR retained,} \\ \mathrm{NEQ}(\mathbf{R}) & \Longrightarrow & \text{equivalent ideal-electron fluence.} \end{array}} \tag{6.1397}$$

Single-electron interpretation. The single-electron-response formulation developed earlier gives an even more microscopic expression. In the absence of additional additive noise, Eq. (6.1322) gives

$$\mathrm{DQE}_{\mathrm{electron}}(\mathbf{R}) = \frac{\left|\mathbb{E}[\widehat{q}_\omega(\mathbf{R})]\right|^2}{\mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right]}. \tag{6.1398}$$

Therefore,

$$\boxed{\mathrm{NEQ}_{\mathrm{electron}}(\mathbf{R}) = \Phi_0 \frac{\left|\mathbb{E}[\widehat{q}_\omega(\mathbf{R})]\right|^2}{\mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right]}.} \tag{6.1399}$$

This expression has a simple interpretation. The incident fluence Φ_0 is reduced by the ratio between the coherent squared mean response and the total second moment of the random single-electron response.

If statistically independent additive detector noise is also present, Eq. (6.1336) gives

$$\boxed{\mathrm{NEQ}(\mathbf{R}) = \frac{\Phi_0^2 \left|\mathbb{E}[\widehat{q}_\omega(\mathbf{R})]\right|^2}{\Phi_0 \mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right] + \mathrm{NPS}_{\mathrm{add}}(\mathbf{R})}.} \tag{6.1400}$$

Equation (6.1400) also shows that NEQ does not necessarily scale linearly with exposure under all operating conditions. When additive readout

noise is important, increasing the electron fluence improves the relative importance of electron signal. Conversely, in counting mode the DQE itself may decrease with dose rate because of coincidence loss. Thus, more generally,

$$\mathrm{NEQ} = \mathrm{NEQ}(\mathbf{R}; \Phi_0, \text{dose rate}, \text{operating mode}), \tag{6.1401}$$

and DQE and NEQ should be quoted for specified detector conditions.

Zero spatial frequency. At zero spatial frequency,

$$\boxed{\mathrm{NEQ}(\mathbf{0}) = \Phi_0 \, \mathrm{DQE}(\mathbf{0}).} \tag{6.1402}$$

For an integrating detector whose only non-ideality is the single-electron amplitude fluctuation, Eq. (6.1341) gives

$$\boxed{\mathrm{NEQ}(\mathbf{0}) = \frac{\Phi_0}{\mathcal{F}_{\mathrm{SER}}}.} \tag{6.1403}$$

For the idealized counting detector that records each incident electron with probability η and assigns every successful event one standardized count,

$$\mathrm{DQE}(\mathbf{0}) = \eta, \tag{6.1404}$$

and hence

$$\boxed{\mathrm{NEQ}(\mathbf{0}) = \eta \Phi_0.} \tag{6.1405}$$

Only in this special case does NEQ coincide directly with the mean fluence of successfully detected primary electrons.

Information interpretation The most useful interpretation of NEQ arises when we consider a weak spatial modulation of the incident electron fluence.

Suppose that the mean incident fluence is Φ_0 and that the specimen introduces a small fractional modulation $m(\boldsymbol{\rho})$:

$$\Phi(\boldsymbol{\rho}) = \Phi_0 \left[1 + m(\boldsymbol{\rho})\right], \qquad |m(\boldsymbol{\rho})| \ll 1. \tag{6.1406}$$

For an ideal Poisson detector, the SNR carried by a weak Fourier component is proportional to the incident electron fluence. This gives a particularly direct interpretation of the NEQ.

NEQ as the effective fluence of a weak Fourier component.

Consider an observation region of area A with incident fluence

$$\Phi(\boldsymbol{\rho}) = \Phi_0 \left[1 + m(\boldsymbol{\rho})\right], \qquad |m| \ll 1. \tag{6.1407}$$

Suppose that $m(\boldsymbol{\rho})$ contains a Fourier component of complex amplitude $\widehat{m}(\mathbf{R})$ at spatial frequency $\mathbf{R}$.

With the Fourier-transform convention used here, the corresponding Fourier amplitude of the *mean number of incident electrons* over the finite observation region is

$$S_{\text{in}}(\mathbf{R}) = A\Phi_0 \widehat{m}(\mathbf{R}). \tag{6.1408}$$

For an ideal Poisson electron field, the shot-noise NPS is Φ_0. Over a finite area A, the variance of the corresponding Fourier coefficient is therefore

$$\sigma^2_{\text{in}}(\mathbf{R}) = A\Phi_0. \tag{6.1409}$$

Using the power definition of SNR adopted throughout this book,

$$\begin{aligned}
\text{SNR}_{\text{in}}(\mathbf{R}) &= \frac{|S_{\text{in}}(\mathbf{R})|^2}{\sigma^2_{\text{in}}(\mathbf{R})} && (6.1410)\\
&= \frac{A^2\Phi_0^2 |\widehat{m}(\mathbf{R})|^2}{A\Phi_0} && (6.1411)\\
&= \boxed{A\Phi_0 |\widehat{m}(\mathbf{R})|^2.} && (6.1412)
\end{aligned}$$

The detector reduces this input SNR according to Eq. (6.1286),

$$\text{SNR}_{\text{out}}(\mathbf{R}) = \text{DQE}(\mathbf{R})\, \text{SNR}_{\text{in}}(\mathbf{R}). \tag{6.1413}$$

Therefore,

$$\begin{aligned}
\text{SNR}_{\text{out}}(\mathbf{R}) &= A\Phi_0\, \text{DQE}(\mathbf{R}) |\widehat{m}(\mathbf{R})|^2 && (6.1414)\\
&= \boxed{A\ \text{NEQ}(\mathbf{R}) |\widehat{m}(\mathbf{R})|^2,} && (6.1415)
\end{aligned}$$

where Eq. (6.1380) has been used.

Equation (6.1415) makes the interpretation of NEQ particularly transparent. For a fixed fractional specimen modulation $\widehat{m}(\mathbf{R})$ and observation

area A, the output SNR is proportional directly to NEQ($\mathbf{R}$). Thus NEQ plays exactly the role that the physical incident fluence Φ_0 would play for an ideal Poisson detector.

The proportionality factor depends only on the Fourier and finite-area normalization adopted. The important point is that the physical fluence Φ_0 has disappeared in favor of NEQ($\mathbf{R}$).

Consequently,

NEQ($\mathbf{R}$) is the incident fluence that an ideal detector would need to provide the same SNR at spatial frequency $\mathbf{R}$ as the real detector provides from the actual exposure.

(6.1416)

This is why NEQ is especially natural in cryo-EM. The resource being spent is electron exposure, and radiation damage limits how many electrons can be used. NEQ translates detector performance back into that same physical currency.

Connection with detectability. The information interpretation can be made more precise for a specified weak imaging task. Consider, for simplicity, a finite sampled image with mean incident exposure μ_0 electrons per pixel and a small fractional signal m_p. Let a unitary discrete Fourier transform represent this signal by coefficients $\widehat{m}_{\mathbf{k}}$.

For a specified detection task, the connection between NEQ and information can be made more precise. Consider two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, whose mean detector images differ by

$$\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0. \tag{6.1417}$$

Assume that the noise has the same covariance matrix $\mathbf{K}$ under both hypotheses. We first determine the linear combination of the image pixels that maximizes their separation relative to the noise.

Optimal linear detectability for stationary noise.

Let the measured image, written as a column vector, be $\mathbf{I}$. A linear observer forms the scalar decision variable

$$t = \mathbf{w}^T\mathbf{I}, \tag{6.1418}$$

where $\mathbf{w}$ contains the weights assigned to the image pixels.

Under the two hypotheses, the difference between the mean values of t is

$$\begin{aligned}\Delta\bar{t} &= \mathbb{E}[t \mid \mathcal{H}_1] - \mathbb{E}[t \mid \mathcal{H}_0] \\ &= \mathbf{w}^T \Delta\boldsymbol{\mu}.\end{aligned} \tag{6.1419}$$

If the noise covariance is $\mathbf{K}$, the variance of the decision variable is

$$\sigma_t^2 = \mathbf{w}^T \mathbf{K} \mathbf{w}. \tag{6.1420}$$

Using the power-SNR convention adopted throughout this book, the detectability of this particular linear observer is therefore

$$d_{\mathbf{w}}'^2 = \frac{\left(\mathbf{w}^T \Delta\boldsymbol{\mu}\right)^2}{\mathbf{w}^T \mathbf{K} \mathbf{w}}. \tag{6.1421}$$

We now seek the weights $\mathbf{w}$ that maximize this quantity. Introduce

$$\mathbf{u} = \mathbf{K}^{1/2}\mathbf{w}, \qquad \mathbf{v} = \mathbf{K}^{-1/2}\Delta\boldsymbol{\mu}. \tag{6.1422}$$

Then

$$\mathbf{w}^T \Delta\boldsymbol{\mu} = \mathbf{u}^T \mathbf{v}, \tag{6.1423}$$

while

$$\mathbf{w}^T \mathbf{K} \mathbf{w} = \mathbf{u}^T \mathbf{u}. \tag{6.1424}$$

Hence

$$d_{\mathbf{w}}'^2 = \frac{(\mathbf{u}^T \mathbf{v})^2}{\mathbf{u}^T \mathbf{u}}. \tag{6.1425}$$

By the Cauchy–Schwarz inequality,

$$(\mathbf{u}^T \mathbf{v})^2 \leq (\mathbf{u}^T \mathbf{u})(\mathbf{v}^T \mathbf{v}), \tag{6.1426}$$

so

$$d_{\mathbf{w}}'^2 \leq \mathbf{v}^T \mathbf{v} = \Delta\boldsymbol{\mu}^T \mathbf{K}^{-1} \Delta\boldsymbol{\mu}. \tag{6.1427}$$

Equality is reached when

$$\mathbf{u} \propto \mathbf{v}, \tag{6.1428}$$

or equivalently

$$\boxed{\mathbf{w}_{\text{opt}} \propto \mathbf{K}^{-1} \Delta\boldsymbol{\mu}.} \tag{6.1429}$$

The maximum linear detectability is therefore

$$\boxed{d'^2 = \Delta\boldsymbol{\mu}^T \mathbf{K}^{-1} \Delta\boldsymbol{\mu}.} \tag{6.1430}$$

We now use the assumption that the noise is stationary. For a finite sampled image, neglecting boundary effects or equivalently adopting a periodic representation, stationarity makes $\mathbf{K}$ a block-circulant covariance matrix. The discrete Fourier basis therefore diagonalizes it.

Let $\mathbf{F}$ denote the unitary discrete Fourier transform. Then

$$\boxed{\mathbf{K} = \mathbf{F}^\dagger \boldsymbol{\Lambda} \mathbf{F},} \tag{6.1431}$$

where $\boldsymbol{\Lambda}$ is diagonal. Its diagonal elements are the noise powers of the individual Fourier modes:

$$\Lambda_{\mathbf{k}} = \mathrm{NPS}_{\mathbf{k}}, \tag{6.1432}$$

with the NPS normalization chosen consistently with the unitary Fourier transform.

Consequently,

$$\mathbf{K}^{-1} = \mathbf{F}^\dagger \boldsymbol{\Lambda}^{-1} \mathbf{F}. \tag{6.1433}$$

Substituting this result into Eq. (6.1430) gives

$$d'^2 = \Delta\boldsymbol{\mu}^T \mathbf{F}^\dagger \boldsymbol{\Lambda}^{-1} \mathbf{F} \Delta\boldsymbol{\mu} \tag{6.1434}$$

$$= \Delta\widehat{\mathbf{I}}^\dagger \boldsymbol{\Lambda}^{-1} \Delta\widehat{\mathbf{I}}, \tag{6.1435}$$

where

$$\Delta\widehat{\mathbf{I}} = \mathbf{F}\Delta\boldsymbol{\mu} \tag{6.1436}$$

is the Fourier transform of the difference between the mean images of the two hypotheses.

Since $\boldsymbol{\Lambda}$ is diagonal,

$$\boxed{d'^2 = \sum_{\mathbf{k}} \frac{|\Delta\widehat{I}_{\mathbf{k}}|^2}{\mathrm{NPS}_{\mathbf{k}}}.} \tag{6.1437}$$

Equation (6.1437) shows why the Fourier representation is especially natural for stationary detector noise. Stationarity makes the Fourier modes mutually uncorrelated, so each frequency contributes independently to the total detectability. A signal component at frequency $\mathbf{R}_{\mathbf{k}}$ contributes in pro-

portion to its squared amplitude and inversely in proportion to the noise power present at that same frequency.

The optimal observer therefore gives greater weight to frequencies for which the expected signal is strong and the NPS is small, and suppresses frequencies dominated by noise.

For a linear detector,

$$\Delta \widehat{I}_{\mathbf{k}} = \mu_0 G_{\mathbf{k}} \widehat{m}_{\mathbf{k}}, \tag{6.1438}$$

where $G_{\mathbf{k}}$ is the mean detector transfer function.

NEQ as frequency-dependent detectability weight.

Substitution into Eq. (6.1437) gives

$$d'^2 = \sum_{\mathbf{k}} \frac{\mu_0^2 |G_{\mathbf{k}}|^2 |\widehat{m}_{\mathbf{k}}|^2}{\mathrm{NPS}_{\mathbf{k}}} \tag{6.1439}$$

$$= \sum_{\mathbf{k}} \left[\frac{\mu_0^2 |G_{\mathbf{k}}|^2}{\mathrm{NPS}_{\mathbf{k}}} \right] |\widehat{m}_{\mathbf{k}}|^2. \tag{6.1440}$$

But the quantity in square brackets is precisely the pixel-based NEQ. Therefore,

$$\boxed{d'^2 = \sum_{\mathbf{k}} \mathrm{NEQ}_{\mathrm{pix}}(\mathbf{R}_{\mathbf{k}}) \, |\widehat{m}_{\mathbf{k}}|^2 \, .} \tag{6.1441}$$

Equation (6.1441) gives a precise meaning to the information carried by the NEQ spectrum. Spatial frequencies for which NEQ is large contribute strongly to the detectability of specimen features containing those frequencies; frequencies for which NEQ is small contribute little even if the detector formally samples them.

This is closely related to the distinction introduced previously between a sampling limit and an information limit. A super-resolution output grid may represent spatial frequencies beyond the physical-pixel Nyquist frequency, but those frequencies are useful only to the extent that their NEQ remains appreciable.

NEQ is not itself an information measure in bits. The word "information" should nevertheless be used carefully. NEQ($\mathbf{R}$) is not a Shannon information in bits, nor is it a literal count of statistically independent

electrons stored at each spatial frequency. The same physical electrons contribute simultaneously to many Fourier components of the image.

NEQ is instead a frequency-dependent *SNR-equivalent fluence*. Only after a particular imaging task or signal is specified does it translate into a scalar detectability through an expression such as Eq. (6.1441). Consequently, the values of NEQ($\mathbf{R}$) should not simply be integrated over spatial frequency and interpreted as a total number of electrons.

The hierarchy can therefore be summarized as

$$\boxed{\begin{aligned} \Phi_0 &\implies \text{physical incident-electron fluence,} \\ \mathrm{DQE}(\mathbf{R}) &\implies \text{fraction of input SNR retained,} \\ \mathrm{NEQ}(\mathbf{R}) &\implies \text{equivalent ideal-electron fluence,} \\ d'^2 &\implies \text{detectability for a specified imaging task.} \end{aligned}} \tag{6.1442}$$

For dose-limited cryo-EM, NEQ is therefore a particularly intuitive bridge between detector physics and image information: it expresses the frequency-dependent performance of the detector in the same currency that ultimately limits the experiment—the number of available electrons.

6.11 Nonideal detector behavior

The MTF, NPS, DQE, and NEQ developed in the preceding sections provide a powerful description of detector performance, but they rely on important approximations. In their simplest form, the detector is assumed to be linear and shift invariant, while its stochastic fluctuations are assumed to be stationary.

Under these assumptions, the mean detector response can be described by a single point-spread function,

$$\overline{y}(\boldsymbol{\rho}) = \int h_{\mathrm{det}}(\boldsymbol{\rho} - \boldsymbol{\rho}')x(\boldsymbol{\rho}')\, d^2\boldsymbol{\rho}', \tag{6.1443}$$

and the noise covariance depends only on the displacement between two positions,

$$C_n(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) = C_n(\boldsymbol{\rho}_2 - \boldsymbol{\rho}_1). \tag{6.1444}$$

These properties allow the Fourier basis to diagonalize both the mean response and the noise covariance, leading naturally to scalar frequency-dependent quantities such as

$$\mathrm{MTF}(\mathbf{R}), \qquad \mathrm{NPS}(\mathbf{R}), \qquad \mathrm{DQE}(\mathbf{R}). \tag{6.1445}$$

Real detectors may violate one or more of these assumptions. Their response can depend on signal level, on previous electron events, on position within the sensor, or on the accumulated irradiation history. These effects are not adequately described as additional stationary noise terms. Instead, they require modifications of the detector response model itself.

The most important examples are considered below.

6.11.1 Saturation

An integrating pixel can store only a finite amount of charge, and the analog readout and ADC also have finite dynamic ranges. Once one of these limits is reached, the detector response ceases to be linear.

A simple model is

$$\boxed{y = \min\left(gx + b, y_{\text{sat}}\right),} \tag{6.1446}$$

where x is the incident or collected signal, g is the linear gain, b is the offset, and y_{sat} is the maximum measurable output.

Below saturation,

$$\frac{dy}{dx} = g, \tag{6.1447}$$

whereas above saturation,

$$\frac{dy}{dx} = 0. \tag{6.1448}$$

Thus an additional incident signal produces progressively less measured signal as the detector approaches saturation.

Saturation is fundamentally different from additive noise. It changes the mean response itself:

$$\mathbb{E}[y \mid x] \tag{6.1449}$$

is no longer proportional to x. Consequently, a modulation transfer function measured in the linear regime need not describe the response close to saturation.

Nonlinearity also generates new spatial frequencies. If an input contains a sinusoidal modulation,

$$x(\boldsymbol{\rho}) = x_0 + A\cos(2\pi\mathbf{R}\cdot\boldsymbol{\rho}), \tag{6.1450}$$

clipping the peaks does not merely reduce the amplitude at $\mathbf{R}$; it also generates harmonics at $2\mathbf{R}$, $3\mathbf{R}$, and higher frequencies.

The concept of one global MTF is therefore meaningful only in a regime where the response can be linearized around the operating point.

Electron-counting detectors exhibit an analogous loss of linearity at high event density, although its microscopic origin is different. There the limitation is not necessarily the full-well capacity of a pixel, but the inability to distinguish overlapping electron events. That effect was treated statistically in Sec. 6.6.5 and should not be confused with analog saturation.

6.11.2 Dead time

A counting system may also become temporarily unable to register a new event immediately after recording a previous one. The interval during which this occurs is called the *dead time*, denoted here by τ_{d}.

The simplest model assumes that, after one event is recorded, the detector is insensitive for a fixed interval τ_{d}. If the incident event rate is r_{in}, a non-paralyzable dead-time model gives

$$\boxed{r_{\mathrm{out}} = \frac{r_{\mathrm{in}}}{1 + r_{\mathrm{in}}\tau_{\mathrm{d}}}.} \tag{6.1451}$$

The origin of this expression is particularly simple. Each recorded event occupies, on average, a time

$$\tau_{\mathrm{d}} + \frac{1}{r_{\mathrm{in}}}, \tag{6.1452}$$

consisting of the dead interval followed by the mean waiting time for the next acceptable incident event. The inverse of this mean cycle time gives Eq. (6.1451).

Another standard idealization is the *paralyzable* model, in which every incident event restarts the dead-time interval, even if it arrives while the detector is already insensitive. An event is recorded only if no other event has arrived during the preceding interval τ_{d}. For Poisson arrivals this probability is

$$\exp\left(-r_{\mathrm{in}}\tau_{\mathrm{d}}\right), \tag{6.1453}$$

giving

$$\boxed{r_{\mathrm{out}} = r_{\mathrm{in}} \exp\left(-r_{\mathrm{in}}\tau_{\mathrm{d}}\right).} \tag{6.1454}$$

These models are useful conceptual limits rather than literal descriptions of every cryo-EM camera. In a modern direct detector, the relevant temporal exclusion can arise from the internal frame time, charge-cloud duration, event-detection window, or digital clustering algorithm.

Dead time is closely related to coincidence loss but emphasizes the temporal aspect of the problem. Coincidence loss occurs when events overlap

within both a sufficiently small spatial neighbourhood and a sufficiently short time interval. Dead time may therefore be regarded as one mechanism that creates temporal exclusion between recorded events.

Because the probability of recording an event depends on recent event history, the output is no longer a simple thinned Poisson process and successive counts are temporally correlated.

6.11.3 Persistence

The opposite temporal effect can also occur: instead of becoming temporarily insensitive after an event, part of the detector response may persist into subsequent frames.

Possible microscopic origins include incomplete charge removal, trapping and delayed release of carriers, slowly relaxing electronic offsets, or memory introduced by the readout circuitry.

A simple linear model writes the recorded signal in frame t as

$$\boxed{y_t = \sum_{\ell=0}^{\infty} k_\ell x_{t-\ell} + n_t,} \tag{6.1455}$$

where k_ℓ is the temporal impulse response of the detector.

For an ideal detector with no persistence,

$$k_0 = 1, \qquad k_{\ell>0} = 0. \tag{6.1456}$$

With persistence, some coefficients with $\ell > 0$ are non-zero. For example,

$$y_t = x_t + \alpha x_{t-1} + \alpha^2 x_{t-2} + \cdots, \qquad 0 < \alpha < 1, \tag{6.1457}$$

describes an exponentially decaying memory.

An electron arriving in one physical exposure interval can then contribute to several recorded frames.

This effect has two consequences. First, the frames are no longer statistically independent. Second, the temporal sequence is filtered by a transfer function

$$K(\nu_t) = \sum_{\ell} k_\ell e^{-j2\pi\nu_t \ell \Delta t}, \tag{6.1458}$$

where ν_t is temporal frequency and Δt is the frame interval.

Thus a complete description of a movie may require a *spatio-temporal* spectrum,

$$\mathrm{NPS}(\mathbf{R}, \nu_t), \tag{6.1459}$$

rather than only the spatial NPS($\mathbf{R}$) considered previously.

This distinction can matter in cryo-EM movie processing because motion correction, dose weighting, and frame averaging all operate on the temporal sequence of detector outputs.

6.11.4 Temporal response and temporal sampling

Persistence, discussed in Sec. 6.11.3, is one particular manifestation of a more general fact: the response of a detector need not be instantaneous. An electron arriving at time t_0 may contribute to the measured signal over a finite interval of time.

In the linear time-invariant regime, this temporal behavior can be described in exactly the same way as the spatial response of the detector. If an impulsive input at time t_0 produces the temporal response

$$h_t(t - t_0), \tag{6.1460}$$

then a general time-dependent input $x(t)$ produces

$$\boxed{y(t) = \int_{-\infty}^{\infty} h_t(t - t')x(t')\,dt' = h_t * x.} \tag{6.1461}$$

Thus the temporal impulse response $h_t(t)$ plays the same role in time that the detector point-spread function $h_{\mathrm{det}}(\boldsymbol{\rho})$ plays in space.

The correspondence is

$$\boxed{\begin{aligned} h_{\mathrm{det}}(\boldsymbol{\rho}) &\longleftrightarrow h_t(t), \\ H_{\mathrm{det}}(\mathbf{R}) = \mathcal{F}_{\boldsymbol{\rho}}\left\{h_{\mathrm{det}}\right\} &\longleftrightarrow H_t(\nu_t) = \mathcal{F}_t\left\{h_t\right\}, \\ \mathbf{R}\ [\mathrm{cycles\,m^{-1}}] &\longleftrightarrow \nu_t\ [\mathrm{Hz}]. \end{aligned}} \tag{6.1462}$$

A broad spatial response suppresses rapidly varying spatial structure. Similarly, a broad temporal response suppresses rapidly varying temporal structure. Persistence corresponds to the particular case in which $h_t(t)$ has a non-negligible tail extending after the primary response.

The temporal-convolution model is nevertheless restricted to *linear* temporal memory. Dead time and coincidence loss are different: the probability of recording one event depends on the presence of other events and therefore cannot, in general, be described by convolution with a fixed $h_t(t)$. These nonlinear effects were treated in Sec. 6.11.2 and Sec. 6.6.5.

Frame integration. Even if the intrinsic temporal response $h_t(t)$ were arbitrarily narrow, the formation of a conventional movie frame introduces an additional temporal response because the detector output is integrated over a finite acquisition interval.

$$y_k = \frac{1}{T}\int_{t_k-T/2}^{t_k+T/2} x(t)\,dt. \tag{6.1463}$$

This is a temporal convolution with the normalized rectangular kernel

$$h_T(t) = \frac{1}{T}\,\mathrm{rect}\left(\frac{t}{T}\right). \tag{6.1464}$$

Temporal transfer function of frame integration.

The Fourier transform of $h_T(t)$ is

$$H_T(\nu_t) = \int_{-\infty}^{\infty} h_T(t)e^{-j2\pi\nu_t t}\,dt \tag{6.1465}$$

$$= \frac{1}{T}\int_{-T/2}^{T/2} e^{-j2\pi\nu_t t}\,dt \tag{6.1466}$$

$$= \boxed{\mathrm{sinc}(\nu_t T),} \tag{6.1467}$$

where

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}. \tag{6.1468}$$

Finite frame integration is therefore a temporal low-pass filter, with

$$\boxed{\mathrm{MTF}_t(\nu_t) = |\mathrm{sinc}(\nu_t T)|\,.} \tag{6.1469}$$

This is the temporal analogue of the finite pixel aperture discussed for spatial sampling.

Temporal sampling. Frame integration and frame sampling are distinct operations. If successive frame values are separated by Δt, the temporal sampling frequency is

$$\nu_{\mathrm{s}} = \frac{1}{\Delta t}, \tag{6.1470}$$

and the corresponding temporal Nyquist frequency is

$$\boxed{\nu_{\mathrm{N},t} = \frac{1}{2\Delta t}.} \tag{6.1471}$$

The analogy with spatial acquisition is therefore

$$\boxed{\begin{array}{rcl} \text{finite pixel aperture} & \longleftrightarrow & \text{finite frame duration,} \\ \text{spatial sampling lattice} & \longleftrightarrow & \text{frame-time sampling lattice,} \\ R_{\mathrm{N}} & \longleftrightarrow & \nu_{\mathrm{N},t}. \end{array}} \tag{6.1472}$$

Temporal frequencies above $\nu_{\mathrm{N},t}$ cannot be represented uniquely in the sampled movie and are aliased into the fundamental temporal-frequency interval.

Intrinsic temporal response. The rectangular integration window describes the temporal filtering introduced deliberately by frame formation. The detector itself may have an additional intrinsic temporal impulse response $h_{\mathrm{int}}(t)$.

If the two effects are linear, their combined response is

$$h_{\mathrm{eff}}(t) = h_T * h_{\mathrm{int}}(t), \tag{6.1473}$$

and therefore

$$\boxed{H_{\mathrm{eff}}(\nu_t) = H_T(\nu_t) H_{\mathrm{int}}(\nu_t).} \tag{6.1474}$$

A slowly decaying intrinsic response causes signal from one exposure interval to leak into subsequent frames. The corresponding persistence and temporal memory effects were discussed in Sec. 6.11.3.

Temporal noise correlations. The same distinction between independent and correlated fluctuations that led to the spatial NPS also applies in time. For a temporally and spatially stationary random field, define the spatio-temporal covariance

$$C_n(\Delta\boldsymbol{\rho}, \Delta t) = \mathbb{E}\left[n(\boldsymbol{\rho}, t) n(\boldsymbol{\rho} + \Delta\boldsymbol{\rho}, t + \Delta t)\right]. \tag{6.1475}$$

Its Fourier transform defines the spatio-temporal noise power spectrum,

$$\boxed{\mathrm{NPS}(\mathbf{R}, \nu_t) = \mathcal{F}_{\Delta\boldsymbol{\rho}, \Delta t}\left\{C_n(\Delta\boldsymbol{\rho}, \Delta t)\right\}.} \tag{6.1476}$$

The purely spatial NPS($\mathbf{R}$) used in the preceding sections is therefore a reduced description that does not retain the full temporal correlation structure.

One practical consequence is that averaging frames reduces noise as $1/M$ only when the frame fluctuations are mutually uncorrelated. In general,

$$\boxed{\mathrm{Var}\left(\frac{1}{M}\sum_{k=1}^{M} n_k\right) = \frac{1}{M^2}\sum_{k=1}^{M}\sum_{\ell=1}^{M}\mathrm{Cov}(n_k, n_\ell).} \tag{6.1477}$$

Event timing. Electron-counting detectors may preserve temporal information at a finer scale than the final movie frames. Conceptually, the raw detector output may contain events

$$\boxed{\left\{(\widehat{\boldsymbol{\rho}}_i, \widehat{t}_i)\right\}_{i=1}^{N},} \tag{6.1478}$$

where $\widehat{\boldsymbol{\rho}}_i$ and $\widehat{t}_i$ denote the estimated position and time of event i.

The temporal coordinate need not be a continuously measured timestamp; it may simply identify the internal detector frame or temporal group in which the event occurred. Nevertheless, if this temporal resolution is finer than the final movie-frame duration, the detected events can subsequently be regrouped into different temporal intervals.

This is closely analogous to spatial super-resolution: information is preserved at a finer sampling scale before the final image representation is constructed.

Unified spatio-temporal response. When the spatial and temporal detector responses are approximately separable,

$$h_{\mathrm{det}}(\boldsymbol{\rho}, t) = h_s(\boldsymbol{\rho})h_t(t), \tag{6.1479}$$

the corresponding transfer function factorizes:

$$\boxed{H_{\mathrm{det}}(\mathbf{R}, \nu_t) = H_s(\mathbf{R})H_t(\nu_t).} \tag{6.1480}$$

This provides a compact unified description of the linear spatial and temporal response of the detector. Nonlinear mechanisms such as saturation, dead time, and coincidence loss remain outside this convolutional framework.

6.11.5 Bad pixels and columns

Some detector elements may respond abnormally or fail completely. Examples include dead pixels, hot pixels, pixels with anomalous gain or dark current, and entire defective rows or columns.

A dead pixel may be represented ideally as

$$g_p \simeq 0, \tag{6.1481}$$

whereas a hot pixel can have an abnormally large offset or dark-current rate.

These defects are not naturally described as stationary random noise. A permanently dead pixel at a known location is a deterministic property of the detector.

If m_p is a binary validity mask,

$$m_p = \begin{cases} 1, & \text{usable pixel}, \\ 0, & \text{defective pixel}, \end{cases} \tag{6.1482}$$

then the simplest observation model is

$$\boxed{y_p = m_p y_p^{\text{ideal}}.} \tag{6.1483}$$

Multiplication by the mask in real space becomes convolution in Fourier space. A defective pixel pattern can therefore mix spatial frequencies rather than merely add a smooth noise spectrum.

A defective column is particularly anisotropic and can generate characteristic line-like structures in the two-dimensional Fourier transform.

In practical data acquisition, bad pixels and columns are usually identified through calibration and subsequently masked, replaced, or interpolated. Such correction may be entirely appropriate, but the corrected values should not be interpreted as independent detector measurements: interpolation introduces correlations with neighbouring pixels.

For a small fraction of defective elements, a global MTF, NPS, and DQE remain useful summaries after appropriate masking. For a large or structured defect pattern, however, the detector is no longer well described as spatially stationary.

6.11.6 Radiation damage of the detector

The detector itself is continuously irradiated by the same high-energy electrons that form the image. Unlike the biological specimen, the semiconductor is designed to tolerate very large accumulated doses, but its properties are not perfectly invariant under irradiation.

Energetic electrons can alter a semiconductor through several classes of process. Atomic displacements can create defects in the crystal lattice, while ionization in insulating oxides and interfaces can produce trapped charge. These changes can modify leakage current, charge collection, transistor thresholds, local gain, dark current, and the density of defective pixels.

The important distinction from specimen radiation damage is therefore one of scale rather than principle:

$$\boxed{\text{accumulated detector dose} \quad \Longrightarrow \quad \text{slow evolution of detector parameters.}} \tag{6.1484}$$

A detector parameter such as the pixel gain should then be written more generally as

$$g_p = g_p(D_{\mathrm{acc}}), \tag{6.1485}$$

where D_{acc} is the accumulated detector exposure.

Similarly,

$$b_p = b_p(D_{\mathrm{acc}}), \qquad r_{\mathrm{d},p} = r_{\mathrm{d},p}(D_{\mathrm{acc}}), \tag{6.1486}$$

may evolve with detector age.

Radiation damage therefore does not usually appear as a new frame-to-frame noise term. It manifests primarily as *drift in the statistical properties and calibration parameters of the detector*. Flat-field, dark-current, bad-pixel, MTF, and DQE calibrations can consequently become outdated as the detector ages.

This is one reason why detector calibration must be regarded as a time-dependent characterization rather than a once-and-for-all measurement.

6.11.7 Spatially varying response

The most general consequence of the preceding effects is that the detector response may depend on position.

For a shift-invariant detector, the response to an electron arriving at $\boldsymbol{\rho}_0$ is

$$h_{\mathrm{det}}(\boldsymbol{\rho} - \boldsymbol{\rho}_0). \tag{6.1487}$$

The same response is merely translated when the electron arrival position changes.

A spatially varying detector instead requires

$$\boxed{h(\boldsymbol{\rho}; \boldsymbol{\rho}_0),} \tag{6.1488}$$

where both the observation position and the electron entrance position must be retained separately.

The mean detector output is then

$$\boxed{\overline{y}(\boldsymbol{\rho}) = \int h(\boldsymbol{\rho};\boldsymbol{\rho}_0)x(\boldsymbol{\rho}_0)\,d^2\boldsymbol{\rho}_0.} \tag{6.1489}$$

Because the kernel is no longer a function only of $\boldsymbol{\rho}-\boldsymbol{\rho}_0$, the operation is not a convolution. Consequently, there is no single transfer function $H_{\mathrm{det}}(\mathbf{R})$ that exactly describes the complete sensor.

The noise covariance must likewise be written in its general form,

$$\boxed{K_n(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2) = \mathbb{E}\left[n(\boldsymbol{\rho}_1)n(\boldsymbol{\rho}_2)\right],} \tag{6.1490}$$

rather than as a function only of the displacement $\boldsymbol{\rho}_2-\boldsymbol{\rho}_1$.

The Fourier transform no longer diagonalizes this covariance exactly. Noise at different spatial frequencies can therefore become correlated:

$$\mathbb{E}\left[\widehat{n}(\mathbf{R})\widehat{n}^*(\mathbf{R}')\right] \neq 0 \qquad \text{for } \mathbf{R}\neq\mathbf{R}'. \tag{6.1491}$$

This is the mathematical point at which the scalar quantities developed earlier cease to provide a complete description.

Local characterization. If the detector properties vary slowly across the field of view, the shift-invariant approximation may still be applied locally. Around a position $\boldsymbol{\rho}_0$, one may define an approximate local PSF

$$h_{\mathrm{loc}}(\Delta\boldsymbol{\rho};\boldsymbol{\rho}_0), \tag{6.1492}$$

and hence a local transfer function

$$H_{\mathrm{loc}}(\mathbf{R};\boldsymbol{\rho}_0). \tag{6.1493}$$

Similarly, local flat-field regions can be used to estimate

$$\mathrm{NPS}(\mathbf{R};\boldsymbol{\rho}_0) \tag{6.1494}$$

and consequently a position-dependent

$$\mathrm{DQE}(\mathbf{R};\boldsymbol{\rho}_0). \tag{6.1495}$$

A single global MTF, NPS, or DQE then represents some effective average over the detector area rather than an exact property of every pixel.

When is the global description adequate? The global stationary description remains extremely useful when

$$\boxed{\begin{array}{c}\text{the detector is operated well inside its linear range,}\\ \text{temporal memory is negligible,}\\ \text{defective pixels occupy only a small masked fraction,}\\ \text{and spatial calibration residuals are small compared with}\\ \text{the stochastic fluctuations of interest.}\end{array}} \tag{6.1496}$$

Under these conditions, the detector can be approximated by one frequency-dependent transfer function and one stationary NPS, making the MTF–NPS–DQE–NEQ framework developed in the preceding sections both compact and physically meaningful.

Outside this regime, however, the more general description is

$$\boxed{\begin{array}{c}\text{space- and time-dependent mean response}\\ +\\ \text{full spatio-temporal covariance.}\end{array}} \tag{6.1497}$$

This distinction will be important when the detector is incorporated into the complete cryo-EM forward model: a single multiplicative MTF($\mathbf{R}$) and an additive stationary noise term are excellent first approximations, but they are not universal laws of the detector.

6.12 Experimental characterization of an electron detector

The quantities introduced in the preceding sections—gain, MTF, NPS, DQE, and NEQ—are not merely theoretical constructs. They can be measured experimentally and are routinely used to characterize electron detectors.

Interpreting such measurements nevertheless requires care. Detector performance depends on the accelerating voltage, incident electron fluence, dose rate, frame rate, detector temperature, readout mode, counting algorithm, gain correction, binning or super-resolution representation, and other acquisition parameters. Consequently, an MTF or DQE curve is not an immutable property of the sensor alone: it characterizes the complete detector system under specified operating conditions.

A typical characterization therefore requires three experimentally distinct measurements:

$$\boxed{\begin{array}{ccc} \text{mean response and gain calibration,} & & \\ \text{response to a known spatial structure} & \Longrightarrow & \text{MTF,} \\ \text{fluctuations under uniform illumination} & \Longrightarrow & \text{NPS}\,. \end{array}} \tag{6.1498}$$

The DQE is then obtained by combining the measured MTF and NPS with an independent measurement of the incident electron fluence (McMullan et al., 2009a; Ruskin et al., 2013b).

6.12.1 Measurement of the gain

The word *gain* can refer to several different stages of the detector. The conversion gain of the sensing node relates collected charge to voltage, the electronic gain relates voltage to ADC units, and the complete system gain relates incident or detected electrons to the final numerical image values.

For detector characterization it is useful to define a mean system conversion factor

$$g_{\text{sys}} = \frac{\text{mean detector output above offset}}{\text{mean number of incident electrons}}. \tag{6.1499}$$

If a uniform exposure contains on average μ_0 incident electrons per output pixel and produces a mean detector value $\overline{I}$ with dark offset $\overline{b}$, then

$$\boxed{g_{\text{sys}} = \frac{\overline{I} - \overline{b}}{\mu_0}.} \tag{6.1500}$$

The units might, for example, be analog-to-digital units per incident electron (ADU/e$^-$), where an ADU is one unit of the numerical output produced by the analog-to-digital converter. If some incident electrons are not detected, this system gain contains both the mean response of a detected event and the detection efficiency. It should therefore not automatically be identified with the microscopic charge-to-voltage gain of the sensor.

Relative pixel-gain calibration. A flat-field exposure can also be used to determine the relative sensitivity of individual pixels. Let $\overline{I}_p$ be the average signal of pixel p over many uniform exposures and let $\overline{b}_p$ be its

mean dark offset. A relative gain map may be estimated as

$$\boxed{G_p = \frac{\overline{I}_p - \overline{b}_p}{\left\langle \overline{I}_p - \overline{b}_p \right\rangle_p}.} \tag{6.1501}$$

For a perfectly uniform detector,

$$G_p = 1 \tag{6.1502}$$

for every pixel.

Many frames should be averaged so that the Poisson fluctuations of the calibration exposure become small compared with the actual pixel-to- pixel gain variation. The gain reference itself therefore has a finite statistical uncertainty, as discussed previously when considering residual flat-field noise.

The variance–mean method and its limitation. For detectors in which each incident quantum produces a deterministic output g, the gain can also be estimated from a series of uniform exposures. Let

$$Y = b + gN + n_{\mathrm{el}}, \tag{6.1503}$$

where

$$N \sim \mathrm{Poisson}(\mu) \tag{6.1504}$$

and the electronic noise has variance σ_{el}^2.

Then

$$\mathbb{E}[Y] - b = g\mu \tag{6.1505}$$

and

$$\mathrm{Var}(Y) = g^2\mu + \sigma_{\mathrm{el}}^2. \tag{6.1506}$$

Eliminating μ gives

$$\boxed{\mathrm{Var}(Y) = g\left(\mathbb{E}[Y] - b\right) + \sigma_{\mathrm{el}}^2.} \tag{6.1507}$$

Thus the slope of the variance–mean relation is g.

For an integrating electron detector, however, the response of one primary electron is not deterministic. Let its total response be the random variable Q, with mean $\overline{Q}$.

What a variance–mean measurement measures for a random single-electron response.

For a Poisson number of incident electrons,

$$S = \sum_{i=1}^{N} Q_i, \qquad N \sim \text{Poisson}(\mu). \tag{6.1508}$$

The compound-Poisson results derived previously give

$$\mathbb{E}[S] = \mu \overline{Q} \tag{6.1509}$$

and

$$\text{Var}(S) = \mu \mathbb{E}[Q^2]. \tag{6.1510}$$

Eliminating μ,

$$\boxed{\text{Var}(S) = \frac{\mathbb{E}[Q^2]}{\mathbb{E}[Q]} \mathbb{E}[S].} \tag{6.1511}$$

Using the single-electron excess-noise factor,

$$\mathcal{F}_{\text{SER}} = \frac{\mathbb{E}[Q^2]}{\mathbb{E}[Q]^2}, \tag{6.1512}$$

the measured slope is

$$\boxed{g_{\text{var/mean}} = \overline{Q} \mathcal{F}_{\text{SER}}.} \tag{6.1513}$$

Thus a variance–mean measurement does not, in general, return the mean single-electron response $\overline{Q}$. Event-to-event fluctuations increase the slope.

At the individual-pixel level the situation is still more complicated because charge sharing introduces covariance between neighboring pixels. Absolute gain calibration is therefore preferably based on a known incident electron fluence or, where possible, on measurements of isolated single-electron events rather than on an unqualified variance–mean slope.

6.12.2 Measurement of the MTF

The MTF is measured by presenting the detector with a spatial signal whose input structure is known and comparing the amplitude of the recorded mod-

ulation with that of the incident modulation.

In principle, one could measure the detector point-spread function directly. If a sufficiently localized electron input could be placed at a known position $\boldsymbol{\rho}_0$, the mean recorded response would give

$$h_{\mathrm{det}}(\boldsymbol{\rho} - \boldsymbol{\rho}_0), \tag{6.1514}$$

and its Fourier transform would give the detector transfer function.

For direct electron detectors, isolated electron events can provide useful information about the spatial extent of the single-electron response. However, the exact entrance position of each electron is not normally known, which makes a direct PSF measurement more difficult than it first appears.

A more practical method is to image a sharp edge.

Knife-edge and silhouette methods Consider an ideal straight edge parallel to the y axis. The incident intensity is a step,

$$x(x, y) = I_0 H(x), \tag{6.1515}$$

where $H(x)$ is the *Heaviside step function*,

$$\boxed{H(x) = \begin{cases} 0, & x < 0, \\ 1, & x > 0. \end{cases}} \tag{6.1516}$$

Thus one side of the edge receives no intensity, whereas the other side receives the uniform intensity I_0. The value assigned to $H(0)$ is irrelevant here because changing the function at a single point does not affect the continuous convolution or its Fourier transform.

Let the two-dimensional detector PSF be $h(x, y)$ and define its one-dimensional line-spread function perpendicular to the edge as

$$\boxed{L(x) = \int_{-\infty}^{\infty} h(x, y)\, dy.} \tag{6.1517}$$

The measured profile across the edge is called the *edge-spread function*, or ESF. Since the incident intensity is $I_0 H(x)$, the recorded profile is the convolution of the step with the line-spread function defined in Eq. (6.1517):

$$\begin{aligned} \mathrm{ESF}(x) &= I_0 (H * L)(x) \\ &= I_0 \int_{-\infty}^{\infty} H(x - x') L(x')\, dx' \\ &= I_0 \int_{-\infty}^{x} L(x')\, dx'. \end{aligned} \tag{6.1518}$$

Differentiating with respect to x therefore gives

$$\boxed{\frac{d}{dx}\,\mathrm{ESF}(x) = I_0 L(x).} \tag{6.1519}$$

Thus, the derivative of the edge-spread function is proportional to the line-spread function. This relation provides a convenient experimental route to the detector MTF.

MTF from the edge-spread function.

From Eq. (6.1519), taking Fourier transforms gives

$$\mathcal{F}\left\{\frac{d}{dx}\,\mathrm{ESF}(x)\right\} = I_0 \widehat{L}(R_x). \tag{6.1520}$$

Using the differentiation property of the Fourier transform,

$$\boxed{\mathcal{F}\left\{\frac{df}{dx}\right\}(R_x) = j2\pi R_x \widehat{f}(R_x),} \tag{6.1521}$$

we obtain

$$j2\pi R_x \widehat{\mathrm{ESF}}(R_x) = I_0 \widehat{L}(R_x). \tag{6.1522}$$

Since

$$\widehat{L}(R_x) = H_{\mathrm{det}}(R_x, 0), \tag{6.1523}$$

the MTF is the modulus of this transfer function normalized at zero frequency:

$$\boxed{\mathrm{MTF}(R_x) = \frac{\left|\mathcal{F}\left\{d\ \mathrm{ESF}\,/dx\right\}(R_x)\right|}{\left|\mathcal{F}\left\{d\ \mathrm{ESF}\,/dx\right\}(0)\right|}.} \tag{6.1524}$$

For the ideal edge considered here,

$$\int_{-\infty}^{\infty} \frac{d}{dx}\,\mathrm{ESF}(x)\,dx = I_0, \tag{6.1525}$$

so the denominator is simply I_0. Therefore, for $R_x \neq 0$,

$$\boxed{\mathrm{MTF}(R_x) = \frac{2\pi|R_x|\left|\widehat{\mathrm{ESF}}(R_x)\right|}{I_0}.} \tag{6.1526}$$

The value at $R_x = 0$ is defined by continuity as

$$\mathrm{MTF}(0) = 1. \tag{6.1527}$$

For an isotropic detector this slice depends only on $R = |R_x|$ and can be interpreted as the radial MTF. For an anisotropic detector, edges with different orientations are needed to characterize different directions.

In practice, the edge is often placed at a small angle relative to the pixel rows or columns. Different detector rows then sample the edge at different subpixel phases, allowing a finely sampled ESF to be reconstructed even though the physical detector lattice is comparatively coarse. This is usually called a *slanted-edge* method.

The edge itself must be sufficiently sharp, straight, and opaque at the electron energy used. Edge roughness, scattering from the test object, mechanical drift, or an imperfect knowledge of the input edge can all bias the inferred MTF.

In transmission electron microscopy, a practical alternative is to use the silhouette of a microscope beam stop or another strongly absorbing object and infer the detector blur from the difference between the known binary silhouette and the measured image (Ruskin et al., 2013b). This avoids the requirement for a specially manufactured microscopic knife edge. Such beam-stop methods have been used to measure MTF and DQE directly on installed TEM detectors.

6.12.3 Measurement of the NPS

The NPS is measured from nominally uniform detector exposures recorded under the same conditions for which detector performance is to be characterized.

Let a flat-field image contain $N_x \times N_y$ pixels with spacings a_x and a_y. After removing the mean, define

$$n_{mn} = I_{mn} - \overline{I}. \tag{6.1528}$$

With the discrete Fourier-transform convention

$$\hat{n}_{kl} = \sum_{m=0}^{N_x-1} \sum_{n=0}^{N_y-1} n_{mn} e^{-j2\pi(km/N_x + ln/N_y)}, \tag{6.1529}$$

an NPS estimate is

$$\boxed{\widehat{\mathrm{NPS}}_{kl} = \frac{a_x a_y}{N_x N_y} \left|\hat{n}_{kl}\right|^2 .} \tag{6.1530}$$

A single periodogram has large statistical fluctuations. In practice, many regions and preferably many independent flat-field images are analyzed and their periodograms averaged.

Care must be taken with removal of the image mean or of slowly varying illumination gradients. Subtracting a fitted trend suppresses genuine low-frequency power as well as unwanted illumination variation. Consequently, the preprocessing applied before NPS estimation must always be stated.

Difference-image method. Stable gain and offset patterns can contaminate the NPS even though they are not stochastic fluctuations. A useful way to suppress such fixed structure is to record two statistically equivalent flat-field images, I_1 and I_2, and form

$$D = I_1 - I_2. \tag{6.1531}$$

If the stochastic fluctuations in the two images are independent and have the same NPS,

$$\mathrm{NPS}_D = 2\,\mathrm{NPS}_I\,. \tag{6.1532}$$

Hence

$$\boxed{\mathrm{NPS}_I = \frac{1}{2}\,\mathrm{NPS}_D\,.} \tag{6.1533}$$

This technique is widely used in experimental detector characterization because stable fixed-pattern structure cancels in the subtraction (Ruskin et al., 2013b).

The assumption of independence is important. If successive frames are temporally correlated because of persistence or shared electronic fluctuations, subtraction can also cancel part of the stochastic noise and thereby underestimate the true NPS.

The two-dimensional NPS should normally be inspected before radial averaging. Row noise, column noise, readout blocks, gain-correction artifacts, or other anisotropies can be obvious in two dimensions but almost invisible after radial averaging.

6.12.4 Measurement of the DQE

The most fundamental definition of the DQE is the SNR ratio derived in Sec. 6.9. Experimentally, however, input and output SNR values are inconvenient to measure directly for every spatial frequency. Instead, DQE is normally calculated from the measured MTF and NPS.

Using Eq. (6.1305),

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{\Phi_0 \overline{Q}^{\,2}\,\mathrm{MTF}^2(\mathbf{R})}{\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}.} \tag{6.1534}$$

Since the mean flat-field output is

$$\overline{I} = \Phi_0 \overline{Q}, \tag{6.1535}$$

the experimentally convenient form is

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{\overline{I}^{\,2}\,\mathrm{MTF}^2(\mathbf{R})}{\Phi_0\,\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}.} \tag{6.1536}$$

Equivalently, with

$$\mathrm{NNPS}(\mathbf{R}) = \frac{\mathrm{NPS}_{\mathrm{out}}(\mathbf{R})}{\overline{I}^{\,2}}, \tag{6.1537}$$

we obtain

$$\boxed{\mathrm{DQE}(\mathbf{R}) = \frac{\mathrm{MTF}^2(\mathbf{R})}{\Phi_0\,\mathrm{NNPS}(\mathbf{R})}.} \tag{6.1538}$$

The MTF and NPS must refer to the *same detector output representation and operating conditions*. An MTF measured in super-resolution mode cannot, for example, be combined consistently with an NPS measured after physical-pixel binning.

Measurement of the incident fluence. The incident fluence Φ_0 is an input quantity and should be determined independently of the detector response being characterized. Otherwise a detector inefficiency can be inadvertently absorbed into the dose calibration and the resulting DQE becomes circular.

If a beam current I_{b} illuminates an area A for an exposure time T, the number of incident electrons is

$$N_{\mathrm{inc}} = \frac{I_{\mathrm{b}} T}{e}, \tag{6.1539}$$

and therefore

$$\boxed{\Phi_0 = \frac{I_{\mathrm{b}} T}{e A}.} \tag{6.1540}$$

In practice, electron-microscope dose calibrations may use a Faraday cup, a calibrated beam-current measurement, or a calibrated reference system. Independent verification of the incident dose is an important part of a quantitative DQE measurement (Ruskin et al., 2013b). Ruskin and colleagues, for example, verified microscope dose calibrations with a Faraday cup.

The special difficulty of DQE(0)**.** The zero-frequency DQE deserves special attention. Formally,

$$\mathrm{DQE}(0) \tag{6.1541}$$

describes the efficiency with which very slowly varying modulation is transmitted.

Experimentally, the exact DC component is often unusable because the mean has been removed from the flat field. Furthermore, only a small number of Fourier samples are available near the origin, so estimates of the very-low-frequency NPS have comparatively large uncertainty.

Consequently, DQE(0) is frequently estimated by extrapolation or by averaging over a small low-frequency interval rather than by using the single Fourier coefficient at $\mathbf{R} = 0$ (McMullan et al., 2009a; Ruskin et al., 2013b). Experimental studies of direct detectors have explicitly noted this difficulty.

For an idealized counting detector at negligible coincidence loss, DQE(0) approaches the electron-detection efficiency, as derived previously. This interpretation does not extend automatically to integrating detectors or to counting detectors operated in a nonlinear regime.

DQE at a specified spatial frequency. A complete detector characterization is the curve

$$\mathrm{DQE}(\mathbf{R}), \tag{6.1542}$$

not one scalar number.

Three values are nevertheless quoted particularly often:

$$\boxed{\begin{aligned} \mathrm{DQE}(0) &: \text{ low-frequency dose efficiency,} \\ \mathrm{DQE}(R) &: \text{ efficiency at a specified spatial frequency,} \\ \mathrm{DQE}(R_{\mathrm{N}}) &: \text{ efficiency at the Nyquist frequency.} \end{aligned}} \tag{6.1543}$$

These quantities can be very different for the same detector. In particular, a detector may have

$$\mathrm{DQE}(0) \simeq 0.9 \tag{6.1544}$$

while retaining only a few tenths of the input SNR near Nyquist.

The meaning of R_{N} must also be specified. For a conventional output it is the Nyquist frequency of the physical-pixel lattice, whereas for a super-resolution or finely sampled event representation it may refer to the Nyquist frequency of the finer output grid.

6.12.5 Dose-rate dependence

Detector characterization must distinguish *total exposure* from *dose rate*. Two measurements can contain the same total number of electrons but produce different detector statistics if those electrons arrive at different rates.

If the incident fluence is Φ_0 accumulated during time T, the average fluence rate is

$$\dot{\Phi} = \frac{\Phi_0}{T}. \tag{6.1545}$$

For a physical pixel of area $a_x a_y$, the corresponding mean incident rate per physical pixel is

$$r_{\mathrm{pix}} = \dot{\Phi} a_x a_y. \tag{6.1546}$$

In integrating detectors, high dose rate can lead to nonlinear response or saturation. In counting detectors, high dose rate increases the probability that two electron events overlap in the same spatial and temporal neighborhood and therefore increases coincidence loss.

Both effects modify the quantities entering the DQE:

$$\mathrm{MTF} = \mathrm{MTF}(\mathbf{R}; \dot{\Phi}), \tag{6.1547}$$

and

$$\mathrm{NPS} = \mathrm{NPS}(\mathbf{R}; \dot{\Phi}). \tag{6.1548}$$

Consequently,

$$\boxed{\mathrm{DQE} = \mathrm{DQE}(\mathbf{R}; \dot{\Phi}).} \tag{6.1549}$$

For counting detectors, increasing dose rate can suppress the low-frequency NPS because coincidence introduces short-range exclusion between recorded events. At the same time, counts and modulation are lost. A naïve insertion of a depressed NPS into the linear DQE formula can therefore produce a misleading result if the detector is already strongly nonlinear. This is why low-occupancy measurements are commonly used to determine the intrinsic counting performance, while characterization at realistic dose rate is needed to describe actual acquisition conditions (Li et al., 2013).

Integrating detectors can exhibit a different failure mode: saturation can decrease the measured variance and thereby produce an apparent increase in DQE even though information is actually being lost. Such behavior has been observed experimentally in direct-detector characterization.

A useful experimental protocol is therefore to hold the *total fluence* approximately constant while varying the exposure duration. Changes in MTF, NPS, and DQE can then be attributed primarily to dose-rate-dependent detector behavior rather than to a change in total counting statistics.

Small-signal characterization in a nonlinear regime. When dose-rate dependence is appreciable, the detector is not globally linear. An MTF and DQE can nevertheless be defined locally around a particular operating point by considering a small modulation $\delta\Phi$ around the mean fluence rate $\dot{\Phi}_0$:

$$\dot{\Phi} = \dot{\Phi}_0 + \delta\dot{\Phi}. \tag{6.1550}$$

The relevant gain and MTF are then differential, small-signal quantities evaluated around $\dot{\Phi}_0$. The NPS must be measured at the same mean dose rate. Thus a DQE curve obtained at one operating point should not automatically be applied to another.

6.12.6 How to interpret a reported detector specification

A statement such as

$$\text{“ DQE} = 0.8'' \tag{6.1551}$$

is incomplete.

At minimum, a quantitative detector-performance statement should identify

$$\boxed{\begin{array}{c}\text{accelerating voltage,}\\ \text{spatial frequency,}\\ \text{physical or super-resolution sampling,}\\ \text{dose rate and total exposure,}\\ \text{counting or integrating mode,}\\ \text{frame rate and relevant temporal processing,}\\ \text{gain, dark, and defect corrections,}\\ \text{and the method used to estimate MTF and NPS.}\end{array}} \tag{6.1552}$$

A more complete notation would therefore be something like

$$\boxed{\mathrm{DQE}\left(\mathbf{R}; E_0, \dot{\Phi}, \Phi_0, \mathrm{mode}, a_{\mathrm{out}}, \mathrm{processing}\right),} \tag{6.1553}$$

although such cumbersome notation is rarely written explicitly.

Manufacturers commonly report DQE as a function of normalized spatial frequency,

$$\xi = \frac{R}{R_{\mathrm{N}}}, \tag{6.1554}$$

with values such as

$$\mathrm{DQE}(0), \qquad \mathrm{DQE}\left(\frac{1}{2}R_{\mathrm{N}}\right), \qquad \mathrm{DQE}(R_{\mathrm{N}}). \tag{6.1555}$$

These numbers summarize different aspects of the detector. The zero-frequency value is dominated by the efficiency with which electron statistics are preserved at large spatial scales. The Nyquist value tests the combined effects of localization or charge spreading, pixel integration, noise correlations, and sampling at the highest frequency supported by the stated output representation.

Thus neither MTF nor DQE should be reduced to one headline number. Their frequency dependence, together with the conditions under which they were measured, is the experimentally meaningful description of detector performance.

6.13 Detector contribution to the complete cryo-EM forward model

We can now return to the image-formation model developed in the preceding chapters and include explicitly the final stage of the experiment: electron detection.

In Chap. 5, the weak-phase specimen model was combined with the electron optics of Chap. 4 to obtain the standard single-particle Fourier-space forward model

$$\widehat{y}_i(\mathbf{R}) = a_i \,\mathrm{CTF}_i(\mathbf{R}) e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i} \widehat{\mathcal{P}_{\mathcal{R}_i} V_0}(\mathbf{R}) + \widehat{n}_i(\mathbf{R}), \tag{6.1556}$$

as derived in Eq. (5.608) and summarized in Eq. (5.2222).

At that stage, however, the detector was represented only implicitly. The term $\widehat{n}_i$ collected all fluctuations not described by the deterministic image-formation model, while the finite spatial response of the detector was not written explicitly. The present chapter allows us to refine both aspects.

6.13.1 The mean image incident on the detector

Let

$$\boxed{\widehat{x}_i(\mathbf{R}) = a_i \,\mathrm{CTF}_i(\mathbf{R}) e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i} \widehat{\mathcal{P}_{\mathcal{R}_i} V_0}(\mathbf{R})} \tag{6.1557}$$

denote the mean electron-optical image reaching the detector plane, before the spatial response and stochasticity of the detector are included.

The function $\mathrm{CTF}_i(\mathbf{R})$ in Eq. (6.1557) is the *complete* electron-optical CTF derived in Sec. 4.15. In particular, Eq. (4.2117) gives

$$\mathrm{CTF}_i(\mathbf{R}) = E_i(\mathbf{R}) P_i(\mathbf{R}) e^{-j\chi_{\mathrm{asym},i}(\mathbf{R})} \left[\sqrt{1 - A_i^2} \sin \chi_{\mathrm{sym},i}(\mathbf{R}) + A_i \cos \chi_{\mathrm{sym},i}(\mathbf{R})\right], \tag{6.1558}$$

where E_i represents the combined attenuation envelopes.

It is important not to multiply Eq. (6.1557) by a second envelope E_i. The envelope is already contained in the definition of the complete CTF.

If instead one adopts the alternative convention

$$\boxed{\mathrm{CTF}_i(\mathbf{R}) = E_i(\mathbf{R}) \,\mathrm{CTF}_i^{(0)}(\mathbf{R}),} \tag{6.1559}$$

where $\mathrm{CTF}_i^{(0)}$ denotes the CTF before attenuation envelopes, then Eq. (6.1557) may equivalently be written as

$$\widehat{x}_i(\mathbf{R}) = a_i E_i(\mathbf{R}) \,\mathrm{CTF}_i^{(0)}(\mathbf{R}) e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i} \widehat{\mathcal{P}_{\mathcal{R}_i} V_0}(\mathbf{R}). \tag{6.1560}$$

The two conventions are equivalent, but they must not be mixed.

6.13.2 Mean detector transfer

The continuous detector response was described in Sec. 6.5. For a linear and shift-invariant detector, Eq. (6.531) gives

$$I_{\mathrm{det}} = h_{\mathrm{det}} * I_{\mathrm{in}}, \tag{6.1561}$$

and therefore, from Eq. (6.534),

$$\widehat{I}_{\mathrm{det}}(\mathbf{R}) = H_{\mathrm{det}}(\mathbf{R}) \widehat{I}_{\mathrm{in}}(\mathbf{R}). \tag{6.1562}$$

Finite pixel integration introduces an additional transfer function. As derived in Eq. (6.566),

$$\boxed{H_{\mathrm{pre}}(\mathbf{R}) = H_{\mathrm{det}}(\mathbf{R}) P_a(\mathbf{R}),} \tag{6.1563}$$

where P_a is the transfer function of the pixel aperture. Thus H_{pre} describes the complete continuous spatial response immediately before discrete sampling.

The mean presampled cryo-EM image is consequently

$$\boxed{\begin{aligned}\mathbb{E}\left[\widehat{y}_i^{\text{pre}}(\mathbf{R}) \mid V_0, \mathcal{R}_i, \mathbf{t}_i\right] = a_i H_{\text{pre},i}(\mathbf{R})\,\text{CTF}_i(\mathbf{R}) \\ \times e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i}\widehat{\mathcal{P}_{\mathcal{R}_i}V_0}(\mathbf{R}).\end{aligned}} \tag{6.1564}$$

Equation (6.1564) completes the mean linear image-formation model developed throughout the preceding chapters.

The order of the factors has a direct physical interpretation:

$$\boxed{\begin{array}{c} V_0 \\ \Downarrow \\ \mathcal{P}_{\mathcal{R}_i}V_0 \\ \Downarrow \\ \text{CTF}_i \\ \Downarrow \\ H_{\text{pre},i} \\ \Downarrow \\ \text{presampled detector image.}\end{array}} \tag{6.1565}$$

The CTF describes how the electron microscope converts specimen structure into intensity contrast. The detector transfer function describes how much of that already formed intensity pattern survives the spatial response of the detector.

These are physically different operations.

The detector transfer function is not, in general, identical to the MTF. The distinction between H_{det} and the MTF is important. From Eq. (6.535),

$$\text{MTF}_{\text{det}}(\mathbf{R}) = \frac{|H_{\text{det}}(\mathbf{R})|}{|H_{\text{det}}(\mathbf{0})|}. \tag{6.1566}$$

With the normalization adopted in Sec. 6.5.1, $H_{\text{det}}(\mathbf{0}) = 1$, and therefore

$$\boxed{\text{MTF}_{\text{det}}(\mathbf{R}) = |H_{\text{det}}(\mathbf{R})|.} \tag{6.1567}$$

Thus the MTF is the *magnitude* of the normalized detector transfer function. If the detector PSF is centrosymmetric and introduces no relevant

spatial phase shift, then H_{det} is real and one may often identify it numerically with the MTF over the frequency range of interest. In the general case, however,

$$H_{\mathrm{det}} \neq \mathrm{MTF}_{\mathrm{det}}, \tag{6.1568}$$

because H_{det} may contain a phase.

The complex transfer function must therefore be retained in the Fourier-space forward model Eq. (6.1564). In a power spectrum, by contrast, only its squared modulus appears.

The same distinction applies to the presampling transfer function. Using Eq. (6.583),

$$|H_{\mathrm{pre}}(\mathbf{R})| = \mathrm{MTF}_{\mathrm{pre}}(\mathbf{R}) \tag{6.1569}$$

under the normalization used here.

6.13.3 The recorded image is a stochastic quantity

Equation (6.1564) describes only the *mean* image. A real exposure is one stochastic realization of the detection process.

It is therefore more appropriate to write

$$\boxed{\widehat{y}_i^{\mathrm{pre}}(\mathbf{R}) = \widehat{\overline{y}}_i^{\mathrm{pre}}(\mathbf{R}) + \widehat{\eta}_i(\mathbf{R}),} \tag{6.1570}$$

where

$$\widehat{\overline{y}}_i^{\mathrm{pre}}(\mathbf{R}) = a_i H_{\mathrm{pre},i}(\mathbf{R})\,\mathrm{CTF}_i(\mathbf{R}) e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i}\widehat{\mathcal{P}_{\mathcal{R}_i}V_0}(\mathbf{R}) \tag{6.1571}$$

and

$$\mathbb{E}\left[\widehat{\eta}_i(\mathbf{R}) \mid \overline{y}_i\right] = 0. \tag{6.1572}$$

The apparently innocuous replacement

$$\widehat{n}_i \longrightarrow \widehat{\eta}_i \tag{6.1573}$$

contains much of the physics developed in this chapter.

In particular, η_i should *not* in general be interpreted as independent Gaussian noise in every pixel. Incident-electron shot noise is Poissonian; the response to one electron may fluctuate in amplitude and shape; neighbouring pixels may be correlated; electronic readout may introduce anisotropic correlations; and coincidence loss can make the complete mapping nonlinear.

For a wide-sense stationary noise field, its second-order statistics are described by the NPS introduced in Sec. 6.7. In Fourier space, schematically,

$$\boxed{\mathbb{E}\left[\widehat{\eta}_i(\mathbf{R})\widehat{\eta}_i^*(\mathbf{R}')\right] = \mathrm{NPS}_i(\mathbf{R})\,\delta^{(2)}\left(\mathbf{R}-\mathbf{R}'\right),} \tag{6.1574}$$

up to the normalization associated with the adopted Fourier convention.

Equation (6.1574) does not imply that the noise is white. White noise would additionally require

$$\mathrm{NPS}_i(\mathbf{R}) = \text{constant}. \tag{6.1575}$$

Nor does it imply Gaussianity: the NPS specifies second-order statistics, not the complete probability distribution.

If detector response varies across the sensor, or if strong nonstationarities are present, even Eq. (6.1574) is insufficient and the complete spatial covariance must be retained.

Sampling completes the digital image formation process. Equation (6.1564) is a *presampling* model. The final digital image is obtained only after sampling on the detector lattice.

As shown in Sec. 6.5.4, sampling generates translated spectral replicas according to Eq. (6.607). Consequently, the simple multiplicative Fourier-space model Eq. (6.1564) may be used directly within the Nyquist region only when aliasing from frequencies outside that region is negligible.

Thus the final stages are

$$\boxed{\begin{array}{c}\text{electron-optical image}\\ \Downarrow\\ \text{stochastic detector response}\\ \Downarrow\\ \text{pixel integration}\\ \Downarrow\\ \text{discrete sampling}\\ \Downarrow\\ \text{recorded digital cryo-EM image.}\end{array}} \tag{6.1576}$$

6.14 The power spectrum of a recorded cryo-EM image

The previous section completed the image-amplitude forward model. We can now complete the corresponding power-spectrum description.

In Sec. 4.16, the experimentally observed power spectrum was represented phenomenologically as

$$S_{\text{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})|\,\text{CTF}(\mathbf{R})|^2, \tag{6.1577}$$

as given in Eq. (4.2162). That representation was deliberately operational: $B(\mathbf{R})$ was defined in Eq. (4.2165) as the power not explained by the chosen CTF-modulated structural model.

We can now give this decomposition a more detailed physical interpretation.

6.14.1 Structural signal power

Let $S_s(\mathbf{R})$ denote the PSD of the particle-dependent structural component before electron-optical and detector transfer. The translation of particle i contributes only the phase factor

$$e^{-j2\pi\mathbf{R}\cdot\mathbf{t}_i}, \tag{6.1578}$$

whose modulus is one. It therefore disappears from the power spectrum.

Applying first the complete electron-optical CTF and then the presampling detector transfer gives

$$\boxed{S_{\text{struct,det}}(\mathbf{R}) = |H_{\text{pre}}(\mathbf{R})|^2|\,\text{CTF}(\mathbf{R})|^2 S_s(\mathbf{R}).} \tag{6.1579}$$

Under the normalized detector convention,

$$|H_{\text{pre}}(\mathbf{R})|^2 = \text{MTF}^2_{\text{pre}}(\mathbf{R}), \tag{6.1580}$$

so Eq. (6.1579) may equivalently be written as

$$S_{\text{struct,det}}(\mathbf{R}) = \text{MTF}^2_{\text{pre}}(\mathbf{R})|\,\text{CTF}(\mathbf{R})|^2 S_s(\mathbf{R}). \tag{6.1581}$$

If the alternative convention of Eq. (6.1559) is used, this becomes

$$\boxed{S_{\text{struct,det}}(\mathbf{R}) = |H_{\text{pre}}(\mathbf{R})|^2|E(\mathbf{R})|^2|\,\text{CTF}^{(0)}(\mathbf{R})|^2 S_s(\mathbf{R}).} \tag{6.1582}$$

This is the origin of the frequently encountered product

$$\text{MTF}^2\, E^2|\,\text{CTF}^{(0)}\,|^2. \tag{6.1583}$$

With the definition of the complete CTF used in Sec. 4.15, however, the factor E^2 is already contained in $|\,\text{CTF}\,|^2$.

6.14.2 Specimen background power

Not all specimen-generated power corresponds to the particle of interest.

In Sec. 4.16, the particle signal and the irreproducible elastic structural background from vitreous ice, supporting material, contaminants, and other specimen components were distinguished. Their pre-microscope PSDs were denoted $S_s(\mathbf{R})$ and $S_{\mathrm{el}}(\mathbf{R})$, respectively, and Eq. (4.2157) gave

$$S_{\mathrm{spec}}(\mathbf{R}) = S_s(\mathbf{R}) + S_{\mathrm{el}}(\mathbf{R}) \tag{6.1584}$$

when the two fields are uncorrelated.

The crucial point is that the elastic structural background is produced *before* image formation. It is therefore transferred by the CTF and by the detector in exactly the same way as other elastic structural information:

$$\boxed{S_{\mathrm{el,det}}(\mathbf{R}) = |H_{\mathrm{pre}}(\mathbf{R})|^2 |\,\mathrm{CTF}(\mathbf{R})|^2 S_{\mathrm{el}}(\mathbf{R}).} \tag{6.1585}$$

Combining Eqs. (6.1579) and (6.1585),

$$\boxed{\begin{aligned} S_{\mathrm{elastic,det}}(\mathbf{R}) = |H_{\mathrm{pre}}(\mathbf{R})|^2 |\,\mathrm{CTF}(\mathbf{R})|^2 \\ \times \left[S_s(\mathbf{R}) + S_{\mathrm{el}}(\mathbf{R})\right]. \end{aligned}} \tag{6.1586}$$

This is the detector-level version of Eq. (4.2159).

It also explains an important conceptual point: specimen *background* is not necessarily unmodulated background in the power spectrum. Elastic fluctuations of vitreous ice or supporting material are genuine specimen structure and therefore generate Thon rings.

Other specimen contributions cannot in general be represented by the same simple factor. Inelastically scattered electrons, for example, may undergo energy-dependent chromatic transfer and partial decoherence, as discussed in Sec. 5.12.12 of Chap. 5. Multiple scattering and other departures from the weak-object model likewise need not obey the same linear CTF.

We shall denote the recorded power from such specimen-side processes that is not represented explicitly by Eq. (6.1586) as

$$S_{\mathrm{spec,other,det}}(\mathbf{R}). \tag{6.1587}$$

6.14.3 Electron shot-noise power

The finite number of imaging electrons introduces fluctuations even for a perfect detector. Before detection, a homogeneous Poisson electron fluence Φ_0 has the white PSD derived in Eq. (6.1124),

$$\mathrm{NPS}_{\mathrm{shot,in}}(\mathbf{R}) = \Phi_0. \tag{6.1588}$$

For a deterministic integrating detector response, the spatial spreading of each electron filters these fluctuations in the same way that it filters the mean signal. Including finite pixel integration, Eq. (6.1127) gives

$$\boxed{S_{\text{shot,det}}(\mathbf{R}) = \Phi_0 \overline{Q}^2 |H_{\text{pre}}(\mathbf{R})|^2.} \tag{6.1589}$$

Thus white incident shot noise does not necessarily remain white after an integrating detector. Spatial spreading correlates neighbouring output values and reshapes its NPS.

Equation (6.1589) must nevertheless be interpreted as the deterministic-response limit. A real primary electron produces a stochastic response q_ω, and the exact electron-induced NPS derived in Eq. (6.1242) is

$$\boxed{\text{NPS}_{\text{electron}}(\mathbf{R}) = \Phi_0 \mathbb{E}\left[|\widehat{q}_\omega(\mathbf{R})|^2\right].} \tag{6.1590}$$

Using the mean–fluctuation decomposition derived in Eq. (6.1246),

$$\boxed{\begin{aligned}\text{NPS}_{\text{electron}}(\mathbf{R}) = & \underbrace{\Phi_0 \left|\overline{\widehat{q}}(\mathbf{R})\right|^2}_{\text{propagated incident shot noise}} \\ & + \underbrace{\Phi_0 \mathbb{E}\left[|\delta\widehat{q}_\omega(\mathbf{R})|^2\right]}_{\text{single-electron-response fluctuations}}.\end{aligned}} \tag{6.1591}$$

This distinction prevents double counting. Energy-deposition fluctuations, charge-generation fluctuations, charge sharing, and event-shape variations are not independent additional primary-electron processes: they are fluctuations of the response to the same incident electron.

6.14.4 Detector-added noise power

Additional fluctuations may arise in the detector and readout chain independently of the incident-electron point process.

Examples developed in this chapter include electronic readout noise, Sec. 6.8.6; dark-current fluctuations, Sec. 6.8.7; residual calibration effects, Sec. 6.8.8; and quantization, Sec. 6.8.9.

If these contributions are mutually uncorrelated, their NPS values add according to Eq. (6.1241). We may therefore write schematically

$$\boxed{\begin{aligned}\text{NPS}_{\text{det,add}}(\mathbf{R}) = \Phi_0 \mathbb{E}\left[|\delta\widehat{q}_\omega(\mathbf{R})|^2\right] \\ + \text{NPS}_{\text{elec}}(\mathbf{R}) + \text{NPS}_{\text{dark}}(\mathbf{R}) \\ + \text{NPS}_{\text{quant}}(\mathbf{R}) + \cdots.\end{aligned}} \tag{6.1592}$$

The first term is included here as detector-added stochasticity because it describes fluctuations around the mean response to one incident electron. It must not be added again if the full $\mathrm{NPS}_{\mathrm{electron}}$ of Eq. (6.1590) is used.

Not every detector imperfection fits into Eq. (6.1592). Fixed-pattern errors may behave as spatially varying deterministic structure rather than independent noise. More importantly, coincidence loss, Sec. 6.8.5, changes the point process itself and cannot in general be represented as an independent additive NPS.

6.14.5 The complete observed power spectrum

We can now assemble the different contributions.

Under the weak-object, linear detector, shift-invariant, and approximately stationary assumptions used above, and neglecting aliasing for the moment, the observed presampling PSD may be written as

$$\boxed{\begin{aligned} S_{\mathrm{obs}}(\mathbf{R}) = {} & \underbrace{|H_{\mathrm{pre}}(\mathbf{R})|^2 |\,\mathrm{CTF}(\mathbf{R})|^2 S_s(\mathbf{R})}_{\text{particle structural signal}} \\ & + \underbrace{|H_{\mathrm{pre}}(\mathbf{R})|^2 |\,\mathrm{CTF}(\mathbf{R})|^2 S_{\mathrm{el}}(\mathbf{R}) + S_{\mathrm{spec,other,det}}(\mathbf{R})}_{\text{specimen background}} \\ & + \underbrace{S_{\mathrm{shot,det}}(\mathbf{R})}_{\text{incident-electron shot noise}} \\ & + \underbrace{\mathrm{NPS}_{\mathrm{det,add}}(\mathbf{R})}_{\text{detector-added fluctuations}} . \end{aligned}} \tag{6.1593}$$

This is the microscopic version of the conceptual decomposition

$$\boxed{\begin{aligned} S_{\mathrm{obs}} = {} & S_{\mathrm{structural,det}} + S_{\mathrm{specimen\,background,det}} \\ & + S_{\mathrm{shot,det}} + S_{\mathrm{detector}} . \end{aligned}} \tag{6.1594}$$

Equation (6.1594) is useful for conceptual bookkeeping, but Eq. (6.1593) shows why its components must be interpreted carefully. Different physical contributions enter at different stages of the experiment and are therefore transferred differently.

Power spectrum of the detected image.

Consider the generic presampling model

$$\widehat{y}(\mathbf{R}) = H_{\text{pre}}(\mathbf{R})\widehat{x}(\mathbf{R}) + \widehat{\eta}(\mathbf{R}), \tag{6.1595}$$

with

$$\mathbb{E}[\widehat{\eta}(\mathbf{R}) \mid x] = 0. \tag{6.1596}$$

The squared modulus is

$$\begin{aligned} |\widehat{y}|^2 = |H_{\text{pre}}|^2|\widehat{x}|^2 + |\widehat{\eta}|^2 \\ + H_{\text{pre}}\widehat{x}\widehat{\eta}^* + H_{\text{pre}}^*\widehat{x}^*\widehat{\eta}. \end{aligned} \tag{6.1597}$$

Taking expectations, the two cross terms vanish because

$$\mathbb{E}\left[\widehat{x}\widehat{\eta}^*\right] = \mathbb{E}\left[\widehat{x}\,\mathbb{E}\left[\widehat{\eta}^* \mid x\right]\right] \tag{6.1598}$$

$$= 0. \tag{6.1599}$$

Therefore

$$\boxed{S_y(\mathbf{R}) = |H_{\text{pre}}(\mathbf{R})|^2 S_x(\mathbf{R}) + \text{NPS}_\eta(\mathbf{R}).} \tag{6.1600}$$

For the elastic specimen contribution, Eq. (4.2158) gives

$$S_x(\mathbf{R}) = |\,\text{CTF}(\mathbf{R})|^2 \left[S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R})\right]. \tag{6.1601}$$

Substitution into Eq. (6.1600) gives

$$\boxed{\begin{aligned} S_y(\mathbf{R}) = |H_{\text{pre}}(\mathbf{R})|^2|\,\text{CTF}(\mathbf{R})|^2 \\ \times \left[S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R})\right] + \text{NPS}_\eta(\mathbf{R}). \end{aligned}} \tag{6.1602}$$

The remaining task is precisely the one developed throughout this chapter: determine the physical composition and spatial-frequency dependence of NPS_η.

6.14.6 Recovering the empirical CTF-plus-background model

We can now return to the phenomenological model introduced in Sec. 4.16,

$$S_{\text{obs}}(\mathbf{R}) \simeq B(\mathbf{R}) + A(\mathbf{R})|\,\text{CTF}(\mathbf{R})|^2. \tag{6.1603}$$

For the elastic structural part, Eq. (6.1586) identifies

$$\boxed{A(\mathbf{R}) = |H_{\text{pre}}(\mathbf{R})|^2 \left[S_s(\mathbf{R}) + S_{\text{el}}(\mathbf{R})\right],} \tag{6.1604}$$

which is the presampling form of Eq. (4.2161).

The remaining terms contribute operationally to the fitted background:

$$\boxed{\begin{aligned} B(\mathbf{R}) \simeq\ & S_{\mathrm{spec,other,det}}(\mathbf{R}) + S_{\mathrm{shot,det}}(\mathbf{R}) \\ & + \mathrm{NPS}_{\mathrm{det,add}}(\mathbf{R}) + \cdots . \end{aligned}} \tag{6.1605}$$

Equations (6.1604) and (6.1605) explain why neither $A(\mathbf{R})$ nor $B(\mathbf{R})$ has a unique microscopic origin.

The function $A(\mathbf{R})$ contains not only the particle signal but also CTF-modulated elastic specimen structure, such as vitreous ice, and it is attenuated by the detector MTF.

The function $B(\mathbf{R})$ may contain inelastic and other non-CTF-like specimen contributions, electron shot noise, single-electron-response fluctuations, electronic noise, dark-current noise, quantization noise, residual calibration structure, and other effects.

Moreover, this division is not exact in all experimental regimes. Coincidence loss changes both the mean signal transfer and the noise statistics; some inelastic scattering may retain partial coherent transfer; calibration errors may be multiplicative rather than additive; and aliasing redistributes power between spatial frequencies.

Consequently,

$$\boxed{S_{\mathrm{obs}}(\mathbf{R}) \simeq A(\mathbf{R})|\,\mathrm{CTF}(\mathbf{R})|^2 + B(\mathbf{R})} \tag{6.1606}$$

should be understood as a highly useful phenomenological reduction of the complete physical model, not as a fundamental partition of the experiment into one "signal" process and one "background" process.

6.14.7 Sampling, aliasing, and nonlinear detection

Two final qualifications delimit the simple form of Eq. (6.1593).

First, the equation is a presampling expression. After discrete sampling, both signal and noise outside the Nyquist region can be folded back into the measured band. For noise, this effect was summarized in Eq. (6.1271); for the image itself, the corresponding spectral replicas were derived in Eq. (6.607). The PSD measured close to Nyquist can therefore contain aliased power from spatial frequencies that were present before sampling.

Second, the factorized model assumes that the detector can be treated linearly around the operating point. At sufficiently high dose rate, coincidence loss in counting detectors, Sec. 6.8.5, and saturation or other nonlinear

behavior, Sec. 6.11, violate this assumption. In that regime there is no globally valid transfer function that can be applied independently of the incident image. MTF, NPS, and DQE must instead be understood as operating-point-dependent or small-signal quantities, as discussed in Sec. 6.12.5.

6.14.8 The complete hierarchy of cryo-EM image formation

The sequence developed over the preceding chapters can now be viewed as one connected physical model:

$$
\boxed{
\begin{array}{c}
\text{three-dimensional specimen potential } V_0 \\
\Downarrow \\
\text{electron–specimen interaction} \\
\Downarrow \\
\mathcal{P}_{\mathcal{R}_i} V_0 \quad \text{under the projection and weak-phase approximations} \\
\Downarrow \\
\mathrm{CTF}_i(\mathbf{R}) \quad \text{electron-optical transfer} \\
\Downarrow \\
H_{\mathrm{pre},i}(\mathbf{R}) \quad \text{mean detector transfer} \\
\Downarrow \\
\text{shot noise and detector stochasticity} \\
\Downarrow \\
\text{discrete spatial and temporal sampling} \\
\Downarrow \\
y_i \quad \text{recorded cryo-EM image.}
\end{array}
}
\tag{6.1607}
$$

The CTF, MTF, NPS, DQE, and NEQ introduced in the preceding sections are therefore not independent detector or microscope specifications. They describe different aspects of successive stages of the same image-

formation process:

CTF	:	transfer of specimen contrast by the electron optics,
MTF	:	transfer of spatial modulation by the detector,
NPS	:	spatial-frequency distribution of recorded fluctuations,
DQE	:	fraction of input SNR preserved by detection,
NEQ	:	equivalent number of ideal electrons carrying that information.

(6.1608)

With this final step, the detector is no longer an implicit device placed after an otherwise complete imaging model. It becomes an explicit part of the cryo-EM forward model: it modifies the mean spatial signal, introduces its own stochastic physics, and determines how much of the information carried by the electrons arriving at the detector is ultimately preserved in the recorded image.

6.15 Summary

This chapter has completed the cryo-EM image-formation model by adding the final physical stage of the experiment: the conversion of the electron intensity distribution arriving at the detector plane into a sampled digital image.

The essential hierarchy developed in the preceding chapters can now be

written as

$$\boxed{\begin{array}{c} V_0(\mathbf{r}) \\ \Downarrow \\ \mathcal{P}_{\mathcal{R}_i} V_0 \\ \Downarrow \\ \mathrm{CTF}_i(\mathbf{R}) \\ \Downarrow \\ H_{\mathrm{pre},i}(\mathbf{R}) \\ \Downarrow \\ \text{stochastic electron detection} \\ \Downarrow \\ \text{spatial and temporal sampling} \\ \Downarrow \\ y_i \quad \text{recorded cryo-EM image.} \end{array}} \tag{6.1609}$$

The specimen determines the wave emerging from the sample, Chap. 5; the electron optics transform that wave into image contrast, Chap. 4; and the detector determines how much of that contrast and of the accompanying electron statistics survives into the digital measurement. This last connection was made explicit in Sec. 6.13.

The detector modifies the mean image through its spatial transfer function. The spatial response of the detector was introduced in Sec. 6.5. For a linear, shift-invariant detector, the incident intensity is convolved with the detector point-spread function, Eq. (6.531), and therefore multiplied in Fourier space by the detector transfer function $H_{\mathrm{det}}(\mathbf{R})$.

Finite pixel integration introduces a further aperture transfer, Sec. 6.5.3, so that the complete presampling transfer function is

$$\boxed{H_{\mathrm{pre}}(\mathbf{R}) = H_{\mathrm{det}}(\mathbf{R}) P_a(\mathbf{R}),} \tag{6.1610}$$

as derived in Eq. (6.566).

The corresponding mean cryo-EM image is therefore

$$\boxed{\begin{aligned} \mathbb{E}\left[\widehat{y}_i^{\mathrm{pre}}(\mathbf{R}) \mid V_0, \mathcal{R}_i, \mathbf{t}_i\right] = a_i H_{\mathrm{pre},i}(\mathbf{R})\, \mathrm{CTF}_i(\mathbf{R}) \\ \times e^{-j2\pi \mathbf{R}\cdot\mathbf{t}_i} \widehat{\mathcal{P}_{\mathcal{R}_i} V_0}(\mathbf{R}), \end{aligned}} \tag{6.1611}$$

which is Eq. (6.1564).

The CTF and detector transfer have different physical meanings. The CTF describes how the microscope converts specimen structure into image contrast, whereas H_{pre} describes how much of that already formed image modulation survives detection and pixel integration.

The MTF introduced in Sec. 6.5.2 is the normalized modulus of this spatial transfer. It therefore describes loss of mean spatial modulation, not noise by itself.

The detector response has a microscopic physical origin. The detector PSF and its fluctuations originate in the interaction of individual high-energy electrons with the sensor. The imaging electron loses energy while propagating through the detector material, Sec. 6.3, producing a spatially extended track of excitations. Secondary-electron transport, carrier generation and diffusion, charge sharing, and the final pixel readout determine where the event is ultimately recorded, Sec. 6.4.

Thus the response to one primary electron is not, in general, a deterministic point. It is more accurately represented by the random single-electron response

$$q_\omega(\boldsymbol{\rho}), \tag{6.1612}$$

introduced in Sec. 6.6.2.

Its expectation determines the mean detector response, while its event-to-event variability contributes directly to the detector noise.

The recorded image is stochastic, and its noise is generally neither white nor independent between pixels. Incident electrons arrive with shot noise, and each detected electron can produce a fluctuating spatial response. Additional fluctuations can arise from electronic readout, dark current, gain variations, quantization, and other detector processes, as developed in Sec. 6.6.

Consequently, replacing all detector fluctuations by independent Gaussian pixel noise is not a fundamental physical model.

The natural description of second-order detector statistics is the noise power spectrum introduced in Sec. 6.7. For a homogeneous Poisson fluence Φ_0, the complete electron-induced contribution is

$$\boxed{\mathrm{NPS}_{\text{electron}}(\mathbf{R}) = \Phi_0 \mathbb{E}\left[\left|\widehat{q}_\omega(\mathbf{R})\right|^2\right],} \tag{6.1613}$$

as derived in Eq. (6.1242).

This expression contains both propagated incident-electron shot noise and fluctuations of the response to one electron. Their exact separation was given in Eq. (6.1246). The different microscopic noise mechanisms and their spectral signatures were developed in Sec. 6.8.

The important lesson is that noise sources must be classified according to *where they enter the image-formation chain.* Noise generated before or during spatial transfer may itself be filtered, whereas noise added later in the readout chain need not undergo the same transfer.

The central result is the complete power spectrum of the recorded image. The most useful synthesis of the chapter is obtained at the power-spectrum level. Combining the specimen contribution, the electron-optical CTF, the detector spatial transfer, electron shot noise, and detector-added fluctuations gives

$$\boxed{\begin{aligned} S_{\mathrm{obs}}(\mathbf{R}) = {} & \underbrace{|H_{\mathrm{pre}}(\mathbf{R})|^2 |\,\mathrm{CTF}(\mathbf{R})|^2 S_s(\mathbf{R})}_{\text{particle structural signal}} \\ & + \underbrace{|H_{\mathrm{pre}}(\mathbf{R})|^2 |\,\mathrm{CTF}(\mathbf{R})|^2 S_{\mathrm{el}}(\mathbf{R}) + S_{\mathrm{spec,other,det}}(\mathbf{R})}_{\text{specimen background}} \\ & + \underbrace{S_{\mathrm{shot,det}}(\mathbf{R})}_{\text{incident-electron shot noise}} \\ & + \underbrace{\mathrm{NPS}_{\mathrm{det,add}}(\mathbf{R})}_{\text{detector-added fluctuations}} \quad . \end{aligned}} \tag{6.1614}$$

Equation (6.1614) reproduces the central result derived in Eq. (6.1593) of Sec. 6.14.5.

It is worth emphasizing its physical meaning.

The first term is the particle-dependent structural information that we ultimately wish to recover. It is modulated by the electron-optical CTF and subsequently attenuated by the detector transfer.

The second term contains specimen-generated background. Elastic structural fluctuations, such as those produced by vitreous ice, are formed before the microscope and are therefore also modulated by both the CTF and the detector. Other specimen contributions, including inelastic or more complicated scattering channels, need not follow the same simple transfer.

The third term is the finite-electron shot noise associated with the imaging beam.

The fourth term contains fluctuations added by the detector itself.

The observed power spectrum is therefore not simply

$$\text{CTF-modulated signal} + \text{white noise}. \tag{6.1615}$$

Different components are generated at different stages of the physical experiment and consequently have different spatial-frequency dependences.

The familiar CTF-plus-background model is an effective description of this more complete physics. The empirical representation introduced previously in Sec. 4.16,

$$\boxed{S_{\text{obs}}(\mathbf{R}) \simeq A(\mathbf{R})|\,\text{CTF}(\mathbf{R})|^2 + B(\mathbf{R}),} \tag{6.1616}$$

is recovered in Sec. 6.14.6.

The present chapter shows, however, that neither $A(\mathbf{R})$ nor $B(\mathbf{R})$ has a unique microscopic origin.

The factor $A(\mathbf{R})$ can contain particle signal, elastic specimen background, envelopes, and detector MTF attenuation. The apparent background $B(\mathbf{R})$ can contain inelastic specimen contributions, shot noise, fluctuations of the single-electron response, electronic noise, dark-current noise, quantization noise, residual detector structure, and other effects.

Thus Eq. (6.1616) is an extremely useful phenomenological model, but it is not a fundamental separation of the experiment into one signal process and one noise process.

MTF, NPS, DQE, and NEQ describe complementary aspects of the same detection process. The quantities developed in the middle part of this chapter should not be viewed as unrelated detector specifications.

The MTF measures preservation of mean spatial modulation, Sec. 6.5.2. The NPS measures the spatial-frequency distribution of detector fluctuations, Sec. 6.7. Together they determine the preservation of signal-to-noise ratio.

The detective quantum efficiency, Sec. 6.9, was defined as

$$\boxed{\text{DQE}(\mathbf{R}) = \frac{\text{SNR}_{\text{out}}(\mathbf{R})}{\text{SNR}_{\text{in}}(\mathbf{R})},} \tag{6.1617}$$

in Eq. (6.1283). It therefore measures what fraction of the information carried by a dose-limited electron field survives the detector.

The noise-equivalent quanta introduced in Sec. 6.10 make the same statement in units of ideal electrons:

$$\boxed{\mathrm{NEQ}(\mathbf{R}) = \Phi_0\,\mathrm{DQE}(\mathbf{R}).} \tag{6.1618}$$

This is Eq. (6.1380).

The interpretation can therefore be summarized as

$$\boxed{\begin{array}{rcl} \mathrm{MTF} & : & \text{how much mean spatial signal is transferred,} \\ \mathrm{NPS} & : & \text{how detector fluctuations are distributed in frequency,} \\ \mathrm{DQE} & : & \text{what fraction of input SNR survives,} \\ \mathrm{NEQ} & : & \text{how many ideal electrons would carry the same information.} \end{array}} \tag{6.1619}$$

Sampling and nonlinearities delimit the simple linear model. The equations above describe the presampling linear detector model. The final digital image is obtained only after spatial sampling.

As developed in Sec. 6.5.4, frequencies above the detector Nyquist frequency generate spectral replicas and can alias back into the recorded band. Both signal and noise may therefore be redistributed near the sampling limit.

The linear model can also fail when the detector operating point changes its response. Coincidence loss in counting detectors, Sec. 6.6.5, saturation, dead time, persistence, spatial nonuniformity, and other departures were discussed in Sec. 6.11. In these regimes, MTF, NPS, and DQE can depend on exposure rate and detector operating conditions, as emphasized experimentally in Sec. 6.12.

The image-formation model at the end of this chapter. The main conceptual result of the chapter is that the detector is not a passive device that merely adds an independent noise term after an otherwise complete cryo-EM image has been formed.

Instead,

$$\boxed{\begin{array}{c} \text{the detector modifies the mean spatial signal,} \\ \text{transforms the statistics of the incident electrons,} \\ \text{introduces additional stochastic fluctuations,} \\ \text{and finally samples the resulting continuous image.} \end{array}} \tag{6.1620}$$

The complete cryo-EM forward model must therefore account jointly for

$$\boxed{\begin{array}{c}\text{specimen physics}\\ +\ \text{electron-optical transfer}\\ +\ \text{detector spatial transfer}\\ +\ \text{electron and detector noise}\\ +\ \text{sampling.}\end{array}} \tag{6.1621}$$

Within the linear, shift-invariant, and approximately stationary regime, Eq. (6.1611) gives the mean image and Eq. (6.1614) gives its power spectrum. Together they complete the image-formation model developed throughout the preceding chapters.

Appendix A

Minimal quantum mechanics for electron microscopy

Quantum mechanics enters this book only at a few specific points. We need it to understand why an electron is described by a wave function, why an electrostatic potential acts on that wave function, and, later, how an electron can transfer energy to a specimen and leave it in a different internal state.

A complete course in quantum mechanics would require much more material than is needed here. We shall therefore introduce only the concepts and notation required elsewhere in this book:

- quantum states and wave functions;
- probability amplitudes and probabilities;
- bra–ket notation;
- inner products and orthogonality;
- operators, eigenstates, and expectation values;
- the Hamiltonian and time evolution;
- composite systems;
- matrix elements and transitions;
- why distinguishable inelastic channels add as probabilities rather than as amplitudes.

We shall not discuss the hydrogen atom, angular-momentum algebra, perturbation theory in general, quantum statistics, quantum field theory, or other topics that are not required for the image-formation model.

Throughout the book we use j for the imaginary unit,

$$j^2 = -1. \tag{A.1}$$

Most quantum-mechanics textbooks use i instead. There is no difference in meaning.

A.1 Quantum mechanics as geometry in a Hilbert space

Readers familiar with functional analysis, signal processing, inverse problems, or numerical linear algebra may find that much of the mathematical structure of quantum mechanics is already familiar.

The central mathematical object is a *Hilbert space*. Very roughly, a Hilbert space is a vector space in which vectors can be added, multiplied by complex numbers, compared through an inner product, and expanded in suitable bases. The additional technical requirement of completeness ensures that limits of convergent sequences of vectors remain inside the space.

Quantum mechanics can then be summarized, at a first mathematical level, by the statement

$$\boxed{\text{a quantum state is represented by a vector in a complex Hilbert space.}} \tag{A.2}$$

This statement provides a useful bridge between ordinary linear algebra and the bra–ket notation that will be introduced in the following sections.

A.1.1 From ordinary vectors to Hilbert spaces

Consider first the familiar finite-dimensional complex vector space

$$\mathbb{C}^N. \tag{A.3}$$

A vector has the form

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix}, \qquad v_n \in \mathbb{C}. \tag{A.4}$$

Two vectors may be added,

$$\mathbf{u} + \mathbf{v}, \tag{A.5}$$

and multiplied by a complex scalar,

$$\alpha\mathbf{v}, \qquad \alpha \in \mathbb{C}. \tag{A.6}$$

The usual complex inner product is

$$\boxed{\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{n=1}^{N} u_n^* v_n,} \tag{A.7}$$

where * denotes complex conjugation.

This inner product induces the norm

$$\boxed{\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}.} \tag{A.8}$$

A finite-dimensional complex vector space equipped with such an inner product is already a Hilbert space.

Quantum mechanics generalizes this idea to spaces that may have an infinite number of dimensions.

A.1.2 Functions can also be vectors

A particularly important example is the space of square-integrable complex functions,

$$\boxed{L^2(\mathbb{R}^3) = \left\{ \psi : \int_{\mathbb{R}^3} |\psi(\mathbf{r})|^2 \, d^3r < \infty \right\}.} \tag{A.9}$$

Although its elements are functions rather than columns of numbers, $L^2(\mathbb{R}^3)$ behaves much like an ordinary vector space.

Two functions can be added,

$$(\psi + \phi)(\mathbf{r}) = \psi(\mathbf{r}) + \phi(\mathbf{r}), \tag{A.10}$$

and multiplied by complex scalars,

$$(\alpha\psi)(\mathbf{r}) = \alpha\psi(\mathbf{r}). \tag{A.11}$$

The analogue of the dot product is

$$\boxed{\langle \phi, \psi \rangle = \int_{\mathbb{R}^3} \phi^*(\mathbf{r})\psi(\mathbf{r}) \, d^3r.} \tag{A.12}$$

The associated norm is

$$\boxed{\|\psi\| = \left[\int_{\mathbb{R}^3} |\psi(\mathbf{r})|^2 \, d^3r\right]^{1/2}.} \tag{A.13}$$

This structure should already be familiar from signal processing. A one-dimensional signal $f(t)$, a two-dimensional image $f(x, y)$, and a three-dimensional volume $f(x, y, z)$ can all be regarded as vectors in function spaces equipped with an inner product.

Thus, much of quantum mechanics is not based on an unfamiliar kind of mathematics. It is linear algebra extended from finite-dimensional vectors to spaces of complex-valued functions.

A.1.3 The abstract state and its representations

In quantum mechanics it is useful to distinguish the physical state from the coordinates used to describe it.

The abstract state is written

$$\boxed{|\psi\rangle,} \tag{A.14}$$

using Dirac's ket notation.

The symbol $|\psi\rangle$ should initially be thought of simply as a vector in a Hilbert space.

If we choose a basis, the same vector can be represented by a set of coordinates. For example, in a finite-dimensional basis

$$|u_1\rangle, \ldots, |u_N\rangle, \tag{A.15}$$

the state may be written

$$\boxed{|\psi\rangle = \sum_{n=1}^{N} c_n |u_n\rangle.} \tag{A.16}$$

The complex numbers c_n are the coordinates of the state in this basis.

This is completely analogous to an ordinary Euclidean vector,

$$\mathbf{v} = v_x \mathbf{e}_x + v_y \mathbf{e}_y + v_z \mathbf{e}_z. \tag{A.17}$$

For an electron described in position space, the basis is continuously labelled by position $\mathbf{r}$, and the coordinates become a function:

$$\boxed{\psi(\mathbf{r}) = \langle \mathbf{r} | \psi \rangle.} \tag{A.18}$$

The familiar electron wave function $\psi(\mathbf{r})$ is therefore not a different physical object from the ket $|\psi\rangle$. It is the representation of the same Hilbert-space vector in the position basis.

Similarly, the momentum-space representation is

$$\boxed{\widetilde{\psi}(\mathbf{p}) = \langle \mathbf{p}|\psi\rangle.} \tag{A.19}$$

We may summarize these different descriptions as

$$\boxed{\begin{array}{rcl} |\psi\rangle & : & \text{abstract state,} \\ \psi(\mathbf{r}) = \langle \mathbf{r}|\psi\rangle & : & \text{position representation,} \\ \widetilde{\psi}(\mathbf{p}) = \langle \mathbf{p}|\psi\rangle & : & \text{momentum representation.} \end{array}} \tag{A.20}$$

The state is the same; only the coordinate system used to describe it has changed.

A.1.4 A connection with Fourier analysis

This change of representation is particularly familiar in the context of electron microscopy.

The position-space wave function and momentum-space wave function are related by a Fourier transform. Using

$$\mathbf{p} = \hbar\mathbf{k} = h\mathbf{R}, \tag{A.21}$$

the plane-wave basis functions are

$$\exp(j\mathbf{k}\cdot\mathbf{r}) = \exp(j2\pi\mathbf{R}\cdot\mathbf{r}), \tag{A.22}$$

exactly as in Chap. 3.

Thus,

$$\boxed{\begin{array}{ccc} \text{position representation} & \xleftrightarrow{\text{Fourier transform}} & \text{momentum representation} \\ \psi(\mathbf{r}) & & \widetilde{\psi}(\mathbf{p}). \end{array}} \tag{A.23}$$

From the Hilbert-space point of view, the Fourier transform is simply a change of basis.

This interpretation is useful throughout this book. When an electron wave is propagated in real space or multiplied by a transfer function in Fourier space, we are not switching between different physical objects. We are describing the same state using different representations in which different operations become convenient.

A.1.5 Superposition is vector addition

One of the fundamental principles of quantum mechanics is the *superposition principle*. In Hilbert-space language, it is simply the linear structure of the state space.

If

$$|\psi_1\rangle \qquad \text{and} \qquad |\psi_2\rangle \tag{A.24}$$

are allowed states, then a linear combination

$$\boxed{|\psi\rangle = \alpha|\psi_1\rangle + \beta|\psi_2\rangle} \tag{A.25}$$

is also a vector in the same Hilbert space.

In position representation,

$$\psi(\mathbf{r}) = \alpha\psi_1(\mathbf{r}) + \beta\psi_2(\mathbf{r}). \tag{A.26}$$

The possibility of adding amplitudes in this way is the mathematical origin of interference.

If the two contributions lead to the same final quantum state, their amplitudes must be combined before calculating the probability:

$$|\psi_1 + \psi_2|^2 = |\psi_1|^2 + |\psi_2|^2 + 2\,\mathrm{Re}\left\{\psi_1^*\psi_2\right\}. \tag{A.27}$$

The cross term is the interference contribution that ultimately underlies coherent electron imaging.

A.1.6 Normalization and physical states

A quantum state used to describe a particle is normally represented by a normalized vector,

$$\boxed{\langle\psi|\psi\rangle = 1.} \tag{A.28}$$

In position representation this becomes

$$\boxed{\int |\psi(\mathbf{r})|^2\, d^3r = 1.} \tag{A.29}$$

There is one subtlety. Multiplying the complete state by a complex number of unit magnitude,

$$|\psi\rangle \longrightarrow e^{j\phi}|\psi\rangle, \tag{A.30}$$

does not change any physical prediction, because

$$|e^{j\phi}|^2 = 1. \tag{A.31}$$

Therefore, strictly speaking, a physical pure state is associated not with one unique normalized vector but with all vectors differing only by a global phase:

$$\boxed{|\psi\rangle \sim e^{j\phi}|\psi\rangle.} \tag{A.32}$$

For the purposes of this book, it is nevertheless convenient to speak simply of the normalized vector $|\psi\rangle$ as "the state."

A.1.7 Operators are transformations of states

A linear operator

$$\hat{A} \tag{A.33}$$

maps vectors in the Hilbert space into other vectors:

$$\boxed{|\phi\rangle = \hat{A}|\psi\rangle.} \tag{A.34}$$

Linearity means

$$\boxed{\hat{A}\left(\alpha|\psi_1\rangle + \beta|\psi_2\rangle\right) = \alpha\hat{A}|\psi_1\rangle + \beta\hat{A}|\psi_2\rangle.} \tag{A.35}$$

This concept is again familiar from signal processing. Differentiation, Fourier transformation, convolution, and multiplication by a transfer function are all examples of linear operations on suitable function spaces.

Quantum mechanics assigns particular operators to physical quantities. For example,

$$\hat{\mathbf{r}} \tag{A.36}$$

represents position,

$$\hat{\mathbf{p}} = -j\hbar\nabla \tag{A.37}$$

represents momentum, and

$$\hat{H} \tag{A.38}$$

is the Hamiltonian, representing the energy and generating the time evolution of the system.

The detailed meaning of operators, eigenstates, eigenvalues, and measurements will be developed in Sec. A.6 and Sec. A.7.5.

A.1.8 Eigenvectors are special directions in Hilbert space

The equation

$$\boxed{\hat{A}|u_n\rangle = a_n|u_n\rangle} \tag{A.39}$$

has exactly the same mathematical meaning as an eigenvalue problem in ordinary linear algebra.

The object

$$|u_n\rangle \tag{A.40}$$

is an eigenvector, or eigenstate, whereas

$$a_n \tag{A.41}$$

is the corresponding eigenvalue, an ordinary scalar number.

This distinction will be maintained throughout this appendix:

$$\boxed{|u_n\rangle \quad \text{state/vector}, \qquad a_n \quad \text{eigenvalue/number.}} \tag{A.42}$$

For Hermitian operators, eigenstates belonging to different eigenvalues are orthogonal and the eigenvalues are real. These properties will be derived explicitly later rather than assumed.

A.1.9 Inner products measure overlap

The inner product between two normalized states,

$$\langle\phi|\psi\rangle, \tag{A.43}$$

is a complex number that quantifies their overlap.

Two states are called *orthogonal* when

$$\boxed{\langle\phi|\psi\rangle = 0.} \tag{A.44}$$

An orthonormal discrete basis satisfies

$$\langle u_m|u_n\rangle = \delta_{mn}. \tag{A.45}$$

Once an orthonormal basis has been chosen, the coefficients in the expansion

$$|\psi\rangle = \sum_n c_n|u_n\rangle \tag{A.46}$$

are obtained from the inner products

$$\boxed{c_n = \langle u_n|\psi\rangle.} \tag{A.47}$$

A.1.10 The continuous position basis requires one qualification

The position-space description introduces a subtle mathematical point.

Ordinary wave functions belonging to $L^2(\mathbb{R}^3)$ are normalizable vectors in the Hilbert space. However, an idealized state of exactly known position,

$$|\mathbf{r}\rangle, \tag{A.48}$$

is not itself an ordinary square-integrable vector. Instead, position states satisfy the generalized normalization

$$\boxed{\langle \mathbf{r}'|\mathbf{r}\rangle = \delta(\mathbf{r}-\mathbf{r}').} \tag{A.49}$$

They are therefore called *generalized eigenstates*.

The same qualification applies to ideal plane waves of exactly known momentum. A plane wave extends throughout all space and cannot be normalized to unit integral in the ordinary sense.

A fully rigorous mathematical treatment uses an extension of the Hilbert-space framework, sometimes called a *rigged Hilbert space*. We shall not need that formalism here. The Dirac-delta normalization used throughout physics provides all the machinery required for the image-formation theory developed in this book.

A.1.11 Why Hilbert spaces are useful for the rest of the book

The Hilbert-space viewpoint unifies several ideas that otherwise may appear unrelated:

quantum state	$\longleftrightarrow$	vector $|\psi\rangle$,
wave function	$\longleftrightarrow$	coordinates $\psi(\mathbf{r})$ in the position basis,
Fourier transform	$\longleftrightarrow$	change from position to momentum basis,
superposition	$\longleftrightarrow$	vector addition,
overlap	$\longleftrightarrow$	inner product,
physical quantity	$\longleftrightarrow$	operator,
definite value of an observable	$\longleftrightarrow$	eigenstate of its operator.

(A.50)

For electron microscopy, this viewpoint will eventually allow us to describe an interaction such as

$$\langle f|\hat{H}_{\text{int}}|i\rangle \tag{A.51}$$

in ordinary geometric terms: start from a vector $|i\rangle$, apply the interaction operator $\hat{H}_{\text{int}}$, and measure the component of the resulting vector along the possible final state $|f\rangle$.

Similarly, when the specimen can finish in different mutually orthogonal states, Hilbert-space geometry will explain why the corresponding inelastic channels do not generally form one coherent electron amplitude.

The following sections develop the individual ingredients of this framework more slowly. Readers already familiar with Hilbert spaces may regard much of the appendix as a translation between this familiar mathematical language and the standard notation of quantum mechanics.

A.2 The state of a quantum system

In Sec. A.1 we introduced the central mathematical idea: a quantum state is represented by a vector in a complex Hilbert space. We now introduce the notation that will be used throughout the rest of this appendix and in the treatment of electron–matter interactions.

A state vector is conventionally written

$$\boxed{|\psi\rangle.} \tag{A.52}$$

The symbol $|\psi\rangle$ is called a *ket*, and this notation is part of the *Dirac notation*, also known as *bra–ket notation*.

The symbol ψ inside the ket is simply a label identifying the state. The ket $|\psi\rangle$ is the abstract Hilbert-space vector; it should not be confused with the complex number $\psi(\mathbf{r})$.

A.2.1 The position representation of a state

The abstract ket $|\psi\rangle$ represents the quantum state without specifying any particular coordinate system. To obtain a concrete representation of that state, we must choose a basis.

It is useful to begin with a finite-dimensional example. Suppose that

$$|u_1\rangle, \ldots, |u_N\rangle \tag{A.53}$$

form an orthonormal basis of a Hilbert space, so that

$$\boxed{\langle u_m|u_n\rangle = \delta_{mn}.} \tag{A.54}$$

Any state $|\psi\rangle$ can then be decomposed in this basis:

$$\boxed{|\psi\rangle = \sum_{n=1}^{N} c_n |u_n\rangle.} \tag{A.55}$$

The numbers c_n are the coordinates, or components, of $|\psi\rangle$ in this particular basis.

How can we obtain one of these components from the state itself? Multiply Eq. (A.55) from the left by $\langle u_m|$:

$$\langle u_m|\psi\rangle = \sum_{n=1}^{N} c_n \langle u_m|u_n\rangle \tag{A.56}$$

$$= \sum_{n=1}^{N} c_n \delta_{mn} \tag{A.57}$$

$$= c_m. \tag{A.58}$$

Therefore,

$$\boxed{c_m = \langle u_m|\psi\rangle.} \tag{A.59}$$

Thus the inner product with a basis vector extracts the component of the state along that basis vector.

Substituting Eq. (A.59) back into Eq. (A.55), we can write the decomposition as

$$\boxed{|\psi\rangle = \sum_{n} |u_n\rangle\langle u_n|\psi\rangle.} \tag{A.60}$$

This equation is the key to understanding the position-space wave function.

Choosing position as the basis

For the position representation, instead of using a discrete set of basis states $|u_n\rangle$, we use a continuously labelled family of states

$$\boxed{|\mathbf{r}\rangle,} \tag{A.61}$$

one for every possible position

$$\mathbf{r} = (x, y, z). \tag{A.62}$$

The state $|\mathbf{r}\rangle$ represents an idealized state of exact position $\mathbf{r}$. As discussed in Sec. A.1, these are generalized basis states rather than ordinary normalizable Hilbert-space vectors.

For the discrete basis, orthonormality was expressed by Eq. (A.54),

$$\langle u_m | u_n \rangle = \delta_{mn}. \tag{A.63}$$

For the continuous position basis, the corresponding relation is

$$\boxed{\langle \mathbf{r}' | \mathbf{r} \rangle = \delta(\mathbf{r} - \mathbf{r}').} \tag{A.64}$$

Similarly, the discrete sum over basis states becomes an integral over all possible positions:

$$\boxed{\sum_n \quad \longrightarrow \quad \int_{\mathbb{R}^3} d^3 r.} \tag{A.65}$$

Thus the continuous analogue of Eq. (A.60) is

$$\boxed{|\psi\rangle = \int_{\mathbb{R}^3} |\mathbf{r}\rangle \langle \mathbf{r} | \psi \rangle \, d^3 r.} \tag{A.66}$$

This is the decomposition of the abstract state $|\psi\rangle$ in the position basis.

Comparing it with the discrete expansion

$$|\psi\rangle = \sum_n c_n |u_n\rangle, \tag{A.67}$$

we see the correspondence

$$\boxed{\begin{array}{ccc} \text{discrete basis} & & \text{position basis} \\ |u_n\rangle & \longleftrightarrow & |\mathbf{r}\rangle, \\ c_n & \longleftrightarrow & \langle \mathbf{r} | \psi \rangle, \\ \displaystyle\sum_n & \longleftrightarrow & \displaystyle\int d^3 r. \end{array}} \tag{A.68}$$

Therefore,

$$\boxed{\langle \mathbf{r} | \psi \rangle} \tag{A.69}$$

is simply the coefficient, or coordinate, multiplying the basis state $|\mathbf{r}\rangle$ in the position-basis decomposition of $|\psi\rangle$.

We give this coefficient a special name:

$$\boxed{\psi(\mathbf{r}) \equiv \langle \mathbf{r}|\psi\rangle.} \tag{A.70}$$

It is the *position-space wave function.*

Using this definition, Eq. (A.66) becomes

$$\boxed{|\psi\rangle = \int_{\mathbb{R}^3} \psi(\mathbf{r})|\mathbf{r}\rangle \, d^3r.} \tag{A.71}$$

The meaning of this equation is directly analogous to an ordinary vector decomposition. For example,

$$\mathbf{v} = v_x\mathbf{e}_x + v_y\mathbf{e}_y + v_z\mathbf{e}_z \tag{A.72}$$

expresses a vector as a sum of basis vectors weighted by their coordinates. Likewise,

$$|\psi\rangle = \int \psi(\mathbf{r})|\mathbf{r}\rangle \, d^3r \tag{A.73}$$

expresses a quantum state as a continuous superposition of position basis states weighted by the coefficients $\psi(\mathbf{r})$.

Extracting the position-space coefficient.

We can verify directly that $\langle \mathbf{r}'|\psi\rangle$ extracts the coefficient associated with the position $\mathbf{r}'$.

Start from the position-basis decomposition, Eq. (A.71),

$$|\psi\rangle = \int \psi(\mathbf{r})|\mathbf{r}\rangle \, d^3r. \tag{A.74}$$

Multiply from the left by $\langle \mathbf{r}'|$:

$$\langle \mathbf{r}'|\psi\rangle = \int \psi(\mathbf{r})\langle \mathbf{r}'|\mathbf{r}\rangle \, d^3r. \tag{A.75}$$

Using the delta-normalization Eq. (A.64),

$$\langle \mathbf{r}'|\psi\rangle = \int \psi(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}') \, d^3r. \tag{A.76}$$

The sifting property of the Dirac delta gives

$$\boxed{\langle \mathbf{r}'|\psi\rangle = \psi(\mathbf{r}').} \tag{A.77}$$

Thus taking the inner product with the position basis state $|\mathbf{r}'\rangle$ selects precisely the coefficient of the quantum state associated with that position.

It is therefore useful to distinguish carefully between

$$\boxed{\begin{array}{rcl} |\psi\rangle & : & \text{the abstract quantum state,} \\ |\mathbf{r}\rangle & : & \text{a position basis state,} \\ \langle\mathbf{r}|\psi\rangle = \psi(\mathbf{r}) & : & \text{the coordinate of } |\psi\rangle \text{ along } |\mathbf{r}\rangle. \end{array}} \tag{A.78}$$

In this sense, the wave function is analogous to the list of coordinates of an ordinary vector. The difference is that the position basis is continuous, so instead of a finite or countable list of numbers c_n, the coordinates form a function $\psi(\mathbf{r})$.

For much of the diffraction and electron-optics theory developed in this book, the position-space representation $\psi(\mathbf{r})$ is the most convenient description. The abstract notation becomes particularly useful when several different states of the specimen must be considered simultaneously, as will occur in the description of elastic and inelastic scattering.

A.3 Probability amplitudes and probabilities

The Hilbert-space structure introduced in Sec. A.1 is mathematical. Quantum mechanics adds a physical rule that connects the wave function with experimental probabilities.

Consider a normalized single-particle state with position-space wave function

$$\psi(\mathbf{r}) = \langle\mathbf{r}|\psi\rangle. \tag{A.79}$$

The *Born rule* states that

$$\boxed{p(\mathbf{r}) = |\psi(\mathbf{r})|^2 = \psi^*(\mathbf{r})\psi(\mathbf{r})} \tag{A.80}$$

is the probability density for position.

It is important to distinguish a probability density from a probability. For a continuous position variable, the probability that the particle lies inside some volume Ω is

$$\boxed{P(\mathbf{r} \in \Omega) = \int_\Omega |\psi(\mathbf{r})|^2 \, d^3r.} \tag{A.81}$$

Consequently, for a normalized state, the probability of finding the particle somewhere in space must be one:

$$\boxed{\int_{\mathbb{R}^3} |\psi(\mathbf{r})|^2 \, d^3r = 1.} \tag{A.82}$$

For a three-dimensional normalized wave function, $p(\mathbf{r})$ therefore has units of

$$\mathrm{m}^{-3}, \tag{A.83}$$

while the integral in Eq. (A.81) is dimensionless.

The probability of finding the particle at one exact mathematical point is not $|\psi(\mathbf{r})|^2$; for a continuous position variable, the probability of any single point is zero. The quantity $|\psi(\mathbf{r})|^2$ specifies how probability is distributed per unit volume around that point.

A.3.1 The wave function is an amplitude

A wave function is generally complex. We may write it in polar form as

$$\boxed{\psi(\mathbf{r}) = A(\mathbf{r}) \exp[j\phi(\mathbf{r})],} \tag{A.84}$$

where

$$A(\mathbf{r}) \geq 0 \tag{A.85}$$

is its magnitude and

$$\phi(\mathbf{r}) \tag{A.86}$$

is its phase.

Equation (A.80) then gives

$$p(\mathbf{r}) = A^2(\mathbf{r}). \tag{A.87}$$

The wave function itself is therefore not a probability. It is a *probability amplitude*. The probability density is obtained only after taking its squared modulus.

At first sight, this might suggest that the phase $\phi(\mathbf{r})$ is irrelevant, since it disappears from $|\psi(\mathbf{r})|^2$. The phase becomes observable indirectly when several amplitudes contribute to the same final state, because those amplitudes must first be added and only then squared.

A.3.2 Interference of probability amplitudes

Suppose that the same final state can be reached through two indistinguishable alternatives with amplitudes

$$\psi_1 = A_1 e^{j\phi_1}, \qquad \psi_2 = A_2 e^{j\phi_2}. \tag{A.88}$$

Because both alternatives lead to the same final state, their amplitudes add:

$$\boxed{\psi = \psi_1 + \psi_2.} \tag{A.89}$$

The corresponding probability is therefore

$$\begin{aligned} |\psi|^2 &= |\psi_1 + \psi_2|^2 && \text{(A.90)} \\ &= (\psi_1 + \psi_2)^*(\psi_1 + \psi_2) && \text{(A.91)} \\ &= |\psi_1|^2 + |\psi_2|^2 + \psi_1^*\psi_2 + \psi_2^*\psi_1 && \text{(A.92)} \\ &= |\psi_1|^2 + |\psi_2|^2 + 2\,\mathrm{Re}\left\{\psi_1^*\psi_2\right\}. && \text{(A.93)} \end{aligned}$$

Using the amplitude–phase representation,

$$\psi_1^*\psi_2 = A_1 A_2 e^{j(\phi_2 - \phi_1)}, \tag{A.94}$$

so

$$\boxed{|\psi|^2 = A_1^2 + A_2^2 + 2A_1A_2\cos(\phi_2 - \phi_1).} \tag{A.95}$$

The last term is the *interference term.* It depends on the relative phase

$$\Delta\phi = \phi_2 - \phi_1. \tag{A.96}$$

For example, if the two amplitudes have equal magnitude, $A_1 = A_2 = A$, then

$$|\psi|^2 = 2A^2\left[1 + \cos(\Delta\phi)\right]. \tag{A.97}$$

When

$$\Delta\phi = 0, \tag{A.98}$$

the amplitudes interfere constructively and

$$|\psi|^2 = 4A^2. \tag{A.99}$$

When

$$\Delta\phi = \pi, \tag{A.100}$$

they interfere destructively and

$$|\psi|^2 = 0. \tag{A.101}$$

This is the same interference mechanism already encountered throughout the wave-diffraction theory of Chap. 3. Quantum mechanics gives it an additional interpretation: the complex quantities being added are probability amplitudes.

Much of phase-contrast electron microscopy ultimately relies on this principle. Different coherent electron-wave contributions acquire different phases during propagation through the specimen and the microscope, and the detector records the squared modulus of their combined amplitude.

The qualification that the alternatives must lead to the same, indistinguishable final state will become particularly important when we discuss inelastic scattering. If the specimen is left in different orthogonal final states, the corresponding electron amplitudes do not in general contribute to the same interference term.

A.4 Bras, kets, and inner products

The Hilbert-space description introduced in Sec. A.1 already provides the concepts of vectors, inner products, norms, and orthogonality. In Sec. A.2.1 we also saw how a state can be decomposed in a basis and why

$$\psi(\mathbf{r}) = \langle \mathbf{r} | \psi \rangle \tag{A.102}$$

is the coefficient of the state $|\psi\rangle$ in the position basis.

We now collect the corresponding Dirac notation more systematically.

A.4.1 From a ket to a bra

A state vector is written as a ket,

$$|\psi\rangle. \tag{A.103}$$

Associated with it is a *bra*,

$$\boxed{\langle \psi |.} \tag{A.104}$$

In a finite-dimensional orthonormal basis, if

$$|\psi\rangle = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_N \end{pmatrix}, \tag{A.105}$$

then the corresponding bra is represented by the complex-conjugate transpose,

$$\boxed{\langle\psi| = \begin{pmatrix}\psi_1^* & \psi_2^* & \cdots & \psi_N^*\end{pmatrix}.} \tag{A.106}$$

Thus, in finite-dimensional matrix notation,

$$\boxed{\langle\psi| = |\psi\rangle^\dagger,} \tag{A.107}$$

where † denotes the complex-conjugate transpose, or *adjoint*.

The important point is that a ket and its corresponding bra play different roles. The ket is a vector; the bra acts on another vector and produces a complex number.

A.4.2 The inner product in Dirac notation

If $|\phi\rangle$ and $|\psi\rangle$ are two states, their inner product is written

$$\boxed{\langle\phi|\psi\rangle.} \tag{A.108}$$

In a finite-dimensional orthonormal basis,

$$|\phi\rangle = \begin{pmatrix}\phi_1\\ \vdots \\ \phi_N\end{pmatrix}, \qquad |\psi\rangle = \begin{pmatrix}\psi_1\\ \vdots \\ \psi_N\end{pmatrix}, \tag{A.109}$$

and therefore

$$\boxed{\langle\phi|\psi\rangle = \sum_{n=1}^{N} \phi_n^* \psi_n.} \tag{A.110}$$

This is exactly the complex inner product introduced in Sec. A.1; Dirac notation merely expresses it in a compact basis-independent form.

The result $\langle\phi|\psi\rangle$ is an ordinary complex number, not another state.

A.4.3 Conjugation and linearity

Two properties of the inner product will be used repeatedly.

First, exchanging the two states complex-conjugates the result:

$$\boxed{\langle\phi|\psi\rangle = \langle\psi|\phi\rangle^*.} \tag{A.111}$$

Second, the inner product is linear in its second argument. Thus,

$$\boxed{\langle\phi|\alpha\psi + \beta\chi\rangle = \alpha\langle\phi|\psi\rangle + \beta\langle\phi|\chi\rangle.} \tag{A.112}$$

In ket notation, the same relation is

$$\langle\phi|\left(\alpha|\psi\rangle + \beta|\chi\rangle\right) = \alpha\langle\phi|\psi\rangle + \beta\langle\phi|\chi\rangle. \tag{A.113}$$

In the first argument, the scalars are complex-conjugated:

$$\boxed{\langle\alpha\phi + \beta\eta|\psi\rangle = \alpha^*\langle\phi|\psi\rangle + \beta^*\langle\eta|\psi\rangle.} \tag{A.114}$$

This follows from the fact that

$$\boxed{\left(\alpha|\phi\rangle + \beta|\eta\rangle\right)^\dagger = \alpha^*\langle\phi| + \beta^*\langle\eta|.} \tag{A.115}$$

The inner product is therefore said to be *conjugate-linear* in its first argument and linear in its second.

A.4.4 Norm and orthogonality in Dirac notation

The norm and orthogonality introduced geometrically in Sec. A.1 take a particularly compact form in Dirac notation.

The squared norm of a state is

$$\boxed{\|\,|\psi\rangle\,\|^2 = \langle\psi|\psi\rangle.} \tag{A.116}$$

For a normalized state,

$$\boxed{\langle\psi|\psi\rangle = 1.} \tag{A.117}$$

Two states are orthogonal when

$$\boxed{\langle\phi|\psi\rangle = 0.} \tag{A.118}$$

Normalization and orthogonality are distinct concepts: two states can be orthogonal without either of them having unit norm.

For normalized states, the Cauchy–Schwarz inequality gives

$$0 \leq |\langle\phi|\psi\rangle| \leq 1. \tag{A.119}$$

Thus the magnitude of the inner product provides a quantitative measure of the overlap between two normalized states.

A.4.5 The inner product in the position representation

We can now connect the abstract inner product directly with the wave functions used throughout this book.

From Eq. (A.70),

$$\psi(\mathbf{r}) = \langle \mathbf{r}|\psi\rangle. \tag{A.120}$$

Using the conjugate-symmetry relation Eq. (A.111),

$$\begin{aligned}\langle \psi|\mathbf{r}\rangle &= \langle \mathbf{r}|\psi\rangle^* \\ &= \psi^*(\mathbf{r}).\end{aligned} \tag{A.121}$$

Hence,

$$\boxed{\langle \psi|\mathbf{r}\rangle = \psi^*(\mathbf{r}).} \tag{A.122}$$

The familiar integral expression for the inner product now follows directly from the position-basis decomposition derived in Sec. A.2.1.

The inner product in the position basis.

In Sec. A.2.1, we decomposed an arbitrary state in the position basis as

$$|\psi\rangle = \int_{\mathbb{R}^3} |\mathbf{r}\rangle\langle \mathbf{r}|\psi\rangle \, d^3r, \tag{A.123}$$

see Eq. (A.66).

Take the inner product of this state with an arbitrary state $|\phi\rangle$:

$$\langle \phi|\psi\rangle = \left\langle \phi \middle| \int_{\mathbb{R}^3} |\mathbf{r}\rangle\langle \mathbf{r}|\psi\rangle \, d^3r \right. \tag{A.124}$$

$$= \int_{\mathbb{R}^3} \langle \phi|\mathbf{r}\rangle\langle \mathbf{r}|\psi\rangle \, d^3r. \tag{A.125}$$

The two factors inside the integral have already been identified. From the definition of the position-space wave function, Eq. (A.70),

$$\langle \mathbf{r}|\psi\rangle = \psi(\mathbf{r}). \tag{A.126}$$

From conjugate symmetry of the inner product, Eq. (A.111),

$$\langle \phi|\mathbf{r}\rangle = \langle \mathbf{r}|\phi\rangle^* = \phi^*(\mathbf{r}). \tag{A.127}$$

Substituting these two expressions gives

$$\boxed{\langle\phi|\psi\rangle = \int_{\mathbb{R}^3} \phi^*(\mathbf{r})\psi(\mathbf{r})\, d^3r.} \tag{A.128}$$

Eq. (A.128) implies

$$\boxed{\langle\psi|\psi\rangle = \int_{\mathbb{R}^3} |\psi(\mathbf{r})|^2\, d^3r.} \tag{A.129}$$

Consequently, the abstract normalization condition Eq. (A.82) implies

$$\langle\psi|\psi\rangle = 1. \tag{A.130}$$

This provides an important connection between the Hilbert-space geometry and the probabilistic interpretation of the wave function: the squared norm of a normalized state is its total probability.

A.4.6 Outer products

There is one further combination of bras and kets that will be useful later.

Whereas

$$\langle\phi|\psi\rangle \tag{A.131}$$

is a complex number, reversing the order produces

$$\boxed{|\phi\rangle\langle\psi|,} \tag{A.132}$$

which is an *operator*.

To see this, let it act on an arbitrary state $|\chi\rangle$:

$$\boxed{(|\phi\rangle\langle\psi|)\,|\chi\rangle = |\phi\rangle\langle\psi|\chi\rangle.} \tag{A.133}$$

The inner product

$$\langle\psi|\chi\rangle \tag{A.134}$$

is a complex number, so the result is that number multiplied by the state $|\phi\rangle$.

A particularly important case is

$$\boxed{|u\rangle\langle u|,} \tag{A.135}$$

for a normalized state $|u\rangle$. Acting on $|\psi\rangle$,

$$|u\rangle\langle u|\psi\rangle \tag{A.136}$$

extracts the component $\langle u|\psi\rangle$ along $|u\rangle$ and returns that component in the direction $|u\rangle$. Such an operator is called a *projector*.

Projectors will appear naturally later when we discuss measurements and degenerate eigenspaces.

A.4.7 How to read bra–ket expressions

With these definitions, longer bra–ket expressions can be read systematically from right to left.

Consider

$$\boxed{\langle\psi_f|\hat{A}|\psi_i\rangle.} \tag{A.137}$$

The expression represents the following sequence:

1. begin with the initial state $|\psi_i\rangle$;
2. apply the operator $\hat{A}$,
$$|\chi\rangle = \hat{A}|\psi_i\rangle; \tag{A.138}$$
3. take the inner product of the resulting state with $|\psi_f\rangle$,
$$\langle\psi_f|\chi\rangle. \tag{A.139}$$

Thus,

$$\langle\psi_f|\hat{A}|\psi_i\rangle \tag{A.140}$$

is an ordinary complex number. If $|\psi_f\rangle$ is normalized, it is the component of the transformed state $\hat{A}|\psi_i\rangle$ along the direction $|\psi_f\rangle$.

The shorter notation

$$\langle f|\hat{A}|i\rangle \tag{A.141}$$

means exactly the same thing, with i and f serving simply as labels for the initial and final states.

Similarly,

$$\boxed{\langle\psi|\hat{A}|\psi\rangle} \tag{A.142}$$

means that $\hat{A}$ acts on $|\psi\rangle$ and that the resulting state is then projected back onto $|\psi\rangle$ through the inner product.

At this stage this is only a mathematical construction. When $\hat{A}$ represents an observable, the additional measurement postulates introduced in Sec. A.7.5 will give $\langle\psi|\hat{A}|\psi\rangle$ its physical interpretation as an expectation value.

A.4.8 A compact dictionary

The essential grammar of Dirac notation can therefore be summarized as

$$\boxed{\begin{array}{rcl} |\psi\rangle & : & \text{vector (ket)}, \\ \langle\psi| & : & \text{corresponding dual vector (bra)}, \\ \langle\phi|\psi\rangle & : & \text{complex number (inner product)}, \\ |\phi\rangle\langle\psi| & : & \text{operator (outer product)}, \\ \hat{A}|\psi\rangle & : & \text{vector obtained by applying an operator}, \\ \langle\phi|\hat{A}|\psi\rangle & : & \text{complex number (matrix element)}. \end{array}} \tag{A.143}$$

No new algebra is hidden in these symbols. They provide a compact language for the Hilbert-space operations already introduced: vectors, inner products, projections, and linear operators.

A.5 Momentum states, Fourier transforms, and uncertainty

In Sec. A.2.1 we described the abstract state $|\psi\rangle$ using the position basis. The position-space wave function

$$\psi(\mathbf{r}) = \langle\mathbf{r}|\psi\rangle \tag{A.144}$$

is the coefficient of the state along the position basis state $|\mathbf{r}\rangle$.

Position, however, is only one possible representation of the same quantum state. We may instead choose a basis whose elements correspond to definite momentum. This leads to the momentum representation and, as we shall see, directly to the Fourier transform already used throughout this book.

A.5.1 The momentum basis

Let

$$\boxed{|\mathbf{p}\rangle} \tag{A.145}$$

denote an idealized state having a definite momentum

$$\mathbf{p} = (p_x, p_y, p_z). \tag{A.146}$$

Just as exact position states are generalized states, exact momentum states are not ordinary normalizable Hilbert-space vectors. They are delta-normalized:

$$\boxed{\langle \mathbf{p}'|\mathbf{p}\rangle = \delta(\mathbf{p}-\mathbf{p}').} \tag{A.147}$$

The momentum states form a continuous basis, so the identity can be written

$$\boxed{\hat{I} = \int_{\mathbb{R}^3} |\mathbf{p}\rangle\langle\mathbf{p}|\, d^3p.} \tag{A.148}$$

Applying this identity to an arbitrary state gives its decomposition in the momentum basis:

$$|\psi\rangle = \hat{I}|\psi\rangle \tag{A.149}$$

$$= \int_{\mathbb{R}^3} |\mathbf{p}\rangle\langle\mathbf{p}|\psi\rangle\, d^3p. \tag{A.150}$$

Therefore,

$$\boxed{|\psi\rangle = \int_{\mathbb{R}^3} \widetilde{\psi}(\mathbf{p})|\mathbf{p}\rangle\, d^3p,} \tag{A.151}$$

where we define

$$\boxed{\widetilde{\psi}(\mathbf{p}) \equiv \langle\mathbf{p}|\psi\rangle.} \tag{A.152}$$

The analogy with the position representation is exact:

position representation	momentum representation
$\lvert\mathbf{r}\rangle$	$\lvert\mathbf{p}\rangle$
$\psi(\mathbf{r}) = \langle\mathbf{r}\lvert\psi\rangle$	$\widetilde{\psi}(\mathbf{p}) = \langle\mathbf{p}\lvert\psi\rangle$
$\lvert\psi\rangle = \int \psi(\mathbf{r})\lvert\mathbf{r}\rangle\, d^3r$	$\lvert\psi\rangle = \int \widetilde{\psi}(\mathbf{p})\lvert\mathbf{p}\rangle\, d^3p.$

(A.153)

The two functions $\psi(\mathbf{r})$ and $\widetilde{\psi}(\mathbf{p})$ are therefore two different sets of coordinates of the same abstract state $|\psi\rangle$.

A.5.2 What does a state of definite momentum look like in real space?

The de Broglie relation connects momentum and wavevector:

$$\boxed{\mathbf{p} = \hbar\mathbf{k}.} \tag{A.154}$$

A state of definite momentum is consequently represented in position space by a plane wave.

With the delta-normalization of Eq. (A.147), the properly normalized generalized plane-wave state is

$$\boxed{\langle \mathbf{r}|\mathbf{p}\rangle = \frac{1}{(2\pi\hbar)^{3/2}} \exp\left(\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}\right).} \tag{A.155}$$

The quantity on the left,

$$\langle \mathbf{r}|\mathbf{p}\rangle, \tag{A.156}$$

has a simple interpretation: it is the position-space representation of the momentum state $|\mathbf{p}\rangle$.

Normalization of a momentum eigenstate.

Suppose

$$\langle \mathbf{r}|\mathbf{p}\rangle = C \exp\left(\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}\right), \tag{A.157}$$

where C is a normalization constant.

The inner product between two momentum states is

$$\langle \mathbf{p}'|\mathbf{p}\rangle = \int \langle \mathbf{p}'|\mathbf{r}\rangle\langle \mathbf{r}|\mathbf{p}\rangle \, d^3r \tag{A.158}$$

$$= |C|^2 \int \exp\left[\frac{j}{\hbar}(\mathbf{p}-\mathbf{p}')\cdot\mathbf{r}\right] d^3r. \tag{A.159}$$

Using the Fourier representation of the Dirac delta,

$$\int \exp\left[\frac{j}{\hbar}(\mathbf{p}-\mathbf{p}')\cdot\mathbf{r}\right] d^3r = (2\pi\hbar)^3\delta(\mathbf{p}-\mathbf{p}'), \tag{A.160}$$

we obtain

$$\langle \mathbf{p}'|\mathbf{p}\rangle = |C|^2(2\pi\hbar)^3\delta(\mathbf{p}-\mathbf{p}'). \tag{A.161}$$

Requiring the normalization Eq. (A.147) gives

$$|C|^2 = \frac{1}{(2\pi\hbar)^3}, \tag{A.162}$$

and therefore

$$\boxed{C = \frac{1}{(2\pi\hbar)^{3/2}},} \tag{A.163}$$

up to an irrelevant constant phase.

A.5.3 The position and momentum representations are Fourier transforms

We can now derive the relation between $\psi(\mathbf{r})$ and $\widetilde{\psi}(\mathbf{p})$.

Starting from the momentum-basis decomposition Eq. (A.151) and projecting onto $\langle\mathbf{r}|$,

$$\psi(\mathbf{r}) = \langle\mathbf{r}|\psi\rangle \tag{A.164}$$

$$= \int \langle\mathbf{r}|\mathbf{p}\rangle\widetilde{\psi}(\mathbf{p})\, d^3p. \tag{A.165}$$

Using Eq. (A.155),

$$\boxed{\psi(\mathbf{r}) = \frac{1}{(2\pi\hbar)^{3/2}} \int_{\mathbb{R}^3} \widetilde{\psi}(\mathbf{p}) \exp\left(\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}\right)\, d^3p.} \tag{A.166}$$

Conversely,

$$\boxed{\widetilde{\psi}(\mathbf{p}) = \frac{1}{(2\pi\hbar)^{3/2}} \int_{\mathbb{R}^3} \psi(\mathbf{r}) \exp\left(-\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}\right)\, d^3r.} \tag{A.167}$$

Thus the position and momentum representations of a quantum state are Fourier-transform pairs.

A.5.4 Connection with the spatial-frequency notation of this book

Throughout the diffraction and electron-optics chapters we have used the spatial-frequency vector

$$\mathbf{R} \tag{A.168}$$

measured in cycles per metre, and the wavevector

$$\mathbf{k} \tag{A.169}$$

measured in radians per metre. They are related by

$$\mathbf{k} = 2\pi\mathbf{R}. \tag{A.170}$$

Combining this relation with Eq. (A.154),

$$\boxed{\mathbf{p} = \hbar\mathbf{k} = h\mathbf{R}.} \tag{A.171}$$

Therefore, a momentum state can equivalently be labelled by its spatial frequency.

It is useful to define a spatial-frequency basis $|\mathbf{R}\rangle$ such that

$$\boxed{\langle\mathbf{r}|\mathbf{R}\rangle = \exp\left(j2\pi\mathbf{R}\cdot\mathbf{r}\right).} \tag{A.172}$$

The normalization is then

$$\boxed{\langle\mathbf{R}'|\mathbf{R}\rangle = \delta(\mathbf{R}-\mathbf{R}').} \tag{A.173}$$

The coefficient of the state in this basis is

$$\boxed{\widehat{\psi}(\mathbf{R}) = \langle\mathbf{R}|\psi\rangle.} \tag{A.174}$$

Consequently,

$$\boxed{\widehat{\psi}(\mathbf{R}) = \int_{\mathbb{R}^3} \psi(\mathbf{r}) \exp\left(-j2\pi\mathbf{R}\cdot\mathbf{r}\right)\, d^3r,} \tag{A.175}$$

and

$$\boxed{\psi(\mathbf{r}) = \int_{\mathbb{R}^3} \widehat{\psi}(\mathbf{R}) \exp\left(j2\pi\mathbf{R}\cdot\mathbf{r}\right)\, d^3R.} \tag{A.176}$$

These are exactly the Fourier-transform conventions used previously in this book.

The momentum-space amplitude $\widetilde{\psi}(\mathbf{p})$ and the spatial-frequency amplitude $\widehat{\psi}(\mathbf{R})$ describe the same representation with different coordinates. Because

$$\mathbf{p} = h\mathbf{R}, \qquad d^3p = h^3 d^3R, \tag{A.177}$$

their normalized amplitudes are related by

$$\boxed{\widehat{\psi}(\mathbf{R}) = h^{3/2}\widetilde{\psi}(h\mathbf{R}).} \tag{A.178}$$

This factor is important: momentum and spatial frequency differ not only by a change of variable but also by the normalization associated with the corresponding integration measure.

A.5.5 Position and momentum states are opposite limiting cases

The symmetry between the two representations becomes particularly clear if we compare states of exactly defined position and exactly defined momentum.

Consider first an exact position state $|\mathbf{r}_0\rangle$. Its position representation is

$$\boxed{\langle \mathbf{r}|\mathbf{r}_0\rangle = \delta(\mathbf{r}-\mathbf{r}_0).} \tag{A.179}$$

It is therefore perfectly localized in position.

Its momentum representation is obtained from Eq. (A.155):

$$\boxed{\langle \mathbf{p}|\mathbf{r}_0\rangle = \frac{1}{(2\pi\hbar)^{3/2}} \exp\left(-\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}_0\right).} \tag{A.180}$$

Its magnitude is independent of $\mathbf{p}$. Thus an ideally localized position state contains all momenta with equal magnitude.

Now consider an exact momentum state $|\mathbf{p}_0\rangle$. Its momentum representation is

$$\boxed{\langle \mathbf{p}|\mathbf{p}_0\rangle = \delta(\mathbf{p}-\mathbf{p}_0),} \tag{A.181}$$

so its momentum is perfectly defined.

In position space, however,

$$\boxed{\langle \mathbf{r}|\mathbf{p}_0\rangle = \frac{1}{(2\pi\hbar)^{3/2}} \exp\left(\frac{j}{\hbar}\mathbf{p}_0\cdot\mathbf{r}\right),} \tag{A.182}$$

whose magnitude is constant throughout space.

Thus,

$$\boxed{\begin{array}{rcl} \text{exact position} & \Longrightarrow & \text{completely spread momentum,} \\ \text{exact momentum} & \Longrightarrow & \text{completely spread position.} \end{array}} \tag{A.183}$$

Neither limiting state is an ordinary normalizable wave function. They are generalized states, but they already reveal a fundamental property of Fourier-transform pairs: localization in one domain requires delocalization in the conjugate domain.

A.5.6 Quantifying localization

The words "localized" and "spread out" can be made quantitative.

For clarity, consider one spatial coordinate x. Let the normalized position-space wave function be $\psi(x)$ and its Fourier transform, using the convention of this book, be

$$\widehat{\psi}(R_x) = \int \psi(x) e^{-j2\pi R_x x}\, dx. \tag{A.184}$$

The mean position is

$$\boxed{\overline{x} = \int x|\psi(x)|^2\, dx,} \tag{A.185}$$

and its standard deviation is

$$\boxed{(\Delta x)^2 = \int (x - \overline{x})^2 |\psi(x)|^2\, dx.} \tag{A.186}$$

Similarly, define the mean spatial frequency

$$\boxed{\overline{R}_x = \int R_x |\widehat{\psi}(R_x)|^2\, dR_x} \tag{A.187}$$

and its standard deviation

$$\boxed{(\Delta R_x)^2 = \int (R_x - \overline{R}_x)^2 |\widehat{\psi}(R_x)|^2\, dR_x.} \tag{A.188}$$

These quantities measure the widths of the wave function in the two Fourier-conjugate domains.

A.5.7 The Fourier uncertainty relation

The two widths cannot be made arbitrarily small simultaneously. For the Fourier-transform convention used in this book,

$$\boxed{\Delta x\, \Delta R_x \geq \frac{1}{4\pi}.} \tag{A.189}$$

This is a mathematical property of Fourier transforms. It is not specifically a quantum-mechanical postulate.

Uncertainty of a Fourier-transform pair.

Translations in x and shifts in spatial frequency do not change the widths Δx and ΔR_x. We may therefore choose the origins so that

$$\overline{x} = 0, \qquad \overline{R}_x = 0. \tag{A.190}$$

Then

$$(\Delta x)^2 = \int x^2 |\psi(x)|^2 \, dx, \tag{A.191}$$

and

$$(\Delta R_x)^2 = \int R_x^2 |\widehat{\psi}(R_x)|^2 \, dR_x. \tag{A.192}$$

The Fourier transform of the derivative is

$$\mathcal{F}\left\{\frac{d\psi}{dx}\right\} = j2\pi R_x \widehat{\psi}(R_x). \tag{A.193}$$

Using Parseval's relation,

$$\int \left|\frac{d\psi}{dx}\right|^2 dx = 4\pi^2 \int R_x^2 |\widehat{\psi}(R_x)|^2 \, dR_x \tag{A.194}$$

$$= 4\pi^2 (\Delta R_x)^2. \tag{A.195}$$

We shall assume that the wave function is sufficiently localized that

$$\boxed{\lim_{x \to \pm\infty} x|\psi(x)|^2 = 0.} \tag{A.196}$$

This condition is satisfied by the localized wave packets considered here; for example, it is satisfied by any Gaussian wave packet.

Consider now the function

$$x|\psi(x)|^2. \tag{A.197}$$

Integrating its derivative over the real line and using the fundamental theorem of calculus gives

$$\int_{-\infty}^{\infty} \frac{d}{dx}\left[x|\psi(x)|^2\right] dx = \left[x|\psi(x)|^2\right]_{-\infty}^{+\infty} \tag{A.198}$$

$$= \lim_{x\to+\infty} x|\psi(x)|^2 - \lim_{x\to-\infty} x|\psi(x)|^2 \tag{A.199}$$

$$= 0. \tag{A.200}$$

We can now expand the derivative. Since

$$|\psi(x)|^2 = \psi^*(x)\psi(x), \tag{A.201}$$

the product rule gives

$$\frac{d}{dx}\left[x|\psi(x)|^2\right] = |\psi(x)|^2 + x\frac{d}{dx}|\psi(x)|^2 \tag{A.202}$$

$$= |\psi(x)|^2 + x\left[\psi^*(x)\frac{d\psi(x)}{dx} + \psi(x)\frac{d\psi^*(x)}{dx}\right]. \tag{A.203}$$

The two terms

$$\psi^*(x)\frac{d\psi(x)}{dx} \qquad \text{and} \qquad \psi(x)\frac{d\psi^*(x)}{dx} \tag{A.204}$$

are complex conjugates of one another. Their sum is therefore twice the real part of either one:

$$\psi^*\frac{d\psi}{dx} + \psi\frac{d\psi^*}{dx} = 2\,\mathrm{Re}\left\{\psi^*\frac{d\psi}{dx}\right\}. \tag{A.205}$$

Hence the vanishing boundary integral obtained above becomes

$$0 = \int_{-\infty}^{\infty} |\psi(x)|^2\, dx + 2\,\mathrm{Re}\left\{\int_{-\infty}^{\infty} x\psi^*(x)\frac{d\psi(x)}{dx}\, dx\right\}. \tag{A.206}$$

Since ψ is normalized,

$$\int_{-\infty}^{\infty} |\psi(x)|^2\, dx = 1, \tag{A.207}$$

and therefore

$$\boxed{\mathrm{Re}\left\{\int_{-\infty}^{\infty} x\psi^*(x)\frac{d\psi(x)}{dx}\,dx\right\} = -\frac{1}{2}.} \tag{A.208}$$

To simplify the next step, define

$$I \equiv \int_{-\infty}^{\infty} x\psi^*(x)\frac{d\psi(x)}{dx}\,dx. \tag{A.209}$$

Equation (A.208) states that

$$\mathrm{Re}\{I\} = -\frac{1}{2}. \tag{A.210}$$

For any complex number z,

$$|z| \geq |\,\mathrm{Re}\{z\}|. \tag{A.211}$$

Consequently,

$$\boxed{|I| \geq \frac{1}{2},} \tag{A.212}$$

or, equivalently,

$$|I|^2 \geq \frac{1}{4}. \tag{A.213}$$

We now obtain an upper bound for the same quantity $|I|^2$. Write

$$I = \int_{-\infty}^{\infty} [x\psi(x)]^* \frac{d\psi(x)}{dx}\,dx. \tag{A.214}$$

This has the form of an inner product between the two functions

$$f(x) = x\psi(x), \qquad g(x) = \frac{d\psi(x)}{dx}. \tag{A.215}$$

The Cauchy–Schwarz inequality therefore gives

$$|I|^2 \leq \left(\int_{-\infty}^{\infty} x^2|\psi(x)|^2\,dx\right)\left(\int_{-\infty}^{\infty}\left|\frac{d\psi(x)}{dx}\right|^2 dx\right). \tag{A.216}$$

We chose the origin so that $\overline{x} = 0$. Therefore, from the definition of the position variance,

$$\int_{-\infty}^{\infty} x^2 |\psi(x)|^2 \, dx = (\Delta x)^2. \tag{A.217}$$

Moreover, from the Fourier derivative property and Parseval's relation derived above,

$$\int_{-\infty}^{\infty} \left| \frac{d\psi(x)}{dx} \right|^2 dx = 4\pi^2 (\Delta R_x)^2. \tag{A.218}$$

Hence,

$$\boxed{|I|^2 \leq 4\pi^2 (\Delta x)^2 (\Delta R_x)^2.} \tag{A.219}$$

We now have a lower and an upper bound for the same quantity:

$$\frac{1}{4} \leq |I|^2 \leq 4\pi^2 (\Delta x)^2 (\Delta R_x)^2. \tag{A.220}$$

Therefore,

$$\frac{1}{4} \leq 4\pi^2 (\Delta x)^2 (\Delta R_x)^2, \tag{A.221}$$

and, since Δx and ΔR_x are non-negative,

$$\boxed{\Delta x \, \Delta R_x \geq \frac{1}{4\pi}.} \tag{A.222}$$

A.5.8 Heisenberg's position–momentum uncertainty principle

For an electron, spatial frequency and momentum are related by Eq. (A.171),

$$p_x = h R_x. \tag{A.223}$$

Consequently,

$$\Delta p_x = h \Delta R_x. \tag{A.224}$$

Multiplying Eq. (A.189) by h gives

$$\Delta x \, \Delta p_x \geq \frac{h}{4\pi}. \tag{A.225}$$

Since

$$\hbar = \frac{h}{2\pi}, \tag{A.226}$$

we obtain

$$\boxed{\Delta x\,\Delta p_x \geq \frac{\hbar}{2}.} \tag{A.227}$$

This is Heisenberg's position–momentum uncertainty relation.

The same argument applies independently to the other Cartesian components:

$$\boxed{\Delta x\,\Delta p_x \geq \frac{\hbar}{2}, \qquad \Delta y\,\Delta p_y \geq \frac{\hbar}{2}, \qquad \Delta z\,\Delta p_z \geq \frac{\hbar}{2}.} \tag{A.228}$$

Using the wavevector

$$\mathbf{k} = 2\pi\mathbf{R}, \tag{A.229}$$

the equivalent form is

$$\boxed{\Delta x\,\Delta k_x \geq \frac{1}{2}.} \tag{A.230}$$

Thus the three equivalent versions are

$$\boxed{\begin{aligned} \Delta x\,\Delta R_x &\geq \frac{1}{4\pi}, \\ \Delta x\,\Delta k_x &\geq \frac{1}{2}, \\ \Delta x\,\Delta p_x &\geq \frac{\hbar}{2}. \end{aligned}} \tag{A.231}$$

They express exactly the same Fourier limitation, written in cycles per metre, radians per metre, or momentum units.

A.5.9 What the uncertainty principle means

Equation (A.227) is sometimes described informally by saying that "position and momentum cannot both be known exactly." While useful as a first statement, this wording can obscure the mathematical origin of the result.

The uncertainty relation says more precisely that, for a given quantum state, the probability distributions of position and momentum cannot both have arbitrarily small standard deviations.

It is therefore not merely a statement about imperfections of the measuring instrument. The limitation is already present in the Fourier relationship between the two representations of the state.

A wave function tightly localized in position must contain a broad range of spatial frequencies and therefore a broad range of momenta. Conversely,

a wave function containing only a narrow range of momenta must extend over a large region of real space.

This is precisely the same phenomenon encountered in signal processing:

$$\boxed{\begin{array}{rcl} \text{narrow signal in real space} & \Longrightarrow & \text{broad Fourier spectrum,} \\ \text{narrow Fourier spectrum} & \Longrightarrow & \text{broad signal in real space.} \end{array}} \tag{A.232}$$

The quantum-mechanical position–momentum uncertainty relation is this Fourier principle together with the physical identification

$$\mathbf{p} = h\mathbf{R}. \tag{A.233}$$

A.5.10 The Gaussian reaches the minimum uncertainty

The lower bound in the Fourier uncertainty relation, Eq. (A.189), is attained by a Gaussian wave packet. This example is also useful for distinguishing carefully between the width of a wave-function amplitude and the width of the corresponding probability density.

Consider the one-dimensional wave packet

$$\boxed{\psi(x) = \frac{1}{(2\pi\sigma_x^2)^{1/4}} \exp\left[-\frac{(x-x_0)^2}{4\sigma_x^2}\right] \exp(j2\pi R_0 x).} \tag{A.234}$$

The parameters have a direct interpretation:

$$x_0 = \text{mean position}, \qquad R_0 = \text{mean spatial frequency}, \tag{A.235}$$

while σ_x will be shown below to be the standard deviation of the *position probability distribution*.

The two factors in Eq. (A.234) play different roles. The Gaussian

$$\exp\left[-\frac{(x-x_0)^2}{4\sigma_x^2}\right] \tag{A.236}$$

determines the spatial envelope of the wave function, whereas

$$\exp(j2\pi R_0 x) \tag{A.237}$$

is a plane-wave modulation that shifts its Fourier spectrum so that it is centred at R_0.

Position-space probability distribution

The wave function itself is a probability amplitude. The position probability density is obtained from its squared modulus:

$$|\psi(x)|^2 = \psi^*(x)\psi(x) \tag{A.238}$$

$$= \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{(x-x_0)^2}{2\sigma_x^2}\right]. \tag{A.239}$$

Therefore,

$$\boxed{|\psi(x)|^2 = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{(x-x_0)^2}{2\sigma_x^2}\right].} \tag{A.240}$$

This is the usual normalized Gaussian probability density with mean x_0 and variance σ_x^2. Hence,

$$\boxed{\overline{x} = x_0, \qquad \Delta x = \sigma_x.} \tag{A.241}$$

The factor $4\sigma_x^2$ in the exponent of the wave function in Eq. (A.234) is therefore intentional. Squaring the wave-function amplitude doubles its exponent:

$$\left|\exp\left[-\frac{(x-x_0)^2}{4\sigma_x^2}\right]\right|^2 = \exp\left[-\frac{(x-x_0)^2}{2\sigma_x^2}\right]. \tag{A.242}$$

Thus σ_x is the standard deviation of the physically relevant probability density $|\psi(x)|^2$. If the Gaussian envelope of $\psi(x)$ itself were compared with a function of the conventional form $\exp[-(x-x_0)^2/(2s^2)]$, its scale parameter would instead be $s = \sqrt{2}\sigma_x$. This scale should not be confused with the position uncertainty Δx.

Fourier-space probability distribution

Using the Fourier-transform convention of this book,

$$\widehat{\psi}(R_x) = \int_{-\infty}^{\infty} \psi(x) \exp(-j2\pi R_x x)\, dx, \tag{A.243}$$

the Fourier transform of Eq. (A.234) is another Gaussian:

$$\boxed{\widehat{\psi}(R_x) = (8\pi\sigma_x^2)^{1/4} \exp\left[-4\pi^2\sigma_x^2(R_x-R_0)^2\right] \exp\left[-j2\pi(R_x-R_0)x_0\right].} \tag{A.244}$$

Several features of this expression are worth noticing. The factor $\exp(j2\pi R_0 x)$ in real space shifts the Fourier-space Gaussian from the origin to $R_x = R_0$. Conversely, the displacement x_0 in real space appears only as a phase factor in Fourier space and therefore does not affect the magnitude of the Fourier spectrum.

As in real space, the Fourier amplitude $\widehat{\psi}(R_x)$ is not itself a probability density. The probability density for spatial frequency is

$$|\widehat{\psi}(R_x)|^2 = \sqrt{8\pi\sigma_x^2}\exp\left[-8\pi^2\sigma_x^2(R_x - R_0)^2\right]. \tag{A.245}$$

Hence,

$$\boxed{|\widehat{\psi}(R_x)|^2 = \sqrt{8\pi\sigma_x^2}\exp\left[-8\pi^2\sigma_x^2(R_x - R_0)^2\right].} \tag{A.246}$$

To identify its standard deviation, compare its exponent with the standard Gaussian probability density

$$\exp\left[-\frac{(R_x - R_0)^2}{2\sigma_R^2}\right]. \tag{A.247}$$

Equating the coefficients of $(R_x - R_0)^2$ gives

$$\frac{1}{2\sigma_R^2} = 8\pi^2\sigma_x^2, \tag{A.248}$$

and therefore

$$\boxed{\sigma_R = \Delta R_x = \frac{1}{4\pi\sigma_x}.} \tag{A.249}$$

Thus the two probability distributions have standard deviations

$$\boxed{\Delta x = \sigma_x, \qquad \Delta R_x = \frac{1}{4\pi\sigma_x}.} \tag{A.250}$$

Their product is

$$\boxed{\Delta x\,\Delta R_x = \sigma_x\frac{1}{4\pi\sigma_x} = \frac{1}{4\pi}.} \tag{A.251}$$

The Gaussian therefore attains equality in the Fourier uncertainty relation of Eq. (A.189).

Position–momentum uncertainty

For an electron, Eq. (A.171) gives

$$p_x = hR_x. \tag{A.252}$$

A linear change of variable scales the standard deviation by the same factor, so

$$\Delta p_x = h\Delta R_x. \tag{A.253}$$

Using Eq. (A.249),

$$\Delta p_x = \frac{h}{4\pi\sigma_x} \tag{A.254}$$

$$= \frac{\hbar}{2\sigma_x}. \tag{A.255}$$

Since $\Delta x = \sigma_x$,

$$\boxed{\Delta x\,\Delta p_x = \frac{\hbar}{2}.} \tag{A.256}$$

The Gaussian wave packet therefore also attains equality in Heisenberg's position–momentum uncertainty relation.

This example makes the reciprocal character of the two representations particularly clear. Decreasing σ_x produces a more localized wave packet in real space but increases ΔR_x and therefore Δp_x. Increasing σ_x has the opposite effect. The Gaussian is special because, among wave functions with finite variances, it achieves the smallest possible product of these two widths.

A.5.11 Does an electron have a preferred wave-packet shape?

The Gaussian wave packet studied above is particularly useful, but it is important not to draw the wrong conclusion from this example:

$$\boxed{\text{an electron does not have an intrinsic Gaussian “shape”.}} \tag{A.257}$$

More generally, it is better not to interpret the shape of $\psi(\mathbf{r})$ as the geometrical shape of the electron itself. The function $\psi(\mathbf{r})$ is the position-space representation of the electron's quantum state,

$$\psi(\mathbf{r}) = \langle \mathbf{r}|\psi\rangle, \tag{A.258}$$

and its form depends on how the electron has been prepared and on how it has subsequently evolved.

A Gaussian is only one possible quantum state.

A free electron does not have a predefined spatial form

The word *free* has a precise meaning: the electron is not interacting with an external potential. It specifies the dynamics of the electron, but it does not specify a unique initial quantum state.

As established in Eq. (A.151), an arbitrary electron state can be written as a superposition of momentum states,

$$\boxed{|\psi\rangle = \int \widetilde{\psi}(\mathbf{p})|\mathbf{p}\rangle \, d^3p.} \tag{A.259}$$

The function $\widetilde{\psi}(\mathbf{p})$ determines how much of each momentum component is present. Subject to the usual physical requirements on the state, it need not be Gaussian.

Consequently, a free electron may be described by many different wave packets. Their spatial forms are determined by the corresponding superpositions of momentum states.

Plane waves are a special limiting case

A particularly important free-electron state is one with exactly defined momentum $\mathbf{p}_0$. From Eq. (A.154), its position representation is a plane wave,

$$\boxed{\langle \mathbf{r}|\mathbf{p}_0\rangle \propto \exp\left(\frac{j}{\hbar}\mathbf{p}_0 \cdot \mathbf{r}\right).} \tag{A.260}$$

Thus a plane wave is not the compulsory shape of a free electron either. It is the special idealized state corresponding to an exactly defined momentum.

As discussed in Eq. (A.147), such a state is not normalizable in the ordinary sense: it extends over all space and is delta-normalized. A physically localized electron must instead contain a range of momenta and is therefore represented by a wave packet.

The distinction can be summarized as

$$\boxed{\begin{array}{lcl} \text{exact momentum} & \longleftrightarrow & \text{plane wave,} \\ \text{range of momenta} & \longleftrightarrow & \text{localized wave packet.} \end{array}} \tag{A.261}$$

Why Gaussian wave packets are so often used

The Gaussian is nevertheless a particularly convenient wave packet.

First, as shown in Eq. (A.256), it reaches the smallest possible position–momentum uncertainty product,

$$\Delta x \, \Delta p_x = \frac{\hbar}{2}. \tag{A.262}$$

Second, its Fourier transform is another Gaussian, as seen in Eq. (A.244). Its localization in position and momentum can therefore be described by simple analytical expressions.

Third, under non-relativistic free propagation, an initially Gaussian packet remains Gaussian: its centre moves and its width changes, but its functional form remains particularly simple. More complicated initial wave packets generally evolve into more complicated spatial patterns.

These properties make Gaussian packets extremely useful analytical models. They do not, however, represent a fundamental requirement of quantum mechanics.

The distinction is useful in electron microscopy

Different idealizations are useful in different electron-microscopy problems.

For approximately uniform illumination, the incoming electron is often represented by a plane wave,

$$\psi(\mathbf{r}) \propto \exp(j\mathbf{k}_0 \cdot \mathbf{r}), \tag{A.263}$$

because over the region of interest its momentum is strongly concentrated around one propagation direction.

A focused electron probe, by contrast, necessarily contains a range of transverse momenta and is naturally described as a superposition of plane waves. Its spatial profile need not be Gaussian.

Thus neither a plane wave nor a Gaussian wave packet should be regarded as the intrinsic form of an electron. They are different quantum states—or useful idealizations of quantum states—chosen according to how the electron has been prepared and to the physical problem being described.

The important distinction is therefore

$$\boxed{\begin{array}{c} \text{“free electron”} \\ \Downarrow \\ \text{specifies the absence of an external interaction} \\ \\ \neq \\ \text{a prescribed wave-function shape.} \end{array}} \tag{A.264}$$

A.5.12 Connection with electron microscopy

The same Fourier trade-off appears repeatedly in electron microscopy.

A highly localized electron probe requires a broad range of transverse spatial frequencies,

$$R_x,\ R_y, \tag{A.265}$$

and therefore a broad range of transverse momenta

$$p_x = hR_x, \qquad p_y = hR_y. \tag{A.266}$$

In geometrical terms, this corresponds to a broad range of propagation angles.

Conversely, restricting the angular spectrum with a narrow aperture reduces the available transverse momenta and necessarily broadens the electron distribution in real space. This is the same Fourier principle responsible for the finite width of a diffraction-limited point-spread function.

Likewise, a sharply localized feature of a specimen requires broad Fourier content for its representation. Removing high spatial frequencies from an image necessarily spreads sharp features over a larger region of real space.

Thus the uncertainty principle should not appear as a foreign concept introduced only for quantum mechanics. It is one manifestation of a Fourier-space trade-off that has already appeared throughout the diffraction and imaging theory of this book.

A.5.13 Summary: three descriptions of the same state

We can now summarize the relation among the abstract state, its position representation, and its momentum or spatial-frequency representation:

$$\boxed{\begin{array}{rcl} |\psi\rangle & : & \text{abstract quantum state,} \\ \psi(\mathbf{r}) = \langle \mathbf{r}|\psi\rangle & : & \text{position representation,} \\ \widetilde{\psi}(\mathbf{p}) = \langle \mathbf{p}|\psi\rangle & : & \text{momentum representation,} \\ \widehat{\psi}(\mathbf{R}) = \langle \mathbf{R}|\psi\rangle & : & \text{spatial-frequency representation,} \\ \mathbf{p} = \hbar\mathbf{k} = h\mathbf{R} & : & \text{relation between their coordinates.} \end{array}} \tag{A.267}$$

The physical state $|\psi\rangle$ has not changed. Position space and momentum space are simply two different bases in which the same state is represented, and the Fourier transform is the change of basis between them.

A.6 Operators

A physical operation performed on a quantum state is represented by an *operator*. We denote a generic operator by

$$\hat{A}. \tag{A.268}$$

It acts on a ket to produce another ket,

$$\boxed{|\phi\rangle = \hat{A}|\psi\rangle.} \tag{A.269}$$

The hat distinguishes an operator from an ordinary scalar quantity when the distinction might otherwise be ambiguous. Fourier transforms in this book are also indicated with a hat, for example $\widehat{f}(\mathbf{R})$, but the meaning is always clear from context: $\hat{A}$ acts on a state whereas $\widehat{f}$ is a transformed function.

In the position representation, many operators become familiar mathematical operations. The position operator acts by multiplication,

$$\boxed{\hat{\mathbf{r}}\,\psi(\mathbf{r}) = \mathbf{r}\,\psi(\mathbf{r}),} \tag{A.270}$$

whereas the momentum operator is

$$\boxed{\hat{\mathbf{p}} = -j\hbar\nabla.} \tag{A.271}$$

Indeed, acting on the plane wave of Eq. (A.155),

$$-j\hbar\nabla \exp\left(\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}\right) = \mathbf{p}\exp\left(\frac{j}{\hbar}\mathbf{p}\cdot\mathbf{r}\right). \tag{A.272}$$

Thus the plane wave is a state of definite momentum.

A.7 Eigenstates, observables, and measurement

We have already seen that a linear operator acts on a state to produce another state. Among all possible states, some have a particularly simple relation with a given operator.

Let $\hat{A}$ be an operator. A nonzero state $|u_n\rangle$ is called an *eigenstate* of $\hat{A}$ if

$$\boxed{\hat{A}|u_n\rangle = a_n|u_n\rangle.} \tag{A.273}$$

The scalar a_n is the corresponding *eigenvalue*.

We shall use this notation consistently:

$$\boxed{|u_n\rangle \quad \text{is a state,} \qquad a_n \quad \text{is a number.}} \tag{A.274}$$

Equation (A.273) means that applying $\hat{A}$ to the eigenstate does not change its direction in Hilbert space; it only multiplies it by the scalar a_n.

This is exactly the familiar eigenvalue problem of linear algebra.

A.7.1 Hermitian operators

Physical observables such as position, momentum, and energy are represented in quantum mechanics by Hermitian operators.

The adjoint $\hat{A}^\dagger$ of an operator $\hat{A}$ is defined by

$$\boxed{\langle\phi|\hat{A}|\psi\rangle = \langle\hat{A}^\dagger\phi|\psi\rangle} \tag{A.275}$$

for arbitrary states $|\phi\rangle$ and $|\psi\rangle$.

In a finite-dimensional orthonormal basis, this is the same adjoint operation introduced previously for vectors and matrices: complex conjugation followed by transposition.

An operator is *Hermitian* when

$$\boxed{\hat{A}^\dagger = \hat{A}.} \tag{A.276}$$

In matrix notation this condition becomes

$$\boxed{A_{mn} = A^*_{nm}.} \tag{A.277}$$

Thus a real symmetric matrix is a particular example of a Hermitian matrix.

A.7.2 Eigenvalues of a Hermitian operator are real

Hermiticity has an immediate consequence for the eigenvalues.

Why the eigenvalues of a Hermitian operator are real.

Let $|u_n\rangle$ be an eigenstate of $\hat{A}$ with eigenvalue a_n,

$$\hat{A}|u_n\rangle = a_n|u_n\rangle. \tag{A.278}$$

Taking the inner product with $\langle u_n|$ gives

$$\langle u_n|\hat{A}|u_n\rangle = a_n\langle u_n|u_n\rangle. \tag{A.279}$$

Now take the complex conjugate:

$$\Big(\langle u_n|\hat{A}|u_n\rangle\Big)^* = a_n^*\langle u_n|u_n\rangle. \tag{A.280}$$

From the definition of the adjoint, Eq. (A.275),

$$\Big(\langle u_n|\hat{A}|u_n\rangle\Big)^* = \langle u_n|\hat{A}^\dagger|u_n\rangle. \tag{A.281}$$

If $\hat{A}$ is Hermitian, Eq. (A.276),

$$\Big(\langle u_n|\hat{A}|u_n\rangle\Big)^* = \langle u_n|\hat{A}|u_n\rangle. \tag{A.282}$$

Comparing Eqs. (A.279) and (A.280),

$$a_n\langle u_n|u_n\rangle = a_n^*\langle u_n|u_n\rangle. \tag{A.283}$$

Since $|u_n\rangle$ is a nonzero state,

$$\langle u_n|u_n\rangle > 0, \tag{A.284}$$

and therefore

$$\boxed{a_n = a_n^*.} \tag{A.285}$$

Hence

$$\boxed{a_n \in \mathbb{R}.} \tag{A.286}$$

Thus Hermiticity guarantees that the eigenvalues are real numbers. This mathematical property is one reason Hermitian operators are appropriate for representing measurable physical quantities.

A.7.3 Eigenstates with different eigenvalues are orthogonal

A second important consequence of Hermiticity concerns the geometry of the eigenstates.

Why different eigenvalues imply orthogonal eigenstates.

Let

$$\hat{A}|u_m\rangle = a_m|u_m\rangle \tag{A.287}$$

and

$$\hat{A}|u_n\rangle = a_n|u_n\rangle. \tag{A.288}$$

Starting from the second equation and multiplying from the left by $\langle u_m|$,

$$\langle u_m|\hat{A}|u_n\rangle = a_n\langle u_m|u_n\rangle. \tag{A.289}$$

Using Hermiticity,

$$\langle u_m|\hat{A}|u_n\rangle = \langle \hat{A}u_m|u_n\rangle \tag{A.290}$$

$$= a_m^*\langle u_m|u_n\rangle. \tag{A.291}$$

Since the eigenvalues are real, Eq. (A.286),

$$\langle u_m|\hat{A}|u_n\rangle = a_m\langle u_m|u_n\rangle. \tag{A.292}$$

Comparing Eqs. (A.289) and (A.292),

$$(a_m - a_n)\langle u_m|u_n\rangle = 0. \tag{A.293}$$

If

$$a_m \neq a_n, \tag{A.294}$$

then

$$\boxed{\langle u_m|u_n\rangle = 0.} \tag{A.295}$$

Thus eigenstates belonging to different eigenvalues of a Hermitian operator are orthogonal.

If several independent eigenstates share the same eigenvalue, the eigenvalue is said to be *degenerate*. Hermiticity does not force states inside the same degenerate subspace to be orthogonal, but an orthonormal basis can always be chosen within that subspace.

A.7.4 Expanding a state in eigenstates

Earlier in this appendix we saw that an arbitrary state can be expanded in any complete orthonormal basis. If the eigenstates of $\hat{A}$ form such a discrete

basis, we may therefore write

$$\boxed{|\psi\rangle = \sum_n c_n |u_n\rangle,} \tag{A.296}$$

with coefficients

$$\boxed{c_n = \langle u_n|\psi\rangle.} \tag{A.297}$$

Nothing specifically quantum mechanical has been added in these two equations: they are simply the basis decomposition already developed previously, now using the eigenstates of $\hat{A}$ as the basis.

For a normalized state,

$$\boxed{\sum_n |c_n|^2 = 1.} \tag{A.298}$$

A.7.5 From eigenvalues to measurement outcomes

Up to this point, the discussion has been mathematical. Hermiticity tells us that the eigenvalues a_n are real and that eigenspaces belonging to different eigenvalues are orthogonal. It does not, by itself, tell us what an experiment will measure.

Quantum mechanics adds a physical postulate connecting the operator $\hat{A}$ with the observable A that it represents.

For an ideal measurement of A, the *measurement postulate* states that the possible measurement outcomes are the eigenvalues of $\hat{A}$:

$$\boxed{\hat{A}|u_n\rangle = a_n|u_n\rangle \qquad \Longrightarrow \qquad A = a_n \quad \text{is a possible measurement result.}} \tag{A.299}$$

This connection between eigenvalues and measurement outcomes is a physical postulate; it is not a consequence of Hermiticity.

A.7.6 The Born rule

The measurement postulate tells us which values may occur. The *Born rule* assigns their probabilities.

For the state Eq. (A.296),

$$|\psi\rangle = \sum_n c_n |u_n\rangle, \tag{A.300}$$

the probability of obtaining the eigenvalue a_n is

$$\boxed{p_n \equiv P(A = a_n) = |\langle u_n|\psi\rangle|^2 = |c_n|^2.} \tag{A.301}$$

Thus the coefficient c_n is a complex probability amplitude, while $p_n = |c_n|^2$ is the corresponding probability.

The notation used throughout the remainder of this appendix is therefore

$$\boxed{\begin{array}{rcl} A & : & \text{physical observable,} \\ \hat{A} & : & \text{operator representing } A, \\ |u_n\rangle & : & \text{eigenstate,} \\ a_n & : & \text{eigenvalue,} \\ c_n = \langle u_n|\psi\rangle & : & \text{probability amplitude,} \\ p_n = |c_n|^2 & : & \text{probability.} \end{array}} \tag{A.302}$$

For example, if

$$|\psi\rangle = \frac{\sqrt{3}}{2}|u_1\rangle + \frac{1}{2}|u_2\rangle, \tag{A.303}$$

then

$$P(A = a_1) = \frac{3}{4}, \qquad P(A = a_2) = \frac{1}{4}. \tag{A.304}$$

If instead

$$|\psi\rangle = |u_k\rangle, \tag{A.305}$$

then

$$\boxed{P(A = a_k) = 1.} \tag{A.306}$$

An eigenstate therefore has a definite value of the corresponding observable in the sense that every ideal measurement of that observable returns the same eigenvalue.

A.7.7 Expected value

Suppose that the measurement of A is repeated on many independently prepared systems, all in the same state $|\psi\rangle$.

The outcomes form the ordinary discrete probability distribution

$$\{a_n, p_n\}. \tag{A.307}$$

Its mean is

$$\boxed{\langle A\rangle = \sum_n a_n p_n = \sum_n a_n |c_n|^2.} \tag{A.308}$$

This same quantity can be written directly in operator notation.

Expectation value in operator form.

Using Eq. (A.296),

$$|\psi\rangle = \sum_n c_n |u_n\rangle, \tag{A.309}$$

and

$$\langle\psi| = \sum_m c_m^* \langle u_m|, \tag{A.310}$$

we obtain

$$\langle\psi|\hat{A}|\psi\rangle = \sum_{m,n} c_m^* c_n \langle u_m|\hat{A}|u_n\rangle \tag{A.311}$$

$$= \sum_{m,n} c_m^* c_n a_n \langle u_m|u_n\rangle \tag{A.312}$$

$$= \sum_n a_n |c_n|^2 \tag{A.313}$$

$$= \langle A\rangle. \tag{A.314}$$

Therefore,

$$\boxed{\langle A\rangle = \langle\psi|\hat{A}|\psi\rangle.} \tag{A.315}$$

The quantity $\langle A\rangle$ is therefore the statistical mean of the measurement outcomes. It should not be confused with one particular eigenvalue a_n.

A.7.8 Spread of the measurement outcomes

The same probability distribution also determines the spread of the possible outcomes. Its variance is

$$\boxed{(\Delta A)^2 = \sum_n (a_n - \langle A\rangle)^2 p_n.} \tag{A.316}$$

Thus ΔA is simply the standard deviation of the results of repeated ideal measurements of A on identically prepared systems.

As for any probability distribution,

$$\boxed{(\Delta A)^2 = \langle A^2\rangle - \langle A\rangle^2,} \tag{A.317}$$

where

$$\boxed{\langle A^2\rangle = \sum_n a_n^2 p_n.} \tag{A.318}$$

Using the Born rule,

$$p_n = |c_n|^2, \tag{A.319}$$

this becomes

$$\langle A^2 \rangle = \sum_n a_n^2 |c_n|^2. \tag{A.320}$$

We now want to express the same quantity directly in terms of the operator $\hat{A}$.

First note that, if $|u_n\rangle$ is an eigenstate of $\hat{A}$,

$$\hat{A}|u_n\rangle = a_n|u_n\rangle, \tag{A.321}$$

then applying $\hat{A}$ a second time gives

$$\hat{A}^2|u_n\rangle = \hat{A}\left(\hat{A}|u_n\rangle\right) \tag{A.322}$$

$$= \hat{A}\left(a_n|u_n\rangle\right). \tag{A.323}$$

Since a_n is an ordinary scalar number, it can be taken outside the action of the operator:

$$\hat{A}^2|u_n\rangle = a_n\hat{A}|u_n\rangle \tag{A.324}$$

$$= a_n\left(a_n|u_n\rangle\right) \tag{A.325}$$

$$= a_n^2|u_n\rangle. \tag{A.326}$$

Therefore,

$$\boxed{\hat{A}^2|u_n\rangle = a_n^2|u_n\rangle.} \tag{A.327}$$

Thus $|u_n\rangle$ is also an eigenstate of $\hat{A}^2$, and the corresponding eigenvalue is a_n^2.

We can now evaluate the expectation value of $\hat{A}^2$ in the state

$$|\psi\rangle = \sum_n c_n|u_n\rangle. \tag{A.328}$$

Why $\langle A^2 \rangle = \langle\psi|\hat{A}^2|\psi\rangle$.

Using

$$|\psi\rangle = \sum_n c_n|u_n\rangle, \tag{A.329}$$

and

$$\langle\psi| = \sum_m c_m^*\langle u_m|, \tag{A.330}$$

we obtain

$$\langle\psi|\hat{A}^2|\psi\rangle = \sum_{m,n} c_m^* c_n \langle u_m|\hat{A}^2|u_n\rangle. \tag{A.331}$$

Using Eq. (A.327),

$$\hat{A}^2|u_n\rangle = a_n^2|u_n\rangle, \tag{A.332}$$

so

$$\langle\psi|\hat{A}^2|\psi\rangle = \sum_{m,n} c_m^* c_n a_n^2 \langle u_m|u_n\rangle. \tag{A.333}$$

Because the eigenstates are orthonormal,

$$\langle u_m|u_n\rangle = \delta_{mn}, \tag{A.334}$$

all terms with $m \neq n$ vanish:

$$\langle\psi|\hat{A}^2|\psi\rangle = \sum_{n} a_n^2 |c_n|^2. \tag{A.335}$$

Comparing with Eq. (A.320), we obtain

$$\boxed{\langle A^2\rangle = \langle\psi|\hat{A}^2|\psi\rangle.} \tag{A.336}$$

We can now return to the variance. From ordinary probability theory,

$$\boxed{(\Delta A)^2 = \langle A^2\rangle - \langle A\rangle^2.} \tag{A.337}$$

Using Eq. (A.315) and Eq. (A.336),

$$\langle A\rangle = \langle\psi|\hat{A}|\psi\rangle, \tag{A.338}$$

and

$$\langle A^2\rangle = \langle\psi|\hat{A}^2|\psi\rangle. \tag{A.339}$$

Therefore,

$$(\Delta A)^2 = \langle\psi|\hat{A}^2|\psi\rangle - \langle A\rangle^2. \tag{A.340}$$

The same result can be written in a more compact form by introducing the operator

$$\boxed{\hat{A} - \langle A\rangle\hat{I},} \tag{A.341}$$

where $\hat{I}$ is the identity operator.

The identity operator is needed because $\hat{A}$ is an operator, whereas $\langle A \rangle$ is an ordinary number. Thus

$$\langle A \rangle \hat{I} \tag{A.342}$$

is the operator that multiplies every state by the scalar $\langle A \rangle$.

Why the variance is the expectation of the squared deviation operator.

Consider

$$\left(\hat{A} - \langle A \rangle \hat{I}\right)^2. \tag{A.343}$$

Expanding the square,

$$\left(\hat{A} - \langle A \rangle \hat{I}\right)^2 = \hat{A}^2 - 2\langle A \rangle \hat{A} + \langle A \rangle^2 \hat{I}. \tag{A.344}$$

Now take the expectation value in the normalized state $|\psi\rangle$:

$$\langle \psi | \left(\hat{A} - \langle A \rangle \hat{I}\right)^2 |\psi\rangle \tag{A.345}$$

$$= \langle \psi | \hat{A}^2 | \psi \rangle - 2\langle A \rangle \langle \psi | \hat{A} | \psi \rangle + \langle A \rangle^2 \langle \psi | \hat{I} | \psi \rangle. \tag{A.346}$$

Using

$$\langle \psi | \hat{A}^2 | \psi \rangle = \langle A^2 \rangle, \tag{A.347}$$

$$\langle \psi | \hat{A} | \psi \rangle = \langle A \rangle, \tag{A.348}$$

and, because $|\psi\rangle$ is normalized,

$$\langle \psi | \hat{I} | \psi \rangle = \langle \psi | \psi \rangle = 1, \tag{A.349}$$

we obtain

$$\langle \psi | \left(\hat{A} - \langle A \rangle \hat{I}\right)^2 |\psi\rangle \tag{A.350}$$

$$= \langle A^2 \rangle - 2\langle A \rangle^2 + \langle A \rangle^2 \tag{A.351}$$

$$= \langle A^2 \rangle - \langle A \rangle^2. \tag{A.352}$$

Using Eq. (A.317), this is precisely the variance:

$$\boxed{(\Delta A)^2 = \langle \psi | \left(\hat{A} - \langle A \rangle \hat{I}\right)^2 |\psi\rangle.} \tag{A.353}$$

The operator

$$\hat{A} - \langle A \rangle \hat{I} \tag{A.354}$$

therefore plays the same role as the deviation

$$a_n - \langle A \rangle \tag{A.355}$$

in ordinary statistics. Squaring it and taking its expectation value produces the variance of the measurement outcomes.

If $|\psi\rangle = |u_k\rangle$, then

$$p_k = 1, \tag{A.356}$$

so

$$\langle A \rangle = a_k, \qquad \Delta A = 0. \tag{A.357}$$

This is the general operator version of the idea already encountered for position and momentum: zero uncertainty means that the state has a definite value of the observable.

A.7.9 Degenerate eigenvalues

Several different eigenstates may correspond to the same eigenvalue. Let

$$|u_{n,\mu}\rangle \tag{A.358}$$

denote the different eigenstates associated with the eigenvalue a_n:

$$\boxed{\hat{A}|u_{n,\mu}\rangle = a_n|u_{n,\mu}\rangle.} \tag{A.359}$$

The index n identifies the eigenvalue, while μ distinguishes states within the corresponding eigenspace.

If these states are chosen orthonormal, the probability of obtaining the value a_n is the sum of the probabilities of all components in that eigenspace:

$$\boxed{P(A = a_n) = \sum_\mu |\langle u_{n,\mu}|\psi\rangle|^2 .} \tag{A.360}$$

Using the projector

$$\hat{P}_n = \sum_\mu |u_{n,\mu}\rangle\langle u_{n,\mu}|, \tag{A.361}$$

this can also be written

$$\boxed{P(A = a_n) = \langle\psi|\hat{P}_n|\psi\rangle.} \tag{A.362}$$

A.7.10 What is derived and what is postulated

The logical structure of this section is worth keeping explicit:

$$\boxed{\begin{aligned} \hat{A}^\dagger = \hat{A} \quad &\Longrightarrow \quad \text{real eigenvalues and orthogonal eigenspaces,} \\ \text{measurement postulate} \quad &\Longrightarrow \quad \text{possible ideal-measurement outcomes are the } a_n, \\ \text{Born rule} \quad &\Longrightarrow \quad P(A = a_n) = p_n, \\ \text{ordinary probability theory} \quad &\Longrightarrow \quad \langle A \rangle = \sum_n a_n p_n, \quad (\Delta A)^2 = \sum_n (a_n - \langle A \rangle)^2 p_n. \end{aligned}} \tag{A.363}$$

The first line is a mathematical consequence of Hermiticity. The next two lines are physical postulates connecting the mathematical formalism to measurement. Once the probabilities p_n are known, the mean and variance follow from ordinary probability theory.

A.8 Discrete and continuous measurement probabilities

In Sec. A.7.5 we considered an observable A whose operator $\hat{A}$ has a discrete set of orthonormal eigenstates $|u_n\rangle$,

$$\hat{A}|u_n\rangle = a_n|u_n\rangle. \tag{A.364}$$

For a state expanded in this basis,

$$|\psi\rangle = \sum_n c_n|u_n\rangle, \qquad c_n = \langle u_n|\psi\rangle, \tag{A.365}$$

the Born rule, Eq. (A.301), gives

$$\boxed{P(A = a_n) = |\langle u_n|\psi\rangle|^2 = |c_n|^2.} \tag{A.366}$$

There is no need for a separate rule when the observable has a continuous spectrum. The structure is the same, but the discrete basis is replaced by a continuously labelled basis.

Position provides the most important example for this book. As shown in Sec. A.2.1,

$$\psi(\mathbf{r}) = \langle \mathbf{r}|\psi\rangle \tag{A.367}$$

is the coefficient of $|\psi\rangle$ in the position basis.

Because position is continuous,

$$|\psi(\mathbf{r})|^2 \tag{A.368}$$

must be interpreted as a *probability density*, rather than as the probability of obtaining one exact value of $\mathbf{r}$. Consequently,

$$\boxed{P(\mathbf{r} \in \Omega) = \int_\Omega |\langle \mathbf{r}|\psi\rangle|^2 \, d^3r = \int_\Omega |\psi(\mathbf{r})|^2 \, d^3r.} \tag{A.369}$$

Thus the discrete and continuous forms of the Born rule have the same geometrical meaning:

$$\boxed{\begin{aligned} \text{discrete basis} \quad &: \quad P(A = a_n) = |\langle u_n|\psi\rangle|^2, \\ \text{continuous position basis} \quad &: \quad p(\mathbf{r}) = |\langle \mathbf{r}|\psi\rangle|^2. \end{aligned}} \tag{A.370}$$

In both cases, one first projects the state onto the basis state associated with a possible outcome and then takes the squared modulus of the resulting amplitude. The only difference is that a continuous observable requires a probability density and an integral over a finite region.

A.9 Expectation values in a chosen representation

The expectation value of an observable has already been derived in Sec. A.7.7. For a normalized state $|\psi\rangle$,

$$\boxed{\langle A\rangle = \langle\psi|\hat{A}|\psi\rangle.} \tag{A.371}$$

This expression is basis independent. It refers to the abstract state $|\psi\rangle$ and the abstract operator $\hat{A}$, without choosing position, momentum, or any other representation.

The quantity $\langle A\rangle$ is a fixed number determined by the probability distribution of the possible measurement outcomes. In general, it need not coincide with any particular eigenvalue a_n.

For calculations in electron microscopy, however, the position representation is particularly useful. We therefore derive the form of Eq. (A.371) in that representation.

Expectation value in the position representation.

From the position-basis decomposition, Eq. (A.66),

$$|\psi\rangle = \int |\mathbf{r}\rangle\langle\mathbf{r}|\psi\rangle \, d^3r. \tag{A.372}$$

Using Eq. (A.70),

$$\langle\mathbf{r}|\psi\rangle = \psi(\mathbf{r}), \tag{A.373}$$

we may write

$$|\psi\rangle = \int \psi(\mathbf{r})|\mathbf{r}\rangle \, d^3r. \tag{A.374}$$

Now consider the state obtained after applying the operator:

$$\hat{A}|\psi\rangle. \tag{A.375}$$

Its position-space representation is

$$\boxed{\langle\mathbf{r}|\hat{A}|\psi\rangle = \left[\hat{A}\psi\right](\mathbf{r}).} \tag{A.376}$$

The notation on the right means: represent the state $\hat{A}|\psi\rangle$ as a function of position.

Using the position-space form of the inner product, Eq. (A.128), we therefore obtain

$$\langle A\rangle = \langle\psi|\hat{A}|\psi\rangle \tag{A.377}$$

$$= \int_{\mathbb{R}^3} \psi^*(\mathbf{r}) \, \left[\hat{A}\psi\right](\mathbf{r}) \, d^3r. \tag{A.378}$$

Hence,

$$\boxed{\langle A\rangle = \int_{\mathbb{R}^3} \psi^*(\mathbf{r}) \, \hat{A}\psi(\mathbf{r}) \, d^3r.} \tag{A.379}$$

Here the conventional notation $\hat{A}\psi(\mathbf{r})$ means the position-space representation of the state $\hat{A}|\psi\rangle$.

A.9.1 Example: mean position

For the position operator,

$$\hat{\mathbf{r}}\psi(\mathbf{r}) = \mathbf{r}\psi(\mathbf{r}). \tag{A.380}$$

Substituting this expression into Eq. (A.379),

$$\langle \mathbf{r} \rangle = \int \psi^*(\mathbf{r}) \mathbf{r} \psi(\mathbf{r}) \, d^3r \tag{A.381}$$

$$= \int \mathbf{r} |\psi(\mathbf{r})|^2 \, d^3r. \tag{A.382}$$

Therefore,

$$\boxed{\langle \mathbf{r} \rangle = \int_{\mathbb{R}^3} \mathbf{r} |\psi(\mathbf{r})|^2 \, d^3r.} \tag{A.383}$$

This is precisely the ordinary mean of the position probability density

$$p(\mathbf{r}) = |\psi(\mathbf{r})|^2. \tag{A.384}$$

Similarly, for any observable whose position-space operator is known, Eq. (A.379) provides its expectation value.

A.9.2 The imaging electron interacting with the specimen

Until now, most of our discussion has concerned a single quantum system described by one state $|\psi\rangle$. Electron microscopy, however, involves at least two physically distinct systems:

1. the high-energy electron travelling through the microscope, which we shall call the *imaging electron*;
2. the specimen with which that electron interacts.

We therefore need a notation that distinguishes the state of the imaging electron from the state of the specimen.

Let

$$|\psi\rangle_{\mathrm{e}} \tag{A.385}$$

denote the state of the imaging electron, and let

$$\boxed{|\Xi_0\rangle_{\mathrm{s}}} \tag{A.386}$$

denote the initial internal quantum state of the specimen.

The subscript e stands for *electron* and the subscript s for *specimen*. We shall use them in this subsection only when they help to avoid ambiguity.

What does the specimen state contain?

The state $|\Xi_0\rangle_{\mathrm{s}}$ does not refer to one particular electron of the specimen. In principle, it describes all the internal degrees of freedom of the specimen that are included in the physical model.

These may include, for example, the positions and spins of all bound electrons and the positions of the nuclei. In practical models some of these degrees of freedom may later be approximated as fixed or treated in an averaged way, but conceptually $|\Xi_0\rangle_{\mathrm{s}}$ represents the state of the specimen as a whole.

To make this more concrete, let

$$\mathbf{q} \tag{A.387}$$

stand collectively for all the internal coordinates of the specimen. Then the position-space representation of the specimen state can be written schematically as

$$\boxed{\Xi_0(\mathbf{q}) = \langle \mathbf{q} | \Xi_0 \rangle_{\mathrm{s}}.} \tag{A.388}$$

The symbol $\mathbf{q}$ should not be interpreted as the position of one particle. It is shorthand for the possibly very large collection of variables needed to specify the internal configuration of the specimen.

The combined electron–specimen system

Before the imaging electron interacts with the specimen, we may describe the two systems separately by

$$|\psi\rangle_{\mathrm{e}} \qquad \text{and} \qquad |\Xi_0\rangle_{\mathrm{s}}. \tag{A.389}$$

The state of the combined system is written

$$\boxed{|\psi\rangle_{\mathrm{e}} \otimes |\Xi_0\rangle_{\mathrm{s}}.} \tag{A.390}$$

The symbol

$$\otimes \tag{A.391}$$

is called the *tensor product*. For our present purposes, it can be read simply as

$$\boxed{\text{state of the imaging electron} \quad + \quad \text{state of the specimen} \quad \longrightarrow \quad \text{state of the complete system.}} \tag{A.392}$$

It is not an ordinary multiplication of two numbers. It indicates that both sets of variables are needed to describe the complete physical system.

For example, in coordinate representation the product state of Eq. (A.390) becomes

$$\boxed{\Psi(\mathbf{r}, \mathbf{q}) = \psi(\mathbf{r})\Xi_0(\mathbf{q}),} \tag{A.393}$$

where $\mathbf{r}$ is the position of the imaging electron and $\mathbf{q}$ represents the internal coordinates of the specimen.

Equation (A.393) makes the meaning of the tensor product particularly transparent: the complete wave function depends both on the coordinates of the imaging electron and on those of the specimen.

The interaction Hamiltonian

The two systems would evolve independently if they did not interact. Their interaction is described by an operator that we denote

$$\boxed{\hat{H}_{\text{int}}.} \tag{A.394}$$

The symbol H is used because this operator contributes to the Hamiltonian, that is, to the operator governing the time evolution of the complete system.

Schematically, the total Hamiltonian can be written as

$$\boxed{\hat{H} = \hat{H}_{\text{e}} + \hat{H}_{\text{s}} + \hat{H}_{\text{int}},} \tag{A.395}$$

where

$$\begin{aligned} \hat{H}_{\text{e}} &: \text{ describes the imaging electron in the absence of the specimen,} \\ \hat{H}_{\text{s}} &: \text{ describes the internal dynamics of the specimen,} \\ \hat{H}_{\text{int}} &: \text{ describes their mutual interaction.} \end{aligned} \tag{A.396}$$

The important point is that $\hat{H}_{\text{int}}$ acts on the *combined* electron–specimen state. It may therefore depend on both sets of coordinates, $\mathbf{r}$ and $\mathbf{q}$.

What does it mean for the specimen to remain in the same internal state?

We are particularly interested in elastic scattering. The statement that the specimen remains in the same internal state does *not* mean that no interaction has taken place.

Let

$$|\Xi_0\rangle_{\mathrm{s}} \tag{A.397}$$

denote the initial internal quantum state of the specimen. This state is the complete quantum description, within the chosen model, of the internal degrees of freedom of the specimen. It is not merely a probability distribution of particle positions. Different representations of the same state can be used to calculate probabilities for position, momentum, energy, and other observables.

For example, in a coordinate representation,

$$\Xi_0(\mathbf{q}) = \langle \mathbf{q} | \Xi_0 \rangle_{\mathrm{s}}, \tag{A.398}$$

where $\mathbf{q}$ collectively denotes the internal coordinates of the specimen. The quantity

$$|\Xi_0(\mathbf{q})|^2 \tag{A.399}$$

is then related to probabilities for those coordinates, but it is only one representation of the abstract state $|\Xi_0\rangle_{\mathrm{s}}$.

Suppose that before the interaction the complete electron–specimen state is

$$|\psi_i\rangle_{\mathrm{e}} |\Xi_0\rangle_{\mathrm{s}}. \tag{A.400}$$

In the elastic channel, after the interaction the specimen is left in the same internal state:

$$\boxed{|\psi_i\rangle_{\mathrm{e}} |\Xi_0\rangle_{\mathrm{s}} \longrightarrow |\psi_f\rangle_{\mathrm{e}} |\Xi_0\rangle_{\mathrm{s}}.} \tag{A.401}$$

Thus the imaging electron may change its state,

$$|\psi_i\rangle_{\mathrm{e}} \longrightarrow |\psi_f\rangle_{\mathrm{e}}, \tag{A.402}$$

even though the internal state of the specimen is unchanged.

For example, the electron momentum may change,

$$\mathbf{p}_i \longrightarrow \mathbf{p}_f, \tag{A.403}$$

which corresponds to scattering into a different direction. The electron wave may also acquire a phase shift.

Therefore,

$$\boxed{\text{elastic scattering} \neq \text{absence of interaction.}} \tag{A.404}$$

Rather, elastic scattering means that, after the interaction, the specimen is left in the same internal quantum state in which it started. Here "internal state" refers to the complete quantum state of the specimen degrees

of freedom included in the model—for example, its electronic, vibrational, rotational, and spin state—and not merely to its internal energy.

If the internal states are energy eigenstates,

$$\hat{H}_{\mathrm{s}}|\Xi_n\rangle = E_n|\Xi_n\rangle, \tag{A.405}$$

then remaining in the same internal state automatically implies remaining at the same internal energy. The converse need not be true, because different internal states may have the same energy.

Thus, in the elastic channel,

$$\boxed{|\psi_i\rangle_{\mathrm{e}}|\Xi_0\rangle_{\mathrm{s}} \longrightarrow |\psi_f\rangle_{\mathrm{e}}|\Xi_0\rangle_{\mathrm{s}},} \tag{A.406}$$

whereas an inelastic transition leaves the specimen in a different internal state,

$$\boxed{|\psi_i\rangle_{\mathrm{e}}|\Xi_0\rangle_{\mathrm{s}} \longrightarrow |\psi_f\rangle_{\mathrm{e}}|\Xi_m\rangle_{\mathrm{s}}, \qquad m \neq 0.} \tag{A.407}$$

Internal motion versus motion of the specimen as a whole

The statement that the specimen remains in the same internal state does not imply that every physical property of the specimen remains unchanged. In particular, its momentum as a whole may change.

For a many-particle object, it is useful to distinguish between two different kinds of motion: motion of the centre of mass and internal motion relative to the centre of mass.

The first describes translation of the specimen as a whole. The second describes its internal structure and dynamics.

For example, for two particles of masses m_1 and m_2, instead of their individual coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$, we may introduce the centre-of-mass coordinate

$$\mathbf{R}_{\mathrm{CM}} = \frac{m_1\mathbf{r}_1 + m_2\mathbf{r}_2}{m_1 + m_2}, \tag{A.408}$$

and the relative coordinate

$$\boldsymbol{\rho} = \mathbf{r}_1 - \mathbf{r}_2. \tag{A.409}$$

Changing $\mathbf{R}_{\mathrm{CM}}$ moves the complete object through space without changing the relative arrangement of its particles. By contrast, changes in $\boldsymbol{\rho}$ correspond to internal motion.

The same distinction applies to a many-particle specimen. When the centre-of-mass motion can be separated from the internal degrees of freedom, its Hamiltonian can be written schematically as

$$\boxed{\hat{H}_{\mathrm{s}} = \frac{\hat{\mathbf{P}}^{\,2}}{2M} + \hat{H}_{\mathrm{internal}},} \tag{A.410}$$

where M is the total mass of the specimen, $\hat{\mathbf{P}}$ describes its centre-of-mass momentum, and $\hat{H}_{\mathrm{internal}}$ describes its internal degrees of freedom.

A state of the specimen may therefore be labelled as

$$\boxed{|\mathbf{P}, \Xi_n\rangle_{\mathrm{s}},} \tag{A.411}$$

where $\mathbf{P}$ specifies the momentum of the specimen as a whole and $|\Xi_n\rangle_{\mathrm{s}}$ specifies its internal state.

These are different pieces of information. The specimen may change its centre-of-mass momentum while remaining in exactly the same internal state:

$$\boxed{|\mathbf{P}_i, \Xi_0\rangle_{\mathrm{s}} \longrightarrow |\mathbf{P}_f, \Xi_0\rangle_{\mathrm{s}}.} \tag{A.412}$$

Thus an elastic electron–specimen scattering event may be written more precisely as

$$\boxed{|\mathbf{p}_i\rangle_{\mathrm{e}}|\mathbf{P}_i, \Xi_0\rangle_{\mathrm{s}} \longrightarrow |\mathbf{p}_f\rangle_{\mathrm{e}}|\mathbf{P}_f, \Xi_0\rangle_{\mathrm{s}}.} \tag{A.413}$$

Momentum conservation requires

$$\boxed{\mathbf{p}_i + \mathbf{P}_i = \mathbf{p}_f + \mathbf{P}_f.} \tag{A.414}$$

Therefore the change in momentum of the imaging electron is balanced by a change in the centre-of-mass momentum of the specimen. This does not require a change of its internal state.

Energy conservation gives

$$E_{\mathrm{e},i} + \frac{P_i^2}{2M} + E_0 = E_{\mathrm{e},f} + \frac{P_f^2}{2M} + E_0, \tag{A.415}$$

where E_0 is the internal energy associated with $|\Xi_0\rangle_{\mathrm{s}}$. Since the internal state is the same on both sides, the internal energy cancels:

$$\boxed{E_{\mathrm{e},i} - E_{\mathrm{e},f} = \frac{P_f^2 - P_i^2}{2M}.} \tag{A.416}$$

If the specimen is initially at rest and receives the momentum transfer

$$\mathbf{q} = \mathbf{p}_i - \mathbf{p}_f, \tag{A.417}$$

then

$$\mathbf{P}_f = \mathbf{q} \tag{A.418}$$

and the recoil energy is

$$\boxed{E_{\text{recoil}} = \frac{q^2}{2M}.} \tag{A.419}$$

For a massive specimen this energy is generally extremely small. Thus, in the elastic-scattering approximation used in electron microscopy, we usually focus on the change in direction and phase of the imaging electron while regarding the internal state of the specimen as unchanged.

Inelastic scattering: changing the internal state of the specimen

We can now contrast elastic scattering with an *inelastic* interaction.

In the elastic case discussed above, the imaging electron may change its momentum and phase, and the specimen may recoil as a whole, but the specimen is left in the same internal quantum state $|\Xi_0\rangle_{\text{s}}$.

In an inelastic event, the interaction leaves the specimen in a different internal state. Including explicitly the centre-of-mass momentum of the specimen, the process can be written as

$$\boxed{|\mathbf{p}_i\rangle_{\text{e}}|\mathbf{P}_i, \Xi_0\rangle_{\text{s}} \longrightarrow |\mathbf{p}_f\rangle_{\text{e}}|\mathbf{P}_f, \Xi_m\rangle_{\text{s}}, \qquad m \neq 0.} \tag{A.420}$$

Here,

$$|\Xi_0\rangle_{\text{s}} \qquad \text{and} \qquad |\Xi_m\rangle_{\text{s}} \tag{A.421}$$

represent two different internal quantum states of the specimen.

The word *internal* is important. The change need not correspond to translation of the specimen as a whole. It may instead correspond to a change in its electronic configuration, molecular vibration, spin configuration, collective electronic motion, or some other degree of freedom internal to the specimen.

Example: excitation of a bound electron. Consider, for simplicity, one bound electron of the specimen that is initially in an orbital φ_a with energy ε_a. During the interaction, the imaging electron may transfer part

of its energy to this bound electron, promoting it to a higher-energy orbital φ_b,

$$\varphi_a \longrightarrow \varphi_b, \qquad \varepsilon_b > \varepsilon_a. \tag{A.422}$$

The change does not mean that the complete specimen state consists of only this one electron. Rather, the state $|\Xi_0\rangle_{\mathrm{s}}$ describes the specimen as a whole. If one of its bound electrons changes orbital, the many-particle state of the specimen changes accordingly:

$$\boxed{|\Xi_0\rangle_{\mathrm{s}} \longrightarrow |\Xi_m\rangle_{\mathrm{s}}.} \tag{A.423}$$

The imaging electron supplies the required excitation energy. In the simplest picture,

$$E_m - E_0 \simeq \varepsilon_b - \varepsilon_a. \tag{A.424}$$

Thus the imaging electron emerges with less kinetic energy than it had before the interaction.

This is an example of inelastic scattering even though the specimen as a whole may remain essentially at the same position.

Other examples of internal changes. The internal state may change in several different ways.

For example, the interaction may excite a molecular vibration. In that case the electronic configuration may remain essentially unchanged, while the nuclei acquire additional vibrational energy.

The imaging electron may also transfer enough energy to remove a bound electron completely from the specimen. This is *ionization*. The final state then contains an ionized specimen together with the ejected electron and belongs to a continuum of possible final states.

Another important possibility in electron microscopy is a collective excitation of many specimen electrons, such as a *plasmon*. In that case it is not useful to picture one particular bound electron as being promoted from one orbital to another. Instead, the electronic density of the material undergoes a collective excitation.

Thus the notation

$$|\Xi_m\rangle_{\mathrm{s}} \tag{A.425}$$

is deliberately general. It may represent an electronic excitation, a vibrational excitation, an ionized state, a collective excitation, or another internal state of the specimen.

Energy conservation. Let E_0 and E_m denote the internal energies associated with $|\Xi_0\rangle_{\rm s}$ and $|\Xi_m\rangle_{\rm s}$, respectively.

The full energy-conservation equation, including centre-of-mass recoil of the specimen, is

$$\boxed{E_{{\rm e},i} + \frac{P_i^2}{2M} + E_0 = E_{{\rm e},f} + \frac{P_f^2}{2M} + E_m.} \tag{A.426}$$

Rearranging,

$$\boxed{E_{{\rm e},i} - E_{{\rm e},f} = (E_m - E_0) + \frac{P_f^2 - P_i^2}{2M}.} \tag{A.427}$$

Thus the energy lost by the imaging electron can be divided into two parts:

$$\boxed{\text{electron energy loss} = \text{internal excitation energy} + \text{recoil energy}.} \tag{A.428}$$

For a massive specimen, the recoil contribution is usually very small. Neglecting it gives the commonly used approximation

$$\boxed{E_{{\rm e},i} - E_{{\rm e},f} \simeq E_m - E_0.} \tag{A.429}$$

This equation explains the term *energy-loss* scattering: the energy lost by the imaging electron appears predominantly as internal energy of the specimen.

Elastic versus inelastic scattering. The distinction can now be stated more precisely.

For elastic scattering,

$$\boxed{|\mathbf{p}_i\rangle_{\rm e}|\mathbf{P}_i, \Xi_0\rangle_{\rm s} \longrightarrow |\mathbf{p}_f\rangle_{\rm e}|\mathbf{P}_f, \Xi_0\rangle_{\rm s}.} \tag{A.430}$$

The electron and the centre-of-mass momentum of the specimen may change, but the internal state remains $|\Xi_0\rangle_{\rm s}$.

For inelastic scattering,

$$\boxed{|\mathbf{p}_i\rangle_{\rm e}|\mathbf{P}_i, \Xi_0\rangle_{\rm s} \longrightarrow |\mathbf{p}_f\rangle_{\rm e}|\mathbf{P}_f, \Xi_m\rangle_{\rm s}, \qquad m \neq 0.} \tag{A.431}$$

The interaction has left a record in the internal degrees of freedom of the specimen: an excitation, ionization, vibration, collective mode, or another internal change has occurred.

Thus,

$$\boxed{\begin{array}{rcl} \text{elastic scattering} & : & \text{same internal specimen state,} \\ \text{inelastic scattering} & : & \text{different internal specimen state.} \end{array}} \tag{A.432}$$

From the electron–specimen interaction to an effective elastic potential

We have described the imaging electron and the specimen as two parts of one combined quantum system. We now want to understand how their mutual interaction eventually gives rise to the familiar static potential

$$-eV(\mathbf{r}) \tag{A.433}$$

that appears in electron-optical image formation.

To do this, we first need to explain why the interaction is represented by an operator denoted $\hat{H}_{\text{int}}$.

Why does the interaction appear in the Hamiltonian? In quantum mechanics, the evolution of a system is governed by an operator called the *Hamiltonian*. Its role will be discussed systematically later in this appendix. For the moment, we only need one central idea:

$$\boxed{\text{the Hamiltonian determines how the quantum state changes with time.}} \tag{A.434}$$

For the complete system consisting of the imaging electron and the specimen, it is useful to separate the Hamiltonian schematically into three contributions:

$$\boxed{\hat{H} = \hat{H}_{\text{e}} + \hat{H}_{\text{s}} + \hat{H}_{\text{int}}.} \tag{A.435}$$

Here,

$$\begin{array}{rcl} \hat{H}_{\text{e}} & : & \text{describes the imaging electron when considered by itself,} \\ \hat{H}_{\text{s}} & : & \text{describes the internal dynamics of the specimen,} \\ \hat{H}_{\text{int}} & : & \text{describes the coupling between the imaging electron and the specimen.} \end{array} \tag{A.436}$$

The first two terms act separately on the two subsystems. They cannot, by themselves, describe scattering of the imaging electron by the specimen.

The term $\hat{H}_{\text{int}}$ is different because it depends on variables belonging to *both* systems.

If

$$\hat{H}_{\text{int}} = 0, \tag{A.437}$$

the imaging electron and the specimen do not interact: each evolves independently.

If

$$\hat{H}_{\text{int}} \neq 0, \tag{A.438}$$

the evolution of one can depend on the state of the other. The imaging electron may change direction or phase, the specimen may recoil, and the specimen may undergo an internal transition.

This is why the operator describing the interaction appears as one part of the Hamiltonian: the Hamiltonian governs the evolution, and $\hat{H}_{\text{int}}$ is the part of that evolution produced by the coupling between the two systems.

The interaction in coordinate representation. Let

$$\mathbf{r} \tag{A.439}$$

denote the position of the imaging electron, and let

$$\mathbf{q} \tag{A.440}$$

collectively denote all the internal specimen coordinates included in our model.

The combined wave function may therefore depend on both sets of variables,

$$\Psi(\mathbf{r}, \mathbf{q}). \tag{A.441}$$

The interaction operator acts on this complete wave function:

$$\hat{H}_{\text{int}} \Psi(\mathbf{r}, \mathbf{q}). \tag{A.442}$$

At this point it is important to distinguish an *operator* from an ordinary function. In general, knowing a function $H_{\text{int}}(\mathbf{r}, \mathbf{q})$ is not the same thing as specifying an arbitrary operator $\hat{H}_{\text{int}}$.

What does it mean for the interaction to act by multiplication? For the electrostatic interaction considered in this book, the interaction has a particularly simple form in coordinate representation. It acts by multiplying the wave function:

$$\boxed{\left[\hat{H}_{\text{int}}\Psi\right](\mathbf{r},\mathbf{q}) = H_{\text{int}}(\mathbf{r},\mathbf{q})\Psi(\mathbf{r},\mathbf{q}).} \tag{A.443}$$

This equation explains precisely what is meant when we say that $\hat{H}_{\text{int}}$ is a *multiplication operator*.

At each configuration $(\mathbf{r},\mathbf{q})$, the operator takes the value of the wave function at that same configuration and multiplies it by the interaction energy $H_{\text{int}}(\mathbf{r},\mathbf{q})$. No derivative is taken and no value of the wave function at another point is required. Such an interaction is called *local* in the coordinate representation.

What other kinds of operators are possible? Not every physical interaction has this form. For example, the momentum operator does not multiply the wave function by an ordinary function of position. Instead,

$$\hat{\mathbf{p}} = -j\hbar\nabla, \tag{A.444}$$

so it acts through derivatives.

Interactions involving magnetic fields may contain the momentum operator and therefore also contain derivatives. Spin-dependent interactions may act on spin components of the state rather than merely multiplying its spatial wave function.

Still more generally, a non-local operator may have the form

$$\left[\hat{K}\Psi\right](\mathbf{r},\mathbf{q}) = \int K(\mathbf{r},\mathbf{q};\mathbf{r}',\mathbf{q}')\Psi(\mathbf{r}',\mathbf{q}')\,d\mathbf{r}'\,d\mathbf{q}', \tag{A.445}$$

so that the result at $(\mathbf{r},\mathbf{q})$ depends on the wave function at other configurations $(\mathbf{r}',\mathbf{q}')$.

None of these complications is needed for the electrostatic interaction used in the basic theory of image formation. We shall therefore work with the local scalar interaction of Eq. (A.443).

The approximation being made is thus not that the interaction is "weak" or unimportant. Rather, we are assuming that the relevant electron–specimen coupling can be represented by an instantaneous electrostatic potential energy depending only on the particle coordinates, without explicit momentum-, spin-, magnetic-, or non-local terms.

Why does a potential-energy function become a multiplication operator? We have said that a local interaction energy

$$H_{\text{int}}(\mathbf{r}, \mathbf{q}) \tag{A.446}$$

acts on the joint wave function by multiplication. It is useful to see why this is the natural operator associated with such an interaction.

Let

$$|\mathbf{r}, \mathbf{q}\rangle \tag{A.447}$$

denote the configuration state in which the imaging electron has position $\mathbf{r}$ and the specimen has internal coordinates $\mathbf{q}$.

The complete electron–specimen state can be expanded in this continuous configuration basis as

$$\boxed{|\Psi\rangle = \int \Psi(\mathbf{r}, \mathbf{q}) |\mathbf{r}, \mathbf{q}\rangle \, d^3r \, d\mathbf{q}.} \tag{A.448}$$

Here

$$\boxed{\Psi(\mathbf{r}, \mathbf{q}) = \langle \mathbf{r}, \mathbf{q} | \Psi \rangle} \tag{A.449}$$

is the coordinate representation of the complete quantum state.

Now suppose that, for each instantaneous configuration $(\mathbf{r}, \mathbf{q})$, classical electrostatics assigns a definite interaction energy

$$H_{\text{int}}(\mathbf{r}, \mathbf{q}). \tag{A.450}$$

The corresponding quantum operator is defined so that the configuration state $|\mathbf{r}, \mathbf{q}\rangle$ has precisely that interaction energy:

$$\boxed{\hat{H}_{\text{int}} |\mathbf{r}, \mathbf{q}\rangle = H_{\text{int}}(\mathbf{r}, \mathbf{q}) |\mathbf{r}, \mathbf{q}\rangle.} \tag{A.451}$$

Thus the configuration states are eigenstates of the interaction operator, and the corresponding eigenvalue is the interaction energy for that configuration.

This is analogous to a diagonal matrix in ordinary linear algebra. If

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & \cdots \\ 0 & d_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}, \tag{A.452}$$

then

$$\mathbf{D} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} d_1 c_1 \\ d_2 c_2 \\ \vdots \end{pmatrix}. \tag{A.453}$$

Each component is multiplied by the value associated with its basis vector.

The same happens here, except that the basis is continuous. Applying $\hat{H}_{\text{int}}$ to Eq. (A.448),

$$\hat{H}_{\text{int}}|\Psi\rangle = \int \Psi(\mathbf{r},\mathbf{q})\hat{H}_{\text{int}}|\mathbf{r},\mathbf{q}\rangle\, d^3r\, d\mathbf{q} \tag{A.454}$$

$$= \int H_{\text{int}}(\mathbf{r},\mathbf{q})\Psi(\mathbf{r},\mathbf{q})|\mathbf{r},\mathbf{q}\rangle\, d^3r\, d\mathbf{q}. \tag{A.455}$$

Therefore, in coordinate representation,

$$\boxed{\left[\hat{H}_{\text{int}}\Psi\right](\mathbf{r},\mathbf{q}) = H_{\text{int}}(\mathbf{r},\mathbf{q})\Psi(\mathbf{r},\mathbf{q}).} \tag{A.456}$$

This is what is meant by saying that the interaction is represented by a *multiplication operator*.

It is important not to interpret

$$\hat{H}_{\text{int}}|\Psi\rangle \tag{A.457}$$

as a new normalized physical state obtained after the interaction.

The Hamiltonian has units of energy, so $\hat{H}_{\text{int}}|\Psi\rangle$ has units of energy times state amplitude and is generally not normalized. Rather, it is a vector in the same state space that describes how the interaction contributes to the evolution of $|\Psi\rangle$.

Later, when the Schrödinger equation is introduced, this role will be made explicit:

$$j\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle. \tag{A.458}$$

Thus the Hamiltonian acting on the state determines its rate of change; it does not by itself represent the final state after the interaction.

The Coulomb interaction is a local multiplication operator. We can now apply the previous construction to the electrostatic interaction relevant for electron microscopy.

For a specified configuration of all particles, Coulomb's law assigns a definite interaction energy. Therefore the corresponding interaction operator is diagonal in the configuration basis $|\mathbf{r},\mathbf{q}\rangle$ and acts by multiplication, as derived in Eq. (A.443).

To see this, suppose that the specimen contains nuclei at positions $\mathbf{R}_A$, with nuclear charges $Z_A e$, and bound electrons at positions $\mathbf{r}_j$. These coordinates are part of the collective set $\mathbf{q}$.

For a particular instantaneous configuration of the specimen, the Coulomb interaction energy of an imaging electron at position $\mathbf{r}$ is

$$\boxed{H_{\text{int}}(\mathbf{r},\mathbf{q}) = -\sum_A \frac{Z_A e^2}{4\pi\varepsilon_0|\mathbf{r}-\mathbf{R}_A|} + \sum_j \frac{e^2}{4\pi\varepsilon_0|\mathbf{r}-\mathbf{r}_j|}.} \tag{A.459}$$

The first term is attractive because the imaging electron has charge $-e$ and the nuclei are positively charged. The second term is repulsive because both the imaging electron and the specimen electrons have negative charge.

Equation (A.459) depends on the imaging-electron coordinate $\mathbf{r}$ and on the instantaneous internal configuration $\mathbf{q}$ of the specimen. It is therefore an interaction between the two systems.

Interactions among the particles *inside* the specimen are not included in Eq. (A.459); those belong to $\hat{H}_{\text{s}}$. Likewise, the kinetic energy of the imaging electron belongs to $\hat{H}_{\text{e}}$.

Eliminating the specimen variables in the elastic channel. The interaction operator

$$\hat{H}_{\text{int}} \tag{A.460}$$

acts on the complete electron–specimen system. Therefore, before we can use it in an equation describing only the imaging electron, we must somehow remove the specimen degrees of freedom.

Let

$$|\psi\rangle_{\text{e}} \tag{A.461}$$

be an arbitrary state of the imaging electron, and let

$$|\Xi_0\rangle_{\text{s}} \tag{A.462}$$

be the internal state of the specimen.

Before the interaction, the corresponding product state is

$$|\psi\rangle_{\text{e}} \otimes |\Xi_0\rangle_{\text{s}}. \tag{A.463}$$

The operator $\hat{H}_{\text{int}}$ acts on this complete state:

$$\hat{H}_{\text{int}}\left(|\psi\rangle_{\text{e}} \otimes |\Xi_0\rangle_{\text{s}}\right). \tag{A.464}$$

This result generally contains many possible specimen components. For the elastic channel, however, we are interested only in the component for which the specimen finishes again in the state $|\Xi_0\rangle_{\text{s}}$.

To extract this component, we project the specimen part of the result onto

$$ {}_{\mathrm{s}}\langle \Xi_0 | . \tag{A.465} $$

Thus, for any imaging-electron state $|\psi\rangle_{\mathrm{e}}$, we define

$$ \boxed{\hat{U}_{\mathrm{el}} |\psi\rangle_{\mathrm{e}} \equiv {}_{\mathrm{s}}\langle \Xi_0 | \hat{H}_{\mathrm{int}} \left(|\psi\rangle_{\mathrm{e}} \otimes |\Xi_0\rangle_{\mathrm{s}} \right) .} \tag{A.466} $$

This equation is the most explicit definition of the effective elastic interaction operator $\hat{U}_{\mathrm{el}}$.

Because Eq. (A.466) must hold for every imaging-electron state $|\psi\rangle_{\mathrm{e}}$, it is customary to write the definition more compactly as

$$ \boxed{\hat{U}_{\mathrm{el}} \equiv {}_{\mathrm{s}}\langle \Xi_0 | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle_{\mathrm{s}} .} \tag{A.467} $$

The bra and ket in this expression refer only to the specimen. The imaging-electron part has not been projected out. For this reason, the result is not generally a number: it is still an operator acting on the imaging electron.

Thus $\hat{U}_{\mathrm{el}}$ is not a new complete Hamiltonian. Rather, it is the interaction contribution that remains in the effective Hamiltonian of the imaging electron:

$$ \boxed{\hat{H}_{\mathrm{e}}^{\mathrm{eff}} = \hat{H}_{\mathrm{e}} + \hat{U}_{\mathrm{el}} .} \tag{A.468} $$

The general role of the Hamiltonian will be developed later. At this point it is sufficient to understand that $\hat{U}_{\mathrm{el}}$ represents the part of the imaging electron's energy and evolution that is caused by its elastic interaction with the specimen.

Coordinate form of the elastic interaction. Because the electrostatic interaction is a multiplication operator, Eq. (A.467) has a particularly simple coordinate representation.

Let

$$ \Xi_0(\mathbf{q}) = \langle \mathbf{q} | \Xi_0 \rangle_{\mathrm{s}} \tag{A.469} $$

be the specimen wave function in the internal coordinates $\mathbf{q}$. Then

$$ \boxed{U_{\mathrm{el}}(\mathbf{r}) = \int \Xi_0^*(\mathbf{q})\, H_{\mathrm{int}}(\mathbf{r}, \mathbf{q})\, \Xi_0(\mathbf{q})\, d\mathbf{q} .} \tag{A.470} $$

Since $H_{\text{int}}(\mathbf{r},\mathbf{q})$ is an ordinary multiplicative function, this can equivalently be written

$$\boxed{U_{\text{el}}(\mathbf{r}) = \int |\Xi_0(\mathbf{q})|^2 H_{\text{int}}(\mathbf{r},\mathbf{q})\, d\mathbf{q}.} \tag{A.471}$$

This expression has a straightforward interpretation. For each fixed position $\mathbf{r}$ of the imaging electron, the microscopic Coulomb interaction depends on the possible internal configurations $\mathbf{q}$ of the specimen. Equation (A.471) averages that interaction over the quantum probability distribution of those configurations in the specimen state $|\Xi_0\rangle_{\text{s}}$.

The integration is only over $\mathbf{q}$. The coordinate $\mathbf{r}$ is not integrated. Consequently, the result is still a function of the imaging-electron position:

$$U_{\text{el}}(\mathbf{r}). \tag{A.472}$$

Schematically,

$$\boxed{\begin{array}{ccc} H_{\text{int}}(\mathbf{r},\mathbf{q}) & \xrightarrow{\langle\Xi_0|\,\cdot\,|\Xi_0\rangle_{\text{s}}} & U_{\text{el}}(\mathbf{r}) \\ \text{electron + specimen variables} & & \text{electron variable only.} \end{array}} \tag{A.473}$$

Why the result is still an operator. Although $U_{\text{el}}(\mathbf{r})$ is an ordinary function, it represents the position-space form of the operator $\hat{U}_{\text{el}}$. Acting on the imaging-electron wave function,

$$\boxed{\left[\hat{U}_{\text{el}}\psi\right](\mathbf{r}) = U_{\text{el}}(\mathbf{r})\psi(\mathbf{r}).} \tag{A.474}$$

This is different from a complete expectation value such as

$$\langle\psi|\hat{U}_{\text{el}}|\psi\rangle. \tag{A.475}$$

The latter also averages over the imaging-electron state and therefore produces a single number:

$$\langle\psi|\hat{U}_{\text{el}}|\psi\rangle = \int |\psi(\mathbf{r})|^2 U_{\text{el}}(\mathbf{r})\, d^3r. \tag{A.476}$$

By contrast,

$${}_{\text{s}}\langle\Xi_0|\hat{H}_{\text{int}}|\Xi_0\rangle_{\text{s}} \tag{A.477}$$

averages only over the specimen state, so the dependence on the imaging electron remains.

Connection with the static electrostatic potential. The result can be expressed more familiarly in terms of the mean charge density of the specimen.

Let

$$\rho_{\mathrm{s}}(\mathbf{r}') \tag{A.478}$$

denote the expectation value of the specimen charge density in the internal state $|\Xi_0\rangle_{\mathrm{s}}$. The electrostatic potential generated by this charge distribution is

$$\boxed{V(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho_{\mathrm{s}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d^3r'.} \tag{A.479}$$

Because the charge of the imaging electron is $-e$, its potential energy is

$$\boxed{U_{\mathrm{el}}(\mathbf{r}) = -eV(\mathbf{r}).} \tag{A.480}$$

We have therefore arrived at the connection

$$\boxed{\hat{H}_{\mathrm{int}} \xrightarrow{\text{select elastic specimen state}} \hat{U}_{\mathrm{el}} \xrightarrow{\text{electrostatic approximation}} -eV(\mathbf{r}).} \tag{A.481}$$

The static electrostatic potential used in electron-optical image formation is therefore the effective interaction experienced by the imaging electron in the channel in which the specimen is not left in a different internal quantum state.

A.10 The Hamiltonian

The Hamiltonian has already appeared in our discussion of the interaction between the imaging electron and the specimen. We now introduce its meaning more systematically.

The Hamiltonian,

$$\hat{H}, \tag{A.482}$$

has two closely related roles in quantum mechanics:

1. it is the operator associated with the total energy of the system;
2. it determines how the quantum state changes with time.

These two roles are fundamental and explain why the interaction between the imaging electron and the specimen was introduced earlier as a term $\hat{H}_{\mathrm{int}}$ in the Hamiltonian.

A.10.1 The Hamiltonian determines time evolution

Let

$$|\psi(t)\rangle \tag{A.483}$$

denote the state of a quantum system at time t.

Its time evolution is governed by the time-dependent Schrödinger equation,

$$\boxed{j\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle.} \tag{A.484}$$

Equation (A.484) is one of the fundamental postulates of non-relativistic quantum mechanics. It is not derived from the Hilbert-space formalism introduced earlier.

Solving for the time derivative,

$$\boxed{\frac{\partial}{\partial t}|\psi(t)\rangle = -\frac{j}{\hbar}\hat{H}|\psi(t)\rangle.} \tag{A.485}$$

Thus $\hat{H}|\psi\rangle$ is not the state of the system "after the Hamiltonian has acted". Rather, it determines the instantaneous rate at which the state changes.

For a very short time interval Δt, Eq. (A.485) gives, to first order,

$$\boxed{|\psi(t+\Delta t)\rangle \simeq |\psi(t)\rangle - \frac{j\Delta t}{\hbar}\hat{H}|\psi(t)\rangle.} \tag{A.486}$$

This makes the role of the Hamiltonian concrete: changing the Hamiltonian changes the evolution of the state.

It also explains why the interaction between two systems must appear in the Hamiltonian. If the interaction changes how the combined state evolves, its contribution must appear in the operator on the right-hand side of Eq. (A.484).

A.10.2 The Hamiltonian is the energy operator

The Hamiltonian is also the operator associated with the total energy of the system.

Following the notation introduced in Sec. A.7, an energy eigenstate $|u_n\rangle$ satisfies

$$\boxed{\hat{H}|u_n\rangle = E_n|u_n\rangle,} \tag{A.487}$$

where

$$|u_n\rangle \tag{A.488}$$

is the eigenstate and

$$E_n \tag{A.489}$$

is the corresponding numerical energy eigenvalue.

We deliberately use different symbols for the state and the eigenvalue:

$$\boxed{|u_n\rangle \quad \text{state}, \qquad E_n \quad \text{energy.}} \tag{A.490}$$

According to the measurement postulate introduced in Sec. A.7.5, the possible results of an ideal energy measurement are the eigenvalues E_n of $\hat{H}$.

A.10.3 Time evolution of an energy eigenstate

The two roles of the Hamiltonian—energy and time evolution—come together particularly clearly for an energy eigenstate.

Time evolution of an energy eigenstate.

Suppose that

$$\hat{H}|u_n\rangle = E_n|u_n\rangle, \tag{A.491}$$

and that the Hamiltonian does not depend explicitly on time.

We seek a solution of the Schrödinger equation of the form

$$|\psi_n(t)\rangle = f_n(t)|u_n\rangle. \tag{A.492}$$

Substituting into Eq. (A.484),

$$j\hbar\frac{df_n}{dt}|u_n\rangle = f_n(t)\hat{H}|u_n\rangle \tag{A.493}$$

$$= E_n f_n(t)|u_n\rangle. \tag{A.494}$$

Since $|u_n\rangle$ is nonzero,

$$j\hbar\frac{df_n}{dt} = E_n f_n. \tag{A.495}$$

Its solution is

$$f_n(t) = f_n(0) \exp\left(-\frac{jE_n t}{\hbar}\right). \tag{A.496}$$

Choosing $f_n(0) = 1$,

$$\boxed{|\psi_n(t)\rangle = \exp\left(-\frac{jE_n t}{\hbar}\right)|u_n\rangle.} \tag{A.497}$$

The time dependence in Eq. (A.497) is only a phase factor. Consequently, an isolated energy eigenstate preserves all probabilities associated with time-independent measurements.

For a superposition of energy eigenstates,

$$|\psi(0)\rangle = \sum_n c_n |u_n\rangle, \tag{A.498}$$

the different components acquire different phases,

$$|\psi(t)\rangle = \sum_n c_n \exp\left(-\frac{jE_n t}{\hbar}\right)|u_n\rangle. \tag{A.499}$$

Their *relative* phases can therefore change with time and produce observable dynamics.

A.10.4 Kinetic and potential energy

For a single non-relativistic particle, the classical total energy has the familiar form

$$E = \frac{\mathbf{p}^2}{2m} + U(\mathbf{r}), \tag{A.500}$$

where the first term is kinetic energy and the second is potential energy.

In quantum mechanics, momentum becomes the operator $\hat{\mathbf{p}}$ introduced previously. The corresponding kinetic energy operator is therefore

$$\boxed{\hat{T} = \frac{\hat{\mathbf{p}}^2}{2m}.} \tag{A.501}$$

The potential energy is represented by another operator,

$$\hat{U}. \tag{A.502}$$

Thus the Hamiltonian can be written

$$\boxed{\hat{H} = \hat{T} + \hat{U} = \frac{\hat{\mathbf{p}}^{\,2}}{2m} + \hat{U}.} \tag{A.503}$$

The hats are useful here because $\hat{T}$, $\hat{U}$, and $\hat{H}$ are operators, not ordinary numerical energies.

A.10.5 A local scalar potential in position representation

In the electrostatic problems considered in this book, the potential energy operator is local in position space.

As discussed previously for the electron–specimen interaction, this means that there exists an ordinary function

$$U(\mathbf{r}) \tag{A.504}$$

such that

$$\boxed{[\hat{U}\psi](\mathbf{r}) = U(\mathbf{r})\psi(\mathbf{r}).} \tag{A.505}$$

Thus we must distinguish carefully between

$$\boxed{\begin{array}{rcl} \hat{U} & : & \text{potential-energy operator,} \\ U(\mathbf{r}) & : & \text{function representing that operator in position space.} \end{array}} \tag{A.506}$$

For a local scalar potential, the distinction is often suppressed and one writes simply

$$\hat{H} = \frac{\hat{\mathbf{p}}^{\,2}}{2m} + U(\mathbf{r}), \tag{A.507}$$

with the understanding that $U(\mathbf{r})$ acts by multiplication.

A.10.6 The Hamiltonian in position representation

Using the momentum operator

$$\hat{\mathbf{p}} = -j\hbar\nabla, \tag{A.508}$$

we obtain

$$\hat{\mathbf{p}}^{\,2} = -\hbar^2\nabla^2. \tag{A.509}$$

Consequently, when Eq. (A.503) acts on a position-space wave function,

$$\boxed{[\hat{H}\psi](\mathbf{r}) = -\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + U(\mathbf{r})\psi(\mathbf{r}).} \tag{A.510}$$

It is customary to summarize this action by writing

$$\boxed{\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + U(\mathbf{r}).} \tag{A.511}$$

The two terms have very different mathematical actions:

$$\boxed{\begin{array}{rcl} -\dfrac{\hbar^2}{2m}\nabla^2 & : & \text{differential operator describing kinetic energy,} \\ U(\mathbf{r}) & : & \text{multiplication operator describing local potential energy.} \end{array}} \tag{A.512}$$

A.10.7 Relation with the electron–specimen interaction

We can now place the discussion of the previous sections into a more general framework.

For the complete system formed by the imaging electron and the specimen, we wrote schematically

$$\boxed{\hat{H}_{\text{tot}} = \hat{H}_{\text{e}} + \hat{H}_{\text{s}} + \hat{H}_{\text{int}}.} \tag{A.513}$$

Here $\hat{H}_{\text{int}}$ is the part of the total Hamiltonian that couples the imaging electron to the specimen.

For the elastic channel, we previously eliminated the internal specimen variables and obtained the effective interaction operator

$$\boxed{\hat{U}_{\text{el}} = {}_{\text{s}}\langle \Xi_0 | \hat{H}_{\text{int}} | \Xi_0 \rangle_{\text{s}}.} \tag{A.514}$$

As discussed when this expression was introduced, $\hat{U}_{\text{el}}$ is not a measurement and it is not a second complete Hamiltonian. It is the part of the interaction that remains as an operator acting on the imaging electron after the specimen has been restricted to the elastic channel.

The effective Hamiltonian governing the imaging electron can therefore be written schematically as

$$\boxed{\hat{H}_{\text{e}}^{\text{eff}} = \hat{H}_{\text{e}} + \hat{U}_{\text{el}}.} \tag{A.515}$$

This equation gives a precise meaning to the effective interaction operator introduced earlier: it is a contribution to the Hamiltonian of the imaging electron.

A.10.8 Electrostatic interaction with the specimen

For the elastic electrostatic interaction developed previously, $\hat{U}_{\mathrm{el}}$ is a local multiplication operator:

$$[\hat{U}_{\mathrm{el}}\psi](\mathbf{r}) = U_{\mathrm{el}}(\mathbf{r})\psi(\mathbf{r}). \tag{A.516}$$

The effective potential energy is

$$\boxed{U_{\mathrm{el}}(\mathbf{r}) = -eV(\mathbf{r}),} \tag{A.517}$$

where $V(\mathbf{r})$ is the electrostatic potential generated by the specimen and $-e$ is the charge of the imaging electron.

Therefore, in the non-relativistic approximation,

$$\boxed{\hat{H}_{\mathrm{e}}^{\mathrm{eff}} = \frac{\hat{\mathbf{p}}^{\,2}}{2m} + \hat{U}_{\mathrm{el}}.} \tag{A.518}$$

In position representation,

$$\boxed{\hat{H}_{\mathrm{e}}^{\mathrm{eff}} = -\frac{\hbar^2}{2m}\nabla^2 - eV(\mathbf{r}).} \tag{A.519}$$

Thus the appearance of $-eV(\mathbf{r})$ in the electron wave equation is not an unrelated additional assumption. It is the coordinate-space form of the effective elastic electron–specimen interaction derived previously.

A.10.9 Relativistic electrons

The Hamiltonian Eq. (A.519) is based on the non-relativistic relation between momentum and kinetic energy,

$$E_{\mathrm{k}} = \frac{p^2}{2m}. \tag{A.520}$$

This approximation is adequate only when the kinetic energy is much smaller than the electron rest energy,

$$\boxed{E_{\mathrm{k}} \ll m_0 c^2.} \tag{A.521}$$

For an electron,

$$m_0 c^2 \simeq 511 \text{ keV}. \tag{A.522}$$

In a modern cryo-electron microscope, the imaging electron typically has a kinetic energy of order

$$200\text{–}300 \text{ keV}, \tag{A.523}$$

which is not negligible compared with its rest energy. Therefore the non-relativistic Hamiltonian is not quantitatively adequate.

Relativistic energy of a free electron

Special relativity replaces the non-relativistic energy–momentum relation by

$$\boxed{E^2 = p^2c^2 + m_0^2c^4.} \tag{A.524}$$

For the positive-energy electron branch,

$$\boxed{E = \sqrt{p^2c^2 + m_0^2c^4}.} \tag{A.525}$$

The total energy includes the rest energy m_0c^2. The kinetic energy is therefore

$$\boxed{E_\mathrm{k} = E - m_0c^2 = \sqrt{p^2c^2 + m_0^2c^4} - m_0c^2.} \tag{A.526}$$

For $p \ll m_0c$, expanding the square root gives

$$E_\mathrm{k} \simeq \frac{p^2}{2m_0}, \tag{A.527}$$

so the familiar non-relativistic expression is recovered as the low-momentum limit.

Why the Dirac Hamiltonian is introduced

One might try to construct a relativistic quantum theory simply by replacing

$$\frac{\hat{\mathbf{p}}^{\,2}}{2m_0} \tag{A.528}$$

with a square-root operator corresponding to Eq. (A.525). Such a construction can be useful in approximate scalar descriptions, but it is not the fundamental relativistic quantum theory of an electron.

The relativistic electron is described by the Dirac equation. Its state is no longer represented by a single complex scalar wave function, but by a four-component spinor,

$$\boxed{\Psi(\mathbf{r}, t) = \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{pmatrix}.} \tag{A.529}$$

The four components are required by the relativistic theory and contain, among other things, the two possible electron spin states. The detailed interpretation and derivation are developed in Sec. 3.4.

The free Dirac Hamiltonian

For a free relativistic electron, the Hamiltonian is

$$\boxed{\hat{H}_{\mathrm{D},0} = c\boldsymbol{\alpha} \cdot \hat{\mathbf{p}} + \beta m_0 c^2.} \tag{A.530}$$

Here

$$\boldsymbol{\alpha} = (\alpha_x, \alpha_y, \alpha_z) \tag{A.531}$$

and

$$\beta \tag{A.532}$$

are 4×4 Dirac matrices. They act on the four components of the spinor in Eq. (A.529).

The corresponding time-dependent Dirac equation has exactly the same structural form as the Schrödinger equation:

$$\boxed{j\hbar \frac{\partial \Psi}{\partial t} = \hat{H}_{\mathrm{D},0} \Psi.} \tag{A.533}$$

Thus the fundamental role of the Hamiltonian has not changed. It still determines the time evolution of the quantum state. What has changed is the form of the Hamiltonian and the mathematical structure of the state on which it acts.

Connection with the relativistic energy–momentum relation

The Dirac matrices satisfy algebraic relations that ensure that the free Dirac Hamiltonian reproduces the relativistic energy–momentum relation.

In particular,

$$\alpha_i \alpha_j + \alpha_j \alpha_i = 2\delta_{ij} \hat{I}_4, \tag{A.534}$$

$$\alpha_i \beta + \beta \alpha_i = 0, \tag{A.535}$$

and

$$\beta^2 = \hat{I}_4. \tag{A.536}$$

Using these relations, one obtains

$$\boxed{\hat{H}_{\mathrm{D},0}^2 = c^2 \hat{\mathbf{p}}^{\,2} + m_0^2 c^4.} \tag{A.537}$$

Therefore, if a free-particle state has momentum $\mathbf{p}$ and energy E, the Dirac equation implies

$$\boxed{E^2 = p^2 c^2 + m_0^2 c^4,} \tag{A.538}$$

which is precisely Eq. (A.524).

The derivation of this result and its connection with the scalar wave equation are given in Sec. 3.4.

Adding the electrostatic specimen interaction

We can now return to the interaction with the specimen.

In the non-relativistic treatment, the effective elastic interaction was the electrostatic potential energy

$$U_{\mathrm{el}}(\mathbf{r}) = -eV(\mathbf{r}). \tag{A.539}$$

Importantly, this electrostatic interaction does not disappear when we move to the relativistic description.

For an electron of charge $-e$ in a purely electrostatic potential $V(\mathbf{r})$, the Dirac Hamiltonian becomes

$$\boxed{\hat{H}_{\mathrm{D}} = c\boldsymbol{\alpha} \cdot \hat{\mathbf{p}} + \beta m_0 c^2 - eV(\mathbf{r})\hat{I}_4.} \tag{A.540}$$

The identity matrix $\hat{I}_4$ emphasizes that the scalar electrostatic potential multiplies all four spinor components in the same way. It is often omitted, giving the shorter notation

$$\hat{H}_{\mathrm{D}} = c\boldsymbol{\alpha} \cdot \hat{\mathbf{p}} + \beta m_0 c^2 - eV(\mathbf{r}). \tag{A.541}$$

The corresponding Dirac equation is

$$\boxed{j\hbar \frac{\partial \Psi(\mathbf{r}, t)}{\partial t} = \left[c\boldsymbol{\alpha} \cdot \hat{\mathbf{p}} + \beta m_0 c^2 - eV(\mathbf{r})\right] \Psi(\mathbf{r}, t).} \tag{A.542}$$

Thus the transition from the non-relativistic to the relativistic description does *not* require replacing the electrostatic interaction $-eV(\mathbf{r})$. The principal change is in the free-electron part of the Hamiltonian.

What changes and what remains the same?

The comparison is particularly instructive:

non-relativistic	relativistic
$\hat{H} = \dfrac{\hat{\mathbf{p}}^{\,2}}{2m_0} - eV(\mathbf{r})$	$\hat{H}_{\mathrm{D}} = c\boldsymbol{\alpha} \cdot \hat{\mathbf{p}} + \beta m_0 c^2 - eV(\mathbf{r})$
$\psi(\mathbf{r}, t)$ scalar wave function	$\Psi(\mathbf{r}, t)$ four-component spinor
$E_{\mathrm{k}} \simeq \dfrac{p^2}{2m_0}$	$E^2 = p^2c^2 + m_0^2c^4.$

(A.543)

Several elements therefore remain unchanged conceptually:

- the Hamiltonian still determines the time evolution;
- the Hamiltonian is still the energy operator;
- the specimen interaction still contributes an interaction term to the Hamiltonian;
- for a purely electrostatic elastic interaction, this term is still $-eV(\mathbf{r})$.

What changes is principally the description of the free electron: the non-relativistic kinetic-energy operator $\hat{\mathbf{p}}^2/2m_0$ is replaced by the Dirac operator $c\boldsymbol{\alpha}\cdot\hat{\mathbf{p}} + \beta m_0 c^2$, and the scalar state is replaced by a spinor.

More general electromagnetic fields

Equation (A.540) assumes that the interaction is purely electrostatic.

More generally, an electromagnetic field is described by a scalar potential

$$V(\mathbf{r}, t) \tag{A.544}$$

and a vector potential

$$\mathbf{A}(\mathbf{r}, t). \tag{A.545}$$

For a particle of charge q, electromagnetic coupling is introduced through

$$\hat{\mathbf{p}} \longrightarrow \hat{\mathbf{p}} - q\mathbf{A}. \tag{A.546}$$

For an electron, $q = -e$, so the Dirac Hamiltonian becomes

$$\boxed{\hat{H}_{\mathrm{D}} = c\boldsymbol{\alpha}\cdot(\hat{\mathbf{p}} + e\mathbf{A}) + \beta m_0 c^2 - eV.} \tag{A.547}$$

In the electrostatic specimen model used for the basic theory of cryo-electron microscopy,

$$\mathbf{A} \simeq 0, \tag{A.548}$$

and Eq. (A.547) reduces to Eq. (A.540).

Why a scalar electron wave is nevertheless used later

The fundamental relativistic theory therefore describes the imaging electron with a Dirac spinor. Nevertheless, most of the image-formation theory developed later in this book uses a scalar electron wave.

This is not because relativistic effects are neglected. Rather, under the conditions relevant to conventional cryo-electron microscopy—a high-energy electron beam, predominantly electrostatic interaction, and weak

spin-dependent effects—the Dirac description can be reduced to an effective scalar propagation equation while retaining the correct relativistic relation between energy, momentum, and wavelength.

The derivation of this reduction is given in Sec. 3.4. Its important consequence for the present appendix is that the scalar wave used later should be regarded as a reduced description of a relativistic electron, not as evidence that the electron itself has become non-relativistic.

The conceptual chain can therefore be summarized as

$$\boxed{\begin{array}{c}\text{relativistic electron}\\ \Downarrow\\ \hat{H}_{\mathrm{D}} = c\boldsymbol{\alpha}\cdot\hat{\mathbf{p}} + \beta m_0 c^2 - eV(\mathbf{r})\\ \Downarrow\\ \text{high-energy electrostatic approximation}\\ \Downarrow\\ \text{effective scalar fast-electron equation used in image formation.}\end{array}} \tag{A.549}$$

A.11 Unitary time evolution

In Sec. A.10 we introduced the Schrödinger equation,

$$j\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle, \tag{A.550}$$

and saw that the Hamiltonian determines the rate at which the quantum state changes.

We shall not repeat that discussion here. Instead, we now ask a different question:

$$\boxed{\text{given the state at one time, how do we obtain the state at a later time?}} \tag{A.551}$$

A.11.1 The time-evolution operator

Suppose that the Hamiltonian does not depend explicitly on time. We want to determine the state $|\psi(t)\rangle$ from a known initial state $|\psi(t_0)\rangle$.

The solution can be expressed through an operator $\hat{U}(t, t_0)$ acting on the initial state,

$$\boxed{|\psi(t)\rangle = \hat{U}(t, t_0)|\psi(t_0)\rangle.} \tag{A.552}$$

For a time-independent Hamiltonian, this operator is

$$\boxed{\hat{U}(t,t_0) = \exp\left[-\frac{j}{\hbar}\hat{H}(t-t_0)\right].} \tag{A.553}$$

The origin of this expression is shown in the following derivation.

Time-evolution operator for a time-independent Hamiltonian.

The Schrödinger equation is

$$j\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle. \tag{A.554}$$

Because $\hat{H}$ is assumed to be independent of time, this equation has the same mathematical structure as the ordinary linear differential equation

$$\frac{dy}{dt} = ay, \tag{A.555}$$

whose solution is an exponential.

For an operator, the exponential is defined by its power series,

$$\boxed{\exp(\hat{B}) = \hat{I} + \hat{B} + \frac{\hat{B}^2}{2!} + \frac{\hat{B}^3}{3!} + \cdots.} \tag{A.556}$$

Consider therefore

$$\hat{U}(t,t_0) = \exp\left[-\frac{j}{\hbar}\hat{H}(t-t_0)\right]. \tag{A.557}$$

Using Eq. (A.556),

$$\hat{U}(t,t_0) = \hat{I} - \frac{j}{\hbar}\hat{H}(t-t_0) + \frac{1}{2!}\left(-\frac{j}{\hbar}\right)^2\hat{H}^2(t-t_0)^2 + \cdots. \tag{A.558}$$

Differentiate term by term with respect to t:

$$\frac{\partial\hat{U}}{\partial t} = -\frac{j}{\hbar}\hat{H} + \frac{1}{1!}\left(-\frac{j}{\hbar}\right)^2\hat{H}^2(t-t_0) \tag{A.559}$$

$$+ \frac{1}{2!}\left(-\frac{j}{\hbar}\right)^3\hat{H}^3(t-t_0)^2 + \cdots. \tag{A.560}$$

Factoring out $-j\hat{H}/\hbar$,

$$\frac{\partial \hat{U}}{\partial t} = -\frac{j}{\hbar}\hat{H}\left[\hat{I} - \frac{j}{\hbar}\hat{H}(t-t_0) + \frac{1}{2!}\left(-\frac{j}{\hbar}\right)^2 \hat{H}^2(t-t_0)^2 + \cdots\right] \tag{A.561}$$

$$= -\frac{j}{\hbar}\hat{H}\hat{U}(t,t_0). \tag{A.562}$$

Therefore,

$$\boxed{j\hbar\frac{\partial \hat{U}}{\partial t} = \hat{H}\hat{U}.} \tag{A.563}$$

Now define

$$|\psi(t)\rangle = \hat{U}(t,t_0)|\psi(t_0)\rangle. \tag{A.564}$$

Since the initial state $|\psi(t_0)\rangle$ does not depend on t,

$$j\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = j\hbar\frac{\partial \hat{U}}{\partial t}|\psi(t_0)\rangle \tag{A.565}$$

$$= \hat{H}\hat{U}(t,t_0)|\psi(t_0)\rangle \tag{A.566}$$

$$= \hat{H}|\psi(t)\rangle. \tag{A.567}$$

Thus the proposed expression satisfies the Schrödinger equation.

It also satisfies the required initial condition. At $t = t_0$,

$$\hat{U}(t_0,t_0) = \exp(0) = \hat{I}, \tag{A.568}$$

so

$$|\psi(t_0)\rangle = \hat{I}|\psi(t_0)\rangle. \tag{A.569}$$

Hence,

$$\boxed{|\psi(t)\rangle = \exp\left[-\frac{j}{\hbar}\hat{H}(t-t_0)\right]|\psi(t_0)\rangle.} \tag{A.570}$$

The simple exponential form in Eq. (A.570) requires that $\hat{H}$ be independent of time. If the Hamiltonian itself changes with time, the evolution is more complicated because Hamiltonians evaluated at different times need not commute. The corresponding time-ordered exponential is not needed for the developments in this appendix.

Equation (A.553) therefore gives the state at any later time when the Hamiltonian is time independent.

For a very short interval

$$\Delta t = t - t_0, \tag{A.571}$$

the exponential may be approximated to first order as

$$\boxed{\hat{U}(t_0 + \Delta t, t_0) \simeq \hat{I} - \frac{j\Delta t}{\hbar}\hat{H}.} \tag{A.572}$$

Hence,

$$\boxed{|\psi(t_0 + \Delta t)\rangle \simeq |\psi(t_0)\rangle - \frac{j\Delta t}{\hbar}\hat{H}|\psi(t_0)\rangle.} \tag{A.573}$$

This expression makes the role of the Hamiltonian particularly transparent: $\hat{H}|\psi\rangle$ determines the direction in Hilbert space in which the state initially evolves, while the factor $\Delta t/\hbar$ determines the magnitude of the change over a short time interval.

A.11.2 Why quantum evolution preserves normalization

A quantum state normalized at one time must remain normalized as it evolves:

$$\langle\psi(t)|\psi(t)\rangle = 1. \tag{A.574}$$

This property follows from the Hermiticity of the Hamiltonian.

Since

$$\hat{H}^\dagger = \hat{H}, \tag{A.575}$$

the adjoint of the time-evolution operator is

$$\hat{U}^\dagger(t, t_0) = \exp\left[+\frac{j}{\hbar}\hat{H}(t - t_0)\right]. \tag{A.576}$$

Therefore,

$$\boxed{\hat{U}^\dagger(t, t_0)\hat{U}(t, t_0) = \hat{I}.} \tag{A.577}$$

An operator satisfying Eq. (A.577) is called *unitary*.

Conservation of the norm during time evolution.

Using Eq. (A.552),

$$|\psi(t)\rangle = \hat{U}(t, t_0)|\psi(t_0)\rangle, \tag{A.578}$$

the corresponding bra is

$$\langle\psi(t)| = \langle\psi(t_0)|\hat{U}^\dagger(t,t_0). \tag{A.579}$$

The norm at time t is therefore

$$\langle\psi(t)|\psi(t)\rangle = \langle\psi(t_0)|\hat{U}^\dagger(t,t_0)\hat{U}(t,t_0)|\psi(t_0)\rangle. \tag{A.580}$$

Using the unitarity relation Eq. (A.577),

$$\langle\psi(t)|\psi(t)\rangle = \langle\psi(t_0)|\hat{I}|\psi(t_0)\rangle \tag{A.581}$$
$$= \langle\psi(t_0)|\psi(t_0)\rangle. \tag{A.582}$$

Thus,

$$\boxed{\langle\psi(t)|\psi(t)\rangle = \langle\psi(t_0)|\psi(t_0)\rangle.} \tag{A.583}$$

If the state is normalized initially, it remains normalized for all times.

This is essential for the probabilistic interpretation of quantum mechanics. The total probability cannot gradually increase or decrease as an isolated quantum system evolves.

A.11.3 Time evolution in the energy eigenstate basis

Suppose that the Hamiltonian has eigenstates $|u_n\rangle$,

$$\hat{H}|u_n\rangle = E_n|u_n\rangle. \tag{A.584}$$

An arbitrary initial state can be expanded in this basis:

$$|\psi(0)\rangle = \sum_n c_n|u_n\rangle. \tag{A.585}$$

As derived in Sec. A.10, each energy eigenstate acquires the phase

$$\exp\left(-\frac{jE_nt}{\hbar}\right). \tag{A.586}$$

Therefore,

$$\boxed{|\psi(t)\rangle = \sum_n c_n \exp\left(-\frac{jE_nt}{\hbar}\right)|u_n\rangle.} \tag{A.587}$$

For a time-independent Hamiltonian, the magnitudes

$$|c_n|^2 \tag{A.588}$$

do not change. Only the phases of the coefficients evolve.

Thus the probability of obtaining the energy E_n remains

$$P(E = E_n) = |c_n|^2. \tag{A.589}$$

The relative phases between different components, however, do change:

$$\exp\left[-\frac{j(E_n - E_m)t}{\hbar}\right]. \tag{A.590}$$

These changing relative phases can produce observable changes in other quantities even though the probabilities of the individual energy components remain constant.

A.11.4 Why an interaction can produce transitions

This observation helps us understand the role of an interaction.

Suppose that we divide the Hamiltonian schematically as

$$\boxed{\hat{H} = \hat{H}_0 + \hat{H}_{\text{int}},} \tag{A.591}$$

where $\hat{H}_0$ describes the system in the absence of the interaction.

Let $|u_n\rangle$ be eigenstates of $\hat{H}_0$:

$$\hat{H}_0|u_n\rangle = E_n|u_n\rangle. \tag{A.592}$$

If

$$\hat{H}_{\text{int}} = 0, \tag{A.593}$$

each component evolves independently and only accumulates its own phase.

When

$$\hat{H}_{\text{int}} \neq 0, \tag{A.594}$$

the interaction operator may connect different states. The quantities

$$\boxed{\langle u_m|\hat{H}_{\text{int}}|u_n\rangle} \tag{A.595}$$

measure these couplings.

If

$$\langle u_m|\hat{H}_{\text{int}}|u_n\rangle \neq 0, \tag{A.596}$$

the interaction can transfer amplitude between the components $|u_n\rangle$ and $|u_m\rangle$.

This is the origin of quantum transitions.

In the electron–specimen problem, this is precisely the distinction encountered earlier. A diagonal specimen matrix element,

$$_{\mathrm{s}}\langle \Xi_0 | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle_{\mathrm{s}}, \tag{A.597}$$

describes the interaction in the channel in which the specimen remains in the same internal state. An off-diagonal matrix element,

$$_{\mathrm{s}}\langle \Xi_m | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle_{\mathrm{s}}, \qquad m \neq 0, \tag{A.598}$$

couples the initial specimen state to a different internal state and therefore makes an inelastic transition possible.

The quantitative probability of such a transition requires time-dependent perturbation theory and will not be developed here. For the purposes of this book, the important point is the structural one:

$$\boxed{\begin{array}{lcl} \langle \Xi_0 | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle & : & \text{interaction within the same specimen state,} \\ \langle \Xi_m | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle & : & \text{coupling between different specimen states.} \end{array}} \tag{A.599}$$

A.11.5 Why energy eigenstates are called stationary

We have already shown that, for a time-independent Hamiltonian, an energy eigenstate

$$\hat{H} | u_n \rangle = E_n | u_n \rangle \tag{A.600}$$

evolves as

$$| \psi_n(t) \rangle = \exp\left(-\frac{j E_n (t - t_0)}{\hbar} \right) | u_n \rangle. \tag{A.601}$$

Thus the state changes only by a global phase factor. Since a global phase does not change any physical probability, the physical state is unchanged by this evolution.

For example, in the position representation,

$$\psi_n(\mathbf{r}, t) = \exp\left(-\frac{j E_n (t - t_0)}{\hbar} \right) u_n(\mathbf{r}), \tag{A.602}$$

and therefore

$$\boxed{|\psi_n(\mathbf{r}, t)|^2 = |u_n(\mathbf{r})|^2.} \tag{A.603}$$

The position probability density is consequently independent of time.

More generally, for any observable A whose operator $\hat{A}$ has no explicit time dependence,

$$\langle A\rangle_t = \langle\psi_n(t)|\hat{A}|\psi_n(t)\rangle \tag{A.604}$$

$$= \exp\left(+\frac{jE_n(t-t_0)}{\hbar}\right)\exp\left(-\frac{jE_n(t-t_0)}{\hbar}\right)\langle u_n|\hat{A}|u_n\rangle \tag{A.605}$$

$$= \langle u_n|\hat{A}|u_n\rangle. \tag{A.606}$$

Hence,

$$\boxed{\langle A\rangle_t = \text{constant.}} \tag{A.607}$$

This is the reason for the term *stationary state*: although its wave function contains a time-dependent phase, all probabilities and expectation values of time-independent observables remain unchanged.

The situation is different for a superposition of states with different energies,

$$|\psi(t)\rangle = \sum_n c_n \exp\left(-\frac{jE_n t}{\hbar}\right)|u_n\rangle. \tag{A.608}$$

Different energy components acquire different phases, so their relative phases evolve with time. Interference between these components can therefore make probability distributions and other observables time dependent.

Strictly speaking, a stationary state need not correspond to a single eigenvector if the energy is degenerate. Any state lying entirely within one energy eigenspace acquires the same common phase and is therefore stationary.

A.12 Matrix elements and transitions between states

Expressions of the form

$$\langle u_m|\hat{A}|u_n\rangle \tag{A.609}$$

have already appeared several times in this appendix. We now give them a more precise interpretation and explain why they are especially important when $\hat{A}$ is an interaction operator.

As discussed in Sec. A.4.7, the expression

$$\langle u_m|\hat{A}|u_n\rangle \tag{A.610}$$

is an ordinary complex number. It measures the component of the state $\hat{A}|u_n\rangle$ along the state $|u_m\rangle$.

What has not yet been made explicit is why this quantity is called a *matrix element*.

A.12.1 Why they are called matrix elements

Let

$$|u_1\rangle, |u_2\rangle, \ldots \tag{A.611}$$

form a complete orthonormal basis. After an operator $\hat{A}$ acts on one basis state $|u_n\rangle$, the resulting state can itself be expanded in that basis:

$$\boxed{\hat{A}|u_n\rangle = \sum_m A_{mn}|u_m\rangle.} \tag{A.612}$$

The coefficient A_{mn} tells us how much of the basis state $|u_m\rangle$ is present in the result.

Taking the inner product with $\langle u_k|$,

$$\langle u_k|\hat{A}|u_n\rangle = \sum_m A_{mn}\langle u_k|u_m\rangle \tag{A.613}$$

$$= \sum_m A_{mn}\delta_{km} \tag{A.614}$$

$$= A_{kn}. \tag{A.615}$$

Therefore,

$$\boxed{A_{mn} = \langle u_m|\hat{A}|u_n\rangle.} \tag{A.616}$$

Thus the numbers $\langle u_m|\hat{A}|u_n\rangle$ are precisely the entries of the matrix representing $\hat{A}$ in the chosen basis:

$$\boxed{\mathbf{A} = \begin{pmatrix} \langle u_1|\hat{A}|u_1\rangle & \langle u_1|\hat{A}|u_2\rangle & \cdots \\ \langle u_2|\hat{A}|u_1\rangle & \langle u_2|\hat{A}|u_2\rangle & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.} \tag{A.617}$$

This is the origin of the name *matrix element*.

A.12.2 Diagonal and off-diagonal matrix elements

Equation (A.612) also gives a useful physical interpretation of the matrix elements.

A diagonal element,

$$\boxed{A_{nn} = \langle u_n|\hat{A}|u_n\rangle,} \tag{A.618}$$

describes the component produced along the same state $|u_n\rangle$.

An off-diagonal element,

$$\boxed{A_{mn} = \langle u_m|\hat{A}|u_n\rangle, \qquad m \neq n,} \tag{A.619}$$

describes the component produced along a different state $|u_m\rangle$.

Indeed, Eq. (A.612) can be written schematically as

$$\boxed{\hat{A}|u_n\rangle = \underbrace{A_{nn}|u_n\rangle}_{\text{same-state component}} + \underbrace{\sum_{m\neq n} A_{mn}|u_m\rangle}_{\text{components along other states}}.} \tag{A.620}$$

This distinction becomes particularly important for an interaction operator.

A.12.3 Interaction matrix elements

Suppose that, in the absence of an interaction, the system is described by a Hamiltonian $\hat{H}_0$ with eigenstates

$$\boxed{\hat{H}_0|u_n\rangle = E_n|u_n\rangle.} \tag{A.621}$$

Now introduce an interaction $\hat{H}_{\text{int}}$, so that

$$\hat{H} = \hat{H}_0 + \hat{H}_{\text{int}}. \tag{A.622}$$

The interaction generally does not leave an eigenstate of $\hat{H}_0$ unchanged. Instead,

$$\boxed{\hat{H}_{\text{int}}|u_i\rangle = \sum_f \underbrace{\langle u_f|\hat{H}_{\text{int}}|u_i\rangle}_{M_{fi}} |u_f\rangle.} \tag{A.623}$$

We define

$$\boxed{M_{fi} \equiv \langle u_f|\hat{H}_{\text{int}}|u_i\rangle.} \tag{A.624}$$

The number M_{fi} is called the *transition matrix element*, or the *coupling matrix element*, between the initial state $|u_i\rangle$ and the final state $|u_f\rangle$.

Equation (A.623) gives its meaning directly:

$$\boxed{M_{fi} = \text{amplitude with which } \hat{H}_{\text{int}}|u_i\rangle \text{ points along } |u_f\rangle.} \tag{A.625}$$

In particular, if

$$M_{fi} = 0, \tag{A.626}$$

the interaction does not directly couple these two states.

If

$$M_{fi} \neq 0, \tag{A.627}$$

the interaction generates a component along $|u_f\rangle$ and can therefore produce a transition from $|u_i\rangle$ to $|u_f\rangle$.

However, M_{fi} is *not by itself the transition probability*, nor is it the complete transition amplitude. To see what additional information is required, we must consider the time evolution.

A.12.4 From coupling to a transition amplitude

The role of the matrix element becomes clearer using the time-evolution formalism introduced in Sec. A.11.

Expand the time-dependent state in eigenstates of $\hat{H}_0$ as

$$\boxed{|\psi(t)\rangle = \sum_n c_n(t) \exp\left(-\frac{jE_n t}{\hbar}\right) |u_n\rangle.} \tag{A.628}$$

The exponential factors describe the evolution that each state would have under $\hat{H}_0$ alone. Any change in the coefficients $c_n(t)$ is therefore produced by the interaction.

Substituting Eq. (A.628) into the Schrödinger equation gives the following relation.

How the interaction couples the coefficients.

Starting from

$$j\hbar \frac{\partial}{\partial t}|\psi(t)\rangle = \left(\hat{H}_0 + \hat{H}_{\text{int}}\right)|\psi(t)\rangle, \tag{A.629}$$

insert Eq. (A.628). Differentiating,

$$j\hbar \frac{\partial|\psi\rangle}{\partial t} = \sum_n \left[j\hbar \frac{dc_n}{dt} + E_n c_n\right] \exp\left(-\frac{jE_n t}{\hbar}\right)|u_n\rangle. \tag{A.630}$$

On the right-hand side,

$$\hat{H}_0|\psi\rangle = \sum_n E_n c_n \exp\left(-\frac{jE_n t}{\hbar}\right)|u_n\rangle. \tag{A.631}$$

These terms cancel with the corresponding $E_n c_n$ terms on the left, leaving

$$j\hbar \sum_n \frac{dc_n}{dt} \exp\left(-\frac{jE_n t}{\hbar}\right) |u_n\rangle = \hat{H}_{\text{int}} \sum_n c_n \exp\left(-\frac{jE_n t}{\hbar}\right) |u_n\rangle. \quad \text{(A.632)}$$

Multiply from the left by

$$\langle u_f| \exp\left(+\frac{jE_f t}{\hbar}\right). \quad \text{(A.633)}$$

Using orthonormality, we obtain

$$\boxed{j\hbar \frac{dc_f}{dt} = \sum_n c_n(t) M_{fn}(t) \exp\left[\frac{j(E_f - E_n)t}{\hbar}\right],} \quad \text{(A.634)}$$

where

$$M_{fn}(t) = \langle u_f|\hat{H}_{\text{int}}(t)|u_n\rangle. \quad \text{(A.635)}$$

Equation (A.634) makes the role of the matrix elements precise: they determine how rapidly probability amplitude can be transferred between the different states.

A.12.5 Weak-interaction approximation

Suppose that at the initial time t_0 the system is entirely in the state $|u_i\rangle$:

$$c_i(t_0) = 1, \qquad c_{n\neq i}(t_0) = 0. \quad \text{(A.636)}$$

If the interaction is weak enough that the initial state is only slightly depleted, we may approximate

$$c_i(t) \simeq 1 \quad \text{(A.637)}$$

when calculating the first small amount of amplitude transferred to a different state $|u_f\rangle$.

Then Eq. (A.634) becomes

$$\frac{dc_f}{dt} \simeq -\frac{j}{\hbar} M_{fi}(t) \exp\left[\frac{j(E_f - E_i)t}{\hbar}\right]. \quad \text{(A.638)}$$

Integrating from t_0 to t,

$$\boxed{c_f^{(1)}(t) = -\frac{j}{\hbar} \int_{t_0}^{t} M_{fi}(t') \exp\left[\frac{j(E_f - E_i)t'}{\hbar}\right] dt'.} \quad \text{(A.639)}$$

This quantity is the first-order *transition amplitude*. The corresponding transition probability is approximately

$$\boxed{P_{i\to f}(t) \simeq \left|c_f^{(1)}(t)\right|^2.} \tag{A.640}$$

This shows why the transition matrix element is important but is not, by itself, the transition probability. The transition also depends on

- how long the interaction acts;
- the time dependence of the interaction;
- the energy difference $E_f - E_i$.

A.12.6 Fermi's golden rule

Equation (A.639) gives the first-order amplitude for a transition from an initial state $|u_i\rangle$ to a final state $|u_f\rangle$.

For a time-independent weak interaction, let

$$M_{fi} = \langle u_f|\hat{H}_{\text{int}}|u_i\rangle \tag{A.641}$$

be constant during an interval of duration T. Taking the initial time as $t = 0$, Eq. (A.639) gives

$$c_f^{(1)}(T) = -\frac{jM_{fi}}{\hbar}\int_0^T \exp\left[\frac{j(E_f - E_i)t}{\hbar}\right] dt \tag{A.642}$$

$$= -2jM_{fi}\exp\left[\frac{j(E_f - E_i)T}{2\hbar}\right]\frac{\sin\left[\dfrac{(E_f - E_i)T}{2\hbar}\right]}{E_f - E_i}. \tag{A.643}$$

The corresponding first-order transition probability is therefore

$$\boxed{P_{i\to f}(T) = 4|M_{fi}|^2\frac{\sin^2\left[\dfrac{(E_f - E_i)T}{2\hbar}\right]}{(E_f - E_i)^2}.} \tag{A.644}$$

For a long interaction time, the function in Eq. (A.644) becomes sharply concentrated around

$$E_f = E_i. \tag{A.645}$$

Using the distributional limit

$$\boxed{\lim_{T\to\infty}\frac{1}{T}\frac{4\sin^2\left(\dfrac{\Delta E\,T}{2\hbar}\right)}{(\Delta E)^2}=\frac{2\pi}{\hbar}\delta(\Delta E),} \tag{A.646}$$

we obtain the transition rate

$$\boxed{\Gamma_{i\to f}=\frac{2\pi}{\hbar}|M_{fi}|^2\delta(E_f-E_i).} \tag{A.647}$$

For a continuum of final states, the rate into a set of states $d\nu_f$ is

$$\boxed{d\Gamma_{i\to f}=\frac{2\pi}{\hbar}|M_{fi}|^2\delta(E_f-E_i)\,d\nu_f.} \tag{A.648}$$

If the final states are described by a density of states $\rho(E_f)$, integration over them gives the familiar form

$$\boxed{\Gamma_{i\to f}=\frac{2\pi}{\hbar}|M_{fi}|^2\rho(E_i).} \tag{A.649}$$

Fermi's golden rule therefore supplies the missing link between a quantum-mechanical coupling matrix element and a measurable transition rate:

$$\boxed{M_{fi}\longrightarrow|M_{fi}|^2\longrightarrow\text{transition rate.}} \tag{A.650}$$

A.12.7 Connection with elastic and inelastic electron scattering

We can now apply this language to the imaging electron and the specimen.

Let the initial combined state be

$$|\psi_i\rangle_{\mathrm{e}}|\Xi_0\rangle_{\mathrm{s}}, \tag{A.651}$$

and a possible final state be

$$|\psi_f\rangle_{\mathrm{e}}|\Xi_m\rangle_{\mathrm{s}}. \tag{A.652}$$

The complete transition matrix element is

$$\boxed{M_{fm,i0}={}_{\mathrm{e}}\langle\psi_f|_{\mathrm{s}}\langle\Xi_m|\hat{H}_{\mathrm{int}}|\psi_i\rangle_{\mathrm{e}}|\Xi_0\rangle_{\mathrm{s}}.} \tag{A.653}$$

This expression includes both parts of the process:

$$\begin{gathered}\text{change of the imaging electron}\\ \text{and}\\ \text{possible change of the specimen internal state.}\end{gathered} \tag{A.654}$$

As discussed earlier, we can first take the matrix element over the specimen variables and define

$$\boxed{\hat{U}_{m0} \equiv {}_{\mathrm{s}}\langle \Xi_m | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle_{\mathrm{s}}.} \tag{A.655}$$

The result $\hat{U}_{m0}$ is still an operator acting on the imaging electron. The complete matrix element then becomes

$$\boxed{M_{fm,i0} = {}_{\mathrm{e}}\langle \psi_f | \hat{U}_{m0} | \psi_i \rangle_{\mathrm{e}}.} \tag{A.656}$$

This notation unifies the elastic and inelastic cases.

For the elastic channel,

$$m = 0, \tag{A.657}$$

so

$$\boxed{\hat{U}_{00} = {}_{\mathrm{s}}\langle \Xi_0 | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle_{\mathrm{s}}.} \tag{A.658}$$

This is the effective elastic interaction operator discussed earlier; for the electrostatic approximation it becomes

$$U_{00}(\mathbf{r}) = -eV(\mathbf{r}). \tag{A.659}$$

For an inelastic channel,

$$m \neq 0, \tag{A.660}$$

and

$$\boxed{\hat{U}_{m0} = {}_{\mathrm{s}}\langle \Xi_m | \hat{H}_{\mathrm{int}} | \Xi_0 \rangle_{\mathrm{s}}} \tag{A.661}$$

describes the part of the interaction that couples the initial specimen state $|\Xi_0\rangle_{\mathrm{s}}$ to the different internal state $|\Xi_m\rangle_{\mathrm{s}}$.

Thus,

$$\boxed{\begin{array}{lcl} m = 0 & : & \text{elastic channel: specimen remains in the same internal state,} \\ m \neq 0 & : & \text{inelastic channel: specimen is left in a different internal state.} \end{array}} \tag{A.662}$$

A.13 Why orthogonal specimen channels add as probabilities

We have already seen in Eq. (A.93) that indistinguishable alternatives leading to the same final state must be added as amplitudes before taking the squared modulus.

We now consider the complementary situation that is particularly important for inelastic electron scattering: different alternatives leave the specimen in different orthogonal internal states.

Suppose that the specimen is initially in $|\Xi_0\rangle_{\mathrm{s}}$. After the interaction, the combined electron–specimen state may contain several possible outgoing channels,

$$\boxed{|\Psi\rangle = \sum_m a_m |\psi_m\rangle_{\mathrm{e}} \otimes |\Xi_m\rangle_{\mathrm{s}}.} \tag{A.663}$$

Here $|\psi_m\rangle_{\mathrm{e}}$ is the outgoing state of the imaging electron associated with the specimen being left in the internal state $|\Xi_m\rangle_{\mathrm{s}}$.

The two systems are therefore correlated: the electron state depends on which final state of the specimen has been produced.

In a conventional electron-microscopy measurement, however, the detector records the electron and does not determine the final internal state of the specimen. We must therefore calculate the electron probability without distinguishing among the possible specimen states.

Why orthogonal specimen channels add as probabilities.

Consider detecting the electron at position $\mathbf{r}$ while the specimen is found in a particular state $|\Xi_n\rangle_{\mathrm{s}}$.

The corresponding amplitude is

$${}_{\mathrm{e}}\langle \mathbf{r}|_{\mathrm{s}}\langle \Xi_n|\Psi\rangle = \sum_m a_m \, {}_{\mathrm{e}}\langle \mathbf{r}|\psi_m\rangle_{\mathrm{e}} \, {}_{\mathrm{s}}\langle \Xi_n|\Xi_m\rangle_{\mathrm{s}}. \tag{A.664}$$

If the possible specimen final states are orthonormal,

$${}_{\mathrm{s}}\langle \Xi_n|\Xi_m\rangle_{\mathrm{s}} = \delta_{nm}, \tag{A.665}$$

only the term $m = n$ survives:

$${}_{\mathrm{e}}\langle \mathbf{r}|_{\mathrm{s}}\langle \Xi_n|\Psi\rangle = a_n \psi_n(\mathbf{r}). \tag{A.666}$$

Therefore the joint probability density for detecting the electron at $\mathbf{r}$ and the specimen in state $|\Xi_n\rangle_{\mathrm{s}}$ is

$$P(\mathbf{r}, n) = |a_n|^2 |\psi_n(\mathbf{r})|^2. \tag{A.667}$$

If the specimen state is not measured, the alternatives $n = 0, 1, 2, \ldots$ are mutually exclusive outcomes of the unobserved specimen measurement. The electron probability density is therefore obtained by summing over them:

$$\boxed{P_{\mathrm{e}}(\mathbf{r}) = \sum_n |a_n|^2 |\psi_n(\mathbf{r})|^2.} \tag{A.668}$$

No cross terms of the form

$$\psi_m^*(\mathbf{r})\psi_n(\mathbf{r}), \qquad m \neq n, \tag{A.669}$$

remain. Their absence is a consequence of the orthogonality of the corresponding specimen states.

The important result is therefore

$$\boxed{\begin{array}{rcl} \text{same complete final state} & \Rightarrow & \text{add amplitudes,} \\ \text{orthogonal final states of an unobserved system} & \Rightarrow & \text{add probabilities.} \end{array}} \tag{A.670}$$

This is the quantum-mechanical origin of the loss of coherence associated with many inelastic scattering processes. The outgoing electron becomes correlated with a specimen degree of freedom that is not observed by the microscope. Because the different specimen states are orthogonal, the corresponding electron channels do not contribute to one common interference amplitude.

This loss of interference caused by correlation with unobserved degrees of freedom is commonly called *decoherence*.

This result does not mean that scattering amplitudes always add incoherently.

If several alternatives leave the specimen in the same final internal state, they belong to the same complete quantum channel and their electron amplitudes may interfere exactly as described in Eq. (A.93). In particular, the different elastic scattering contributions that leave the specimen in $|\Xi_0\rangle_{\mathrm{s}}$ form one coherent electron wave.

Thus multiple elastic scattering remains a coherent wave phenomenon; the distinction above concerns different *orthogonal specimen final states*, not

different scattering events in general.

A.14 Relativistic electron–electron scattering

The relativistic description developed in Sec. A.10.9 tells us how a free electron propagates and how it couples to an external electromagnetic field. We now consider a different problem: the relativistic interaction between *two* electrons.

This problem appears naturally in electron microscopy whenever a fast imaging electron transfers a sufficiently large amount of energy to an electron in matter. If the transferred energy is large compared with the binding energy of the target electron, the latter can be treated approximately as a free electron initially at rest. The elementary process is then

$$\boxed{e^- + e^- \longrightarrow e^- + e^-,} \tag{A.671}$$

known as *Møller scattering* (Møller, 1932).

In the language of radiation transport and detector physics, the struck electron that emerges from such a collision is a *secondary electron*, or *knock-on electron*; when the transferred energy is sufficiently large, it is commonly referred to as a *delta ray*. This terminology provides a direct connection between the elementary electron–electron scattering process in Eq. (A.671) and the response of an electron detector. An energetic secondary electron can itself travel through the detector material, undergo further collisions, and deposit its energy some distance away from the point at which the primary electron first interacted. In a silicon detector, this lateral transport of secondary electrons is therefore one of the mechanisms that spreads the deposited energy over several pixels, producing spatial cross-talk and contributing to the degradation of the modulation transfer function (MTF).

Thus Møller scattering is not only a fundamental QED process needed to calculate the hard-collision contribution to the stopping power; it also provides a microscopic origin for part of the lateral spreading of the detector signal.

The treatment requires several pieces of relativistic notation that were not needed in the scalar scattering theory developed earlier. We introduce them first and only then calculate the scattering cross section.

A.14.1 Four-momentum and Lorentz-invariant products

The relativistic energy–momentum relation was introduced in Eq. (A.524),

$$E^2 = p^2c^2 + m_0^2c^4. \tag{A.672}$$

Relativity combines the total energy E and the ordinary three-dimensional momentum $\mathbf{p}$ into the *four-momentum*. With an upper index we write

$$\boxed{p^\mu = \left(\frac{E}{c}, p_x, p_y, p_z\right).} \tag{A.673}$$

The index $\mu = 0, 1, 2, 3$ labels the temporal component and the three spatial components.

Four-vectors are combined using the Minkowski metric

$$\boxed{g_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1).} \tag{A.674}$$

The corresponding four-momentum with a lower index is defined by

$$\boxed{p_\mu \equiv g_{\mu\nu}p^\nu.} \tag{A.675}$$

With the metric in Eq. (A.674), this gives

$$\boxed{p_\mu = \left(\frac{E}{c}, -p_x, -p_y, -p_z\right).} \tag{A.676}$$

Thus raising or lowering the index does not introduce a different physical momentum. It changes the signs of the spatial components according to the Minkowski metric.

For two four-vectors a^μ and b^μ, their Minkowski scalar product is

$$\boxed{a^\mu b_\mu = g_{\mu\nu}a^\mu b^\nu = a^0b^0 - \mathbf{a}\cdot\mathbf{b}.} \tag{A.677}$$

Repeated Greek indices are summed from 0 to 3.

Applying Eq. (A.677) to the four-momentum,

$$p^\mu p_\mu = \left(\frac{E}{c}\right)^2 - p_x^2 - p_y^2 - p_z^2 \tag{A.678}$$

$$= \frac{E^2}{c^2} - |\mathbf{p}|^2. \tag{A.679}$$

Using the relativistic energy–momentum relation, Eq. (A.524),

$$\frac{E^2}{c^2} - |\mathbf{p}|^2 = m_0^2 c^2, \tag{A.680}$$

and therefore

$$\boxed{p^\mu p_\mu = m_0^2 c^2.} \tag{A.681}$$

Unlike E and $\mathbf{p}$ separately, the quantity $p^\mu p_\mu$ has the same value in every inertial frame. Such a quantity is called a *Lorentz invariant*.

A.14.2 Covariant Dirac notation

The remaining calculation is substantially clearer in the natural units convention commonly used in relativistic quantum theory,

$$\boxed{\hbar = c = 1.} \tag{A.682}$$

Within this section, energies, masses, and momenta therefore have the same physical dimension. In particular, Eq. (A.673) becomes

$$p^\mu = (E, \mathbf{p}), \tag{A.683}$$

and the mass-shell condition becomes

$$\boxed{p^2 = m^2,} \tag{A.684}$$

where, for compactness, we temporarily write m for the electron rest mass.

The terminology *mass shell* deserves some explanation. Recall that, with the Minkowski metric,

$$p^2 = p^\mu p_\mu = E^2 - |\mathbf{p}|^2 \tag{A.685}$$

in natural units. Therefore, Eq. (A.684) is simply the relativistic energy–momentum relation written in covariant form:

$$\boxed{E^2 = |\mathbf{p}|^2 + m^2.} \tag{A.686}$$

Its physical meaning is important: for a real free electron, the energy and momentum are not independent quantities. Once the three-momentum $\mathbf{p}$ is specified, the energy is fixed to

$$E = \sqrt{|\mathbf{p}|^2 + m^2} \tag{A.687}$$

for the positive-energy electron branch.

Thus, although the four-momentum

$$p^\mu = (E, \mathbf{p}) \tag{A.688}$$

has four components, a free electron cannot occupy an arbitrary point in four-dimensional energy–momentum space. Its allowed four-momenta must satisfy

$$p^2 = m^2. \tag{A.689}$$

The set of points satisfying this equation forms a three-dimensional surface in four-dimensional energy–momentum space. This surface is called the *mass shell*.

This condition will be used repeatedly in the scattering calculation. Each real incoming or outgoing electron satisfies

$$\boxed{p_i^2 = m^2, \qquad i = 1, 2, 3, 4.} \tag{A.690}$$

Consequently, many scalar products between the four-momenta can be rewritten in terms of the Mandelstam variables introduced below.

For example,

$$\mathsf{t} = (p_1 - p_3)^2 \tag{A.691}$$

can be expanded as

$$\mathsf{t} = p_1^2 + p_3^2 - 2p_1 \cdot p_3 \tag{A.692}$$

$$= 2m^2 - 2p_1 \cdot p_3, \tag{A.693}$$

because both the incoming electron 1 and the outgoing electron 3 are on their mass shell. Hence

$$\boxed{2p_1 \cdot p_3 = 2m^2 - \mathsf{t}.} \tag{A.694}$$

This is one reason why the mass-shell condition is so useful in relativistic scattering calculations.

The adjective *on-shell* is used for a real particle whose four-momentum satisfies its physical mass relation. Thus a real electron is on shell when

$$p^2 = m^2. \tag{A.695}$$

By contrast, an intermediate or exchanged quantum in a scattering amplitude need not satisfy the corresponding free-particle relation. Such a

quantum is said to be *off shell* or *virtual*. For example, the virtual photon exchanged in Møller scattering generally has

$$Q^2 \neq 0, \tag{A.696}$$

even though a real photon propagating in vacuum satisfies

$$Q^2 = 0. \tag{A.697}$$

In Møller scattering, the exchanged photon belongs to either the t-channel or the u-channel. Its four-momentum is therefore the four-momentum transferred between the electrons. For example, in the t-channel,

$$Q_t = p_1 - p_3, \tag{A.698}$$

so that

$$\boxed{Q_t^2 = \mathsf{t} < 0} \tag{A.699}$$

for a non-zero physical momentum transfer. Likewise, the exchanged four-momentum in the u-channel satisfies

$$Q_u^2 = \mathsf{u} < 0. \tag{A.700}$$

A four-momentum with negative invariant square is called *space-like*. Thus the virtual photon exchanged between the two electrons carries a space-like four-momentum, in contrast with a real photon in vacuum, whose four-momentum is *light-like*,

$$Q^2 = 0. \tag{A.701}$$

This is another way of seeing why the exchanged photon cannot be interpreted as a real photon propagating freely between the electrons: it is an internal quantum of the interaction and is not constrained to satisfy the real-photon mass-shell condition.

The Dirac matrices $\boldsymbol{\alpha}$ and β were introduced in Eq. (A.530). For relativistic scattering it is more convenient to combine them into the four *gamma matrices*

$$\boxed{\gamma^0 = \beta_{\mathrm{D}}, \qquad \gamma^i = \beta_{\mathrm{D}}\alpha_i, \quad i = 1, 2, 3,} \tag{A.702}$$

where we have written β_{D} for the Dirac matrix denoted by β in Eq. (A.530), in order not to confuse it with the dimensionless velocity $\beta = v/c$.

The algebra of the α_i and β_{D} matrices given below Eq. (A.530) implies

$$\boxed{\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu}\hat{I}_4.} \tag{A.703}$$

This relation is the covariant form of the Dirac-matrix algebra.

Covariant Dirac-matrix algebra.

For $\mu = \nu = 0$,

$$(\gamma^0)^2 = \beta_{\mathrm{D}}^2 = \hat{I}_4. \tag{A.704}$$

For one temporal and one spatial index,

$$\gamma^0\gamma^i + \gamma^i\gamma^0 = \beta_{\mathrm{D}}^2\alpha_i + \beta_{\mathrm{D}}\alpha_i\beta_{\mathrm{D}} \tag{A.705}$$
$$= \alpha_i - \alpha_i \tag{A.706}$$
$$= 0, \tag{A.707}$$

where we used $\alpha_i\beta_{\mathrm{D}} = -\beta_{\mathrm{D}}\alpha_i$.

Finally,

$$\gamma^i\gamma^j + \gamma^j\gamma^i = \beta_{\mathrm{D}}\alpha_i\beta_{\mathrm{D}}\alpha_j + \beta_{\mathrm{D}}\alpha_j\beta_{\mathrm{D}}\alpha_i \tag{A.708}$$
$$= -\left(\alpha_i\alpha_j + \alpha_j\alpha_i\right) \tag{A.709}$$
$$= -2\delta_{ij}\hat{I}_4. \tag{A.710}$$

Since $g^{00} = 1$, $g^{0i} = 0$, and $g^{ij} = -\delta_{ij}$, these three cases combine into Eq. (A.703).

A free positive-energy solution of the Dirac equation, Eq. (A.533), with definite four-momentum p and spin label s can be written as a plane wave whose four-component amplitude is denoted by

$$u_s(p). \tag{A.711}$$

The four-component structure of the Dirac state was introduced in Eq. (A.529), while the corresponding free Dirac Hamiltonian is given in Eq. (A.530). Thus $u_s(p)$ is a four-component Dirac spinor of the same kind as the state introduced in Eq. (A.529); the label p specifies the momentum and s distinguishes the two possible positive-energy electron spin states.

The spinor satisfies the free Dirac equation

$$\boxed{\left(\not{p} - m\right) u_s(p) = 0,} \tag{A.712}$$

where the *slash notation*

$$\boxed{\not{p} \equiv \gamma^\mu p_\mu} \tag{A.713}$$

is a compact way of contracting the four-momentum with the gamma matrices.

For covariant calculations we also introduce the *Dirac adjoint*

$$\boxed{\bar{u}_s(p) \equiv u_s^\dagger(p)\gamma^0.} \tag{A.714}$$

The bar therefore does not denote a new independent state: it is formed from the original spinor by taking its Hermitian conjugate and multiplying by γ^0.

A free positive-energy solution of the Dirac equation with definite four-momentum p and spin label s can be written as a plane wave whose four-component amplitude is denoted by

$$u_s(p). \tag{A.715}$$

Thus $u_s(p)$ is a four-component Dirac spinor of the same kind as the state introduced in Eq. (A.529); the label p specifies the momentum and s distinguishes the two possible positive-energy electron spin states.

The spinor satisfies the free Dirac equation

$$\boxed{\left(\not{p} - m\right) u_s(p) = 0,} \tag{A.716}$$

where the *slash notation*

$$\boxed{\not{p} \equiv \gamma^\mu p_\mu} \tag{A.717}$$

is a compact way of contracting the four-momentum with the gamma matrices.

To describe the electromagnetic interaction between two *quantum* electrons, the electromagnetic field itself must also be quantized. That step leads from relativistic quantum mechanics to quantum electrodynamics (QED). A complete derivation of QED is outside the scope of this appendix.

We shall need only its lowest-order result: two charged Dirac particles can interact by exchanging one virtual photon. In natural units, the corresponding scattering amplitude contains the product of their two Dirac currents divided by the squared four-momentum carried by the virtual photon,

$$\boxed{\mathcal{M} \propto \frac{J_1^\mu J_{2,\mu}}{Q^2}.} \tag{A.718}$$

The factor $1/Q^2$ is the covariant analogue of the $1/q^2$ Fourier-space Coulomb interaction encountered in the scattering theory of Chap. 5.

A.14.3 Invariant kinematics for two-particle scattering

Consider the general two-particle process

$$1 + 2 \longrightarrow 3 + 4, \tag{A.719}$$

with incoming four-momenta p_1, p_2 and outgoing four-momenta p_3, p_4. Four-momentum conservation requires

$$\boxed{p_1 + p_2 = p_3 + p_4.} \tag{A.720}$$

Three particularly useful Lorentz-invariant combinations are

$$\boxed{\begin{aligned} \mathsf{s} &= (p_1 + p_2)^2, \\ \mathsf{t} &= (p_1 - p_3)^2, \\ \mathsf{u} &= (p_1 - p_4)^2. \end{aligned}} \tag{A.721}$$

They are known as the *Mandelstam variables*. Standard texts denote them by s, t, u. We use the typeface $\mathsf{s}, \mathsf{t}, \mathsf{u}$ here to avoid confusing the last of these with the Dirac spinor $u_s(p)$.

Why the Mandelstam variables are Lorentz invariants.

A Lorentz transformation acts linearly on any four-vector. Thus, if

$$p^\mu \longrightarrow p'^\mu = \Lambda^\mu{}_\nu p^\nu, \tag{A.722}$$

then each of the particle four-momenta

$$p_1, \quad p_2, \quad p_3, \quad p_4 \tag{A.723}$$

transforms in this way.

The defining property of a Lorentz transformation is that it preserves the Minkowski scalar product, Eq. (A.677). In components, this means

$$\boxed{g_{\mu\nu}\Lambda^\mu{}_\alpha \Lambda^\nu{}_\beta = g_{\alpha\beta}.} \tag{A.724}$$

Let a^μ be any four-vector. Under a Lorentz transformation,

$$a'^2 = g_{\mu\nu} a'^\mu a'^\nu \tag{A.725}$$
$$= g_{\mu\nu} \Lambda^\mu{}_\alpha a^\alpha \Lambda^\nu{}_\beta a^\beta \tag{A.726}$$
$$= \left(g_{\mu\nu} \Lambda^\mu{}_\alpha \Lambda^\nu{}_\beta\right) a^\alpha a^\beta \tag{A.727}$$
$$= g_{\alpha\beta} a^\alpha a^\beta \tag{A.728}$$
$$= a^2. \tag{A.729}$$

Thus the Minkowski square of any four-vector is Lorentz invariant.

We now apply this to the combinations appearing in Eq. (A.721). Because Lorentz transformations are linear, sums and differences of four-vectors are again four-vectors. For example,

$$p_1 + p_2 \tag{A.730}$$

is a four-vector, since

$$(p_1 + p_2)'^\mu = p_1'^\mu + p_2'^\mu = \Lambda^\mu{}_\nu (p_1^\nu + p_2^\nu). \tag{A.731}$$

Therefore its Minkowski square is invariant:

$$(p_1 + p_2)'^2 = (p_1 + p_2)^2. \tag{A.732}$$

Hence

$$\mathsf{s} = (p_1 + p_2)^2 \tag{A.733}$$

is Lorentz invariant.

Exactly the same argument applies to

$$p_1 - p_3 \qquad \text{and} \qquad p_1 - p_4, \tag{A.734}$$

which are also four-vectors. Their Minkowski squares are therefore invariant:

$$(p_1 - p_3)'^2 = (p_1 - p_3)^2, \tag{A.735}$$

and

$$(p_1 - p_4)'^2 = (p_1 - p_4)^2. \tag{A.736}$$

Consequently,

$$\mathsf{t} = (p_1 - p_3)^2, \qquad \mathsf{u} = (p_1 - p_4)^2 \tag{A.737}$$

are also Lorentz invariants.

We conclude that all three Mandelstam variables are Lorentz invariant:

$$\boxed{\begin{aligned} \mathsf{s} &= (p_1 + p_2)^2, \\ \mathsf{t} &= (p_1 - p_3)^2, \\ \mathsf{u} &= (p_1 - p_4)^2. \end{aligned}} \tag{A.738}$$

Their interpretation is simple. s characterizes the total energy available to the two-particle system, while t and u characterize the two possible four-momentum transfers between incoming and outgoing particles.

For electron–electron scattering all four particles have the same rest mass, so

$$p_1^2 = p_2^2 = p_3^2 = p_4^2 = m^2. \tag{A.739}$$

The three Mandelstam variables then satisfy

$$\boxed{\mathsf{s} + \mathsf{t} + \mathsf{u} = 4m^2.} \tag{A.740}$$

The Mandelstam sum for equal masses.

From Eq. (A.721),

$$\mathsf{s} + \mathsf{t} + \mathsf{u} = (p_1 + p_2)^2 + (p_1 - p_3)^2 + (p_1 - p_4)^2 \tag{A.741}$$

$$= 3p_1^2 + p_2^2 + p_3^2 + p_4^2 \tag{A.742}$$

$$+ 2p_1 \cdot (p_2 - p_3 - p_4). \tag{A.743}$$

Momentum conservation, Eq. (A.720), gives

$$p_2 - p_3 - p_4 = -p_1. \tag{A.744}$$

Hence

$$\mathsf{s} + \mathsf{t} + \mathsf{u} = 3p_1^2 + p_2^2 + p_3^2 + p_4^2 - 2p_1^2 \tag{A.745}$$

$$= p_1^2 + p_2^2 + p_3^2 + p_4^2 \tag{A.746}$$

$$= 4m^2. \tag{A.747}$$

A.14.4 Møller scattering: two indistinguishable alternatives

We now specialize the general relativistic scattering notation to the interaction of two electrons,

$$\boxed{e^-(p_1, s_1) + e^-(p_2, s_2) \longrightarrow e^-(p_3, s_3) + e^-(p_4, s_4),} \tag{A.748}$$

where p_1 and p_2 are the four-momenta of the incoming electrons, p_3 and p_4 those of the outgoing electrons, and $s_1, \ldots, s_4$ denote their spin states.

The electron–photon interaction. The Dirac equation developed above describes a relativistic electron moving in a prescribed electromagnetic field. To describe the interaction of two electrons with one another, however, the electromagnetic field must itself be treated quantum mechanically. This extension of relativistic quantum mechanics is *quantum electrodynamics* (QED).

A complete derivation of QED would require the quantization of both the Dirac field and the electromagnetic field and lies beyond the scope of this appendix. For the present purpose, we need only the lowest-order interaction between them.

In QED, the electron is represented by a Dirac field $\psi(\mathbf{r}, t)$ and the electromagnetic field by the four-potential

$$A^\mu = (\phi, \mathbf{A}) \tag{A.749}$$

in the natural units $\hbar = c = 1$ introduced in Eq. (A.682). The interaction between these two fields is described by the interaction Lagrangian density

$$\boxed{\mathcal{L}_{\text{int}} = -e\,\bar{\psi}\gamma^\mu A_\mu \psi,} \tag{A.750}$$

where $e > 0$ is the magnitude of the electron charge and

$$\bar{\psi} = \psi^\dagger \gamma^0 \tag{A.751}$$

is the Dirac adjoint, in direct analogy with the adjoint spinor introduced above.

The combination

$$\boxed{j^\mu \equiv \bar{\psi}\gamma^\mu \psi} \tag{A.752}$$

is called the *Dirac four-current*. Apart from the electron-charge factor, it is the relativistic quantity that combines the density and the three-dimensional current carried by the Dirac field. Equation (A.750) may therefore be written as

$$\mathcal{L}_{\text{int}} = -e\, j^\mu A_\mu, \tag{A.753}$$

showing that the electromagnetic four-potential couples to the Dirac four-current.

For a scattering calculation, we are interested not in an arbitrary field configuration but in the transition of a free electron from an initial state of four-momentum p and spin s to a final state of four-momentum p' and spin s'. The corresponding free-electron spinors are

$$u_s(p) \qquad \text{and} \qquad u_{s'}(p'). \tag{A.754}$$

The current associated with this particular transition is the matrix element

$$\boxed{J^\mu(p', s'; p, s) \equiv \bar{u}_{s'}(p')\gamma^\mu u_s(p).} \tag{A.755}$$

This is a specific example of the general matrix-element concept introduced in Sec. A.12: the initial and final states are now Dirac spinors and the operator between them is γ^μ.

The perturbative formulation of QED translates Eq. (A.750) into a set of rules for constructing scattering amplitudes. A point at which an electron interacts with the electromagnetic field is called an *electron–photon vertex*. Each such vertex contributes the factor

$$\boxed{-ie\gamma^\mu.} \tag{A.756}$$

When this factor is placed between an initial and a final electron spinor, it produces

$$-\,ie\,\bar{u}_{s'}(p')\gamma^\mu u_s(p) = -ie\,J^\mu, \tag{A.757}$$

so the appearance of the current matrix element in the scattering amplitude follows directly from the electron–photon interaction.

A second ingredient is required because the electromagnetic interaction must propagate from one electron to the other. If the exchanged photon carries four-momentum Q^μ, QED assigns to it, in Feynman gauge, the *photon propagator*

$$\boxed{\frac{-ig_{\mu\nu}}{Q^2 + i0}.} \tag{A.758}$$

The quantity $Q^2 = Q^\mu Q_\mu$ is the Minkowski square of the four-momentum carried between the two interaction vertices. The symbol $+i0$ denotes an infinitesimal imaginary contribution that specifies how the expression is evaluated at the pole $Q^2 = 0$. It will not affect the algebra needed below, and we shall therefore write simply

$$\frac{-ig_{\mu\nu}}{Q^2}. \tag{A.759}$$

The physical meaning of the factor $1/Q^2$ can be connected with the Coulomb scattering developed previously. In the static limit there is no energy transfer, so that

$$Q^\mu = (0, \mathbf{q}), \tag{A.760}$$

and therefore

$$Q^2 = -|\mathbf{q}|^2. \tag{A.761}$$

Thus the photon propagator contains the familiar $1/|\mathbf{q}|^2$ dependence of the Fourier-transformed Coulomb interaction encountered in the first-Born treatment of Chap. 5. The photon propagator is its fully relativistic generalization, allowing both energy and momentum to be transferred.

Finally, the photon exchanged between the two electrons is not a real photon that could be detected as a freely propagating particle. A real photon satisfies the mass-shell condition

$$Q^2 = 0. \tag{A.762}$$

The exchanged photon generally has

$$Q^2 \neq 0, \tag{A.763}$$

and is therefore said to be *virtual* or *off shell*. Such a virtual photon appears only as an intermediate part of the scattering process and provides the quantum-mechanical carrier of the electromagnetic interaction between the two electrons.

The direct scattering alternative. Consider first the association

$$1 \longrightarrow 3, \qquad 2 \longrightarrow 4. \tag{A.764}$$

In this alternative, the electron entering with four-momentum p_1 leaves with four-momentum p_3, while the electron entering with p_2 leaves with p_4.

The four-momentum carried between the two electron currents is therefore

$$\boxed{Q_t = p_1 - p_3.} \tag{A.765}$$

Using the definition of the Mandelstam variables in Eq. (A.721),

$$\boxed{Q_t^2 = (p_1 - p_3)^2 = \mathsf{t}.} \tag{A.766}$$

The two electron currents associated with this process are

$$J_{13}^\mu = \bar{u}_{s_3}(p_3)\gamma^\mu u_{s_1}(p_1), \tag{A.767}$$

and

$$J_{24}^{\nu} = \bar{u}_{s_4}(p_4)\gamma^{\nu}u_{s_2}(p_2). \tag{A.768}$$

The direct amplitude can now be constructed explicitly.

Direct one-photon-exchange amplitude.

The lowest-order process contains three factors:

$$\boxed{\begin{array}{c}\text{electron–photon vertex}\\ \times \\ \text{photon propagator}\\ \times \\ \text{electron–photon vertex.}\end{array}} \tag{A.769}$$

Using Eq. (A.756) at the two vertices and Eq. (A.758) for the virtual photon gives

$$\begin{aligned} i\mathcal{M}_t = & \left[\bar{u}_{s_3}(p_3)(-ie\gamma^{\mu})u_{s_1}(p_1)\right] \\ & \times \left[\frac{-ig_{\mu\nu}}{Q_t^2}\right] \\ & \times \left[\bar{u}_{s_4}(p_4)(-ie\gamma^{\nu})u_{s_2}(p_2)\right]. \end{aligned} \tag{A.770}$$

The numerical factors combine as

$$(-ie)(-i)(-ie) = ie^2, \tag{A.771}$$

so

$$i\mathcal{M}_t = \frac{ie^2}{Q_t^2}\left[\bar{u}_{s_3}(p_3)\gamma^{\mu}u_{s_1}(p_1)\right] g_{\mu\nu}\left[\bar{u}_{s_4}(p_4)\gamma^{\nu}u_{s_2}(p_2)\right]. \tag{A.772}$$

The metric contracts the two four-currents. Since

$$g_{\mu\nu}\gamma^{\nu} = \gamma_{\mu}, \tag{A.773}$$

we obtain

$$\mathcal{M}_t = \frac{e^2}{Q_t^2}\left[\bar{u}_{s_3}(p_3)\gamma^{\mu}u_{s_1}(p_1)\right]\left[\bar{u}_{s_4}(p_4)\gamma_{\mu}u_{s_2}(p_2)\right]. \tag{A.774}$$

Finally, using Eq. (A.766), $Q_t^2 = \mathsf{t}$,

$$\boxed{\mathcal{M}_t = \frac{e^2}{\mathsf{t}} \left[\bar{u}_{s_3}(p_3)\gamma^\mu u_{s_1}(p_1)\right]\left[\bar{u}_{s_4}(p_4)\gamma_\mu u_{s_2}(p_2)\right].} \tag{A.775}$$

The strength of the electron–photon interaction is conventionally expressed through the *fine-structure constant*,

$$\boxed{\alpha \equiv \frac{e^2}{4\pi\varepsilon_0 \hbar c}.} \tag{A.776}$$

Numerically,

$$\alpha \simeq \frac{1}{137.036}. \tag{A.777}$$

Unlike the electron charge e itself, α is dimensionless. It therefore provides a natural measure of the intrinsic strength of the electromagnetic interaction, independent of the particular system of units used.

Its appearance can be understood from the combination of constants in Eq. (A.776). The factor

$$\frac{e^2}{4\pi\varepsilon_0} \tag{A.778}$$

sets the strength of the Coulomb interaction, while

$$\hbar c \tag{A.779}$$

has the same physical dimensions. Their ratio is therefore a pure number:

$$[\alpha] = 1. \tag{A.780}$$

The fine-structure constant first appeared historically in the relativistic description of atomic spectra, hence its name, but its meaning is considerably more general: it is the fundamental dimensionless coupling constant of quantum electrodynamics.

This is precisely why it becomes useful in the present scattering problem. Each electron–photon vertex contributes one factor of the electric charge e. The lowest-order electron–electron scattering process contains two such vertices, so its amplitude contains a factor

$$e^2. \tag{A.781}$$

It is conventional to express this interaction strength in terms of α.

In the natural-unit convention used in this section,

$$\hbar = c = 1, \tag{A.782}$$

together with the rationalized Heaviside–Lorentz convention for the electromagnetic field, the SI definition Eq. (A.776) reduces to

$$\boxed{\alpha = \frac{e^2}{4\pi}.} \tag{A.783}$$

Equivalently,

$$\boxed{e^2 = 4\pi\alpha.} \tag{A.784}$$

Thus the factor e^2 appearing in the one-photon-exchange amplitude can be written as $4\pi\alpha$. For example,

$$\frac{e^2}{\mathsf{t}} = \frac{4\pi\alpha}{\mathsf{t}}. \tag{A.785}$$

The Møller scattering amplitude is therefore proportional to α, while the corresponding cross section, which depends on $|\mathcal{M}|^2$, is proportional at lowest order to

$$\boxed{\alpha^2.} \tag{A.786}$$

The small numerical value of α is also what makes perturbative QED effective for processes such as Møller scattering. Contributions with additional photon interactions introduce additional powers of α and are therefore progressively smaller. The one-photon-exchange calculation used here is the leading, lowest-order term in this expansion.

Therefore Eq. (A.775) becomes

$$\boxed{\begin{aligned}\mathcal{M}_t = \frac{4\pi\alpha}{\mathsf{t}} &\left[\bar{u}_{s_3}(p_3)\gamma^\mu u_{s_1}(p_1)\right] \\ &\times \left[\bar{u}_{s_4}(p_4)\gamma_\mu u_{s_2}(p_2)\right].\end{aligned}} \tag{A.787}$$

The physical structure of this expression can be summarized as

$$\boxed{\mathcal{M}_t \sim J^\mu_{13}\,\frac{1}{\mathsf{t}}\,J_{24,\mu}.} \tag{A.788}$$

The first factor describes the transition current of one electron, the middle factor describes propagation of the electromagnetic interaction, and the final factor describes its coupling to the transition current of the second electron.

The exchange alternative. If the two outgoing particles were distinguishable, the preceding amplitude would describe one particular assignment of the final particles. Electrons, however, are identical. There is no physical measurement that can determine whether the incoming electron labelled 1 became outgoing electron 3 or outgoing electron 4.

We must therefore also include the alternative assignment

$$1 \longrightarrow 4, \qquad 2 \longrightarrow 3. \tag{A.789}$$

The four-momentum carried by the virtual photon is now

$$\boxed{Q_u = p_1 - p_4,} \tag{A.790}$$

and therefore, from Eq. (A.721),

$$\boxed{Q_u^2 = (p_1 - p_4)^2 = \mathsf{u}.} \tag{A.791}$$

The same one-photon-exchange construction used above gives

$$\boxed{\begin{aligned} \mathcal{M}_u = \frac{4\pi\alpha}{\mathsf{u}} &\left[\bar{u}_{s_4}(p_4)\gamma^\mu u_{s_1}(p_1)\right] \\ &\times \left[\bar{u}_{s_3}(p_3)\gamma_\mu u_{s_2}(p_2)\right]. \end{aligned}} \tag{A.792}$$

The two amplitudes $\mathcal{M}_t$ and $\mathcal{M}_u$ therefore describe the two possible ways of associating the identical incoming and outgoing electrons.

Why the two amplitudes subtract. The two alternatives represented by Eqs. (A.787) and (A.792) lead to the same physical final state. Quantum mechanics therefore requires us to combine their *amplitudes*, not their probabilities.

There is, however, an additional property specific to electrons. Electrons have spin 1/2 and are therefore fermions. The quantum state of two identical fermions must be antisymmetric under exchange of the two particles. Thus interchanging the labels of the two outgoing electrons introduces a minus sign.

Consequently, the complete lowest-order Møller amplitude is

$$\boxed{\mathcal{M}_{\mathrm{M}} = \mathcal{M}_t - \mathcal{M}_u.} \tag{A.793}$$

This minus sign has observable consequences. Squaring the total amplitude gives

$$|\mathcal{M}_{\mathrm{M}}|^2 = |\mathcal{M}_t - \mathcal{M}_u|^2 \tag{A.794}$$

$$= |\mathcal{M}_t|^2 + |\mathcal{M}_u|^2 - 2\,\mathrm{Re}\left(\mathcal{M}_t \mathcal{M}_u^*\right). \tag{A.795}$$

The final term is an *exchange-interference* contribution. It would not be present if the two scattered particles were distinguishable.

Thus the exchange contribution is not a classical correction to electron–electron scattering. It follows directly from the indistinguishability of electrons and the antisymmetry required of identical fermions. In the next subsection we shall average Eq. (A.795) over the unobserved initial and final spin states and obtain the corresponding spin-averaged scattering probability.

A.14.5 Unpolarized scattering and spin sums

The Møller amplitudes derived in Sec. A.14.4 still depend explicitly on the spin states of the four electrons,

$$s_1, \quad s_2, \quad s_3, \quad s_4. \tag{A.796}$$

In the applications considered later, the spins of the two incident electrons are not prepared in any preferred state and the spins of the two outgoing electrons are not measured. The experimentally relevant cross section must therefore average over the possible initial spin states and sum over all possible final spin states.

Each electron has two spin states. Hence there are

$$2 \times 2 = 4 \tag{A.797}$$

possible spin configurations for the two incoming electrons. Since none is preferred, they enter with equal statistical weight 1/4. The final spin states, by contrast, correspond to alternative outcomes that are all accepted by the measurement and must therefore be added.

Consequently, the quantity entering the unpolarized scattering cross section is

$$\boxed{\overline{|\mathcal{M}_{\mathrm{M}}|^2} = \frac{1}{4} \sum_{s_1,s_2,s_3,s_4} |\mathcal{M}_{\mathrm{M}}|^2.} \tag{A.798}$$

Using the total Møller amplitude, Eq. (A.793),

$$\mathcal{M}_{\mathrm{M}} = \mathcal{M}_t - \mathcal{M}_u, \tag{A.799}$$

its squared magnitude is

$$\boxed{\begin{aligned} |\mathcal{M}_{\mathrm{M}}|^2 = |\mathcal{M}_t|^2 + |\mathcal{M}_u|^2 \\ - 2\,\mathrm{Re}\left(\mathcal{M}_t \mathcal{M}_u^*\right). \end{aligned}} \tag{A.800}$$

We shall evaluate the spin sum of these three contributions separately.

Completeness of the positive-energy spinors. The general concept of completeness was introduced earlier in this appendix; for example, Eq. (A.148) expressed completeness of the continuous momentum basis. Here we need the corresponding result for the two positive-energy spin states of a Dirac electron with a fixed four-momentum p.

Free Dirac spinors may be assigned different normalization conventions. For relativistic scattering calculations it is convenient to adopt the standard *covariant normalization*

$$\boxed{\bar{u}_s(p)u_{s'}(p) = 2m\,\delta_{ss'}.} \tag{A.801}$$

Here m denotes the electron rest mass in the natural-unit convention of Eq. (A.682). The factor $2m$ is a normalization convention, not an additional physical law.

With this convention, the two positive-energy spinors satisfy the completeness relation

$$\boxed{\sum_s u_s(p)\bar{u}_s(p) = \not{p} + m.} \tag{A.802}$$

Completeness of the positive-energy Dirac spinors.

The momentum-space Dirac equation, Eq. (A.712), is

$$(\not{p} - m)u_s(p) = 0, \tag{A.803}$$

or equivalently

$$\not{p}\,u_s(p) = m\,u_s(p). \tag{A.804}$$

Consider the matrix

$$\boxed{\Lambda_+(p) \equiv \frac{\not{p} + m}{2m}.} \tag{A.805}$$

Acting on a positive-energy spinor and using Eq. (A.804),

$$\Lambda_+(p)u_s(p) = \frac{\not{p} + m}{2m}u_s(p) \tag{A.806}$$

$$= \frac{m + m}{2m}u_s(p) \tag{A.807}$$

$$= u_s(p). \tag{A.808}$$

Thus $\Lambda_+(p)$ leaves every positive-energy solution unchanged.

It is also a projector. Indeed,

$$\Lambda_+^2(p) = \frac{(\not{p} + m)^2}{4m^2} \tag{A.809}$$

$$= \frac{\not{p}\not{p} + 2m\not{p} + m^2}{4m^2}. \tag{A.810}$$

From the gamma-matrix anticommutation relation, Eq. (A.703),

$$\not{p}\not{p} = p^2, \tag{A.811}$$

and for an on-shell electron, Eq. (A.684),

$$p^2 = m^2. \tag{A.812}$$

Therefore,

$$\Lambda_+^2(p) = \frac{2m^2 + 2m\not{p}}{4m^2} \tag{A.813}$$

$$= \frac{\not{p} + m}{2m} \tag{A.814}$$

$$= \Lambda_+(p). \tag{A.815}$$

The same projector can be constructed directly from the two spinors that span the positive-energy subspace. With the normalization Eq. (A.801),

$$\Lambda_+(p) = \frac{1}{2m} \sum_s u_s(p)\bar{u}_s(p). \tag{A.816}$$

Equating Eqs. (A.805) and (A.816) gives

$$\boxed{\sum_s u_s(p)\bar{u}_s(p) = \not{p} + m,} \tag{A.817}$$

which proves Eq. (A.802).

Equation (A.802) is the key to carrying out the spin sums. After an amplitude is multiplied by its complex conjugate, each spin label appears once in a spinor and once in its Dirac adjoint. The corresponding sum

therefore contains precisely the combination

$$\sum_s u_s(p)\bar{u}_s(p), \tag{A.818}$$

which can be replaced by $\not{p} + m$.

From spin sums to traces. We first demonstrate this procedure for the direct contribution $|\mathcal{M}_t|^2$. From Eq. (A.787),

$$\mathcal{M}_t = \frac{4\pi\alpha}{\mathsf{t}} \left[\bar{u}_{s_3}(p_3)\gamma^\mu u_{s_1}(p_1)\right] \left[\bar{u}_{s_4}(p_4)\gamma_\mu u_{s_2}(p_2)\right]. \tag{A.819}$$

To construct its squared magnitude we need the complex conjugate of a Dirac-current matrix element. The gamma matrices satisfy

$$(\gamma^\mu)^\dagger = \gamma^0\gamma^\mu\gamma^0, \tag{A.820}$$

and therefore

$$\boxed{[\bar{u}_a\gamma^\mu u_b]^* = \bar{u}_b\gamma^\mu u_a.} \tag{A.821}$$

Complex conjugate of a Dirac current.
Using $\bar{u}_a = u_a^\dagger\gamma^0$,

$$[\bar{u}_a\gamma^\mu u_b]^* = \left[u_a^\dagger\gamma^0\gamma^\mu u_b\right]^* \tag{A.822}$$

$$= u_b^\dagger(\gamma^\mu)^\dagger(\gamma^0)^\dagger u_a. \tag{A.823}$$

Since $(\gamma^0)^\dagger = \gamma^0$ and using Eq. (A.820),

$$[\bar{u}_a\gamma^\mu u_b]^* = u_b^\dagger\gamma^0\gamma^\mu\gamma^0\gamma^0 u_a \tag{A.824}$$

$$= u_b^\dagger\gamma^0\gamma^\mu u_a \tag{A.825}$$

$$= \bar{u}_b\gamma^\mu u_a. \tag{A.826}$$

Hence

$$|\mathcal{M}_t|^2 = \frac{(4\pi\alpha)^2}{\mathsf{t}^2} \left[\bar{u}_{s_3}(p_3)\gamma^\mu u_{s_1}(p_1)\right] \left[\bar{u}_{s_1}(p_1)\gamma^\nu u_{s_3}(p_3)\right] \times \left[\bar{u}_{s_4}(p_4)\gamma_\mu u_{s_2}(p_2)\right] \left[\bar{u}_{s_2}(p_2)\gamma_\nu u_{s_4}(p_4)\right]. \tag{A.827}$$

We can now perform the spin sums.

Converting the direct spin sum into traces.

Consider first the factors containing s_1 and s_3. Summing over s_1 gives

$$\sum_{s_1} [\bar{u}_{s_3}(p_3)\gamma^\mu u_{s_1}(p_1)] [\bar{u}_{s_1}(p_1)\gamma^\nu u_{s_3}(p_3)] \tag{A.828}$$

$$= \bar{u}_{s_3}(p_3)\gamma^\mu \left[\sum_{s_1} u_{s_1}(p_1)\bar{u}_{s_1}(p_1)\right] \gamma^\nu u_{s_3}(p_3) \tag{A.829}$$

$$= \bar{u}_{s_3}(p_3)\gamma^\mu(\not{p}_1 + m)\gamma^\nu u_{s_3}(p_3), \tag{A.830}$$

where we used Eq. (A.802).

For any 4×4 matrix A,

$$\bar{u}_s A u_s = \mathrm{Tr}\left[A u_s \bar{u}_s\right], \tag{A.831}$$

where Tr denotes the sum of the diagonal elements of a matrix. Therefore,

$$\sum_s \bar{u}_s A u_s = \mathrm{Tr}\left[A \sum_s u_s \bar{u}_s\right] \tag{A.832}$$

$$= \mathrm{Tr}\left[A(\not{p} + m)\right]. \tag{A.833}$$

Applying this result to the sum over s_3 gives

$$\boxed{\begin{aligned}&\sum_{s_1,s_3} [\bar{u}_{s_3}(p_3)\gamma^\mu u_{s_1}(p_1)] [\bar{u}_{s_1}(p_1)\gamma^\nu u_{s_3}(p_3)] \\ &\qquad = \mathrm{Tr}\left[(\not{p}_3 + m)\gamma^\mu(\not{p}_1 + m)\gamma^\nu\right].\end{aligned}} \tag{A.834}$$

Exactly the same argument for s_2 and s_4 gives

$$\boxed{\begin{aligned}&\sum_{s_2,s_4} [\bar{u}_{s_4}(p_4)\gamma_\mu u_{s_2}(p_2)] [\bar{u}_{s_2}(p_2)\gamma_\nu u_{s_4}(p_4)] \\ &\qquad = \mathrm{Tr}\left[(\not{p}_4 + m)\gamma_\mu(\not{p}_2 + m)\gamma_\nu\right].\end{aligned}} \tag{A.835}$$

Including the factor 1/4 that averages over the initial spins, Eq. (A.798), we obtain

$$\boxed{\begin{aligned}\overline{|\mathcal{M}_t|^2} = \frac{(4\pi\alpha)^2}{4\mathsf{t}^2}\,\mathrm{Tr}\left[(\not{p}_3+m)\gamma^\mu(\not{p}_1+m)\gamma^\nu\right] \\ \times\,\mathrm{Tr}\left[(\not{p}_4+m)\gamma_\mu(\not{p}_2+m)\gamma_\nu\right].\end{aligned}} \tag{A.836}$$

Thus the completeness relation has not replaced $|\mathcal{M}_t|^2$ directly. Rather, multiplication by the complex conjugate creates the combinations $u_s\bar{u}_s$, and the sums over the unobserved spin states convert those combinations into $\not{p}+m$. The remaining matrix products are naturally expressed as traces.

Gamma-matrix trace identities. The traces in Eq. (A.836) can be evaluated without choosing an explicit representation of the gamma matrices. We need only their anticommutation relation, Eq. (A.703), and the cyclic property of the trace,

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA). \tag{A.837}$$

The principal identities required below are

$$\boxed{\mathrm{Tr}\,(\gamma^\mu\gamma^\nu) = 4g^{\mu\nu},} \tag{A.838}$$

and

$$\boxed{\begin{aligned}\mathrm{Tr}\,(\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma) = 4(g^{\mu\nu}g^{\rho\sigma} - g^{\mu\rho}g^{\nu\sigma} \\ + g^{\mu\sigma}g^{\nu\rho}).\end{aligned}} \tag{A.839}$$

Traces containing an odd number of gamma matrices vanish.

Two- and four-gamma traces.

Taking the trace of the anticommutation relation Eq. (A.703),

$$\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2g^{\mu\nu}\hat{I}_4, \tag{A.840}$$

gives

$$\mathrm{Tr}(\gamma^\mu\gamma^\nu) + \mathrm{Tr}(\gamma^\nu\gamma^\mu) = 2g^{\mu\nu}\,\mathrm{Tr}(\hat{I}_4). \tag{A.841}$$

By cyclicity the two traces on the left are equal, while $\mathrm{Tr}(\hat{I}_4) = 4$. Hence

$$\mathrm{Tr}(\gamma^\mu\gamma^\nu) = 4g^{\mu\nu}, \tag{A.842}$$

which proves Eq. (A.838).

Now define

$$T = \mathrm{Tr}\left(\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma\right). \tag{A.843}$$

Using $\gamma^\mu\gamma^\nu = 2g^{\mu\nu} - \gamma^\nu\gamma^\mu$,

$$T = 2g^{\mu\nu}\,\mathrm{Tr}(\gamma^\rho\gamma^\sigma) - \mathrm{Tr}(\gamma^\nu\gamma^\mu\gamma^\rho\gamma^\sigma). \tag{A.844}$$

Applying the anticommutation relation twice more and then using cyclicity of the trace gives

$$\begin{aligned} 2T = 8\big(&g^{\mu\nu}g^{\rho\sigma} - g^{\mu\rho}g^{\nu\sigma} \\ &+ g^{\mu\sigma}g^{\nu\rho}\big), \end{aligned} \tag{A.845}$$

and therefore Eq. (A.839).

Expanding a trace of the type appearing in Eq. (A.836) and using Eqs. (A.838) and (A.839) gives the useful identity

$$\boxed{\begin{aligned} \mathrm{Tr}\left[(\not{a}+m)\gamma^\mu(\not{b}+m)\gamma^\nu\right] = 4\big[&a^\mu b^\nu + a^\nu b^\mu \\ &+ (m^2 - a\cdot b)g^{\mu\nu}\big]. \end{aligned}} \tag{A.846}$$

Direct contribution. We can now evaluate Eq. (A.836). The required scalar products are conveniently expressed in terms of the Mandelstam variables introduced in Eq. (A.721). Since every external electron is on shell, Eq. (A.690),

$$\mathsf{s} = (p_1 + p_2)^2 = 2m^2 + 2p_1\cdot p_2, \tag{A.847}$$

$$\mathsf{t} = (p_1 - p_3)^2 = 2m^2 - 2p_1\cdot p_3, \tag{A.848}$$

$$\mathsf{u} = (p_1 - p_4)^2 = 2m^2 - 2p_1\cdot p_4. \tag{A.849}$$

Therefore,

$$\boxed{\begin{aligned} 2p_1\cdot p_2 &= \mathsf{s} - 2m^2, \\ 2p_1\cdot p_3 &= 2m^2 - \mathsf{t}, \\ 2p_1\cdot p_4 &= 2m^2 - \mathsf{u}. \end{aligned}} \tag{A.850}$$

Four-momentum conservation gives the analogous relations for the other pairs.

Direct contribution to the spin-averaged amplitude.

Applying Eq. (A.846) to the two traces in Eq. (A.836) and contracting the indices μ and ν gives

$$\begin{aligned}&\mathrm{Tr}\left[(\not{p}_3+m)\gamma^\mu(\not{p}_1+m)\gamma^\nu\right]\\&\quad\times\mathrm{Tr}\left[(\not{p}_4+m)\gamma_\mu(\not{p}_2+m)\gamma_\nu\right]\\&=8\big[(\mathsf{s}-2m^2)^2+(\mathsf{u}-2m^2)^2+4m^2\mathsf{t}\big].\end{aligned}\tag{A.851}$$

Substitution into Eq. (A.836) yields

$$\overline{|\mathcal{M}_t|^2}=\frac{2(4\pi\alpha)^2}{\mathsf{t}^2}\left[(\mathsf{s}-2m^2)^2+(\mathsf{u}-2m^2)^2+4m^2\mathsf{t}\right].\tag{A.852}$$

Using the equal-mass Mandelstam relation Eq. (A.740),

$$\mathsf{s}+\mathsf{t}+\mathsf{u}=4m^2,\tag{A.853}$$

the expression in brackets becomes

$$\mathsf{s}^2+\mathsf{u}^2+8m^2\mathsf{t}-8m^4.\tag{A.854}$$

Hence

$$\boxed{\overline{|\mathcal{M}_t|^2}=\frac{2(4\pi\alpha)^2}{\mathsf{t}^2}\left(\mathsf{s}^2+\mathsf{u}^2+8m^2\mathsf{t}-8m^4\right).}\tag{A.855}$$

Exchange contribution. The exchange amplitude, Eq. (A.792), is obtained from the direct amplitude by exchanging the two final electrons,

$$p_3\longleftrightarrow p_4.\tag{A.856}$$

From Eq. (A.721), this exchange corresponds to

$$\mathsf{t}\longleftrightarrow\mathsf{u}.\tag{A.857}$$

Therefore Eq. (A.855) immediately gives

$$\boxed{\overline{|\mathcal{M}_u|^2}=\frac{2(4\pi\alpha)^2}{\mathsf{u}^2}\left(\mathsf{s}^2+\mathsf{t}^2+8m^2\mathsf{u}-8m^4\right).}\tag{A.858}$$

Exchange interference. The remaining contribution in Eq. (A.800) is different from the previous two because it contains one direct and one exchange amplitude. After taking the complex conjugate and performing the four spin sums with Eq. (A.802), the four spinor chains join into a single trace:

$$\boxed{\begin{aligned} &-2\,\mathrm{Re}\,\overline{\mathcal{M}_t\mathcal{M}_u^*} \\ &\quad = -\frac{(4\pi\alpha)^2}{2\mathsf{tu}}\,\mathrm{Tr}\Big[(\not p_3+m)\gamma^\mu(\not p_1+m)\gamma^\nu \\ &\qquad\qquad \times(\not p_4+m)\gamma_\mu(\not p_2+m)\gamma_\nu\Big]. \end{aligned}} \tag{A.859}$$

The difference between this expression and Eq. (A.836) is important. In $|\mathcal{M}_t|^2$ the two electron lines remain separate and therefore produce two traces. In $\mathcal{M}_t\mathcal{M}_u^*$, the exchanged assignment connects the two lines, producing one continuous trace.

Using the gamma-matrix algebra of Eq. (A.703), the trace identities above, and the Mandelstam relations Eqs. (A.850) and (A.740), the interference term reduces to

$$\boxed{-2\,\mathrm{Re}\,\overline{\mathcal{M}_t\mathcal{M}_u^*} = \frac{4(4\pi\alpha)^2}{\mathsf{tu}}\left(\mathsf{s}^2-8m^2\mathsf{s}+12m^4\right).} \tag{A.860}$$

Evaluation of the Møller interference trace.

We start from the trace appearing in Eq. (A.859),

$$\begin{aligned} T \equiv \mathrm{Tr}\Big[&(\not p_3+m)\gamma^\mu(\not p_1+m)\gamma^\nu \\ &\times(\not p_4+m)\gamma_\mu(\not p_2+m)\gamma_\nu\Big]. \end{aligned} \tag{A.861}$$

To make the algebra easier to follow, temporarily write

$$A=p_3,\qquad B=p_1,\qquad C=p_4,\qquad D=p_2. \tag{A.862}$$

Then

$$\begin{aligned} T = \mathrm{Tr}\Big[&(\not A+m)\gamma^\mu(\not B+m)\gamma^\nu \\ &\times(\not C+m)\gamma_\mu(\not D+m)\gamma_\nu\Big]. \end{aligned} \tag{A.863}$$

From the gamma-matrix anticommutation relation, Eq. (A.703), one obtains the useful contractions

$$\boxed{\begin{aligned}\gamma^\mu\gamma^\alpha\gamma_\mu &= -2\gamma^\alpha,\\ \gamma^\mu\gamma^\alpha\gamma^\beta\gamma_\mu &= 4g^{\alpha\beta}\hat{I}_4,\\ \gamma^\mu\gamma^\alpha\gamma^\beta\gamma^\gamma\gamma_\mu &= -2\gamma^\gamma\gamma^\beta\gamma^\alpha.\end{aligned}} \tag{A.864}$$

Using these identities,

$$\begin{aligned}&\gamma^\mu(\not{B}+m)\gamma^\nu(\not{C}+m)\gamma_\mu\\ &\quad= \gamma^\mu\not{B}\gamma^\nu\not{C}\gamma_\mu + m\gamma^\mu\gamma^\nu\not{C}\gamma_\mu\\ &\qquad+ m\gamma^\mu\not{B}\gamma^\nu\gamma_\mu + m^2\gamma^\mu\gamma^\nu\gamma_\mu && \text{(A.865)}\\ &\quad= -2\not{C}\gamma^\nu\not{B} + 4mC^\nu + 4mB^\nu - 2m^2\gamma^\nu && \text{(A.866)}\\ &\quad= \boxed{-2\not{C}\gamma^\nu\not{B} + 4m(B^\nu + C^\nu) - 2m^2\gamma^\nu.} && \text{(A.867)}\end{aligned}$$

Substituting Eq. (A.867) into Eq. (A.861) separates the trace into three contributions,

$$T = T_1 + T_2 + T_3, \tag{A.868}$$

with

$$\begin{aligned}T_1 &= -2\,\mathrm{Tr}\left[(\not{A}+m)\not{C}\gamma^\nu\not{B}(\not{D}+m)\gamma_\nu\right], && \text{(A.869)}\\ T_2 &= 4m(B^\nu + C^\nu)\,\mathrm{Tr}\left[(\not{A}+m)(\not{D}+m)\gamma_\nu\right], && \text{(A.870)}\\ T_3 &= -2m^2\,\mathrm{Tr}\left[(\not{A}+m)\gamma^\nu(\not{D}+m)\gamma_\nu\right]. && \text{(A.871)}\end{aligned}$$

We now evaluate these three terms separately.

For T_1, use Eq. (A.864) to write

$$\begin{aligned}\gamma^\nu\not{B}(\not{D}+m)\gamma_\nu &= \gamma^\nu\not{B}\not{D}\gamma_\nu + m\gamma^\nu\not{B}\gamma_\nu && \text{(A.872)}\\ &= 4(B\cdot D)\hat{I}_4 - 2m\not{B}. && \text{(A.873)}\end{aligned}$$

Hence

$$\begin{aligned}T_1 = &-8(B\cdot D)\,\mathrm{Tr}\left[(\not{A}+m)\not{C}\right]\\ &+ 4m\,\mathrm{Tr}\left[(\not{A}+m)\not{C}\not{B}\right]. \end{aligned}\tag{A.874}$$

Using Eq. (A.838) and the fact that traces containing an odd number of gamma matrices vanish,

$$\mathrm{Tr}(\not{A}\not{C}) = 4A \cdot C, \tag{A.875}$$

and

$$\mathrm{Tr}\left[(\not{A} + m)\not{C}\not{B}\right] = 4m\, B \cdot C. \tag{A.876}$$

Therefore,

$$\boxed{T_1 = -32(A \cdot C)(B \cdot D) + 16m^2(B \cdot C).} \tag{A.877}$$

For T_2,

$$\mathrm{Tr}\left[(\not{A} + m)(\not{D} + m)\gamma_\nu\right] = 4m(A_\nu + D_\nu), \tag{A.878}$$

because the terms containing either one or three gamma matrices have zero trace. Thus

$$\boxed{T_2 = 16m^2(B + C) \cdot (A + D).} \tag{A.879}$$

Finally, for T_3,

$$\begin{aligned}\mathrm{Tr}&\left[(\not{A} + m)\gamma^\nu(\not{D} + m)\gamma_\nu\right] \\ &= \mathrm{Tr}\left[\not{A}\gamma^\nu\not{D}\gamma_\nu\right] + m^2\,\mathrm{Tr}\left[\gamma^\nu\gamma_\nu\right].\end{aligned} \tag{A.880}$$

Using

$$\gamma^\nu\not{D}\gamma_\nu = -2\not{D}, \tag{A.881}$$

we obtain

$$\mathrm{Tr}\left[\not{A}\gamma^\nu\not{D}\gamma_\nu\right] = -8A \cdot D. \tag{A.882}$$

Moreover,

$$\gamma^\nu\gamma_\nu = 4\hat{I}_4, \tag{A.883}$$

so that

$$\mathrm{Tr}\left[\gamma^\nu\gamma_\nu\right] = 16. \tag{A.884}$$

Consequently,

$$\boxed{T_3 = 16m^2(A \cdot D) - 32m^4.} \tag{A.885}$$

Combining Eqs. (A.877), (A.879), and (A.885),

$$\boxed{\begin{aligned} T = & -32(A\cdot C)(B\cdot D) \\ & +16m^2\big[A\cdot B + A\cdot C + A\cdot D \\ & \qquad + B\cdot C + B\cdot D + C\cdot D\big] - 32m^4. \end{aligned}} \tag{A.886}$$

We now return to the physical four-momenta using Eq. (A.862). The Mandelstam definitions, Eq. (A.721), together with the on-shell condition for the four external electrons give

$$\begin{aligned} p_1\cdot p_2 = p_3\cdot p_4 &= \frac{\mathsf{s}-2m^2}{2}, \\ p_1\cdot p_3 = p_2\cdot p_4 &= \frac{2m^2-\mathsf{t}}{2}, \\ p_1\cdot p_4 = p_2\cdot p_3 &= \frac{2m^2-\mathsf{u}}{2}. \end{aligned} \tag{A.887}$$

Therefore,

$$(A\cdot C)(B\cdot D) = \frac{(\mathsf{s}-2m^2)^2}{4}, \tag{A.888}$$

while the sum of all six scalar products appearing in Eq. (A.886) is

$$\begin{aligned} & A\cdot B + A\cdot C + A\cdot D + B\cdot C + B\cdot D + C\cdot D \\ & \quad = (\mathsf{s}-2m^2) + (2m^2-\mathsf{t}) + (2m^2-\mathsf{u}) \qquad \text{(A.889)} \\ & \quad = \mathsf{s} + 2m^2 - \mathsf{t} - \mathsf{u}. \qquad \text{(A.890)} \end{aligned}$$

Using the equal-mass Mandelstam relation, Eq. (A.740),

$$\mathsf{s}+\mathsf{t}+\mathsf{u} = 4m^2, \tag{A.891}$$

we obtain

$$\mathsf{s}+2m^2-\mathsf{t}-\mathsf{u} = 2(\mathsf{s}-m^2). \tag{A.892}$$

Substitution into Eq. (A.886) gives

$$\begin{aligned} T &= -8(\mathsf{s}-2m^2)^2 + 32m^2(\mathsf{s}-m^2) - 32m^4 \qquad \text{(A.893)} \\ &= -8\left(\mathsf{s}^2 - 8m^2\mathsf{s} + 12m^4\right). \qquad \text{(A.894)} \end{aligned}$$

Hence

$$\boxed{T = -8\left(\mathsf{s}^2 - 8m^2\mathsf{s} + 12m^4\right).} \tag{A.895}$$

Finally, substituting this result into Eq. (A.859),

$$\begin{aligned} &-2\,\mathrm{Re}\,\overline{\mathcal{M}_t\mathcal{M}_u^*} \\ &\quad = -\frac{(4\pi\alpha)^2}{2\mathsf{tu}}\left[-8\left(\mathsf{s}^2 - 8m^2\mathsf{s} + 12m^4\right)\right], \end{aligned} \tag{A.896}$$

and therefore

$$\boxed{-2\,\mathrm{Re}\,\overline{\mathcal{M}_t\mathcal{M}_u^*} = \frac{4(4\pi\alpha)^2}{\mathsf{tu}}\left(\mathsf{s}^2 - 8m^2\mathsf{s} + 12m^4\right).} \tag{A.897}$$

This proves Eq. (A.860).

Complete spin-averaged Møller amplitude. Substituting Eqs. (A.855), (A.858), and (A.860) into Eq. (A.800) gives

$$\boxed{\begin{aligned} \overline{|\mathcal{M}_{\mathrm{M}}|^2} = 2(4\pi\alpha)^2\Bigg[&\frac{\mathsf{s}^2 + \mathsf{u}^2 + 8m^2\mathsf{t} - 8m^4}{\mathsf{t}^2} \\ &+ \frac{\mathsf{s}^2 + \mathsf{t}^2 + 8m^2\mathsf{u} - 8m^4}{\mathsf{u}^2} \\ &+ \frac{2\left(\mathsf{s}^2 - 8m^2\mathsf{s} + 12m^4\right)}{\mathsf{tu}}\Bigg]. \end{aligned}} \tag{A.898}$$

The first term is the contribution from the direct scattering alternative, the second is the corresponding contribution from the exchanged alternative, and the third is their quantum-mechanical interference. The latter is the explicit signature, at the level of the scattering probability, of the indistinguishability and fermionic character of the two electrons.

Equation (A.898) also shows why exchange effects are usually much less important for the strongly forward-scattered electrons that dominate conventional electron microscopy. In the forward limit,

$$\mathsf{t} \to 0, \tag{A.899}$$

whereas u remains finite. The direct contribution therefore contains the strongest singular behavior,

$$\overline{|\mathcal{M}_t|^2} \propto \frac{1}{\mathsf{t}^2}, \tag{A.900}$$

while the exchange term remains finite as $1/\mathsf{u}^2$, and the exchange-interference contribution grows only as $1/(\mathsf{tu})$. Thus, sufficiently close to the forward direction, the t-channel dominates.

The same conclusion can be expressed in the laboratory frame. As shown later in Eqs. (A.1021) and (A.1026),

$$\mathsf{t} = -2mW, \qquad \mathsf{u} = -2m(K - W). \tag{A.901}$$

For a small transferred energy,

$$W \ll K, \tag{A.902}$$

we therefore have

$$|\mathsf{t}| \ll |\mathsf{u}|. \tag{A.903}$$

The direct t-channel contribution then dominates, and exchange corrections are correspondingly small.

This should not, however, be interpreted as saying that exchange can always be neglected at cryo-EM energies. In hard electron–electron collisions, where the transferred energy is no longer small compared with K, t and u can become comparable. In particular, near the symmetric energy-sharing limit $W \simeq K/2$, the exchange and interference terms are essential. They must therefore be retained in the rigorous Møller derivation used later for the hard-collision stopping power.

A.14.6 From the invariant amplitude to a cross section

The previous subsection determined the spin-averaged invariant scattering amplitude $\overline{|\mathcal{M}_{\mathrm{M}}|^2}$. We must now connect this quantity with an experimentally measurable cross section.

This step introduces one final piece of relativistic scattering theory: the density of available final states and the incident particle flux. We derive the corresponding two-body cross-section formula here, since it will also clarify the normalization conventions used in the Møller calculation.

Invariant transition probability and phase space. In the covariant normalization used in relativistic quantum field theory, a one-particle momentum eigenstate satisfies

$$\boxed{\langle \mathbf{p}', s' | \mathbf{p}, s \rangle = 2E_{\mathbf{p}}(2\pi)^3 \delta^{(3)}(\mathbf{p}' - \mathbf{p}) \delta_{s's}.} \tag{A.904}$$

For the two-body scattering process

$$1 + 2 \longrightarrow 3 + 4, \tag{A.905}$$

the invariant amplitude $\mathcal{M}$ is defined through the connected part of the S-matrix element,

$$\boxed{\langle f|S|i\rangle_{\text{conn}} = i(2\pi)^4\delta^{(4)}(p_1 + p_2 - p_3 - p_4)\mathcal{M}.} \tag{A.906}$$

The four-dimensional delta function enforces conservation of both energy and three-momentum.

The density of final states is most conveniently written in the Lorentz-invariant form

$$\boxed{d\Phi_2 = (2\pi)^4\delta^{(4)}(p_1 + p_2 - p_3 - p_4)\frac{d^3\mathbf{p}_3}{(2\pi)^3\,2E_3}\frac{d^3\mathbf{p}_4}{(2\pi)^3\,2E_4}.} \tag{A.907}$$

This quantity is called the *two-body Lorentz-invariant phase-space element.*

From the invariant amplitude to the differential cross section.

We start from the definition of the invariant amplitude, Eq. (A.906),

$$\langle f|S|i\rangle_{\text{conn}} = i(2\pi)^4\delta^{(4)}(p_1 + p_2 - p_3 - p_4)\mathcal{M}. \tag{A.908}$$

For compactness, define the four-momentum mismatch

$$P \equiv p_1 + p_2 - p_3 - p_4. \tag{A.909}$$

Then energy–momentum conservation corresponds to

$$P = 0. \tag{A.910}$$

The momentum eigenstates in Eq. (A.904) have the covariant normalization

$$\langle \mathbf{p}', s'|\mathbf{p}, s\rangle = 2E_{\mathbf{p}}(2\pi)^3\delta^{(3)}(\mathbf{p}' - \mathbf{p})\delta_{s's}. \tag{A.911}$$

These states are not normalized to unity. To interpret $|\langle f|S|i\rangle|^2$ as a transition probability, it is useful temporarily to place the system in a large finite volume V.

In a finite normalization volume,

$$(2\pi)^3\delta^{(3)}(\mathbf{0}) \longrightarrow V. \tag{A.912}$$

A one-particle state normalized to unity is therefore

$$\boxed{|\mathbf{p}, s\rangle_V = \frac{1}{\sqrt{2EV}}|\mathbf{p}, s\rangle.} \tag{A.913}$$

Indeed,

$$_V\langle \mathbf{p}, s|\mathbf{p}, s\rangle_V = \frac{1}{2EV}\, 2E(2\pi)^3\delta^{(3)}(\mathbf{0}) \tag{A.914}$$

$$= 1. \tag{A.915}$$

For the process

$$1 + 2 \longrightarrow 3 + 4, \tag{A.916}$$

there are four external one-particle states. Thus the connected S-matrix element between unit-normalized box states is

$$\boxed{S_{fi} = \frac{i(2\pi)^4\delta^{(4)}(P)\mathcal{M}}{\sqrt{(2E_1V)(2E_2V)(2E_3V)(2E_4V)}}.} \tag{A.917}$$

We now square this expression. For fixed initial and final spin states,

$$|S_{fi}|^2 = \frac{|\mathcal{M}|^2}{(2E_1V)(2E_2V)(2E_3V)(2E_4V)}\left[(2\pi)^4\delta^{(4)}(P)\right]^2. \tag{A.918}$$

The square of the Dirac delta in Eq. (A.918) requires some care. A Dirac delta is a distribution, so $\delta^{(4)}(P)^2$ is not interpreted as an ordinary product of functions. Instead, the second delta function is regularized by considering a finite spatial volume V and a finite observation time T.

The four-dimensional delta has the Fourier representation

$$(2\pi)^4\delta^{(4)}(P) = \int d^4x\, e^{iP\cdot x}. \tag{A.919}$$

At $P = 0$, and restricting the integration to the spatial normalization volume V and to an observation time interval of duration T,

$$(2\pi)^4\delta^{(4)}(0) \longrightarrow \int_T dt \int_V d^3\mathbf{x} \tag{A.920}$$

$$= VT. \tag{A.921}$$

Thus, in the regulated large-V, large-T limit,

$$\left[(2\pi)^4\delta^{(4)}(P)\right]^2 = (2\pi)^4\delta^{(4)}(P)\left[(2\pi)^4\delta^{(4)}(0)\right] \tag{A.922}$$

$$\longrightarrow \boxed{(2\pi)^4\delta^{(4)}(P)\,VT.} \tag{A.923}$$

Thus one delta function remains to enforce energy–momentum conservation, while the second produces the space–time normalization factor VT.

We next have to count the available final momentum states. In a finite volume V, the number of one-particle momentum states contained in $d^3\mathbf{p}$ is

$$\boxed{\frac{V\,d^3\mathbf{p}}{(2\pi)^3}.} \tag{A.924}$$

Therefore the number of two-particle final states in $d^3\mathbf{p}_3\,d^3\mathbf{p}_4$ is

$$\frac{V\,d^3\mathbf{p}_3}{(2\pi)^3}\frac{V\,d^3\mathbf{p}_4}{(2\pi)^3}. \tag{A.925}$$

Multiplying Eq. (A.918) by this final-state density and using Eq. (A.923), the differential transition probability becomes

$$\begin{aligned} dP = {} & \frac{\overline{|\mathcal{M}|^2}}{(2E_1V)(2E_2V)(2E_3V)(2E_4V)} \\ & \times (2\pi)^4\delta^{(4)}(P)\,VT\frac{V\,d^3\mathbf{p}_3}{(2\pi)^3}\frac{V\,d^3\mathbf{p}_4}{(2\pi)^3}. \end{aligned} \tag{A.926}$$

Here we have already replaced $|\mathcal{M}|^2$ by its appropriate spin average or sum, denoted by $\overline{|\mathcal{M}|^2}$.

Collecting the powers of V and the energy factors gives

$$\begin{aligned} dP = {} & T\frac{\overline{|\mathcal{M}|^2}}{4E_1E_2V} \\ & \times (2\pi)^4\delta^{(4)}(P)\frac{d^3\mathbf{p}_3}{(2\pi)^3\,2E_3}\frac{d^3\mathbf{p}_4}{(2\pi)^3\,2E_4}. \end{aligned} \tag{A.927}$$

The factor multiplying the amplitude is precisely the two-body Lorentz-invariant phase-space element defined in Eq. (A.907),

$$d\Phi_2 = (2\pi)^4 \delta^{(4)}(P) \frac{d^3\mathbf{p}_3}{(2\pi)^3\, 2E_3} \frac{d^3\mathbf{p}_4}{(2\pi)^3\, 2E_4}. \tag{A.928}$$

Hence

$$\boxed{dP = T \frac{\overline{|\mathcal{M}|^2}}{4E_1E_2V} d\Phi_2.} \tag{A.929}$$

A transition rate is the transition probability accumulated per unit time. Dividing by the observation time T therefore gives

$$\boxed{dR = \frac{dP}{T} = \frac{\overline{|\mathcal{M}|^2}}{4E_1E_2V} d\Phi_2.} \tag{A.930}$$

We have obtained a rate, but a cross section is a rate per unit incident flux. We therefore next determine the flux associated with one incoming particle of type 1 and one incoming particle of type 2 in the normalization volume V.

The Lorentz-invariant quantity

$$\boxed{\mathcal{F} \equiv \sqrt{(p_1 \cdot p_2)^2 - m_1^2 m_2^2}} \tag{A.931}$$

provides the required relativistic flux factor.

Its physical meaning is easiest to see in the rest frame of particle 2. In that frame,

$$p_2 = (m_2, \mathbf{0}), \tag{A.932}$$

so

$$p_1 \cdot p_2 = E_1 m_2. \tag{A.933}$$

Therefore,

$$\mathcal{F} = \sqrt{E_1^2 m_2^2 - m_1^2 m_2^2} \tag{A.934}$$

$$= m_2 \sqrt{E_1^2 - m_1^2} \tag{A.935}$$

$$= m_2 |\mathbf{p}_1|. \tag{A.936}$$

Since $E_2 = m_2$ in this frame,

$$\frac{\mathcal{F}}{E_1 E_2} = \frac{|\mathbf{p}_1|}{E_1} = v_1. \tag{A.937}$$

Thus

$$\boxed{v_{\mathrm{M}} \equiv \frac{\mathcal{F}}{E_1 E_2}} \tag{A.938}$$

reduces to the ordinary relative speed when one of the particles is at rest. It is often called the *Møller velocity*.

For one incident particle of each type normalized in a volume V, the number density of either particle is $1/V$. The incident flux for one pair is therefore

$$\boxed{\Phi_{\mathrm{inc}} = \frac{v_{\mathrm{M}}}{V} = \frac{\mathcal{F}}{E_1 E_2 V}.} \tag{A.939}$$

Finally, the differential cross section is the transition rate divided by the incident flux:

$$d\sigma = \frac{dR}{\Phi_{\mathrm{inc}}} \tag{A.940}$$

$$= \frac{\overline{|\mathcal{M}|^2}}{4E_1 E_2 V} d\Phi_2 \frac{E_1 E_2 V}{\mathcal{F}} \tag{A.941}$$

$$= \boxed{\frac{\overline{|\mathcal{M}|^2}}{4\mathcal{F}} d\Phi_2.} \tag{A.942}$$

All dependence on the arbitrary normalization volume V and observation time T has disappeared, as it must for a physical cross section. Equation (A.942) is the Lorentz-invariant relation between the invariant scattering amplitude and the differential two-particle phase space.

Two-body phase space in the centre-of-momentum frame. We now evaluate Eq. (A.942) in the centre-of-momentum frame. Quantities in this frame will be denoted by an asterisk.

The total incoming four-momentum is

$$p_1 + p_2 = \left(\sqrt{\mathsf{s}}, \mathbf{0}\right), \tag{A.943}$$

because, by definition,

$$\mathsf{s} = (p_1 + p_2)^2. \tag{A.944}$$

Momentum conservation requires the two final particles to have opposite three-momenta,

$$\mathbf{p}_{4,*} = -\mathbf{p}_{3,*}. \tag{A.945}$$

Let

$$p_{\mathrm{f},*} = |\mathbf{p}_{3,*}| = |\mathbf{p}_{4,*}|. \tag{A.946}$$

Two-body phase space in the centre-of-momentum frame.
Starting from Eq. (A.907),

$$\begin{aligned} d\Phi_2 = (2\pi)^4 \delta^{(3)}(\mathbf{p}_{3,*} + \mathbf{p}_{4,*}) \\ \times \delta(\sqrt{\mathsf{s}} - E_{3,*} - E_{4,*}) \frac{d^3\mathbf{p}_{3,*}}{(2\pi)^3\, 2E_{3,*}} \frac{d^3\mathbf{p}_{4,*}}{(2\pi)^3\, 2E_{4,*}}. \end{aligned} \tag{A.947}$$

The three-dimensional delta function allows the $\mathbf{p}_{4,*}$ integration to be performed immediately and imposes

$$\mathbf{p}_{4,*} = -\mathbf{p}_{3,*}. \tag{A.948}$$

Hence

$$d\Phi_2 = \frac{1}{16\pi^2} \frac{d^3\mathbf{p}_{3,*}}{E_{3,*}E_{4,*}} \delta(\sqrt{\mathsf{s}} - E_{3,*} - E_{4,*}). \tag{A.949}$$

Writing the remaining momentum element in spherical coordinates,

$$d^3\mathbf{p}_{3,*} = p_*^2\, dp_*\, d\Omega_*, \tag{A.950}$$

gives

$$d\Phi_2 = \frac{1}{16\pi^2} \frac{p_*^2\, dp_*}{E_{3,*}E_{4,*}} \delta(\sqrt{\mathsf{s}} - E_{3,*} - E_{4,*}) d\Omega_*. \tag{A.951}$$

Define

$$f(p_*) = \sqrt{\mathsf{s}} - E_{3,*}(p_*) - E_{4,*}(p_*). \tag{A.952}$$

Since

$$E_{i,*} = \sqrt{p_*^2 + m_i^2}, \tag{A.953}$$

we have

$$\frac{df}{dp_*} = -\frac{p_*}{E_{3,*}} - \frac{p_*}{E_{4,*}} \tag{A.954}$$

$$= -p_* \frac{E_{3,*} + E_{4,*}}{E_{3,*}E_{4,*}}. \tag{A.955}$$

At the physical solution $p_* = p_{\mathrm{f},*}$, energy conservation gives

$$E_{3,*} + E_{4,*} = \sqrt{\mathsf{s}}, \tag{A.956}$$

so

$$\left|\frac{df}{dp_*}\right|_{p_{\mathrm{f},*}} = \frac{p_{\mathrm{f},*}\sqrt{\mathsf{s}}}{E_{3,*}E_{4,*}}. \tag{A.957}$$

Using

$$\delta(f(p_*)) = \frac{E_{3,*}E_{4,*}}{p_{\mathrm{f},*}\sqrt{\mathsf{s}}}\delta(p_* - p_{\mathrm{f},*}), \tag{A.958}$$

the radial momentum integral in Eq. (A.951) becomes

$$\boxed{d\Phi_2 = \frac{1}{16\pi^2}\frac{p_{\mathrm{f},*}}{\sqrt{\mathsf{s}}}d\Omega_*.} \tag{A.959}$$

In the centre-of-momentum frame the incoming momenta are

$$p_{1,*} = (E_{1,*}, \mathbf{p}_*), \qquad p_{2,*} = (E_{2,*}, -\mathbf{p}_*), \tag{A.960}$$

with

$$p_{\mathrm{i},*} = |\mathbf{p}_*|. \tag{A.961}$$

The invariant flux factor Eq. (A.931) then simplifies considerably.

Invariant flux in the centre-of-momentum frame.

Since

$$p_1 \cdot p_2 = E_{1,*}E_{2,*} + p_{\mathrm{i},*}^2, \tag{A.962}$$

we have

$$\mathcal{F}^2 = \left(E_{1,*}E_{2,*} + p_{\mathrm{i},*}^2\right)^2 - m_1^2 m_2^2. \tag{A.963}$$

Using

$$m_i^2 = E_{i,*}^2 - p_{\mathrm{i},*}^2, \tag{A.964}$$

we obtain

$$\mathcal{F}^2 = p_{\mathrm{i},*}^2 \left(E_{1,*} + E_{2,*}\right)^2 . \tag{A.965}$$

But

$$E_{1,*} + E_{2,*} = \sqrt{\mathsf{s}}, \tag{A.966}$$

and therefore

$$\boxed{\mathcal{F} = p_{\mathrm{i},*}\sqrt{\mathsf{s}}.} \tag{A.967}$$

Substituting Eqs. (A.959) and (A.967) into Eq. (A.942) gives

$$d\sigma = \frac{\overline{|\mathcal{M}|^2}}{4p_{\mathrm{i},*}\sqrt{\mathsf{s}}}\frac{1}{16\pi^2}\frac{p_{\mathrm{f},*}}{\sqrt{\mathsf{s}}}d\Omega_* \tag{A.968}$$

$$= \frac{1}{64\pi^2\mathsf{s}}\frac{p_{\mathrm{f},*}}{p_{\mathrm{i},*}}\overline{|\mathcal{M}|^2}\, d\Omega_*. \tag{A.969}$$

Therefore,

$$\boxed{\frac{d\sigma}{d\Omega_*} = \frac{1}{64\pi^2\mathsf{s}}\frac{p_{\mathrm{f},*}}{p_{\mathrm{i},*}}\overline{|\mathcal{M}|^2}.} \tag{A.970}$$

Specialization to elastic electron–electron scattering. For elastic electron–electron scattering the masses before and after the collision are the same, so the magnitude of the centre-of-momentum three-momentum is unchanged:

$$p_{\mathrm{f},*} = p_{\mathrm{i},*} \equiv p_*. \tag{A.971}$$

Equation (A.970) therefore reduces to

$$\boxed{\frac{d\sigma}{d\Omega_*} = \frac{\overline{|\mathcal{M}_{\mathrm{M}}|^2}}{64\pi^2\mathsf{s}}.} \tag{A.972}$$

We now replace the scattering angle by the invariant momentum-transfer variable t. In the centre-of-momentum frame,

$$p_1 = (E_*, \mathbf{p}_*), \qquad p_3 = (E_*, \mathbf{p}'_*), \tag{A.973}$$

with

$$|\mathbf{p}_*| = |\mathbf{p}'_*| = p_*. \tag{A.974}$$

If θ_* is the angle between the initial and final momenta,

$$\begin{aligned}
\mathsf{t} &= (p_1 - p_3)^2 && \text{(A.975)}\\
&= (E_* - E_*)^2 - |\mathbf{p}_* - \mathbf{p}'_*|^2 && \text{(A.976)}\\
&= -\left(p_*^2 + p_*^2 - 2p_*^2 \cos\theta_*\right) && \text{(A.977)}\\
&= \boxed{-2p_*^2\,(1 - \cos\theta_*)\,.} && \text{(A.978)}
\end{aligned}$$

Differentiating with respect to $\cos\theta_*$,

$$d\mathsf{t} = 2p_*^2 d(\cos\theta_*). \tag{A.979}$$

After integration over the azimuth,

$$d\Omega_* = 2\pi\, d(\cos\theta_*), \tag{A.980}$$

where the integration limits are understood with the corresponding orientation. Hence

$$\boxed{\frac{d\Omega_*}{d\mathsf{t}} = \frac{\pi}{p_*^2}.} \tag{A.981}$$

For two equal masses, each electron in the centre-of-momentum frame has energy

$$E_* = \frac{\sqrt{\mathsf{s}}}{2}. \tag{A.982}$$

Using the on-shell relation $E_*^2 = p_*^2 + m^2$,

$$\begin{aligned}
p_*^2 &= \frac{\mathsf{s}}{4} - m^2 && \text{(A.983)}\\
&= \boxed{\frac{\mathsf{s} - 4m^2}{4}.} && \text{(A.984)}
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{d\sigma}{d\mathsf{t}} &= \frac{d\sigma}{d\Omega_*}\frac{d\Omega_*}{d\mathsf{t}} && \text{(A.985)}\\
&= \frac{\overline{|\mathcal{M}_{\mathrm{M}}|^2}}{64\pi^2\mathsf{s}}\frac{\pi}{p_*^2} && \text{(A.986)}\\
&= \boxed{\frac{\overline{|\mathcal{M}_{\mathrm{M}}|^2}}{16\pi\mathsf{s}(\mathsf{s} - 4m^2)}.} && \text{(A.987)}
\end{aligned}$$

Avoiding double counting for identical final electrons. There is one subtlety specific to Møller scattering. The two final particles are identical electrons. The configurations

$$(p_3, s_3; p_4, s_4) \tag{A.988}$$

and

$$(p_4, s_4; p_3, s_3) \tag{A.989}$$

therefore represent the same physical final state.

If the complete two-particle phase space is integrated while both outgoing electrons are treated as separately labelled particles, each physical final state is counted twice. One may correct this by including the symmetry factor

$$\frac{1}{2!}. \tag{A.990}$$

An equivalent procedure is to integrate over only one of the two equivalent halves of the final-state phase space.

We shall use the latter convention in the stopping-power application. There, W will denote the kinetic energy of the less energetic outgoing electron, so that

$$0 < W \leq \frac{K}{2}. \tag{A.991}$$

This restriction counts every physical Møller final state exactly once. With this convention no additional factor 1/2 is required in Eq. (A.987).

A.14.7 Laboratory frame and transferred energy

For stopping-power applications, the natural reference frame is the laboratory frame in which the target electron is initially at rest.

Let electron 1 be the incident electron and electron 2 the target electron. The incident electron has kinetic energy K. Since, in the natural-unit convention $\hbar = c = 1$, the rest energy of an electron is simply m, its total energy is

$$\boxed{E_1 = m + K.} \tag{A.992}$$

This is simply the definition of kinetic energy,

$$K = E_1 - m. \tag{A.993}$$

The four-momentum of the incident electron is therefore

$$p_1 = (E_1, \mathbf{p}_1) = (m + K, \mathbf{p}_1)\,. \tag{A.994}$$

The target electron is initially at rest, so its three-momentum vanishes:

$$\mathbf{p}_2 = \mathbf{0}. \tag{A.995}$$

Its total energy is then only its rest energy,

$$E_2 = m, \tag{A.996}$$

and hence its four-momentum is

$$\boxed{p_2 = (m, \mathbf{0}).} \tag{A.997}$$

All four external particles are real electrons. They therefore satisfy the mass-shell condition derived in Eq. (A.684),

$$\boxed{p_i^2 = m^2, \qquad i = 1, 2, 3, 4.} \tag{A.998}$$

Thus, whenever the square of the four-momentum of one of the electrons appears below, it can be replaced by m^2.

The invariant s. From the definition of the Mandelstam variables, Eq. (A.721),

$$\mathsf{s} = (p_1 + p_2)^2. \tag{A.999}$$

Expanding the Minkowski square,

$$\mathsf{s} = p_1^2 + p_2^2 + 2p_1 \cdot p_2. \tag{A.1000}$$

The first two terms are immediately determined by the mass-shell condition, Eq. (A.998),

$$p_1^2 = p_2^2 = m^2. \tag{A.1001}$$

It remains to evaluate $p_1 \cdot p_2$. From the Minkowski scalar product defined in Eq. (A.677),

$$p_1 \cdot p_2 = E_1 E_2 - \mathbf{p}_1 \cdot \mathbf{p}_2. \tag{A.1002}$$

But the target electron is initially at rest, so

$$\mathbf{p}_2 = \mathbf{0}, \tag{A.1003}$$

and therefore

$$p_1 \cdot p_2 = E_1 E_2 \tag{A.1004}$$

$$= (m + K)m. \tag{A.1005}$$

Substituting these results into Eq. (A.1000),

$$\mathsf{s} = m^2 + m^2 + 2m(m + K) \tag{A.1006}$$

$$= 2m^2 + 2m^2 + 2mK \tag{A.1007}$$

$$= \boxed{4m^2 + 2mK.} \tag{A.1008}$$

The invariant t. Let W denote the kinetic energy of the outgoing electron labelled 4. Its total energy is therefore

$$\boxed{E_4 = m + W.} \tag{A.1009}$$

Four-momentum conservation, Eq. (A.720), gives

$$p_1 + p_2 = p_3 + p_4. \tag{A.1010}$$

Rearranging,

$$\boxed{p_1 - p_3 = p_4 - p_2.} \tag{A.1011}$$

The Mandelstam variable t is defined by Eq. (A.721) as

$$\mathsf{t} = (p_1 - p_3)^2. \tag{A.1012}$$

Using Eq. (A.1011),

$$\mathsf{t} = (p_4 - p_2)^2. \tag{A.1013}$$

Expanding,

$$\mathsf{t} = p_4^2 + p_2^2 - 2p_4 \cdot p_2. \tag{A.1014}$$

Again, both electrons are on shell, so

$$p_4^2 = p_2^2 = m^2. \tag{A.1015}$$

The scalar product is

$$p_4 \cdot p_2 = E_4 E_2 - \mathbf{p}_4 \cdot \mathbf{p}_2. \tag{A.1016}$$

Since the initial target electron has $\mathbf{p}_2 = \mathbf{0}$,

$$p_4 \cdot p_2 = E_4 E_2 \tag{A.1017}$$

$$= (m + W)m. \tag{A.1018}$$

Therefore,

$$\mathsf{t} = m^2 + m^2 - 2m(m + W) \tag{A.1019}$$

$$= 2m^2 - 2m^2 - 2mW \tag{A.1020}$$

$$= \boxed{-2mW.} \tag{A.1021}$$

The invariant u. For four equal masses, the Mandelstam variables satisfy Eq. (A.740),

$$\mathsf{s} + \mathsf{t} + \mathsf{u} = 4m^2. \tag{A.1022}$$

Hence

$$\mathsf{u} = 4m^2 - \mathsf{s} - \mathsf{t}. \tag{A.1023}$$

Substituting Eqs. (A.1008) and (A.1021),

$$\mathsf{u} = 4m^2 - \left(4m^2 + 2mK\right) - (-2mW) \tag{A.1024}$$

$$= -2mK + 2mW \tag{A.1025}$$

$$= \boxed{-2m(K - W).} \tag{A.1026}$$

Energy conservation gives

$$(m + K) + m = E_3 + (m + W). \tag{A.1027}$$

Therefore,

$$E_3 = m + K - W, \tag{A.1028}$$

so the kinetic energy of electron 3 is

$$\boxed{K_3 = K - W.} \tag{A.1029}$$

Changing variable from t to W. Equation (A.1021) establishes a one-to-one relation between the invariant t and the transferred kinetic energy W,

$$\mathsf{t} = -2mW. \tag{A.1030}$$

Differentiating,

$$d\mathsf{t} = -2m\, dW, \tag{A.1031}$$

and hence

$$\left|\frac{d\mathsf{t}}{dW}\right| = 2m. \tag{A.1032}$$

A differential cross section transforms according to

$$d\sigma = \frac{d\sigma}{d\mathsf{t}}\,|d\mathsf{t}| = \frac{d\sigma}{dW}\,dW. \tag{A.1033}$$

Therefore,

$$\frac{d\sigma}{dW} = \left|\frac{d\mathsf{t}}{dW}\right| \frac{d\sigma}{d\mathsf{t}} \tag{A.1034}$$

$$= 2m\frac{d\sigma}{d\mathsf{t}}. \tag{A.1035}$$

Using Eq. (A.987),

$$\boxed{\frac{d\sigma}{dW} = \frac{m\,\overline{|\mathcal{M}_{\mathrm{M}}|^2}}{8\pi\mathsf{s}(\mathsf{s}-4m^2)}.} \tag{A.1036}$$

Relativistic variables in terms of the kinetic energy. It is convenient to express the remaining relativistic factors through the incident kinetic energy K. Define the dimensionless kinetic energy

$$\boxed{\tau \equiv \frac{K}{m}.} \tag{A.1037}$$

The total incident energy is Eq. (A.992),

$$E_1 = m + K = m(1+\tau). \tag{A.1038}$$

For a relativistic electron,

$$E_1 = \gamma m, \tag{A.1039}$$

so

$$\boxed{\gamma = 1 + \tau.} \tag{A.1040}$$

Using the usual relation between the Lorentz factor and velocity,

$$\gamma = \frac{1}{\sqrt{1-\beta^2}}, \tag{A.1041}$$

we obtain

$$\beta^2 = 1 - \frac{1}{\gamma^2} \tag{A.1042}$$

$$= 1 - \frac{1}{(1+\tau)^2}. \tag{A.1043}$$

Since

$$1 + \tau = \frac{m+K}{m}, \tag{A.1044}$$

this can also be written as

$$\beta^2 = 1 - \frac{m^2}{(m+K)^2} \tag{A.1045}$$

$$= \frac{(m+K)^2 - m^2}{(m+K)^2} \tag{A.1046}$$

$$= \boxed{\frac{K(K+2m)}{(K+m)^2}.} \tag{A.1047}$$

A.14.8 Restoring SI units

The Møller calculation in Secs. A.14.5–A.14.7 was carried out in natural units,

$$\hbar = c = 1. \tag{A.1048}$$

Before using the resulting cross section in the stopping-power calculation, we now restore SI units.

The main point to keep in mind is the meaning of the symbol m. In natural units, mass and energy have the same physical dimension because $c = 1$. Thus the electron mass m appearing, for example, in the mass-shell relation Eq. (A.684) is numerically equal to the electron rest energy.

When SI units are restored, we distinguish the electron rest mass m_0, measured in kilograms, from its rest energy m_0c^2. Therefore, whenever the natural-unit quantity m appears as an energy scale, the corresponding SI quantity is

$$\boxed{m \quad \longrightarrow \quad m_0c^2.} \tag{A.1049}$$

The quantities K and W introduced in Sec. A.14.7 are kinetic energies. They therefore remain energies when SI units are restored; no additional factor of c is required.

For example, the dimensionless kinetic-energy variable defined in Eq. (A.1037),

$$\tau = \frac{K}{m}, \tag{A.1050}$$

becomes

$$\boxed{\tau = \frac{K}{m_0c^2}.} \tag{A.1051}$$

Thus τ is simply the incident kinetic energy expressed in units of the electron rest energy.

The Lorentz factor was related to τ in Eq. (A.1040),

$$\boxed{\gamma = 1 + \tau.} \tag{A.1052}$$

Since both γ and τ are dimensionless, this relation has exactly the same form in SI units.

Likewise,

$$\beta = \frac{v}{c} \tag{A.1053}$$

is dimensionless. Its relation to the kinetic energy was derived in Eq. (A.1047); restoring SI units therefore changes only the energy scale $m \to m_0 c^2$, not the value of β itself.

The natural electromagnetic length scale. The fine-structure constant was introduced in Eq. (A.776) as

$$\boxed{\alpha = \frac{e^2}{4\pi\varepsilon_0 \hbar c}.} \tag{A.1054}$$

Its natural-unit form, Eq. (A.783), is

$$\alpha = \frac{e^2}{4\pi}. \tag{A.1055}$$

Because α is dimensionless, the combination α/m has the same dimensions as $1/m$. In the natural-unit convention, inverse energy is also a length. Restoring $\hbar$ and c converts

$$\frac{1}{m} \tag{A.1056}$$

into

$$\frac{\hbar}{m_0 c}. \tag{A.1057}$$

Therefore,

$$\frac{\alpha}{m} \quad \longrightarrow \quad \frac{\alpha\hbar}{m_0 c}. \tag{A.1058}$$

Using the SI definition of α from Eq. (A.776),

$$\frac{\alpha\hbar}{m_0 c} = \frac{1}{m_0 c}\frac{e^2}{4\pi\varepsilon_0 \hbar c}\hbar \tag{A.1059}$$

$$= \frac{e^2}{4\pi\varepsilon_0 m_0 c^2}. \tag{A.1060}$$

This quantity is the *classical electron radius*,

$$\boxed{r_e \equiv \frac{e^2}{4\pi\varepsilon_0 m_0 c^2} = \frac{\alpha\hbar}{m_0 c}.} \tag{A.1061}$$

Thus, using Eq. (A.1061), the corresponding natural-unit relation is

$$\boxed{r_e = \frac{\alpha}{m}.} \tag{A.1062}$$

Restoring the prefactor of the Møller cross section. The laboratory-frame Møller cross section obtained from Eqs. (A.898), (A.1036), (A.1008), (A.1021), and (A.1026) contains the natural-unit dimensional factor

$$\frac{\alpha^2}{m}. \tag{A.1063}$$

Its conversion to SI units is easiest to see by writing

$$\frac{\alpha^2}{m} = \left(\frac{\alpha}{m}\right)^2 m. \tag{A.1064}$$

Using Eq. (A.1062),

$$\frac{\alpha}{m} = r_e, \tag{A.1065}$$

while the remaining factor m converts to the electron rest energy according to Eq. (A.1049). Therefore,

$$\boxed{\frac{\alpha^2}{m} \quad \longrightarrow \quad r_e^2 m_0 c^2.} \tag{A.1066}$$

This conversion also gives the required physical dimensions. Since

$$[r_e^2 m_0 c^2] = \mathrm{m}^2\,\mathrm{J}, \tag{A.1067}$$

a term proportional to $1/W^2$ has dimensions

$$\left[r_e^2 m_0 c^2 \frac{1}{W^2}\right] = \frac{\mathrm{m}^2}{\mathrm{J}}. \tag{A.1068}$$

This is precisely the dimension required for $d\sigma/dW$, namely cross-sectional area per unit transferred energy.

Møller cross section in SI units. We can now write the Møller energy-transfer cross section entirely in SI units. The laboratory-frame relations between the Mandelstam variables and the kinetic energies are Eqs. (A.1008), (A.1021), and (A.1026). The natural-unit dimensional prefactor is converted to SI units according to Eq. (A.1066). Combining these results gives

$$\boxed{\begin{aligned}\frac{d\sigma_{\mathrm{M}}}{dW} = \frac{2\pi r_e^2 m_0 c^2}{\beta^2}\Bigg[&\frac{1}{W^2} + \frac{1}{(K-W)^2} \\ &+ \frac{\tau^2}{(1+\tau)^2}\frac{1}{K^2} \\ &- \frac{2\tau+1}{(1+\tau)^2}\frac{1}{W(K-W)}\Bigg].\end{aligned}} \tag{A.1069}$$

Here K is the kinetic energy of the incident electron and W is the kinetic energy of the outgoing electron labelled 4, as introduced in Sec. A.14.7. The dimensionless kinetic energy is given in SI units by Eq. (A.1051),

$$\tau = \frac{K}{m_0 c^2}. \tag{A.1070}$$

The Lorentz factor satisfies

$$\gamma = 1 + \tau, \tag{A.1071}$$

as derived in Eq. (A.1040). Likewise, Eq. (A.1047) gives the relation between the dimensionless velocity

$$\beta = \frac{v}{c} \tag{A.1072}$$

and the incident kinetic energy. Restoring the SI energy scale $m \to m_0 c^2$ in that equation gives

$$\boxed{\beta^2 = \frac{K(K+2m_0c^2)}{(K+m_0c^2)^2} = 1 - \frac{1}{(1+\tau)^2}.} \tag{A.1073}$$

The form of Eq. (A.1069) makes the symmetry between the two outgoing electrons explicit. Exchanging the two final electrons corresponds to exchanging their kinetic energies,

$$W \longleftrightarrow K - W. \tag{A.1074}$$

Under this transformation, the first two terms in Eq. (A.1069) are interchanged, while the third and fourth terms remain unchanged. The complete

differential cross section is therefore invariant under exchange of the two final electrons. This symmetry is the direct consequence of their indistinguishability, whose implications for the counting of final states were discussed in Sec. A.14.6.

For the stopping-power calculation, it is convenient to make the leading $1/W^2$ dependence explicit. Factoring $1/W^2$ from Eq. (A.1069) gives

$$\boxed{\frac{d\sigma_{\mathrm{M}}}{dW} = \frac{2\pi r_e^2 m_0 c^2}{\beta^2 W^2}\left[1 + \frac{W^2}{(K-W)^2} + \frac{\tau^2}{(1+\tau)^2}\frac{W^2}{K^2} - \frac{2\tau+1}{(1+\tau)^2}\frac{W}{K-W}\right].} \tag{A.1075}$$

Equations (A.1069) and (A.1075) are algebraically identical. The first makes the exchange symmetry of the two outgoing electrons transparent, whereas the second is more convenient for the energy-weighted integration used in the stopping power.

A.15 Summary of notation and main quantum-mechanical results

This appendix has introduced only the quantum mechanics needed for the image-formation theory developed in this book. The purpose of this final section is to collect the notation and the principal results in one place.

The central idea throughout the appendix has been that quantum mechanics is, mathematically, a theory of vectors, inner products, and operators in a complex Hilbert space. Dirac notation provides a compact way of expressing this structure.

A.15.1 Dictionary of notation

Table A.1 summarizes the notation that will occur most frequently later in the book.

A.15.2 The main results to remember

The mathematical developments of this appendix can be condensed into a small number of principles.

Table A.1: Minimal dictionary of quantum-mechanical notation used in this book (states, representations, and operators).

Notation	**Name**	**Meaning**
$\lvert\psi\rangle$	ket	Abstract quantum state.
$\langle\psi\rvert$	bra	Dual vector associated with $\lvert\psi\rangle$.
$\langle\phi\lvert\psi\rangle$	inner product	Complex number measuring the overlap between two states.
$\langle\psi\lvert\psi\rangle = 1$	normalization	The state is normalized so that the total probability is one.
$\langle\phi\lvert\psi\rangle = 0$	orthogonality	The two states have zero overlap.
$\lvert\mathbf{r}\rangle$	position state	Generalized eigenstate of the position operator with position $\mathbf{r}$.
$\psi(\mathbf{r}) = \langle\mathbf{r}\lvert\psi\rangle$	wave function	Coordinates of the abstract state $\lvert\psi\rangle$ in the position basis.
$\lvert\mathbf{p}\rangle$	momentum state	Generalized eigenstate of the momentum operator with momentum $\mathbf{p}$.
$\widetilde{\psi}(\mathbf{p}) = \langle\mathbf{p}\lvert\psi\rangle$	momentum wave function	Coordinates of the same state in the momentum basis.
$\hat{A}$	operator	Linear map acting on quantum states.
$\hat{A}\lvert u_n\rangle = a_n\lvert u_n\rangle$	eigenvalue equation	$\lvert u_n\rangle$ is an eigenstate of $\hat{A}$ with scalar eigenvalue a_n.
$\langle\psi\lvert\hat{A}\rvert\psi\rangle$	expectation value	Mean value of observable A for an ensemble prepared in $\lvert\psi\rangle$.
$\langle u_m\lvert\hat{A}\rvert u_n\rangle$	matrix element	Coefficient measuring the component produced along $\lvert u_m\rangle$ when $\hat{A}$ acts on $\lvert u_n\rangle$.

1. A quantum state is an abstract vector. The state is written

$$|\psi\rangle. \tag{A.1076}$$

A wave function is not a different physical object, but a representation of this same state in a chosen basis. In the position basis,

$$\boxed{\psi(\mathbf{r}) = \langle \mathbf{r}|\psi\rangle.} \tag{A.1077}$$

Similarly, in the momentum basis,

$$\boxed{\widetilde{\psi}(\mathbf{p}) = \langle \mathbf{p}|\psi\rangle.} \tag{A.1078}$$

The two representations are related by a Fourier transform. With the notation used throughout this book,

$$\boxed{\mathbf{p} = \hbar\mathbf{k} = h\mathbf{R}, \qquad \mathbf{k} = 2\pi\mathbf{R}.} \tag{A.1079}$$

2. Probabilities are squared projection amplitudes. If an observable A has orthonormal eigenstates $|u_n\rangle$,

$$\hat{A}|u_n\rangle = a_n|u_n\rangle, \tag{A.1080}$$

then the amplitude for obtaining the value a_n from a state $|\psi\rangle$ is

$$\langle u_n|\psi\rangle, \tag{A.1081}$$

and the Born rule gives

$$\boxed{P(A = a_n) = |\langle u_n|\psi\rangle|^2.} \tag{A.1082}$$

For position, the corresponding probability density is

$$\boxed{p(\mathbf{r}) = |\psi(\mathbf{r})|^2,} \tag{A.1083}$$

so that the probability of finding the particle in a region Ω is

$$P(\mathbf{r} \in \Omega) = \int_\Omega |\psi(\mathbf{r})|^2 \, d^3r. \tag{A.1084}$$

3. Superposition leads to interference. If several indistinguishable alternatives lead to the same complete final state, their amplitudes must be added before taking the squared modulus:

$$\boxed{P = |A_1 + A_2 + \cdots|^2 .} \tag{A.1085}$$

The cross terms generated by the square are the mathematical origin of interference.

By contrast, if the alternatives become correlated with mutually orthogonal final states of another unobserved system, the corresponding probabilities add:

$$\boxed{P = |A_1|^2 + |A_2|^2 + \cdots .} \tag{A.1086}$$

This distinction is central to the treatment of elastic and inelastic electron scattering.

4. Observables are represented by Hermitian operators. An observable A is represented by an operator $\hat{A}$. Its eigenvalue equation is

$$\boxed{\hat{A}|u_n\rangle = a_n|u_n\rangle.} \tag{A.1087}$$

For a Hermitian operator,

$$\hat{A}^\dagger = \hat{A}, \tag{A.1088}$$

the eigenvalues are real, and eigenstates belonging to distinct eigenvalues are orthogonal.

The possible outcomes of an ideal measurement are the eigenvalues a_n, and their probabilities are given by Eq. (A.1082).

5. Expectation values and uncertainties are statistical properties of the measurement distribution. For a normalized state,

$$\boxed{\langle A\rangle = \langle\psi|\hat{A}|\psi\rangle.} \tag{A.1089}$$

The variance is

$$\boxed{(\Delta A)^2 = \langle\psi|\left(\hat{A} - \langle A\rangle\hat{I}\right)^2|\psi\rangle.} \tag{A.1090}$$

For position and momentum, Fourier duality leads to

$$\boxed{\Delta x\,\Delta p_x \geq \frac{\hbar}{2},} \tag{A.1091}$$

or equivalently, in the spatial-frequency convention of this book,

$$\boxed{\Delta x\,\Delta R_x \geq \frac{1}{4\pi}.} \tag{A.1092}$$

6. The Hamiltonian determines time evolution. The state obeys the Schrödinger equation

$$\boxed{j\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle.} \tag{A.1093}$$

For a time-independent Hamiltonian,

$$\boxed{|\psi(t)\rangle = \exp\left[-\frac{j}{\hbar}\hat{H}(t-t_0)\right]|\psi(t_0)\rangle.} \tag{A.1094}$$

Because a physical Hamiltonian is Hermitian, this evolution is unitary and preserves normalization.

If

$$\hat{H}|u_n\rangle = E_n|u_n\rangle, \tag{A.1095}$$

then

$$|\psi_n(t)\rangle = \exp\left[-\frac{jE_n(t-t_0)}{\hbar}\right]|u_n\rangle. \tag{A.1096}$$

An energy eigenstate therefore changes only by a global phase and is a stationary state.

7. Matrix elements describe how operators couple states. In an orthonormal basis,

$$\boxed{A_{mn} = \langle u_m|\hat{A}|u_n\rangle} \tag{A.1097}$$

is the matrix element of $\hat{A}$.

Equivalently,

$$\boxed{\hat{A}|u_n\rangle = \sum_m \langle u_m|\hat{A}|u_n\rangle|u_m\rangle.} \tag{A.1098}$$

Thus

$$\langle u_m|\hat{A}|u_n\rangle \tag{A.1099}$$

measures the component produced along $|u_m\rangle$ when $\hat{A}$ acts on $|u_n\rangle$.

For an interaction Hamiltonian,

$$\boxed{M_{fi} = \langle u_f|\hat{H}_{\text{int}}|u_i\rangle} \tag{A.1100}$$

is the coupling matrix element between the two states. A nonzero value means that the interaction can transfer amplitude between them. The actual transition probability additionally depends on the time evolution and on the energy difference between the states.

8. Electron–specimen scattering is described by a joint quantum state. If the imaging electron is in state $|\psi\rangle_{\mathrm{e}}$ and the specimen is in internal state $|\Xi_n\rangle_{\mathrm{s}}$, an uncorrelated combined state is written

$$\boxed{|\psi\rangle_{\mathrm{e}} \otimes |\Xi_n\rangle_{\mathrm{s}}.} \tag{A.1101}$$

The interaction Hamiltonian $\hat{H}_{\mathrm{int}}$ acts on the combined electron–specimen state.

Selecting a particular final specimen state gives the effective operator

$$\boxed{\hat{U}_{m0} = {}_{\mathrm{s}}\langle\Xi_m|\hat{H}_{\mathrm{int}}|\Xi_0\rangle_{\mathrm{s}},} \tag{A.1102}$$

which is still an operator acting on the imaging electron.

The two cases of greatest importance are

$$\boxed{\begin{array}{lcl} m = 0 & : & \text{specimen remains in the same internal state: elastic channel,} \\ m \neq 0 & : & \text{specimen is left in a different internal state: inelastic channel.} \end{array}} \tag{A.1103}$$

For the elastic electrostatic interaction,

$$\boxed{\hat{U}_{\mathrm{el}} = {}_{\mathrm{s}}\langle\Xi_0|\hat{H}_{\mathrm{int}}|\Xi_0\rangle_{\mathrm{s}},} \tag{A.1104}$$

and, in position representation,

$$\boxed{[\hat{U}_{\mathrm{el}}\psi](\mathbf{r}) = U_{\mathrm{el}}(\mathbf{r})\psi(\mathbf{r}), \qquad U_{\mathrm{el}}(\mathbf{r}) = -eV(\mathbf{r}).} \tag{A.1105}$$

Thus the familiar electrostatic specimen potential is the coordinate-space representation of the elastic component of the full quantum electron–specimen interaction.

9. The imaging electron is relativistic. The non-relativistic Hamiltonian

$$\frac{\hat{\mathbf{p}}^{\,2}}{2m_0} - eV(\mathbf{r}) \tag{A.1106}$$

is useful for introducing the structure of quantum mechanics, but is not quantitatively adequate for the electron energies used in modern cryo-electron microscopy.

The relativistic description uses the Dirac Hamiltonian,

$$\boxed{\hat{H}_{\mathrm{D}} = c\boldsymbol{\alpha} \cdot \hat{\mathbf{p}} + \beta m_0 c^2 - eV(\mathbf{r}).} \tag{A.1107}$$

The electrostatic interaction term remains $-eV(\mathbf{r})$; what changes fundamentally is the free-electron part of the Hamiltonian and the fact that the state is a four-component spinor.

Under the high-energy, predominantly electrostatic conditions relevant to conventional cryo-electron microscopy, this relativistic description can be reduced to the scalar fast-electron wave equation used in the image-formation theory.

A.15.3 A compact conceptual map

The main logical structure of the appendix can finally be summarized as

$$\boxed{\begin{array}{cc} |\psi\rangle & \hat{A} \\ \text{quantum state} & \text{operator} \\ \Downarrow & \Downarrow \\ \langle u|\psi\rangle & \langle u_m|\hat{A}|u_n\rangle \\ \text{projection amplitude} & \text{matrix element} \\ \Downarrow & \Downarrow \\ |\langle u|\psi\rangle|^2 & \text{coupling between states} \\ \text{probability} & \end{array}} \tag{A.1108}$$

and, for electron microscopy,

$$\boxed{\begin{array}{c} \hat{H}_{\mathrm{int}} \\ \text{electron–specimen interaction} \\ \Downarrow \\ {}_{\mathrm{s}}\langle \Xi_m|\hat{H}_{\mathrm{int}}|\Xi_0\rangle_{\mathrm{s}} = \hat{U}_{m0} \\ \Downarrow \\ \begin{cases} m = 0: & \text{elastic interaction,} \\ m \neq 0: & \text{inelastic coupling.} \end{cases} \end{array}} \tag{A.1109}$$

These ideas contain the quantum-mechanical background required for the image-formation theory developed in the remainder of the book. The detailed connection between the relativistic Dirac equation and the scalar fast-electron wave equation is developed in Sec. 3.4, while their application to the interaction of the imaging electron with the specimen is developed in Chap. 5.

Bibliography

Afonine, P. V., Poon, B. K., Read, R. J., Sobolev, O. V., Terwilliger, T. C., Urzhumtsev, A., and Adams, P. D. (2018). Real-space refinement in PHENIX for cryo-em and crystallography. *Acta Crystallographica Section D: Structural Biology*, 74(6):531–544.

Alvarez, S. (2013). A cartography of the van der waals territories. *Dalton Transactions*, 42:8617–8636.

Axelrod, J. J., Zhang, J. T., Petrov, P. N., Glaeser, R. M., and Müller, H. (2024). Modern approaches to improving phase contrast electron microscopy. *Current Opinion in Structural Biology*, 86:102805.

Baker, L. A., Smith, E. A., Bueler, S. A., and Rubinstein, J. L. (2010). The resolution dependence of optimal exposures in liquid nitrogen temperature electron cryomicroscopy of catalase crystals. *J. Structural Biology*, 169(3):431–437.

Bammes, B. E., Jakana, J., Schmid, M. F., and Chiu, W. (2010). Radiation damage effects at four specimen temperatures from 4 to 100 k. *J. Structural Biology*, 169(3):331–341.

Battaglia, M., Contarato, D., Denes, P., Doering, D., Duden, T., Krieger, B., Giubilato, P., Gnani, D., and Radmilovic, V. (2010). Characterisation of a cmos active pixel sensor for use in the team microscope. *Nuclear Instruments and Methods in Physics Research Section A*, 622(3):669–677.

Battaglia, M., Contarato, D., Denes, P., and Giubilato, P. (2009). Cluster imaging with a direct detection cmos pixel sensor in transmission electron microscopy. *Nuclear Instruments and Methods in Physics Research A*, 608(2):363–365.

Baxter, W. T., Grassucci, R. A., Gao, H., and Frank, J. (2009). Determination of signal-to-noise ratios and spectral snrs in cryo-em low-dose imaging of molecules. *Journal of Structural Biology*, 166(2):126–132.

Berger, M. J., Coursey, J. S., Zucker, M. A., and Chang, J. (2017). ESTAR, PSTAR, and ASTAR: Computer programs for calculating stopping-power and range tables for electrons, protons, and helium ions. NIST Standard Reference Database 124. Version 2.0.1, accessed September 2026.

Bharadwaj, A. and Jakobi, A. J. (2022). Electron scattering properties of biological macromolecules and their use for cryo-em map sharpening. *Faraday Discussions*, 240:168–183.

Boukhayma, A., Peizerat, A., and Enz, C. (2016). Temporal readout noise analysis and reduction techniques for low-light cmos image sensors. *IEEE Transactions on Electron Devices*, 63(1).

Brilot, A. F., Chen, J. Z., Cheng, A., Pan, J., Harrison, S. C., Potter, C. S., Carragher, B., Henderson, R., and Grigorieff, N. (2012). Beam-induced motion of vitrified specimen on holey carbon film. *Journal of Structural Biology*, 177(3):630–637.

Buijsse, B., Trompenaars, P., Altin, V., Danev, R., and Glaeser, R. M. (2020). Spectral dqe of the volta phase plate. *Ultramicroscopy*, 218:113079.

Campbell, M. G., Cheng, A., Brilot, A. F., Moeller, A., Lyumkis, D., Veesler, D., Pan, J., Harrison, S. C., Potter, C. S., Carragher, B., and Grigorieff, N. (2012). Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *Structure*, 20(11):1823–1828.

Cordero, B., Gómez, V., Platero-Prats, A. E., Revés, M., Echeverría, J., Cremades, E., Barragán, F., and Alvarez, S. (2008). Covalent radii revisited. *Dalton Transactions*, pages 2832–2838.

Danev, R. and Baumeister, W. (2016). Cryo-em single particle analysis with the volta phase plate. *eLife*, 5:e13046.

Danev, R., Buijsse, B., Khoshouei, M., Plitzko, J. M., and Baumeister, W. (2014). Volta potential phase plate for in-focus phase contrast transmission electron microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 111(44):15635–15640.

de Jonge, N. (2018). Theory of the spatial resolution of (scanning) transmission electron microscopy in liquid water or ice layers. *Ultramicroscopy*, 187:113–125.

Dickerson, J. L., Lu, P.-H., Hristov, D., Dunin-Borkowski, R. E., and Russo, C. J. (2022). Imaging biological macromolecules in thick specimens: The role of inelastic scattering in cryoem. *Ultramicroscopy*, 237:113510.

Dickerson, J. L., Naydenova, K., Peet, M. J., Wilson, H., Nandy, B., McMullan, G., Morrison, R., and Russo, C. J. (2025). Reducing the effects of radiation damage in cryo-em using liquid helium temperatures. *Proceedings of the National Academy of Sciences*, 122(17):e2421538122.

Dourki, I., Westermeier, F., Schopper, F., Richter, R. H., Andricek, L., Ninkovic, J., Treis, J., Koffmane, C., Wassatsch, A., Peric, I., Epp, S. W., and Miller, R. J. D. (2017). Characterization and optimization of a thin direct electron detector for fast imaging applications. *Journal of Instrumentation*, 12:C03047.

Doyle, P. A. and Turner, P. S. (1968). Relativistic hartree–fock x-ray and electron scattering factors. *Acta Crystallographica Section A*, 24:390–397.

Egerton, R. (2013). Control of radiation damage in the TEM. *Ultramicroscopy*, 127:100–108.

Egerton, R. F. (2011). *Electron Energy-Loss Spectroscopy in the Electron Microscope*. Springer, New York, 3 edition.

Egerton, R. F. (2012). Mechanisms of radiation damage in beam-sensitive specimens, for TEM accelerating voltages between 10 and 300 kv. *Microscopy Research and Technique*, 75(11):1550–1556.

Egerton, R. F. (2025). Two- and three-dimensional electron imaging of beam-sensitive specimens. *Micron*, 194:103819.

Egerton, R. F., Li, P., and Malac, M. (2004). Radiation damage in the TEM and SEM. *Micron*, 35:399–409.

Feja, B. and Aebi, U. (1999). Determination of the inelastic mean free path of electrons in vitrified ice layers for on-line thickness measurements by zero-loss imaging. *Journal of Microscopy*, 193(1):15–19.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394.

Glaeser, R. M. (2016). *Methods in Enzymology. The Resolution Revolution: Recent Advances In cryoEM*, chapter Specimen behavior in the electron beam, pages 19–50. Academic Press.

Goodman, J. W. (2017). *Introduction to Fourier Optics*. WH Freeman, 4th edition.

Grant, T. and Grigorieff, N. (2015). Measuring the optimal exposure for single particle cryo-em using a 2.6 å reconstruction of rotavirus VP6. *eLife*, 4:e06980.

Grassucci, R. A., Taylor, D., and Frank, J. (2008). Visualization of macromolecular complexes using cryo-electron microscopy with FEI Tecnai transmission electron microscopes. *Nature Protocols*, 3(2):330–339.

Grimm, R., Typke, D., Bärmann, M., and Baumeister, W. (1996). Determination of the inelastic mean free path in ice by examination of tilted vesicles and automated most probable loss imaging. *Ultramicroscopy*, 63(3–4):169–179.

Gubbens, A., Barfels, M., Trevor, C., Twesten, R., Mooney, P., Thomas, P., Menon, N., Kraus, B., Mao, C., and McGinn, B. (2010). The GIF Quantum, a next generation post-column imaging energy filter. *Ultramicroscopy*, 110(8):962–970.

Guo, H., Franken, E., Deng, Y., Benlekbir, S., Singla Lezcano, G., Janssen, B., Yu, L., Ripstein, Z. A., Tan, Y. Z., and Rubinstein, J. L. (2020). Electron-event representation data enable efficient cryoem file storage with full preservation of spatial and temporal resolution. *IUCrJ*, 7(5):860–869.

Haider, M., Rose, H., Uhlemann, S., Schwan, E., Kabius, B., and Urban, K. (1998). A spherical-aberration-corrected 200 kv transmission electron microscope. *Ultramicroscopy*, 75(1):53–60.

Hattne, J., Shi, D., Glynn, C., Zee, C.-T., Gallagher-Jones, M., Martynowycz, M. W., Rodriguez, J. A., and Gonen, T. (2018). Analysis of global and site-specific radiation damage in Cryo-EM. *Structure*, 26:759–766.e4.

Hawkes, P. and Kasper, E. (2017). *Principles of Electron Optics: Volume One: Basic Geometrical Optics*. Elsevier.

Hawkes, P. W. (2015). The correction of electron lens aberrations. *Ultramicroscopy*, 156:A1–A64.

Hayward, S. B. and Glaeser, R. M. (1979). Radiation damage of purple membrane at low temperature. *Ultramicroscopy*, 04(2):201–210.

Heide, H.-G. (1984). Observations on ice layers. *Ultramicroscopy*, 14(3):271–278.

Hettler, S., Kano, E., Dries, M., Gerthsen, D., Pfaffmann, L., Bruns, M., Beleggia, M., and Malac, M. (2018a). Charging of carbon thin films in scanning and phase-plate transmission electron microscopy. *Ultramicroscopy*, 184:252–266.

Hettler, S., Kano, E., Dries, M., Gerthsen, D., Pfaffmann, L., Bruns, M., Beleggia, M., and Malac, M. (2018b). Charging of carbon thin films in scanning and phase-plate transmission electron microscopy. *Ultramicroscopy*, 184:252–266.

Heymann, J. B. (2022). The progressive spectral signal-to-noise ratio of cryo-electron micrograph movies as a tool to assess quality and radiation damage. *Computer methods and programs in biomedicine*, 220:106799.

Houchmandzadeh, B. (2020). The hamilton-jacobi equation: an intuitive approach. *American J. Physics*, 88:353–359.

Karia, D., Koh, A. F., Yang, W., et al. (2025). Sub-3 å resolution protein structure determination by single-particle cryo-em at 100 kev. *Structure*, 33(10):1717–1727.e4.

Karuppasamy, M., Karimi Nejadasl, F., Vulovic, M., Koster, A. J., and Ravelli, R. B. G. (2011). Radiation damage in single-particle cryo-electron microscopy: effects of dose and dose rate. *Journal of Synchrotron Radiation*, 18(3):398–412.

Kirkland, E. J. (2010). *Advanced Computing in Electron Microscopy*. Springer, New York, 2 edition.

Konnik, M. and Welsh, J. (2014). High-level numerical simulations of noise in ccd and cmos photosensors: Review and tutorial. *arXiv preprint arXiv:1412.4031*.

Krivanek, O. L., Dellby, N., and Lupini, A. R. (1999). Towards sub-å electron beams. *Ultramicroscopy*, 78(1–4):1–11.

Kuijper, M., van Hoften, G., Janssen, B., Geurink, R., De Carlo, S., Vos, M., van Duinen, G., van Haeringen, B., and Storms, M. (2015). Fei's direct

electron detector developments: Embarking on a revolution in cryo-tem. *Journal of Structural Biology*, 192(2):179–187.

Kulik, M., Chodkiewicz, M. L., and Dominiak, P. M. (2022). Theoretical 3D electron diffraction electrostatic potential maps of proteins modeled with a multipolar pseudoatom data bank. *Acta Crystallographica Section D: Structural Biology*, 78(8):1010–1020.

Kulik, M. and Dominiak, P. M. (2025). Protein electrostatic potential fourier maps calculated using the transferable aspherical atom model and the independent atom model across resolutions. *IUCrJ*, 12(6):616–632.

Kusuhara, F., Wakashima, S., Nasuno, S., Kuroda, R., and Sugawa, S. (2016). Analysis and reduction technologies of floating diffusion capacitance in cmos image sensor for photon-countable sensitivity. *ITE Transactions on Media Technology and Applications*, 4(2):91–98.

Langmore, J. P. and Smith, M. F. (1992). Quantitative energy-filtered electron microscopy of biological molecules in ice. *Ultramicroscopy*, 46:349–373.

Leapman, R. D. and Sun, S. (1995). Cryo-electron energy loss spectroscopy: observations on vitrified hydrated specimens and radiation damage. *Ultramicroscopy*, 59(1-4):71–79.

Li, X., Zheng, S. Q., Egami, K., Agard, D. A., and Cheng, Y. (2013). Influence of electron dose rate on electron counting images recorded with the k2 camera. *Journal of Structural Biology*, 184(2):251–260.

Loane, R. F., Xu, P., and Silcox, J. (1991). Thermal vibrations in convergent-beam electron diffraction. *Acta Crystallographica Section A*, 47(3):267–278.

Lobato, I. and Van Dyck, D. (2014). An accurate parameterization for scattering factors, electron densities and electrostatic potentials for neutral atoms that obey all physical constraints. *Acta Crystallographica Section A: Foundations and Advances*, 70(6):636–649.

Lobato, I. and Van Dyck, D. (2015). Multem: A new multislice program to perform accurate and fast electron diffraction and imaging simulations using graphics processing units with cuda. *Ultramicroscopy*, 156:9–17.

Malac, M., Hettler, S., Hayashida, M., Kano, E., Egerton, R. F., and Beleggia, M. (2021). Phase plates in the transmission electron microscope: operating principles and applications. *Microscopy*, 70(1):75–115.

McMullan, G., Chen, S., Henderson, R., and Faruqi, A. R. (2009a). Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy*, 109(9):1126–1143.

McMullan, G., Faruqi, A. R., Clare, D., and Henderson, R. (2014). Comparison of optimal performance at 300 kev of three direct electron detectors for use in low dose electron microscopy. *Ultramicroscopy*, 147:156–163.

McMullan, G., Faruqi, A. R., Henderson, R., Guerrini, N., Turchetta, R., Jacobs, A., and van Hoften, G. (2009b). Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector. *Ultramicroscopy*, 109(9):1144–1147.

McMullan, G., Vinothkumar, K. R., and Henderson, R. (2015). Thon rings from amorphous ice and implications of beam-induced brownian motion in single particle electron cryo-microscopy. *Ultramicroscopy*, 158:26–32.

Meyer, J. C., Eder, F., Kurasch, S., Skakalova, V., Kotakoski, J., Park, H. J., Roth, S., Chuvilin, A., Eyhusen, S., Benner, G., Krasheninnikov, A. V., and Kaiser, U. (2012). Accurate measurement of electron beam induced displacement cross sections for single-layer graphene. *Physical Review Letters*, 108(19):196102.

Møller, C. (1932). Zur theorie des durchgangs schneller elektronen durch materie. *Annalen der Physik*, 406(5):531–585.

Moreno, M. S., Jorissen, K., and Rehr, J. J. (2007). Practical aspects of electron energy-loss spectroscopy (EELS) calculations using FEFF8. *Micron*, 38(1):1–11.

Muller, D. A., Edwards, B., Kirkland, E. J., and Silcox, J. (2001). Simulation of thermal diffuse scattering including a detailed phonon dispersion curve. *Ultramicroscopy*, 86(3–4):371–380.

Murari, K., Etienne-Cummings, R., Thakor, N., and Cauwenberghs, G. (2009). Which photodiode to use: A comparison of cmos-compatible structures. *IEEE Sensors Journal*, 9(7):752–760.

Naydenova, K., Jia, P., and Russo, C. J. (2020). Cryo-em with sub-1 å specimen movement. *Science*, 370(6513):223–226.

Parkhurst, J. M., Cavalleri, A., Dumoux, M., Basham, M., Clare, D., Siebert, C. A., Evans, G., Naismith, J. H., Kirkland, A., and Essex, J. W. (2024). Computational models of amorphous ice for accurate simulation of cryo-em images of biological samples. *Ultramicroscopy*, 256:113882.

Peet, M. J., Henderson, R., and Russo, C. J. (2019). The energy dependence of contrast and damage in electron cryomicroscopy of biological molecules. *Ultramicroscopy*, 203:125–131.

Peng, L. M., Ren, G., Dudarev, S. L., and Whelan, M. J. (1996). Robust parameterization of elastic and absorptive electron atomic scattering factors. *Acta Crystallographica*, A52:257–276.

Peng, R., Fu, X., Mendez, J. H., Randolph, P. S., Bammes, B. E., and Stagg, S. M. (2023). Characterizing the resolution and throughput of the apollo direct electron detector. *Journal of Structural Biology: X*, 7:100080.

Ravikumar, A., Gopnarayan, M. N., Subramaniam, S., and Srinivasan, N. (2022). Comparison of side-chain dispersion in protein structures determined by cryo-EM and X-ray crystallography. *IUCrJ*, 9(1):98–103.

Reimer, L., editor (1995). *Energy-Filtering Transmission Electron Microscopy*, volume 71 of *Springer Series in Optical Sciences*. Springer.

Rice, W. J., Cheng, A., Noble, A. J., Eng, E. T., Kim, L. Y., Carragher, B., and Potter, C. S. (2018). Routine determination of ice thickness for cryo-EM grids. *Journal of Structural Biology*, 204(1):38–44.

Rohou, A. and Grigorieff, N. (2015). Ctffind4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology*, 192(2):216–221.

Rosenthal, P. B. and Henderson, R. (2003). Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of Molecular Biology*, 333(4):721–745.

Ruderman, D. L. and Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817.

Ruskin, A. I., Yu, Z., and Grigorieff, N. (2013a). Quantitative characterization of electron detectors for transmission electron microscopy. *Journal of Structural Biology*, 184(3):385–393.

Ruskin, A. I., Yu, Z., and Grigorieff, N. (2013b). Quantitative characterization of electron detectors for transmission electron microscopy. *Journal of Structural Biology*, 184(3):385–393.

Russo, C. J. and Henderson, R. (2018a). Charge accumulation in electron cryomicroscopy. *Ultramicroscopy*, 187:43–49.

Russo, C. J. and Henderson, R. (2018b). Ewald sphere correction using a single side-band image processing algorithm. *Ultramicroscopy*, 187:26–33.

Russo, C. J. and Henderson, R. (2018c). Microscopic charge fluctuations cause minimal contrast loss in cryoem. *Ultramicroscopy*, 187:56–63.

Russo, C. J. and Passmore, L. A. (2014). Ultrastable gold substrates for electron cryomicroscopy. *Science*, 346(6215):1377–1380.

Russo, C. J. and Passmore, L. A. (2016). Ultrastable gold substrates: Properties of a support for high-resolution electron cryomicroscopy of biological specimens. *Journal of Structural Biology*, 193(1):33–44.

Scholze, F., Henneken, H., Kuschnerus, P., Rabus, H., Richter, M., and Ulm, G. (2000). Determination of the electron–hole pair creation energy for semiconductors from the spectral responsivity of photodiodes. *Nuclear Instruments and Methods in Physics Research A*, 439(2–3):208–215.

Schreiber, M. T., Maigné, A., Beleggia, M., Shibata, S., and Wolf, M. (2023a). Temporal dynamics of charge buildup in cryo-electron microscopy. *J. Structural Biology: X*, 7:100081.

Schreiber, M. T., Maigné, A., Beleggia, M., Shibata, S., and Wolf, M. (2023b). Temporal dynamics of charge buildup in cryo-electron microscopy. *Journal of Structural Biology: X*, 7:100081.

Schwartz, O., Axelrod, J. J., Campbell, S. L., Turnbaugh, C., Glaeser, R. M., and Müller, H. (2019). Laser phase plate for transmission electron microscopy. *Nature Methods*, 16(10):1016–1020.

Shi, H., Wu, C., and Zhang, X. (2023). Addressing compressive deformation of proteins embedded in crystalline ice. *Structure*, 31(2):213–220.

Shtyrov, A., Wilson, H., Slowik, D., Yamashita, K., Li, J., Wojdyr, M., Chen, S., McMullan, G., Short, J. M., Russo, C. J., Henderson, R., and Murshudov, G. N. (2026). Measurement of atomic scattering factors by cryoelectron microscopy. *Proceedings of the National Academy of Sciences*, 123(19):e2528758123.

Singer, A. (2021). Wilson statistics: derivation, generalization and applications to electron cryomicroscopy. *Acta Crystallographica Section A: Foundations and Advances*, 77(5):472–479.

Sorzano, C. O. S., Jonic, S., Núñez Ramírez, R., Boisset, N., and Carazo, J. M. (2007). Fast, robust and accurate determination of transmission electron microscopy contrast transfer function. *J. Structural Biology*, 160:249–262.

Sorzano, C. O. S., Otero, A., Olmos, E. M., et al. (2009). Error analysis in the determination of the electron microscopical contrast transfer function parameters from experimental power spectra. *BMC Structural Biology*, 9:18.

Sorzano, C. O. S., Vargas, J., Otón, J., Abrishami, V., de la Rosa-Trevín, J. M., del Riego, S., Fernández-Alderete, A., Martínez-Rey, C., Marabini, R., and Carazo, J. M. (2015). Fast and accurate conversion of atomic models into electron density maps. *AIMS Biophysics*, 2(1):8–20.

Sorzano, C. O. S., Vargas, J., Oton, J., Abrishami, V., de la Rosa-Trevin, J. M., Gomez-Blanco, J., Vilas, J. L., Marabini, R., and Carazo, J. M. (2017). A review of resolution measures and related aspects in 3D electron microscopy. *Progress in biophysics and molecular biology*, 124:1–30.

Spieler, H. (2005). *Semiconductor Detector Systems*. Oxford University Press.

Thorne, R. E. (2020). Hypothesis for a mechanism of beam-induced motion in cryo-electron microscopy. *IUCrJ*, 7(3):416–421.

van der Schaaf, A. and van Hateren, J. H. (1996). Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770.

Vilas, J. L., Vargas, J., Martínez, M., Ramírez-Aportela, E., Melero, R., Jiménez-Moreno, A., Garduño, E., Conesa, P., Marabini, R., Maluenda, D., Carazo, J. M., and Sorzano, C. O. S. (2020). Re-examining the spectra of macromolecules. current practice of spectral quasi B-factor flattening. *J. Structural Biology*, 209:107447.

Ważny, G., Jaciuk, M., Indyka, P., Glatt, S., Biela, A., and Rawski, M. (2026). Influence of total electron dose on the quality of nucleic acids potential maps in cryo-em. *Ultramicroscopy*, page 114358.

Weickenmeier, A. and Kohl, H. (1991). Computation of absorptive form factors for high-energy electron diffraction. *Acta Crystallographica Section A*, 47:590–597.

Yesibolati, M. N., Laganà, S., Kadkhodazadeh, S., Mikkelsen, E. K., Sun, H., Kasama, T., Hansen, O., Zaluzec, N. J., and Mølhave, K. (2020). Electron inelastic mean free path in water. *Nanoscale*, 12:20649–20657.

Yonekura, K. and Maki-Yonekura, S. (2016). Refinement of cryo-EM structures using scattering factors of charged atoms. *Journal of Applied Crystallography*, 49(5):1517–1523.

Yonekura, K., Matsuoka, R., Yamashita, Y., Yamane, T., Ikeguchi, M., Kidera, A., and Maki-Yonekura, S. (2018). Ionic scattering factors of atoms that compose biological molecules. *IUCrJ*, 5(3):348–353.

Yu, Y., Spoth, K. A., Colletta, M., et al. (2025). Dose-efficient cryo-electron microscopy for thick samples using tilt-corrected scanning transmission electron microscopy. *Nature Methods*, 22:2138–2148.

Zeng, X., Stahlberg, H., and Grigorieff, N. (2007). A maximum likelihood approach to two-dimensional crystals. *Journal of Structural Biology*, 160(3):362–374.

Zhang, Y., van Schayck, J. P., Pedrazo-Tardajos, A., Claes, N., Noteborn, W. E. M., Lu, P.-H., Duimel, H., Dunin-Borkowski, R. E., Bals, S., Peters, P. J., and Ravelli, R. B. G. (2023). Charging of vitreous samples in cryogenic electron microscopy mitigated by graphene. *ACS Nano*, 17(16):15836–15846.

Zhang, Z., Lobato, I., Brown, H., Lamoen, D., Jannis, D., Verbeeck, J., Van Aert, S., and Nellist, P. D. (2025). Relativistic eels scattering cross-sections for microanalysis based on dirac solutions. *Ultramicroscopy*, 269:114083.

Zheng, S. Q., Palovcak, E., Armache, J.-P., Verba, K. A., Cheng, Y., and Agard, D. A. (2017). Motioncor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods*, 14:331–332.

Zivanov, J., Nakane, T., and Scheres, S. H. W. (2020). Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in Relion-3.1. *IUCrJ*, 7:253–267.