INVERSE OPTICAL DESIGN AND ITS APPLICATIONS

by

Julia Angela Sakamoto

Copyright © Julia Angela Sakamoto 2012

A Dissertation Submitted to the Faculty of the

COLLEGE OF OPTICAL SCIENCES

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2012

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Julia A. Sakamoto entitle Inverse Optical Design and Its Applications and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

Harrison H. Barrett Date: 1/6/2012

Russell A. Chipman Date: 1/6/2012

Eric W. Clarkson Date: 1/6/2012

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director: Harrison H. Barrett Date: 1/6/2012

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Request for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: _____
Julia Angela Sakamoto

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all family and friends, near and far, who have offered encouragement and support, and a variety of memorable experiences throughout these past years.

Thank you, Mom, for doing anything and everything to ensure my success and well-being throughout my academic life. No one is more supportive and selfless as you. Dad, I immensely value your advice, guidance, and words of wisdom. You are a model person in my life and have my utmost admiration and respect. The old adage, “I could not have done it without (both of) you,” is absolutely fitting.

Kenneth, I am so lucky to have found you in this phase of my life and look forward to beginning the next one together. Thank you for your unwavering support and devotion, kind and thoughtful spirit, and our treasured conversations during this whole process. You are a gem.

Harry, you have been such a wonderful teacher and mentor, and a terrific role-model. I have fond memories of the chalkboard brainstorming sessions and learning so many fascinating things from you. You have taught me more than mere skills -- you have refined my thinking and expanded my mind. I am very grateful for the many opportunities you have provided over the years.

Thank you also to Dr. Pui Lam for those countless, invaluable problem-solving sessions, meaningful conversations, and your supreme dedication as an educator. You are one of those special teachers who truly make a lifelong impact. And thank you for pushing me out of the nest. ☺

This work was supported by Science Foundation Ireland under grant no. 01/PI.2/B039C and an E.T.S. Walton Fellowship for H. H. Barrett. Development of the basic methodology for parameter estimation was supported in part by the National Institutes of Health under grant numbers R37 EB000803 and P41 EB002035. Further support was received through the Biomedical Imaging and Spectroscopy (BMIS) and Technology and Research Initiative Fund (TRIF) fellowship programs at the University of Arizona, as well as Canon, Inc. Much appreciation to Robin Richards, Eugene Cochran, and Amy Phillips at the Office of Technology Transfer for your instrumental help in patenting Inverse Optical Design.

DEDICATION

To Mom, Dad, Christina, Grandpa, Kenneth, Casey, Aiko, and Mia
for your boundless love and support.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	12
LIST OF TABLES	23
ABSTRACT	26
CHAPTER 1. INTRODUCTION	28
1.1. Application to vision science and ophthalmology	31
1.2. Application to optical shop testing	35
1.3. Dissertation overview.....	42
CHAPTER 2. MAXIMUM-LIKELIHOOD ESTIMATION	45
2.1 Historical background	46
2.2 Statement of the problem	49
2.3 Notation system and terminology.....	51
2.4 Fisher information	55
2.4.1 <i>Score</i>	55
2.4.2 <i>Fisher information matrix</i>	56
2.4.3 <i>Cramér-Rao inequality</i>	57
2.4.4 <i>System design</i>	64
2.5 Properties of ML estimators	64
2.5.1 <i>Bias</i>	65
2.5.2 <i>Variance and covariance</i>	67
2.5.3 <i>Mean-square error</i>	68

TABLE OF CONTENTS – *Continued*

	Page
2.5.4 <i>Asymptotic properties</i>	68
2.5.5 <i>Invariance</i>	71
2.5.6 <i>Sufficiency</i>	73
2.6 Computer simulated experiments	76
2.7 Nuisance parameters	78
2.8 Gaussian distributions and electronic noise	81
2.9 Practical challenges	86
CHAPTER 3. OPTIMIZATION METHODS	88
3.1 Selecting a search algorithm	89
3.2. Global optimization algorithms	91
3.3. Simulated annealing	93
3.3.1 <i>Overview</i>	93
3.3.2. <i>Basic concepts in statistical mechanics</i>	97
3.3.3 <i>The Metropolis algorithm</i>	103
3.3.4. <i>Continuous minimization by simulated annealing</i>	104
CHAPTER 4. PROPAGATION OF LIGHT	118
4.1 The electromagnetic field	119
4.1.1 <i>Maxwell's equations</i>	120
4.1.2 <i>Constitutive relations</i>	122
4.1.3 <i>Time-dependent wave equation</i>	125

TABLE OF CONTENTS – *Continued*

	Page
4.1.4 <i>Time-independent wave equation</i>	129
4.2 Plane waves and spherical waves	131
4.2.1 <i>Plane waves</i>	132
4.2.2 <i>Spherical waves</i>	134
4.3 Geometrical optics	135
4.3.1 <i>The eikonal equation</i>	135
4.3.2 <i>Differential equation of light rays</i>	140
4.3.3 <i>Refraction and reflection</i>	142
4.4 Diffraction by a planar aperture	144
4.4.1 <i>A brief history of diffraction theory</i>	144
4.4.2 <i>Geometry of the problem</i>	146
4.4.3 <i>Huygens' principle</i>	148
4.4.4 <i>Fresnel diffraction</i>	150
4.4.5 <i>Fraunhofer diffraction</i>	153
CHAPTER 5. INVERSE OPTICAL DESIGN OF THE HUMAN EYE USING	
LIKELIHOOD METHODS AND WAVEFRONT SENSING	155
5.1 Basic anatomy of the human eye	156
5.2 Ray-tracing through a schematic eye	160
5.3 Shack-Hartmann wavefront sensors	172
5.3.1 <i>Centroid estimation and Fisher information</i>	176

TABLE OF CONTENTS – *Continued*

	Page
5.4 Data-acquisition system	179
5.4.1 <i>System configuration</i>	179
5.4.2 <i>Optical-design program</i>	181
5.5 Fisher information and Cramér-Rao lower bounds	189
5.6 Likelihood surfaces	199
5.7 Maximum-likelihood estimation of ocular parameters	213
5.8 Summary of Chapter 5	218
CHAPTER 6. MAXIMUM-LIKELIHOOD ESTIMATION OF	
PARAMETERIZED WAVEFRONTS USING MULTIFOCAL DATA.....	222
6.1 Formulation of the problem	222
6.2 Propagation algorithm	224
6.2.1 <i>Diffraction propagation vs. ray-tracing</i>	224
6.2.2 <i>Diffraction equation for a converging spherical wave</i>	226
6.2.3 <i>Parameterized wavefront description</i>	229
6.2.4 <i>Sampling considerations</i>	230
6.2.5 <i>Parallel processing with the graphics processing unit</i>	231
6.3 Numerical studies	234
6.3.1 <i>Test lens description</i>	234
6.3.2 <i>Pupil sampling</i>	238
6.3.3 <i>Fisher information and Cramér-Rao lower bounds</i>	242

TABLE OF CONTENTS – *Continued*

	Page
6.3.4 <i>Likelihood surfaces</i>	249
6.3.5 <i>Maximum-likelihood estimates</i>	255
6.4 Experimental results	261
6.4.1 <i>System configuration</i>	261
6.4.2 <i>Test lens description</i>	263
6.4.3 <i>Experimental data</i>	266
6.4.4 <i>Pupil sampling</i>	267
6.4.5 <i>Huygens' method vs. Fresnel propagation</i>	269
6.4.6 <i>Fisher information and Cramér-Rao lower bounds</i>	273
6.4.7 <i>Likelihood surfaces</i>	275
6.4.8 <i>Nuisance parameters</i>	278
6.4.9 <i>Maximum-likelihood estimates</i>	281
6.5 Summary of Chapter 6	285
CHAPTER 7. INVERSE OPTICAL DESIGN FOR OPTICAL TESTING.....	289
7.1 Inverse optical design of aspheric lenses	289
7.1.1 <i>Optical-design program</i>	290
7.1.2 <i>Test lens description and system configuration</i>	295
7.1.3 <i>Fisher information and Cramér-Rao bounds</i>	301
7.1.4 <i>Likelihood surfaces</i>	302
7.1.5 <i>Maximum-likelihood estimates</i>	307

TABLE OF CONTENTS – *Continued*

	Page
7.2 Inverse optical design of GRIN-rod lenses	310
7.2.1 Ray-tracing through a GRIN-rod lens	310
7.2.2 Test lens description	318
7.2.3 Fisher information and Cramer-Rao bounds	321
7.2.4 Likelihood surfaces	322
7.3 Summary of Chapter 7	324
CHAPTER 8. CONCLUSION AND FUTURE WORK	326
APPENDIX A. FRINGE ZERNIKE POLYNOMIALS	331
APPENDIX B. LIST OF ACRONYMS	336
REFERENCES	338

LIST OF FIGURES

Figure	Page
1.1 System configuration for estimating patient-specific ocular parameters, based on a clinical Shack-Hartmann aberrometer for measurement of aberrations	29
1.2 Anisoplanatism involving pupil aberrations. Image from Stéphane Chamot (National University of Ireland, Galway)	32
1.3 Basic test configuration for performing inverse optical design of a GRIN-rod lens	36
1.4 Basic system configuration for parameterized wavefront measurement with a single source	37
1.5 Basic system configuration for parameterized wavefront measurement with an aspheric test element and multiple point source locations	39
1.6 Basic system configuration for augmented wavefront measurement with a Shack-Hartmann WFS and multiple point source locations	41
2.1 Example of a probability distribution of $\hat{\theta}$ conditioned on θ	66
3.1 A multimodal test function in two dimensions, exhibiting a high degree of nonlinearity and various local minima	91
3.2 Illustration of the travelling salesman problem and its solution	97
3.3 The simulated annealing algorithm implemented by Corana <i>et al.</i> (1987)	106
4.1 Geometry for diffraction by a planar aperture (Barrett & Myers, 2004)	147
5.1 Basic anatomy of the human eye, as seen through a cross-sectional view	157

LIST OF FIGURES - *Continued*

Figure	Page
5.2 Geometrical eye model corresponding to parameters in Table 5.2, with an on-axis source and 8-mm pupil to demonstrate spherical aberration	172
5.3 Shack-Hartmann WFS measuring a perfect incoming wavefront	174
5.4 Shack-Hartmann WFS measuring an aberrated incoming wavefront	174
5.5 Blurred spot profiles in the focal plane of a Shack-Hartmann WFS	177
5.6 Data acquisition system for estimating ocular parameters	181
5.7 Geometrical eye model used to generate WFS data, corresponding to ocular parameters in Table 5.2	183
5.8 WFS data used as input to inverse optical design, for beam angle $\alpha = 0^\circ$	184
5.9 WFS data used as input to inverse optical design, for beam angle $\alpha = 6^\circ$	185
5.10 WFS data used as input to inverse optical design, for beam angle $\alpha = 12^\circ$	186
5.11 Focal spot on the retina for a source beam angle of $\alpha = 0^\circ$	187
5.12 Focal spot on the retina for a source beam angle of $\alpha = 6^\circ$	188
5.13 Focal spot on the retina for a source beam angle of $\alpha = 12^\circ$	188
5.14 FIM for the chosen system configuration (log scale)	191
5.15 Inverse of the FIM for the chosen system configuration (log scale)	192
5.16 FIM for the system after increasing the detector element size	194
5.17 Inverse of the FIM after increasing the detector element size	195
5.18 Detector data for $\alpha = 0^\circ$ after increasing the beam and pupil diameters	196

LIST OF FIGURES - *Continued*

Figure	Page
5.19 FIM for the system after increasing the beam and pupil diameters	197
5.20 Inverse of the FIM after increasing the beam and pupil diameters	197
5.21 FIM for the system after reducing the number of beam angles	198
5.22 Inverse of the FIM after reducing the number of beam angles	199
5.23 Likelihood surface along $R_{\text{cornea,posterior}}$ and $R_{\text{lens,anterior}}$ axes. Final ML estimates indicated by $\times$ sign	201
5.24 Likelihood surface along $R_{\text{cornea,posterior}}$ and $R_{\text{lens,posterior}}$ axes. Final ML estimates indicated by $\times$ sign	202
5.25 Likelihood surface along $R_{\text{cornea,posterior}}$ and Δt_{cornea} axes. Final ML estimates indicated by $\times$ sign	202
5.26 Likelihood surface along $R_{\text{cornea,posterior}}$ and $\Delta t_{\text{ant.chamber}}$ axes. Final ML estimates indicated by $\times$ sign	203
5.27 Likelihood surface along $R_{\text{cornea,posterior}}$ and Δt_{lens} axes. Final ML estimates indicated by $\times$ sign	203
5.28 Likelihood surface along $R_{\text{cornea,posterior}}$ and $\Delta t_{\text{vitreous}}$ axes. Final ML estimates indicated by $\times$ sign	204
5.29 Likelihood surface along $R_{\text{cornea,posterior}}$ and n_{cornea} axes. Final ML estimates indicated by $\times$ sign	204

LIST OF FIGURES - *Continued*

Figure	Page
5.30 Likelihood surface along $R_{\text{cornea,posterior}}$ and $n_{\text{ant.chamber}}$ axes. Final ML estimates indicated by $\times$ sign	205
5.31 Likelihood surface along $R_{\text{cornea,posterior}}$ and n_{lens} axes. Final ML estimates indicated by $\times$ sign	205
5.32 Likelihood surface along $R_{\text{cornea,posterior}}$ and n_{vitreous} axes. Final ML estimates indicated by $\times$ sign	206
5.33 Likelihood surface along Δt_{cornea} and n_{cornea} axes. Final ML estimates indicated by $\times$ sign	207
5.34 Likelihood surface along $\Delta t_{\text{ant.chamber}}$ and $n_{\text{ant.chamber}}$ axes. Final ML estimates indicated by $\times$ sign	207
5.35 Likelihood surface along Δt_{lens} and n_{lens} axes. Final ML estimates indicated by $\times$ sign	208
5.36 Likelihood surface along $\Delta t_{\text{vitreous}}$ and n_{vitreous} axes. Final ML estimates indicated by $\times$ sign	208
5.37 Understanding the likelihood as a function of defocus. P1 corresponds to the true minimum and a myopic eye (focuses before retina); P3, P4, and P5 are high points and correspond to zero defocus; P2 corresponds to a hyperopic eye (focuses behind retina)	210
5.38 Level of defocus at P1 ($R_{\text{cornea,posterior}} = 6.381$ mm, $n_{\text{vitreous}} = 16.40$ mm)	211

LIST OF FIGURES - *Continued*

Figure	Page
5.39 Level of defocus at P2 ($R_{\text{cornea,posterior}} = 6.000$ mm, $n_{\text{vitreous}} = 15.50$ mm)	211
5.40 Level of defocus at P3 ($R_{\text{cornea,posterior}} = 6.188$ mm, $n_{\text{vitreous}} = 15.97$ mm)	212
5.41 Level of defocus at P4 ($R_{\text{cornea,posterior}} = 6.512$ mm, $n_{\text{vitreous}} = 15.85$ mm)	212
5.42 Level of defocus at P5 ($R_{\text{cornea,posterior}} = 5.871$ mm, $n_{\text{vitreous}} = 16.09$ mm)	213
5.43 16 simulated annealing trials for the estimation of ocular parameters	215
5.44 Reconstructed eye model of the estimated parameters, superimposed with the true values underlying the data	218
6.1 Data-acquisition system for collecting multiple irradiance patterns near the focus of an optical element	224
6.2 Focal region of the highly aberrated test lens. Paraxial focal plane is at $z = z_f = 157.8$ mm	236
6.3 Wavefront error in the exit pupil of the highly aberrated test lens as a function of normalized radius. Units are in waves	238
6.4 Detector data at $z = z_1$ for the highly aberrated test lens using a pupil sampling of: (a) $P = 1024$, (b) $P = 512$, and (c) $P = 256$	239
6.5 Detector data at $z = z_2$ for the highly aberrated test lens using a pupil sampling of: (a) $P = 1024$, (b) $P = 512$, and (c) $P = 256$	240
6.6 Detector data for the highly aberrated test lens using a pupil sampling of $P = 1024$ at image plane: (a) $z = z_1$ and (b) $z = z_2$	241

LIST OF FIGURES - *Continued*

Figure	Page
6.7 FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 37\}$ in the exit pupil of the highly aberrated test lens (log scale)	243
6.8 Inverse of the FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 37\}$ in the exit pupil of the highly aberrated test lens (log scale)	245
6.9 FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of The highly aberrated test lens (log scale)	247
6.10 Inverse of the FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the highly aberrated test lens (log scale)	248
6.11 Likelihood surface along the α_4 (defocus) and α_9 (primary spherical aberration) axes for the highly aberrated test lens	250
6.12 Likelihood surface along the α_4 (defocus) and α_{25} (tertiary spherical aberration) axes for the highly aberrated test lens	251
6.13 Likelihood surface along the α_9 (primary spherical aberration) and α_{16} (secondary spherical aberration) axes for the highly aberrated test lens	251
6.14 Likelihood surface along the α_{16} (secondary spherical aberration) and α_{25} (tertiary spherical aberration) axes for the highly aberrated test lens	252
6.15 Likelihood surface along the α_4 (defocus) and α_5 (primary astigmatism at 0°) axes for the highly aberrated test lens	253

LIST OF FIGURES - *Continued*

Figure	Page
6.16 Likelihood surface along the α_{25} (tertiary spherical aberration) and α_7 (primary coma, x-axis) axes for the highly aberrated test lens	253
6.17 Likelihood surface along the α_2 (tilt, y-axis) and α_3 (tilt, x-axis) axes for the highly aberrated test lens	254
6.18 Likelihood surface along the α_5 (primary astigmatism at 0°) and α_7 (primary coma, x-axis) axes for the highly aberrated test lens	254
6.19 Likelihood surface along the α_3 (tilt, y-axis) and α_8 (primary coma, y-axis) axes for the highly aberrated test lens	255
6.20 12 simulated annealing trials for the estimation of wavefront parameters in the exit pupil of the highly aberrated test lens (log-log scale)	257
6.21 Comparison between the true data and estimated irradiance patterns for the highly aberrated test lens	260
6.22 Data-acquisition system for collecting multiple irradiance patterns near the focus of a spherical test lens, including a movable imaging lens	262
6.23 Focal region of the spherical test lens. Paraxial focal plane is at $z = z_f = 90.83$ mm	264
6.24 Theoretical wavefront error in the exit pupil of the spherical lens as a function of normalized radius. Units are in waves	266

LIST OF FIGURES - *Continued*

Figure	Page
6.25 Experimental data for the spherical test lens for image planes: (a) $z = z_1$ and (b) $z = z_2$. Scale bar corresponds to the intermediate image plane just before the imaging lens	267
6.26 Detector data at $z = z_1$ for spherical lens using pupil sampling of: (a) $P = 512$, (b) $P = 256$, and (c) $P = 128$	268
6.27 Detector data at $z = z_2$ for spherical lens using pupil sampling of: (a) $P = 512$, (b) $P = 256$, and (c) $P = 128$	268
6.28 Irradiance data at $z = z_1$ for the spherical lens: (a) Fresnel approximation, (b) Huygens integral, (c) difference	271
6.29 Irradiance data at $z = z_2$ for the spherical lens: (a) Fresnel approximation, (b) Huygens integral, (c) difference	272
6.30 FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the spherical test lens (log scale)	273
6.31 Inverse of the FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the spherical test lens (log scale)	274
6.32 Likelihood surface along α_4 (defocus) and α_{16} (secondary spherical aberration) axes for the spherical test lens	276
6.33 Likelihood surface along α_7 (primary coma, x-axis) and α_9 (primary spherical aberration) axes for the spherical test lens	276

LIST OF FIGURES - *Continued*

Figure	Page
6.34 Likelihood surface along α_2 (tilt, x-axis) and α_4 (defocus) axes for the spherical test lens	277
6.35 Likelihood surface along α_5 (primary astigmatism at 0°) and α_{16} (secondary spherical aberration) axes for the spherical test lens	277
6.36 Likelihood surface along α_2 (tilt, x-axis) and α_7 (primary coma, x-axis) axes for the spherical test lens	278
6.37 Determining the nuisance parameters in the system for image plane $z = z_1$ via a 2D grid search prior to the estimation of wavefront parameters	280
6.38 Determining the nuisance parameters in the system for image plane $z = z_2$ via a 2D grid search prior to the estimation of wavefront parameters	281
6.39 12 simulated annealing trials for the estimation of wavefront parameters in the exit pupil of the spherical test lens	282
6.40 Comparison between the true data and estimated irradiance patterns for the spherical test lens	285
6.41 Data-acquisition system for collecting multiple irradiance patterns near the focus of an optical element, including a movable diffuser and imaging lens.....	287
7.1 Ray-trace data from our CUDA algorithm for the precision asphere	297
7.2 Ray-trace data computed by ZEMAX TM for the precision asphere	297

LIST OF FIGURES - *Continued*

Figure	Page
7.3 Irradiance data computed at: (a) $z = 95$ mm after lens for on-axis source, (b) $z = 100$ mm for same on-axis source, and (c) $z = 90$ mm for off-axis source	299
7.4 Irradiance data computed with ZEMAX TM at: (a) $z = 95$ mm after lens for on-axis source, (b) $z = 100$ mm for same on-axis source, and (c) $z = 90$ mm for off-axis source	300
7.5 (a) FIM and (b) inverse of the FIM for prescription parameters describing the precision asphere (logarithmic scale).....	302
7.6 Likelihood surface along R_C and κ axes. Global minimum is located at center of plot	304
7.7 Likelihood surface along R_C and α_4 axes. Global minimum is located at center of plot. (logarithmic scale)	304
7.8 Likelihood surface along R_C and α_6 axes. Global minimum is located at center of plot. (logarithmic scale)	305
7.9 Likelihood surface along κ and α_4 axes. Global minimum is located at center of plot	305
7.10 Likelihood surface along κ and α_6 axes. Global minimum is located at center of plot	306

LIST OF FIGURES - *Continued*

Figure	Page
7.11 Likelihood surface along κ and α_6 axes. Global minimum is located at center of plot	306
7.12 20 simulated annealing trials for the estimation of prescription parameters describing the precision asphere	308
7.13 Refractive index distribution of the GRIN-rod lens	319
7.14 Real eikonal rays traced through the GRIN-rod lens. Plot is expanded in the transverse direction to show detail	320
7.15 (a) Irradiance distribution in the detector plane and (b) irradiance profile for the GRIN-rod test lens	320
7.16 (a) FIM and (b) inverse of the FIM for the parameters describing the refractive index distribution of the GRIN-rod lens (logarithmic scale).....	321
7.17 Likelihood surface along n_0 and g axes. Global minimum is located at center of plot	323
7.18 Likelihood surface along n_0 and h_4 axes. Global minimum is located at center of plot.....	323
7.19 Likelihood surface along g and h_4 axes. Global minimum is located at center of plot.....	324
A.1 Fringe Zernike Polynomials 2-37.....	335

LIST OF TABLES

Table	Page
5.1 Navarro wide-angle schematic eye model at $\lambda = 780$ nm	162
5.2 Geometry of eye model used to generate WFS data	163
5.3 Square-root of the CRB (standard deviation) for various system configurations	193
5.4 Estimated ocular parameters, including the true values, starting point in the search, upper and lower limits in the search space, and estimated values with standard deviations	217
6.1 Product specifications for NVIDIA® Tesla™ C1060 and C2075 models	233
6.2 System data provided by ZEMAX™ for the highly aberrated test lens at $\lambda = 0.6328$ μm	235
6.3 Fringe Zernike coefficients $\{\alpha_n, n = 1, \dots, 37\}$, peak-to-valley, RMS, and variance, provided by ZEMAX™ for the highly aberrated test lens. Unlisted coefficients are zero	237
6.4 Square-root of the CRB for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 37\}$ in the exit pupil of the highly aberrated test lens	246
6.5 Square-root of the CRB for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the highly aberrated test lens	248
6.6 Range in likelihood surface plots for Fringe Zernike coefficients $\{\alpha_n, n = 4, 9, 16, 25, 36, 37\}$ in the exit pupil of the highly aberrated test lens	249

LIST OF TABLES, *Continued*

Table	Page
6.7 ML estimates of wavefront parameters for the highly aberrated test lens, including their standard deviations and the starting point in the search	258
6.8 System data provided by ZEMAX™ for the spherical test lens at $\lambda = 0.6328 \mu\text{m}$	263
6.9 Fringe Zernike coefficients $\{\alpha_n, n = 1, \dots, 37\}$, peak-to-valley, RMS, and variance, provided by ZEMAX™ for the spherical test lens. Unlisted coefficients are zero	265
6.10 Computation time using Huygens' method for a pupil sampling of 256×256 and various detector grid sizes	270
6.11 Square-root of the CRB for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the spherical test lens	274
6.12 Range in likelihood surface plots for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots,$ $9, 16\}$ in the exit pupil of the spherical test lens	275
6.13 ML estimates of wavefront parameters for the spherical test lens, including their standard deviations. Design values were used as a starting point in the search.	284
7.1 True values of parameters underlying the irradiance data for the precision asphere, and design values of Edmund Optics Precision Asphere NT47-731.....	295

LIST OF TABLES, *Continued*

Table	Page
7.2 System data provided by ZEMAX™ for the precision asphere at $\lambda = 0.6328$ μm	296
7.3 Square-root of the CRB for prescription parameters describing the precision asphere	302
7.4 Range in likelihood surfaces for parameters describing the precision asphere, relative to the true values	303
7.5 ML estimates of prescription parameters describing the precision asphere, including standard deviations. Design values were used as a starting point in the search	309
7.6 Design parameters of the GRIN-rod test lens at an arbitrary design wavelength. Included are the distances in the optical system used in the simulations	318
7.7 Square-root of the CRB for the parameters describing the refractive index distribution of the GRIN-rod lens	321
7.8 Range in likelihood surfaces for parameters describing the GRIN-rod lens, relative to the true values	322
A.1 Fringe Zernike Polynomials $\{Z_n, n = 1, \dots, 37\}$	333

ABSTRACT

We present a new method for determining the complete set of patient-specific ocular parameters, including surface curvatures, asphericities, refractive indices, tilts, decentrations, thicknesses, and index gradients. The data consist of the raw detector outputs of one or more Shack-Hartmann wavefront sensors (WFSs); unlike conventional wavefront sensing, we do not perform centroid estimation, wavefront reconstruction, or wavefront correction. Parameters in the eye model are estimated by maximizing the likelihood. Since a purely Gaussian noise model is used to emulate electronic noise, maximum-likelihood (ML) estimation reduces to nonlinear least-squares fitting between the data and the output of our optical design program. Bounds on the estimate variances are computed with the Fisher information matrix (FIM) for different configurations of the data-acquisition system, thus enabling system optimization. A global search algorithm called simulated annealing (SA) is used for the estimation step, due to multiple local extrema in the likelihood surface. The ML approach to parameter estimation is very time-consuming, so rapid processing techniques are implemented with the graphics processing unit (GPU).

We are leveraging our general method of reverse-engineering optical systems in optical shop testing for various applications. For surface profilometry of aspheres, which involves the estimation of high-order aspheric coefficients, we generated a rapid ray-tracing algorithm that is well-suited to the GPU architecture. Additionally, reconstruction of the index distribution of GRIN lenses is performed using analytic

solutions to the eikonal equation. Another application is parameterized wavefront estimation, in which the pupil phase distribution of an optical system is estimated from multiple irradiance patterns near focus. The speed and accuracy of the forward computations are emphasized, and our approach has been refined to handle large wavefront aberrations and nuisance parameters in the imaging system.

CHAPTER 1

INTRODUCTION

In traditional optical design, a trial configuration of optical components is entered into a computer, rays are traced, and the images of one or more point objects are computed; then the configuration is altered in some way to improve the images. At each step in this iteration, the problem can be stated: given the optical system, find the image. While this process is invaluable for many uses, simply changing our view of the problem has potential for various powerful applications. We have developed a unique method which we refer to as *inverse optical design* (IOD); that is, given the image, find the system. In other words, by obtaining the data at some output plane, we can estimate the set of parameters describing the optical system. The basic method has been patented by Barrett, Sakamoto, and Goncharov (2010).

The original motivation of this research is to develop a new technique for studying the time-varying, optical properties of the eye of an individual patient, either for clinical ophthalmology or basic research. The imaging system is based on a Shack-Hartmann aberrometer for measurement of aberrations in human eyes, in which an incoming light wave provided by a laser diode is distorted as it enters and leaves through the complicated, dynamic optical properties of the eye. The image in this case consists of the output in the focal plane of a Shack-Hartmann *wavefront sensor* (WFS), a device that measures the distortions of a wavefront and provides very useful information about

aberrations in an optical system. In essence, these data are used to estimate surface curvatures, conic constants, refractive indices, thicknesses, tilts and decentrations of all components in the eye, as well as the graded-index (GRIN) distribution of the crystalline lens, which has not previously been achieved using a single ocular diagnostic system. The patient-specific eye model could then be used as a theoretical basis for vision correction of higher-order aberrations, or to develop data bases for the diagnosis of pathologies, facilitate a broad range of critical studies in vision science, or optimize a *multi-conjugate adaptive-optics* (MCAO) system for imaging the entire retina with a substantial improvement in resolution.

Fig. 1.1: System configuration for estimating patient-specific ocular parameters, based on a clinical Shack-Hartmann aberrometer for measurement of aberrations.

Inverse optical design relies on computational methods, incorporating an optical design program and statistical analysis. Data are taken from one or more output planes of the optical system, then entered into a computerized optimization algorithm, invoking statistical approaches like *maximum-likelihood* estimation or *maximum a posteriori* (MAP) estimation. ML estimation essentially perform a search through parameter space to find the set of parameters that maximizes the probability of occurrence of the observed data, by comparing the data to the output of the optical design program. MAP estimation is a generalization of ML estimation which accepts some prior knowledge on the probability distribution of the parameters. We chose to implement ML estimation for this research.

The performance of such estimators can be analyzed with the *Fisher information matrix*, as it provides the theoretical minimum possible variance in those estimates, referred to as the *Cramér-Rao lower bound*. It essentially measures information content in the system in terms of the sensitivity of the data to changes in each parameter, but also reveals any parametric coupling, including coupling in the estimates.

A significant limitation to overcome in inverse optical design is that the ML estimation step is very time-consuming, so making it practical requires the development of a proficient search algorithm, as well as dedicated computer hardware. Improvement of computational time can be achieved by parallelizing the optical design program describing the forward model of the system. Parallel algorithms can be implemented on a variety of hardware platforms, such as the cell processor used in the Sony PlayStation 3 or the graphics processing unit in high-performance video cards.

It should be noted that inverse optical design is a nonlinear estimation problem, since the data depend nonlinearly on the parameters. Similar methods have been applied in astronomy (Redding, 1993), in which image data are used to estimate optical prescription parameters (i.e., first-order geometrical parameters), a process referred to as *prescription retrieval*. The original application was to estimate the conic constant of the Hubble Space Telescope primary mirror, after it was configured incorrectly and produced an unexpected level of spherical aberration.

1.1 Application to vision science and ophthalmology

The availability of a complete patient-specific eye model would provide many key advantages in vision science and ophthalmology. For example, the model could be used as a theoretical basis for vision correction of higher-order aberrations in laser refractive surgery or with corrective lenses, since classical devices only measure and correct for basic refractive errors (i.e., defocus and astigmatism). Accurate determination of optical parameters in normal and abnormal eyes could also be valuable in developing data bases for clinical diagnosis of pathologies, while measurements of ocular surface misalignments would be useful after implantation of intraocular lenses in cataract surgery (Rosales & Marcos, 2006; Rosales, Dubbelman, Marcos, & Van der Heijde, 2006a; Tabernero, Benito, Nourrit, & Artal, 2006). Moreover, knowledge of the refractive index distribution in the lens may be beneficial in optical coherence tomography, which is based on interferometric reflectometry and index changes (Jones, Atchison, Meder, & Pope, 2005; Moffat, Atchison, & Pope, 2002). It could additionally lead to a substantial

improvement in retinal imaging. For instance, current adaptive-optics (AO) ophthalmoscopes incorporating a Shack-Hartmann wavefront sensor and wavefront corrector conjugated to a single surface of the eye offer high resolution (Hofer *et al.*, 2001; Roorda *et al.*, 2002), but over a very limited field-of-view (FOV) (Liang, Williams, & Miller, 2007) due to a form of anisoplanatism involving aberrations of the eye. Aberrations collected over different field positions on the retina result from the passage through different parts of the ocular media so that the AO correction is valid only over a certain field area, referred to as the isoplanatic patch (Fig. 1.2). One solution is to conjugate multiple wavefront sensors and correctors to various refractive surfaces in the eye, thereby increasing the isoplanatic patch size and enabling wide-field measurements, but choice of the optimal planes at which to conjugate the correctors would be facilitated by knowing the real eye structure of the individual.

Fig. 1.2: Anisoplanatism involving pupil aberrations. Image from Stéphane Chamot (National University of Ireland, Galway).

In addition to improvements in vision correction and retinal imaging, the availability of patient-specific parameters could facilitate a broad range of ongoing vision science studies. Of significant interest is the *in vivo* GRIN distribution and *lenticular* geometry of the human crystalline lens as a function of both age and accommodation (Hemenger, Garner, & Ooi, 1995; Rosales *et al.*, 2006a; Smith, Atchison, & Pierscionek, 1992), but this information has been difficult to obtain, and reliable measurements are scarce (Dubbelman & Van der Heijde, 2001; Jones *et al.*, 2005; Liang *et al.*, 1997; Moffat *et al.*, 2002; Navarro, Santamaría, & Bescós, 1985). While previous studies suggested that aspheric surfaces in the anterior segment and an effective refractive index for the lens are sufficient to model spherical aberration, lack of knowledge regarding the GRIN distribution precludes both the prediction of off-axis aberrations and study of dispersion in the lens, so that experimental data are limited (Navarro *et al.*, 1985). A complete mapping of the human eye could also be used to evaluate intersubject variability and statistical variations, as well as vision performance and image quality in the central and peripheral visual fields (Navarro, Moreno, & Dorronsoro, 1998; Sheehan, Goncharov, O'Dwyer, Toal, & Dainty, 2007), which could be enhanced by accurate measurement of the retinal curvature (Escudero-Sanz & Navarro, 1999; Mallen & Kashyap, 2007). Another fundamental study in physiological optics is how individual ocular components factor into the overall performance of the human eye (Artal & Guirao, 1998) and how such performance would change if one or more surfaces are altered, a critical element in surgical procedures. While schematic eyes have been extremely useful for that purpose, they often lack asymmetry such as decentration of the lens or pupil,

which manifest in the fovea as aberrations of non-axially-symmetric systems (e.g., coma, astigmatism, and transverse chromatic aberration) and may have a significant impact on ocular performance (Bará & Navarro, 2003; Rynders, Lidkea, Chisholm, & Thibos, 1995). A patient-specific mapping of the entire eye, including non-axially-symmetric components, would enable further investigations that have been previously unapproachable.

The method of inverse optical design could provide an *in vivo*, non-invasive, and complete mapping of the human eye, including dozens of parameters that are essential to an accurate representation of the eye and its aberrations. Existing *in vivo* methods supply a small subset of ocular parameters. For example, a common technique in phakometry uses Purkinje images of the back reflections from the anterior and posterior surfaces of both the cornea and crystalline lens, providing basic curvatures, tilts, and decentrations (Rosales & Marcos, 2006; Rosales *et al.*, 2006a). However, one difficulty in this approach is that insufficient knowledge of the refractive index distribution of the lens leads to significant measurement errors in the lens posterior radius (Schwiegerling, 2004). Scheimpflug slit imaging is increasingly being used to obtain sharp cross-sectional images of the anterior eye segment, imparting surface shapes, misalignments, and intraocular distances, although accurate determination of these parameters relies on the correction of optical distortions in the imaging system and within the eye itself (Rosales & Marcos, 2006). Distortion due to the geometry of the Scheimpflug camera can be corrected analytically with relative ease, but correction of distortion due to refraction at intermediate ocular surfaces is much less approachable. Measurements of a

particular surface are subjected to refraction at all successive surfaces (Dubbelman & Van der Heijde, 2001; Koretz, Strenk, Strenk, & Semmlow, 2004; Rosales *et al.*, 2006a) and traversal through media of individually varying thickness and curvature. Hence, arbitrary quantification errors in one surface are propagated throughout the system (Dubbelman, Weeber, Van der Heijde, & Völker-Dieben, 2002). Conversely, magnetic resonance imaging has recently been used for *in vivo* visualization of structures in the anterior segment, which eliminates the distortion dilemma (Koretz *et al.*, 2004), but suffers from low resolution, signal-to-noise ratio (SNR) constraints, and eye motion artifacts due to longer acquisition times (Strenk *et al.*, 1999). On the other hand, corneal topography is a rapidly developing technique that provides very detailed and reliable measurements regarding corneal curvature (Schwiegerling, Greivenkamp, & Miller, 1995; Navarro, González, & Hernández, 2006; Zhou, Hong, Miller, Thibos, & Bradley, 2004; Guirao & Artal, 2000), including astigmatism and surface irregularities, although it does not provide information about the remaining ocular surfaces. However, such accurate corneal information could be used to supplement or validate the parameter estimates acquired with our system, or even used as input to inverse optical design to narrow the high-dimensional parameter space.

1.2 Application to optical shop testing

Although inverse optical design has an ophthalmic origin, the basic concept of parameter estimation is much more widely applicable; the same technique could be applied to any situation where the parameters of an optical system are desired. Our method of optical

testing via parametric modeling has applications in precision testing of optical components and systems for commercial, industrial, military, and aerospace purposes. Several examples are coating and surface profilometry of aspheres, measurement of aberrations in intraocular lenses (IOLs) and contact lenses, laser machining, and tomographic reconstruction of the three-dimensional refractive index distribution in GRIN lenses or fiber optic cables (Fig. 1.3).

Fig. 1.3: Basic test configuration for performing inverse optical design of a GRIN-rod lens.

We also need not restrict ourselves to the estimation of geometrical parameters by means of ray-tracing. Another application area is phase-retrieval, the process of trying to recover the wavefront error in an optical system such as a group of lenses, given one or more irradiance measurements near focus. If we consider the coefficients in an arbitrary wavefront expansion as the estimable parameters, we could in principle measure the wavefront error produced by the optical system. Since the wavefront is treated as a

continuous function, this process avoids conventional gridlocks such as aliasing or phase ambiguities and has potential for measuring extremely large aberrations, a problem that confounds traditional interferometry. This technique employs diffraction propagation and is therefore practical on both the microscopic and macroscopic scales from micro-optics to large telescope mirrors, and for reflective or transmissive parts. It is also practical for measuring very large peak-to-valley wavefront errors. Figure 1.4 shows the basic system configuration for our method of parameterized wavefront measurement with a reflective test element, simply requiring a point source and detector array.

Fig. 1.4: Basic system configuration for parameterized wavefront measurement with a single source.

Our method of parameterized wavefront measurement is similar to that of Brady and Fienup (2004, 2005a), but we extend it in various ways. We perform system optimization by investigating the FIM and associated CRB, as well as the probability surface. Furthermore, we refined our approach to deal with large-aberration wavefronts, plus we perform rapid processing on the GPU platform.

We devise methods for dealing with common practical issues. For instance, a single source and detector may be adequate for sensing the focal spots generated by small-scale optical elements, but will be insufficient for dealing with large aspheric telescope mirrors that are designed to image sources at infinity as the nominal spot will not fit onto the detector array. As an example, a parabolic mirror with a 3.5-meter diameter and an f-number of 1.5 (i.e., a high light-collection efficiency) creates a spot size of 100 millimeters, many times larger than most CCD detectors used in scientific cameras. If the entire spot is not detected, parts of the mirror will be invisible to the system and would manifest as phase errors in the estimated aberration function of the mirror. One feasible solution is to use an array of spatially-separated identical point sources that scan the entire mirror as viewed by a single detector at a fixed location (Fig. 1.5). Each source contributes optically to the total phase aberration contained in sequential images, such that accurate recovery of the entire aberration function is possible. In a phase-diversity approach, differences between images are used as input to the nonlinear optimization and common-mode information falls out, which could serve as a technique to reduce the stray light entering the problem. Additionally, this configuration contains no moving parts or intervening optics and is cost-efficient

compared to a multiple-detector system. Another possible solution involves the use of projection and relay optics to transfer the large focal spot to the smaller detector array, which involves placing a large field lens and projection screen at the intermediate focal plane. If the incorporated lenses are placed in respective focal planes of the system, they are not considered intervening optics do not introduce wavefront phase distortions, plus they can easily be built into the forward computation model used in the optimization routine.

Fig. 1.5: Basic system configuration for parameterized wavefront measurement with an aspheric test element and multiple point source locations.

A unique feature of our method involves the examination of augmented systems in which an additional element is introduced into the optical path for sake of greater information yield. A worthy candidate for this purpose is a Shack-Hartmann WFS, containing a two-dimensional array of small lenslets that samples the incoming wavefront and produces an array of blurred spots in the focal plane. In conventional wavefront sensing, an algorithm processes the detected image and computes the centroids of the spots. The centroids are used to estimate the average local wavefront slopes, which are combined to give a rough reconstruction of the wavefront. Discrete samples of the continuous wavefront, which contains an infinite number of points, may work decently when low aberrations are present, but undersampling of a rapidly varying wavefront will lead to aliasing and phase ambiguities. Also, useful information regarding the finer structure of the wavefront is also thrown away during the centroiding process.

In our method, we do not perform centroiding, but instead use the raw detector outputs in the detector focal plane. Even if aliasing occurs, we will know what the aliased wavefront looks like from our computational model, so the statistical estimation of phase polynomial coefficients works fine. Here we use a finite set of data to determine a finite set of polynomial coefficients, but we are able to reconstruct a smooth and continuous wavefront. Additionally, our method permits the splashing of focal spots outside of their territories in the case of large aberrations, a problematic occurrence in the classical centroiding process. This approach to wavefront estimation using WFS data was investigated by Luca Caucci, a graduate student within the research group.

Fig. 1.6: Basic system configuration for augmented wavefront measurement with a Shack-Hartmann WFS and multiple point source locations.

Interferometry has been ubiquitous in surface profilometry, but requires a larger number of optic elements that introduce aberrations into the system and is extremely sensitive to vibrations between the reference and test arms. For testing aspheres, it also requires a null optic as a reference surface which is very difficult and costly to fabricate, and can suffer from common problems such as undersampling and hysteresis.

Additionally, while phase-stepping and phase-shifting interferometers are specifically designed to handle aberrations greater than an optical wavelength, there are phase errors and ambiguities associated with the phase unwrapping process. Our method avoids these problems, since it requires as little as a source and detector, can operate with or without a null component, and is less sensitive to vibrations due to the single optical path. It can

also measure large wavefront errors without the need for phase unwrapping. On the other hand, even if we choose to use a null configuration, ML estimation will still perform well as long as we employ an accurate forward model of the system.

1.3 Dissertation overview

Chapters 2 – 4 provide the theoretical framework needed to perform inverse optical design, which can be used for all of the aforementioned applications. Chapter 2 introduces fundamental concepts in maximum-likelihood estimation in the context of parameter estimation, including various performance metrics and properties of estimators. We provide rigorous likelihood functions for the detector output, which must incorporate all noise sources and factors that influence the data, and we discuss various methods for handling nuisance parameters. The general formulation of the Fisher information and the Cramér-Rao lower bound are provided. We also discuss limitations of the ML approach, including intensive computational requirements.

A necessary component in implementing the ML approach to parameter estimation is an appropriate optimization algorithm, which searches the parameter or configuration space to find the configuration that maximizes the probability of the observed data. In Chapter 3, we discuss various optimization methods and the selection of a suitable search algorithm. Since the likelihood functions in inverse optical design or wavefront estimation tend to be complicated with many local extrema, we are primarily interested in global search algorithms. We give a broad overview of the simulated annealing (SA) algorithm, a feasible candidate for the global optimization problem,

which can process high-dimensional functions with extensive nonlinearities, discontinuities, and randomness. For the applications presented in this dissertation, we implemented an adaptive form of simulated annealing for optimizing multimodal functions in a continuous domain.

Each iteration of an inverse problem requires a solution to the forward problem; we must compute the output of an optical design program, either through ray-tracing or diffraction propagation. Chapter 4 provides a comprehensive review of the propagation of light, including relevant concepts from geometrical optics and diffraction theory. Although we begin with the fundamental Maxwell's equations, we derive many practical expressions that are central to the propagation algorithms developed in this research. We derive the basic equation of geometrical optics, the eikonal equation, which is needed in ray-tracing through GRIN lenses. This leads to a discussion of a useful version of Snell's law in vector form, used throughout this research for the refraction of light rays at optical surfaces. In the wavefront estimation problem, we rely on scalar diffraction theory for modeling the wave propagation from the exit pupil of an optical system to final image plane. Several expressions are developed under various approximations.

In Chapters 5 – 7, we present results obtained for several applications using the theoretical background provided in the preceding chapters. Specific details on the propagation algorithm are outlined for each application. We routinely investigate the FIM and associated CRB, as well as the complicated behavior of the likelihood function. Simulated annealing is used in all cases for the estimation procedure.

Chapter 5 deals with the original motivation of this research, the estimation of patient-specific parameters of the eye using irradiance data from the focal plane a WFS. We present an overview of the human eye and discuss schematic eye models of varying levels of complexity. A description of our optical design program is provided, which involves an algebraic method for non-paraxial ray-tracing through the optical system of the eye.

In Chapter 6, we discuss the estimation of wavefront parameters, including the case of large wavefront errors. Since the computational demands are much higher for diffraction propagation compared to ray-tracing, we stress the complexity and speed of the computations required for accurate determination of wavefront parameters. An introduction to parallel processing with GPUs is provided.

Chapter 7 involves two additional applications of IOD for optical testing, including the testing of precision aspheric lenses and GRIN lenses. To estimate parameters describing high-order aspheric surfaces, we developed a rapid ray-tracing program for implementation on the GPU platform. For ray-tracing through the refractive-index distribution of GRIN-rod lenses, we used analytic solutions to the eikonal equation.

CHAPTER 2

MAXIMUM-LIKELIHOOD ESTIMATION

The method of maximum-likelihood (ML) is one of the oldest and most significant techniques in estimation theory. It is a standard approach in statistical inference, which includes classification tasks and estimation problems. This chapter expounds on the application of ML estimation specifically to parameter estimation.

In Section 2.1, we provide a historical overview of parameter estimation, from the earliest methods of estimation to more modern applications with the birth of computer technology. In the process we discuss the evolution of ML estimation, beginning with its discovery by R. A. Fisher.

We describe the fundamental estimation problem in Section 2.2, while distinguishing between the Bayesian and frequentist approaches in statistical inference, including the branch of parameter estimation. Section 2.3 imparts the essential notation and terminology, while introducing the concept of maximizing the likelihood.

In Section 2.4, we discuss the concept of information in a probability model. We show how to calculate the Fisher information matrix and use it to derive an important performance bound on parameter estimates, the Cramér-Rao lower bound. Alternative bounds are also discussed, along with their advantages and drawbacks. Section 2.5 deals with general performance metrics and properties of estimators, with an emphasis on the

ML estimator. This includes the optimal properties of the ML estimators in the large-sample or asymptotic limit. Section 2.5 parallels the development in Section 2.4.

Section 2.6 proposes realistic methods for dealing with nuisance parameters in a probability model. In Section 2.7, we provide rigorous likelihood functions for modeling electronic noise in detector arrays, described by Gaussian statistics. Finally, we discuss practical challenges of ML estimation in Section 2.8, including the need for an accurate description of random phenomena, as well as computational limitations to be overcome.

This chapter assumes a knowledge of elementary probability and statistics.

2.1 Historical background

The first published proposition of the method of least-squares for estimating coefficients in linear curve fitting was made by Legendre (1805), whose primary interest was in predicting comet orbits. While he suggested the technique merely as a convenient procedure for treating observations, Legendre made no reference to the theory of probability. Meanwhile, Gauss independently discovered the least-squares method and applied it as early as 1795, but it was not until 1809 that he published a comprehensive treatment of the method, formally outlining the theory and mathematical foundation. In this manuscript, Gauss showed that estimates obtained through least-squares fitting maximized the probability density for a normal distribution of errors, which gave way to the development of statistical parameter estimation. It was also a prevision of the method of maximum-likelihood. Continued work into the early twentieth century focused on the

computational side of the least-squares method, involving other mathematicians such as Cauchy, Bienaymé, Chebyshev, Gram, and Schmidt (Seal, 1967). Orthogonal polynomials were also an important development of the work during this era. An overview of the development of the least-squares method is found in Plackett (1972), while Gauss's contribution is examined by Trotter (1957).

The work of Pearson near the turn of the century and R. A. Fisher in subsequent decades provided the underpinnings for advancements in statistical estimation methods. Several of Pearson's contributions to classical statistics are principal component analysis (Pearson, 1901), the chi-square distribution (Pearson, 1900), correlation theory (Pearson, 1898, 1900), and the method of moments (Pearson, 1894, 1936), with the latter providing an early method for the estimation problem.

Fisher based much of his work on that of Pearson, even concurring that the method of moments was superior to the least-squares method – but he had ideas for an even better approach. In 1912, Fisher introduced the principle of inverse probability, from which he derived the “absolute criterion” (Fisher, 1912), although he later discarded the ideas. By 1922 he clarified the difference between “probability” and “likelihood”, thereby completing the basic theory of maximum-likelihood, which he presented in a series of papers (Fisher, 1922, 1925, 1934, 1935). In these papers, he described estimator properties such as sufficiency, efficiency, consistency, and information. Aldrich (1997) gives an excellent, detailed account of the development of ML estimation.

During the next fifteen years, Wald (1939, 1945, 1950) made significant contributions to statistical decision theory, which suggests principles for choosing

estimation criteria in optimal decision-making. It has since played a substantial role in point estimation and hypothesis testing.

Contemporary application of statistical estimation theory began in the 1940s and 1950s by Hood and Koopmans, who used the theory to estimate variables in macroeconomic models. Their work was very important in the progress of econometrics during this era and is outlined in the Cowles Commission Reports (Hood & Koopmans, 1953). Meanwhile, G. E. P. Box and others (1958, 1959, 1962) made significant contributions in the physical sciences by constructing mathematical models and estimating model parameters.

Methods of nonlinear parameter estimation that were already established by mathematicians such as Newton, Gauss, and Cauchy truly did not reach extensive practical application until the advent of computer technology in the 1950s. The first general computer program to determine estimates for nonlinear models was written by Booth and Peterson (1958), which implemented nonlinear least-squares fitting. More specific, it employed Gauss' method with finite difference approximations to solve least-squares problems of a single equation. The earliest computer program using ML for nonlinear parameter estimation was created by Eisenpress, Bomberault, and Greenstadt (1966a, 1966b) for econometric models of multiple equations. Their system applied the full Newton method with rotational discrimination by evaluating analytic derivatives of all orders.

2.2 Statement of the problem

This section describes the fundamental components in classical estimation problems. We refrain from elaborating on notation, which we expound on in Section 2.3.

A general estimation procedure consists of several important factors. There must first be a vector parameter θ describing some source or object, representing a point in *parameter space*. In the case of inverse optical design for the ophthalmic application, for example, the parameters are the curvatures, conic constants, thicknesses, indices, and so on.

The next component is a probabilistic mapping from parameter space to the finite-dimensional *observation space*, in which the observed data $\mathbf{g}$ reside. This is simply the probability law, denoted $\text{pr}(\mathbf{g}|\theta)$, governing the effect of the parameters on the data, including any noise characteristics or random phenomena in the system. In inverse optical design, as well as many other estimation problems, the mapping is nonlinear since the data depend nonlinearly on the parameters.

The final requirement is an estimation rule, or procedure, for mapping the observation space to an estimate, written $\hat{\theta}$. This rule for processing an observation or set of observations to generate an estimate is also referred to as an *estimator*. We treat this rule as deterministic, that is, the same data vector will always produce the same estimate. When the estimate is represented by specific numerical values, the process is referred to as *point estimation*. However, point estimation conveys nothing about the

uncertainty in the estimates. *Interval estimation*, on the other hand, uses sample data to determine an interval of probable values of the unknown parameters (Neyman, 1937).

There are essentially two distinctive approaches in the treatment of statistical inference tasks such as estimation problems. The classical or frequentist method regards the parameter to be estimated as unknown, but not random. This approach considers an ensemble of data vectors that is acquired through sampling of $\text{pr}(\mathbf{g}|\boldsymbol{\theta})$ and computes performance metrics using averages of the estimates. In this sense, the repeated sampling can be used to verify all probabilities and probability laws.

Contrastingly, parameters are treated as random variables in the Bayesian method, so that knowledge of a prior probability $\text{pr}(\boldsymbol{\theta})$ must be assumed. However, this probability is admitted as a “degree of belief” (Ferguson, 1967; Raiffa & Schlaifer, 1961; Savage, 1954) with “subjective choices of plausibility” (Bard, 1974). Both $\text{pr}(\boldsymbol{\theta})$ and $\text{pr}(\mathbf{g}|\boldsymbol{\theta})$ are used in Bayes’s rule to ascribe to the parameter $\boldsymbol{\theta}$ a posterior density, denoted $\text{pr}(\boldsymbol{\theta}|\mathbf{g})$, conditioned on the observed data vector $\mathbf{g}$. Therefore, a Bayesian has no concept of an ensemble of data vectors, and performance metrics are determined solely from the posterior (Barrett, Dainty, & Lara, 2007).

We utilize classical estimation theory in this paper and treat the parameters to be estimated as nonrandom variables.

Estimation procedures typically involve the optimization of some objective function, or in many cases, the minimization of a cost function, denoted $C = (\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$. The cost function assigns a penalty to the point in parameter space $\hat{\boldsymbol{\theta}}$ when the true

underlying parameter is $\boldsymbol{\theta}$. In other words, it measures the departure of the given data from that generated by a proposed system configuration. In the next section, we will describe the quantity that is optimized in ML estimation.

2.3 Notation system and terminology

The notation used throughout this paper will be adopted primarily from Barrett and Myers (2004).

We use $\mathbf{g} = g_1, \dots, g_M$ to represent an $M \times 1$ vector containing random data from some probability law. The probability law itself is described by a $P \times 1$ vector set of parameters $\boldsymbol{\theta} = \theta_1, \dots, \theta_P$. Note that vectors are indicated by boldface lowercase letters, while matrices are denoted as boldface uppercase letters. If the data can take on continuous values, then the probability law is a *probability density function* (PDF), denoted as $\text{pr}(\mathbf{g}|\boldsymbol{\theta})$. Conversely, if the data are discretely-valued, then the probability law is simply a probability, given by $\text{Pr}(\mathbf{g}|\boldsymbol{\theta})$. For the sake of this research, we will consider only continuous random variables.

The PDF $\text{pr}(\mathbf{g}|\boldsymbol{\theta})$ is simply the distribution from which individual samples of $\mathbf{g}$ are drawn. In other words, it represents the probability of obtaining the data vector $\mathbf{g}$ conditional upon the parameter vector $\boldsymbol{\theta}$. The most commonly used distributions in practice are the normal, log-normal, and gamma for continuous variables, and the Poisson, binomial, and multinomial for discrete variables. In the classical approach to

parameter estimation, once we have a particular data vector, we can express the PDF as a function of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{g}$, referred to as the *likelihood*:

$$L(\boldsymbol{\theta} | \mathbf{g}) = \text{pr}(\mathbf{g} | \boldsymbol{\theta}) . \quad (2.1)$$

For a set of M independent and identically distributed (i.i.d.) observations, the likelihood can be written as

$$L(\boldsymbol{\theta} | \mathbf{g}) = \prod_{m=1}^M \text{pr}(g_m | \boldsymbol{\theta}) . \quad (2.2)$$

We must emphasize that $L(\boldsymbol{\theta} | \mathbf{g})$ is not a PDF on $\boldsymbol{\theta}$.

In general, an estimate of the parameter vector is denoted $\hat{\boldsymbol{\theta}}$, and values that maximize the likelihood $\hat{\boldsymbol{\theta}}_{\text{ML}}$ are referred to as ML estimates of $\boldsymbol{\theta}$. If the estimate is a deterministic function of $\mathbf{g}$, which is usually the case, we can express it as $\hat{\boldsymbol{\theta}}(\mathbf{g})$. However, we will often drop the explicit dependence on $\mathbf{g}$ for brevity.

ML estimation essentially returns the $\boldsymbol{\theta}$ argument which maximizes the probability of occurrence of the observed data, defined as

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \equiv \underset{\boldsymbol{\theta}}{\text{argmax}} \text{pr}(\mathbf{g} | \boldsymbol{\theta}) , \quad (2.3)$$

where $\hat{\boldsymbol{\theta}}$ represents an estimate of the vector set of parameters. Since the logarithm increases monotonically with its argument, (2.3) is equivalent to

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln \operatorname{pr}(\mathbf{g} | \boldsymbol{\theta}), \quad (2.4)$$

where $\ln \operatorname{pr}(\mathbf{g} | \boldsymbol{\theta})$ is the *log-likelihood*. Note that the log-likelihood is a random variable due to its dependence on $\mathbf{g}$. For practical purposes, it is typically more convenient to minimize the negative log-likelihood, so that (2.4) then becomes

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} [-\ln \operatorname{pr}(\mathbf{g} | \boldsymbol{\theta})]. \quad (2.5)$$

Furthermore, if the log-likelihood has a continuous first derivative with respect to $\boldsymbol{\theta}$, then this derivative evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{ML}}$ is equal to zero. This is called the *likelihood equation*:

$$\left. \frac{\partial}{\partial \boldsymbol{\theta}} \ln \operatorname{pr}(\mathbf{g} | \boldsymbol{\theta}) \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{ML}}(\mathbf{g})} = 0, \quad (2.6)$$

where $\partial \alpha / \partial \boldsymbol{\theta}$ is the column vector $[\partial \alpha / \partial \boldsymbol{\theta}]_i = \partial \alpha / \partial \boldsymbol{\theta}_i$ for a function $\alpha(\boldsymbol{\theta})$.

Example: Correlated Gaussian noise

Suppose we have a data vector given by $\mathbf{g} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{b} + \mathbf{n}$, where $\mathbf{s}(\boldsymbol{\theta})$ is a signal parameterized by $\boldsymbol{\theta}$, $\mathbf{b}$ is a known background, and $\mathbf{n}$ represents correlated Gaussian noise, with $\mathbf{n} \sim \mathcal{N}_M(\mathbf{0}, \mathbf{K}_n)$. Note that $\mathcal{N}_M(\mathbf{0}, \mathbf{K}_n)$ is the normal distribution with zero mean and covariance matrix $\mathbf{K}_n$, where M samples are drawn. The conditional PDF on the data is written as

$$\text{pr}(\mathbf{g} | \boldsymbol{\theta}) = \frac{1}{(2\pi)^{M/2} [\det(\mathbf{K}_n)]^{1/2}} \exp \left\{ -\frac{1}{2} [\mathbf{g} - \mathbf{b} - \mathbf{s}(\boldsymbol{\theta})]^t \mathbf{K}_n^{-1} [\mathbf{g} - \mathbf{b} - \mathbf{s}(\boldsymbol{\theta})] \right\}, \quad (2.7)$$

and its logarithm is given by

$$\ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) = \ln \left\{ \frac{1}{(2\pi)^{M/2} [\det(\mathbf{K}_n)]^{1/2}} \right\} - \frac{1}{2} [\mathbf{g} - \mathbf{b} - \mathbf{s}(\boldsymbol{\theta})]^t \mathbf{K}_n^{-1} [\mathbf{g} - \mathbf{b} - \mathbf{s}(\boldsymbol{\theta})]. \quad (2.8)$$

According to (2.5), we can minimize the negative log-likelihood with respect to $\boldsymbol{\theta}$ to obtain

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\text{argmin}} [\mathbf{g} - \bar{\mathbf{g}}(\boldsymbol{\theta})]^t \mathbf{K}_n^{-1} [\mathbf{g} - \bar{\mathbf{g}}(\boldsymbol{\theta})], \quad (2.9)$$

where the average data vector is $\bar{\mathbf{g}}(\boldsymbol{\theta}) = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{b}$, since $\mathbf{n}$ is zero-mean. Additionally, $\bar{\mathbf{g}}(\boldsymbol{\theta})$ is the anticipated data vector for a given set of parameters.

2.4 Fisher information and the Cramér-Rao bound

Before discussing the properties of ML estimators, we will first introduce the concept of Fisher information and how it is used to derive the theoretical minimum possible variance on parameter estimates. This is integral to a discussion on the asymptotic theory of ML estimation, as well as various performance metrics, which are covered in Section 2.5.

2.4.1 *Score*

The *score* is a vector that describes the sensitivity of the likelihood to changes in the parameters:

$$\mathbf{s}(\mathbf{g}) = \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \text{pr}(\mathbf{g} | \boldsymbol{\theta})}{\text{pr}(\mathbf{g} | \boldsymbol{\theta})} = \frac{\partial}{\partial \boldsymbol{\theta}} \ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) . \quad (2.10)$$

Mathematically, it is the gradient with respect to $\boldsymbol{\theta}$ of the log-likelihood. It can easily be shown that the expectation of the score with respect to $\text{pr}(\mathbf{g} | \boldsymbol{\theta})$ is zero: $\langle \mathbf{s} \rangle_{\mathbf{g} | \boldsymbol{\theta}} = 0$, where the brackets denote the average. Barrett and Myers (2004) also point out that since the score is the gradient of the log-likelihood, all of its components vanish at the point in parameter space corresponding to the ML estimate, $\mathbf{s}(\mathbf{g} | \hat{\boldsymbol{\theta}}_{\text{ML}}) = 0$, provided there are no constraints such as positivity.

2.4.2 Fisher information matrix

The performance of an ML estimator can be analyzed with the *Fisher information matrix* (FIM), as it describes the ability to estimate a vector set of parameters. If $\mathbf{s}$ has zero mean, then the FIM is simply the covariance matrix of the score, which is expressed in outer-product notation as

$$\mathbf{F} = \langle \mathbf{s} \mathbf{s}^t \rangle_{\mathbf{g}|\boldsymbol{\theta}}, \quad (2.11)$$

with individual components $F_{jk} = \langle s_j s_k \rangle$. Thus, for a vector parameter of P real components, the FIM is a $P \times P$ symmetric matrix with real components given by

$$F_{jk} = \left\langle \left[\frac{\partial}{\partial \theta_j} \ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \right] \left[\frac{\partial}{\partial \theta_k} \ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \right] \right\rangle_{\mathbf{g}|\boldsymbol{\theta}}, \quad (2.12)$$

where the angle brackets denote the average over $\mathbf{g}$ for a given $\boldsymbol{\theta}$. Converted to integral form, (2.9) becomes

$$\begin{aligned} F_{jk} &= \int_{\infty} d^M \mathbf{g} \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \left[\frac{1}{\text{pr}(\mathbf{g} | \boldsymbol{\theta})} \frac{\partial}{\partial \theta_j} \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \right] \left[\frac{1}{\text{pr}(\mathbf{g} | \boldsymbol{\theta})} \frac{\partial}{\partial \theta_k} \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \right] \\ &= - \int_{\infty} d^M \mathbf{g} \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \left[\frac{1}{\text{pr}(\mathbf{g} | \boldsymbol{\theta})} \frac{\partial}{\partial \theta_k} \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \right] \end{aligned}$$

$$= - \int_{\infty} d^M \mathbf{g} \, \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) , \quad (2.13)$$

which can then be expressed as

$$F_{jk} = - \left\langle \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \right\rangle_{\mathbf{g} | \boldsymbol{\theta}} . \quad (2.14)$$

The second derivative in (2.14) indicates that the FIM components represent the average degree of curvature of the log-likelihood, where the average encompasses all data sets for a given parameter vector (Barrett & Myers, 2004).

2.4.3 Cramér-Rao inequality

The degree of dispersion in the sampling distribution of a random vector is conveyed in the *dispersion matrix*, defined as the inverse of the FIM, $\mathbf{D} = \mathbf{F}^{-1}$. We intuitively know that the more disperse the distribution, the more uncertain is the value of any particular realization of the random variable, thereby leading to greater variance in the parameter estimates (Bard, 1974). It is well-documented in the literature that the variance of any unbiased estimate obeys the *Cramér-Rao inequality* (Cramér, 1946; Rao, 1945):

$$[\mathbf{K}_{\hat{\boldsymbol{\theta}}}]_{pp} = \text{Var}\{\hat{\theta}_p\} \geq [\mathbf{F}^{-1}]_{pp} , \quad (2.15)$$

where θ_p is the p^{th} parameter and $\mathbf{K}_{\hat{\theta}}$ is the covariance matrix of the estimates. (See Section 2.5 for the definition of the *bias* in an estimate.) Thus, the variance of the p^{th} parameter cannot be smaller than the p^{th} diagonal entry in the dispersion matrix.

Although the inequality was first stated by Fisher and proved by Dugué (1937), Cramér and Rao are accredited with its discovery.

The theoretical minimum possible variance in (2.15) is referred to as the *Cramér-Rao lower bound* (CRB) of an estimate. An estimate that achieves the CRB is said to be *efficient*. The concept of efficiency will be further discussed in Section 2.5 when we describe the properties of ML estimators.

The relationship in (2.15) can be stated more generally using the notational convention of Loewner ordering for the two positive-definite matrices, $\mathbf{K}_{\hat{\theta}}$ and $\mathbf{F}^{-1}$.

Since $\mathbf{K}_{\hat{\theta}} - \mathbf{F}^{-1}$ is positive-semidefinite, one can prove that the covariance matrix for an unbiased estimator must satisfy

$$\mathbf{K}_{\hat{\theta}} \geq \mathbf{F}^{-1}. \quad (2.16)$$

For any biased estimator, it can be shown that

$$\mathbf{K}_{\hat{\theta}} \geq (\nabla_{\theta} \mathbf{b} + \mathbf{I}) \mathbf{F}^{-1} (\nabla_{\theta} \mathbf{b} + \mathbf{I})^t, \quad (2.17)$$

where $\mathbf{I}$ is the $P \times P$ unit matrix. Therefore, the bias changes the minimum variance, and can even reduce the variance if the gradient is negative (Barrett & Myers, 2004).

For a scalar parameter, the variance in the unbiased case as shown in (2.15) is simply given by

$$\text{Var}\{\hat{\theta}\} \geq \frac{1}{\left\langle -\frac{\partial^2}{\partial \theta^2} \ln \text{pr}(\mathbf{g} | \theta) \right\rangle}, \quad (2.18)$$

or equivalently,

$$\text{Var}\{\hat{\theta}\} \geq \frac{1}{\left\langle \left[\frac{\partial}{\partial \theta} \ln \text{pr}(\mathbf{g} | \theta) \right]^2 \right\rangle}. \quad (2.19)$$

Similarly, the variance in the biased case can be written

$$\text{Var}\{\hat{\theta}\} \geq \frac{\left[\frac{db(\theta)}{d\theta} + 1 \right]^2}{\left\langle \left[\frac{\partial}{\partial \theta} \ln \text{pr}(\mathbf{g} | \theta) \right]^2 \right\rangle}. \quad (2.20)$$

We can prove the equivalence between (2.18) and (2.19) by starting with the fact that the integral of any probability density is 1:

$$\int_{\infty} d^M g \, \text{pr}(\mathbf{g} | \theta) = 1. \quad (2.21)$$

Here we will assume that the first and second derivatives of the log-likelihood exist and are absolutely integrable. Differentiating (2.21) with respect to θ and applying the differentiation rule

$$\frac{\partial \text{pr}(\mathbf{g} | \theta)}{\partial \theta} = \frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} \text{pr}(\mathbf{g} | \theta), \quad (2.22)$$

we have

$$\int_{\infty} d^M g \frac{\partial \text{pr}(\mathbf{g} | \theta)}{\partial \theta} = \int_{\infty} d^M g \, \text{pr}(\mathbf{g} | \theta) \frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} = 0. \quad (2.23)$$

Differentiating again and applying (2.22) gives

$$\int_{\infty} d^M g \, \text{pr}(\mathbf{g} | \theta) \frac{\partial^2 \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta^2} + \int_{\infty} d^M g \, \text{pr}(\mathbf{g} | \theta) \left(\frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} \right)^2 = 0, \quad (2.24)$$

which can be expressed in terms of expected values:

$$\left\langle \frac{\partial^2 \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta^2} \right\rangle = - \left\langle \left[\frac{\partial}{\partial \theta} \ln \text{pr}(\mathbf{g} | \theta) \right]^2 \right\rangle. \quad (2.25)$$

Therefore, (2.18) and (2.19) are equivalent.

Proof of the Cramér-Rao inequality can be obtained through the Schwartz inequality (Van Trees, 1968). Without a loss in generality, we will consider the case of a real scalar parameter.

Since the estimate in (2.15) is unbiased, the expectation of the difference between the estimate and the true value of the parameter vanishes:

$$\langle \hat{\theta}(\mathbf{g}) - \theta \rangle_{\mathbf{g}|\theta} \equiv \int_{\infty} d^M g \, \text{pr}(\mathbf{g} | \theta) [\hat{\theta}(\mathbf{g}) - \theta] = 0, \quad (2.26)$$

where the overbar denotes the mean. Differentiating each side with respect to θ and bringing the differentiation into the integral gives

$$\int_{\infty} d^M g \, \frac{\partial}{\partial \theta} \{ \text{pr}(\mathbf{g} | \theta) [\hat{\theta}(\mathbf{g}) - \theta] \} = 0, \quad (2.27)$$

which in turn leads to

$$-\int_{\infty} d^M g \, \text{pr}(\mathbf{g} | \theta) + \int_{\infty} d^M g \, \frac{\partial \text{pr}(\mathbf{g} | \theta)}{\partial \theta} [\hat{\theta}(\mathbf{g}) - \theta] = 0. \quad (2.28)$$

After substituting (2.21) and (2.22) into (2.28), we have

$$\int_{\infty} d^M g \, \frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} \text{pr}(\mathbf{g} | \theta) [\hat{\theta}(\mathbf{g}) - \theta] = 1, \quad (2.29)$$

which can be rewritten as

$$\int_{\infty} d^M g \left\{ \frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} \sqrt{\text{pr}(\mathbf{g} | \theta)} \right\} \left\{ \sqrt{\text{pr}(\mathbf{g} | \theta)} [\hat{\theta}(\mathbf{g}) - \theta] \right\} = 1. \quad (2.30)$$

Now the Schwartz inequality states that for any two functions $f(x)$ and $g(x)$,

$$\left[\int_a^b f(x)g(x)dx \right]^2 \leq \int_a^b f^2(x)dx \int_a^b g^2(x)dx. \quad (2.31)$$

Applying (2.31) to (2.30) leads to

$$\left\{ \int_{\infty} d^M g \text{pr}(\mathbf{g} | \theta) \left[\frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} \right]^2 \right\} \left\{ \int_{\infty} d^M g \text{pr}(\mathbf{g} | \theta) [\hat{\theta}(\mathbf{g}) - \theta]^2 \right\} \geq 1. \quad (2.32)$$

Thus, we have

$$\text{Var}\{\hat{\theta}\} \equiv \langle [\hat{\theta}(\mathbf{g}) - \theta]^2 \rangle_{\mathbf{g}|\theta} \geq \left\langle \left[\frac{\partial}{\partial \theta} \ln \text{pr}(\mathbf{g} | \theta) \right]^2 \right\rangle^{-1}, \quad (2.33)$$

which is equivalent to (2.19) and proves the Cramér-Rao inequality.

We can also demonstrate that if an efficient estimate exists, it is the ML estimate (Melsa & Cohn, 1978; Van Trees, 1968). From the derivation of the Schwartz inequality, we know that the equality in (2.32) holds if and only if

$$\frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} = \alpha(\theta)[\hat{\theta}(\mathbf{g}) - \theta], \quad (2.34)$$

for all $\mathbf{g}$ and θ , where $\alpha(\theta)$ is a constant that depends on θ . Combining this with the likelihood equation in (2.6), we have

$$0 = \left. \frac{\partial \ln \text{pr}(\mathbf{g} | \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}_{\text{ML}}(\mathbf{g})} = \alpha(\theta)[\hat{\theta}(\mathbf{g}) - \theta] \Big|_{\theta = \hat{\theta}_{\text{ML}}(\mathbf{g})}. \quad (2.35)$$

The only data-dependent solution that equates the right-hand side to zero requires that $\hat{\theta}(\mathbf{g}) = \hat{\theta}_{\text{ML}}$. Therefore, the ML estimate is efficient as long as an efficient estimate exists.

Whenever an efficient estimate does not exist, the Cramér-Rao inequality can be improved by computing a larger bound that more accurately depicts the minimum possible variance. While the CRB as stated in (2.18) involves the second partial derivative of the log-likelihood function, the Bhattacharyya bound incorporates higher partial derivatives (Bhattacharyya, 1946, 1947, 1948). Although this procedure is very straightforward, the apparent downside is its computational exhaustiveness, which is prohibitive in most estimation tasks. A bound that offers more practical value is the Barankin bound, since it does not require a differentiable probability density and yields the greatest lower bound (Barankin, 1949; McAulay & Hofstetter, 1971). One major disadvantage, however, is that it requires a maximization over the function of interest,

which is usually not a trivial task. Due to the complexity and impracticality of these alternative bounds, we will restrict our attention to the CRB.

2.4.4 System design

In general, the FIM (and the dispersion matrix) can be computed for any system configuration and therefore used to design and optimize the system that acquires the data to be used as input to the inverse optical design, prior to practical application. We refer to that system as the *inverse-design system*.

Since the FIM is the covariance matrix of a set of parameters, its off-diagonal entries indicate coupling between different pairs of parameters. Strong coupling can lead to great difficulty in the estimation task, and possibly large errors in the parameter estimates. For an efficient estimator, the inverse of the FIM is essentially the covariance matrix of the estimates. Thus, its off-diagonal elements represent coupling between these estimates. One goal in system design is to find a system configuration that lessens the degree of coupling between parameters, while reducing the number of local minima in the likelihood surface. In Chapters 5, 6, and 7, we will investigate the FIMs and dispersion matrices for various system configurations and types of estimable parameters.

2.5 Properties of ML estimators

We begin with a discussion on performance metrics for general estimators from the classical perspective. Then we describe the many optimal properties of maximum-likelihood estimators, including those according to the asymptotic theory of ML

estimation, such as efficiency, consistency, and unbiasedness. Furthermore, we discuss the invariance of the ML estimator under changes in parameterization, plus its ability to best utilize information in the data.

2.5.1 Bias

In classical estimation theory, the *sampling distribution* $\text{pr}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta})$ (Fig. 2.1) is defined as the distribution of $\hat{\boldsymbol{\theta}}(\mathbf{g})$ that is acquired through repeated sampling of the data vector $\mathbf{g}$ from $\text{pr}(\mathbf{g} | \boldsymbol{\theta})$ for fixed $\boldsymbol{\theta}$, then performing the same estimation rule on each sample (Barrett *et al.*, 2007). Since $\hat{\boldsymbol{\theta}}$ is derived from noisy data, it is a random variable that depends on the true value of the parameter. We can implicitly express the mean of a $P \times 1$ vector of estimates in terms of the sampling distribution:

$$\bar{\hat{\boldsymbol{\theta}}} = \int d^P \hat{\boldsymbol{\theta}} \text{pr}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta}) \hat{\boldsymbol{\theta}} . \quad (2.36)$$

If we also know the sampling distribution and estimation rule on $\mathbf{g}$, we can transform (2.36) into the following explicit form:

$$\bar{\hat{\boldsymbol{\theta}}} = \int d^M g \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \hat{\boldsymbol{\theta}}(\mathbf{g}) \equiv \left\langle \hat{\boldsymbol{\theta}}(\mathbf{g}) \right\rangle_{\mathbf{g} | \boldsymbol{\theta}} . \quad (2.37)$$

The *bias* of an estimate is the discrepancy between the expected value of the estimate and the true value of the parameter, and conveys the amount systematic error in the estimation procedure. For a $P \times 1$ vector parameter, the bias is also a $P \times 1$ vector:

$$\mathbf{b}(\theta) \equiv \bar{\hat{\theta}} - \theta = \int_{\infty} d^M \mathbf{g} \, \text{pr}(\mathbf{g} | \theta) [\hat{\theta}(\mathbf{g}) - \theta], \quad (2.38)$$

or in terms of the sampling distribution,

$$\mathbf{b}(\theta) = \int_{\infty} d^P \theta \, \text{pr}(\hat{\theta} | \theta) (\hat{\theta} - \theta). \quad (2.39)$$

Fig. 2.1: Example of a probability distribution of $\hat{\theta}$ conditioned on θ .

An *unbiased* estimate is one whose bias vanishes for all values of the underlying parameter. The concept of unbiasedness will be discussed in Subsection 2.5.4 when we cover the asymptotic properties of ML estimators.

Bias is certainly not the only error in any given estimator, for even an unbiased estimator can generate a bad estimate from a particular data set. We clearly desire estimators with small bias, but the bias itself is removable, even in cases where it is a complicated function of the parameter being estimated and the suitable correction is not always obvious (Gray & Schucany, 1972; Miller, 1964; Quenouille, 1956; Robson & Whitlock, 1964).

2.5.2 Variance and covariance

Another performance metric for an estimator is the variance, which quantifies the amount of random error in the estimator. It results from fluctuations in the estimate $\hat{\theta}$ over multiple trials. Denoting the p^{th} element of the estimate by $\hat{\theta}_p$, the variance of the p^{th} parameter is written

$$\begin{aligned} \text{Var}\{\hat{\theta}_p\} &\equiv \langle |\hat{\theta}_p(\mathbf{g}) - \bar{\hat{\theta}}_p|^2 \rangle_{\mathbf{g}|\theta} = \int_{\infty} d^M g \text{ pr}(\mathbf{g}|\theta) |\hat{\theta}_p(\mathbf{g}) - \langle \hat{\theta}_p(\mathbf{g}) \rangle|^2 \\ &= \int_{\infty} d^P \theta \text{ pr}(\hat{\theta}|\theta) |\hat{\theta}_p - \langle \hat{\theta}_p \rangle|^2, \end{aligned} \quad (2.40)$$

while elements in the general covariance matrix are given by

$$[\mathbf{K}_{\hat{\theta}}]_{pp'} = \langle [\hat{\theta}_p - \bar{\hat{\theta}}_p][\hat{\theta}_{p'} - \bar{\hat{\theta}}_{p'}]^* \rangle_{\mathbf{g}|\theta}. \quad (2.41)$$

Bear in mind that the variance and covariance are for a particular value of the parameter, although they fluctuate with respect to the average estimate.

2.5.3 Mean-square error

The *mean-square error* (MSE) is similar to the variance, except that fluctuations are measured about the true value of the underlying parameter, not the average estimate.

Therefore, the MSE contains information about both the bias and the variance, that is, the overall fluctuation. For a vector parameter, it is given by

$$\text{MSE} \equiv \langle \|\hat{\boldsymbol{\theta}}(\mathbf{g}) - \boldsymbol{\theta}\|^2 \rangle_{\mathbf{g}|\boldsymbol{\theta}} = \int_{\infty} d^M g \text{ pr}(\mathbf{g}|\boldsymbol{\theta}) \|\hat{\boldsymbol{\theta}}(\mathbf{g}) - \boldsymbol{\theta}\|^2. \quad (2.42)$$

For an unbiased estimator, the MSE is equivalent to the variance.

2.5.4 Asymptotic properties

Although ML estimation has broad practical appeal, it sometimes produces inferior results for small sample sizes. However, for a large number of observations, the method possesses many feasible properties.

The properties of an ML estimator which are valid when the estimation error is small are commonly referred to as asymptotic (Van Trees, 1968). One way to analyze the asymptotic properties is to draw M independent observations of the data $\mathbf{g}$ from the sampling distribution $\text{pr}(\mathbf{g}|\boldsymbol{\theta})$, then let $M \rightarrow \infty$, although they also hold when better data

are acquired, such as by acquiring more photons when Poisson noise is dominant or by letting the variance go to zero for Gaussian noise (Barrett *et al.*, 2007).

As mentioned in Section 2.4, an efficient estimate is one that achieves the CRB. We also demonstrated in (2.35) that if an efficient estimator exists, it is the ML estimate. Moreover, the ML estimate is *asymptotically efficient*; in other words, the minimum variance in (2.15) is obtained as the number of samples increases without bound. Thus, for a vector parameter $\boldsymbol{\theta}$, we have

$$\lim_{M \rightarrow \infty} \frac{\text{Var}\{\hat{\boldsymbol{\theta}}\}}{\left\langle \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \right\rangle^{-1}} = 1. \quad (2.43)$$

Efficient and unbiased estimates are typically not obtained for samples of finite size, but when the sample size approaches infinity, we desire an estimate that converges toward the true parameter value. Consider an estimate based on the data $\mathbf{g}$ from M independent observations, denoted as $\hat{\boldsymbol{\theta}}_M(\mathbf{g})$. We say that the estimate is *conditionally consistent* if, for any positive ε and η , no matter how small, there exists some N such that

$$\Pr [\|\hat{\boldsymbol{\theta}}_M(\mathbf{g}) - \boldsymbol{\theta}\| < \varepsilon | \boldsymbol{\theta}] > 1 - \eta \quad (2.44)$$

for all $M > N$. The estimate is *unconditionally consistent* if (2.44) is satisfied for all $\boldsymbol{\theta}$ (Barrett & Myers, 2004). For any given small value of ε , there is a sufficiently large N

such that, for all larger sample sizes, the probability of the error $\Delta\theta = \|\hat{\theta}_M(\mathbf{g}) - \theta\|$ being less than ε is as close to 1 as we like. The estimate $\hat{\theta}_M$ is said to *converge in probability*, or to *converge stochastically*, to the true value θ (Kendall & Stuart, 1979). Thus, the distribution of a consistent estimate becomes increasingly narrow about the true value of the parameter as the number of observations increases. Cramer (1946) proved that over a broad range of conditions, the ML estimate is consistent.

Note that the property of consistency is concerned with the behavior of an estimator as the number of observations tends to infinity, but requires nothing of the behavior for a finite set of observations. If the mean of $\hat{\theta}_M$, as expressed in (2.36) or (2.37), is equal to θ for all θ and N (Kendall & Stuart, 1979), the estimator is unbiased. The terminology for unbiasedness was introduced by Neyman and Pearson (1936) in the context of hypothesis testing.

An estimable parameter is one for which there exists an unbiased estimator for all true values of the parameter. However, even if an estimate is unbiased for a certain value of the parameter θ , it is not necessarily unbiased under reparameterization for nontrivial functions of θ (Bard, 1974).

The properties of unbiasedness and consistency do not imply each other; that is, an unbiased estimate is not automatically consistent, while a consistent estimator is not automatically unbiased. Nonetheless, a consistent estimator whose asymptotic distribution has a finite mean must also be asymptotically unbiased (Kendall & Stuart, 1979).

Fisher established that the sampling distribution of an efficient estimate approaches a Gaussian distribution with minimum variance as the number of samples increases (Fisher, 1922). It can also be shown under general conditions that the sampling distribution of an ML estimate is asymptotically Gaussian due to the central-limit theorem (Cramér, 1946; Daniels, 1961; Huber, 1967; Lecam, 1970), which states that the distribution of the sum of M independent random variables approaches the Gaussian (or normal), distribution as M is made sufficiently large.

To summarize, ML estimators are asymptotically efficient, unbiased, consistent, and normally distributed. Despite the motivation for using the ML estimator, one might wonder if there exists an estimation technique that outperforms the ML procedure. Even if an efficient estimate does not exist, there could possibly be an unbiased estimate with a lower variance. The caveat is that there is no general rule for discovering one. For a given estimation task, we can make attempts to improve the ML estimator, although the resultant process is typically more complicated and difficult to implement. We therefore embrace the ML approach for its relative simplicity, as well as its optimal use of information in the data, as we shall see in the following sections.

2.5.5 Invariance

A very useful property of the ML estimator is its *invariance* under a change in parameterization (Tan & Drossos, 1975), so that we can estimate some function of the parameter θ , rather than the actual θ . Suppose that $\mathbf{f}(\theta)$ is an invertible single-valued

function defined for all $\boldsymbol{\theta}$, where $\mathbf{f}$ is a vector of functions. It can be shown that the ML estimator of $\mathbf{f}$, denoted $\hat{\mathbf{f}}_{\text{ML}}$, is given by (Melsa & Cohn, 1978)

$$\hat{\mathbf{f}}_{\text{ML}} = \mathbf{f}(\hat{\boldsymbol{\theta}}_{\text{ML}}). \quad (2.44)$$

The proof begins with the inverse of $\mathbf{f}$, denoted $\mathbf{f}^{-1}$, such that $\mathbf{f}^{-1}[\mathbf{f}(\boldsymbol{\theta})] = \boldsymbol{\theta}$ for all $\boldsymbol{\theta}$. The probability density of the data $\mathbf{g}$, conditioned on $\boldsymbol{\theta}$, can then be written as

$$\text{pr}(\mathbf{g} | \mathbf{f}) = \text{pr}[\mathbf{g} | \mathbf{f}^{-1}(\mathbf{f})]. \quad (2.45)$$

If we let $\mathbf{f}^* = \mathbf{f}(\hat{\boldsymbol{\theta}}_{\text{ML}})$, then (2.43) transforms into

$$\text{pr}(\mathbf{g} | \mathbf{f}^*) = \text{pr}(\mathbf{g} | \hat{\boldsymbol{\theta}}_{\text{ML}}). \quad (2.47)$$

Now the definition of the ML estimate $\hat{\boldsymbol{\theta}}_{\text{ML}}$ states that

$$\text{pr}(\mathbf{g} | \hat{\mathbf{f}}_{\text{ML}}) \geq \text{pr}(\mathbf{g} | \mathbf{f}) \quad (2.48)$$

for all $\mathbf{f} \neq \hat{\mathbf{f}}_{\text{ML}}$, or equivalently,

$$\text{pr}(\mathbf{g} | \hat{\boldsymbol{\theta}}_{\text{ML}}) \geq \text{pr}(\mathbf{g} | \boldsymbol{\theta}) \quad (2.49)$$

for all $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}_{\text{ML}}$. Therefore, it must also hold that

$$\text{pr}(\mathbf{g} | \mathbf{f}^*) \geq \text{pr}(\mathbf{g} | \mathbf{f}) \quad (2.50)$$

for all $\mathbf{f} \neq \mathbf{f}^*$, so that $\hat{\mathbf{f}}_{\text{ML}} = \mathbf{f}^* = \mathbf{f}(\hat{\boldsymbol{\theta}}_{\text{ML}})$, thereby proving that the ML estimate of a function is simply the function evaluated at the ML estimate.

2.5.6 Sufficiency

In estimation, a *sufficient statistic* is one that extracts all relevant information from the data and optimizes the performance of a particular estimation task (Fisher, 1922, 1925). The maximum-likelihood estimator is a sufficient statistic, since it makes optimal use of the information in the data (Barrett & Myers, 2004).

A necessary and sufficient condition for $\hat{\boldsymbol{\theta}}$ to be a sufficient estimate is that there exists a factorization

$$L(\boldsymbol{\theta} | \mathbf{g}) \equiv \text{pr}(\mathbf{g} | \boldsymbol{\theta}) = \text{pr}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta}) f(\mathbf{g}), \quad (2.51)$$

where $\text{pr}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta})$ is a function of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ alone and $f(\mathbf{g})$ is independent of $\boldsymbol{\theta}$. We see from (2.51) that the choice of $\hat{\boldsymbol{\theta}}$ to maximize the log-likelihood is equivalent to choosing $\hat{\boldsymbol{\theta}}$ to maximize $\text{pr}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta})$. This is a special case of the *Neyman-Fisher factorization criterion*, originally established by Fisher (1922), after which Neyman (1935) developed a method of finding sufficient statistics. The proof of this criterion is beyond the scope of this paper, but a rigorous proof can be found in Halmos and Savage (1949).

The condition for sufficiency in (2.51) has a very interesting consequence.

Taking the logarithm of both sides and differentiating leads to

$$\frac{\partial \ln L(\boldsymbol{\theta} | \mathbf{g})}{\partial \boldsymbol{\theta}} \equiv \frac{\partial \ln \text{pr}(\mathbf{g} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ln \text{pr}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \quad (2.52)$$

By comparing (2.52) with (2.35), the condition of efficiency for the log-likelihood function, we find that such an estimator can exist if and only if there is a sufficient statistic. In other words, as long as (2.35) is satisfied, (2.52) is also satisfied. Thus, the criterion of efficiency is more restrictive than that for sufficiency. In contrast, even if (2.35) does not hold, we may still have a sufficient estimator.

Example: Normal data with an unknown mean

Consider a data vector $\mathbf{g}$ that contains M i.i.d. samples of a normal process with mean μ and variance σ^2 , where μ is unknown. Our goal is to factor the likelihood function into the form in (2.51).

The conditional PDF of the data is written as

$$L(\mu | \mathbf{g}) \equiv \text{pr}(\mathbf{g} | \mu) = \prod_{m=1}^M \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left[-\frac{1}{2} \frac{(g_m - \mu)^2}{\sigma^2} \right], \quad (2.52)$$

or equivalently,

$$L(\mu | \mathbf{g}) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{M}{2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{m=1}^M (g_m - \mu)^2 \right]. \quad (2.53)$$

We would like to rewrite (2.53) into the form $\text{pr}(\hat{\mu} | \mu) f(\mathbf{g})$ by implementing a trick when manipulating normal densities. The sample mean is defined as

$$\hat{\mu} \equiv \bar{g} = \frac{1}{M} \sum_{m=1}^M g_m, \quad (2.54)$$

which leads to

$$\begin{aligned} L(\mu | \mathbf{g}) &= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{M}{2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{m=1}^M (g_m - \bar{g} + \bar{g} - \mu)^2 \right] \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{M}{2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{m=1}^M [(g_m - \bar{g})^2 + 2(g_m - \bar{g})(\bar{g} - \mu) + (\bar{g} - \mu)^2] \right]. \end{aligned} \quad (2.55)$$

Now observe that the middle term vanishes:

$$\begin{aligned} \sum_{m=1}^M (g_m - \bar{g})(\bar{g} - \mu) &= (\bar{g} - \mu) \sum_{m=1}^M (g_m - \bar{g}) \\ &= M(\bar{g} - \mu)(\bar{g} - \bar{g}) \\ &= 0. \end{aligned} \quad (2.56)$$

The likelihood function becomes

$$L(\mu | \mathbf{g}) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{M}{2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{m=1}^M (g_m - \bar{g})^2 \right] \exp \left[-\frac{1}{2\sigma^2} \sum_{m=1}^M (\bar{g} - \mu)^2 \right], \quad (2.57)$$

where the portion independent of μ is given by

$$f(\mathbf{g}) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{M}{2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{m=1}^M (g_m - \bar{g})^2 \right], \quad (2.58)$$

so that the sampling distribution is

$$\text{pr}(\hat{\mu} | \mu) = \exp \left[-\frac{1}{2\sigma^2} \sum_{m=1}^M (\hat{\mu} - \mu)^2 \right], \quad (2.59)$$

which depends on just $\hat{\mu} \equiv \bar{g}$ and μ . Therefore, the sample mean is a one-dimensional sufficient statistic for the mean.

2.6 Computer simulated experiments

Proof of principle can be demonstrated with data from a real experiment or obtained through numerical simulation of a real physical system. Simulated experiments can help to determine whether a system design or estimation procedure is likely to succeed, while they may be useful in optimizing system configurations prior to practical application.

The following procedure can be used to determine properties of a sampling distribution $\text{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$ through a set of simulated experiments:

1. Define the forward model in the design program, denoted as $f(\boldsymbol{\theta})$, and the probability distribution of the errors. Assign “true” values to the vector set of parameters $\boldsymbol{\theta}_{\text{true}}$.

2. Generate a different vector set of errors $\mathbf{e}_{\mu}$ for each of N experiments, where μ is the experiment index, drawn from the specified probability distribution. Many computers have available routines for producing *pseudorandom numbers*, which are effectively random numbers uniformly distributed between 0 and 1. These are then used to generate random numbers according to any other desired distribution. The simulated data $\mathbf{g}_{\mu}$ are obtained by adding $\mathbf{e}_{\mu}$ to the output of the design program at the true parameters $f(\boldsymbol{\theta}_{\text{true}})$:

$$\mathbf{g}_{\mu} = f(\boldsymbol{\theta}_{\text{true}}) + \mathbf{e}_{\mu}. \quad (2.60)$$

3. For each experiment, apply the estimation procedure to the simulated data as if they were real data. Each replicated experiment yields an estimate $\hat{\boldsymbol{\theta}}_{\mu}$ for the parameters.

4. Properties of the sampling distribution are obtained by averaging over the N replications. The estimated mean and covariance matrix of the sampling distribution are respectively given by

$$\bar{\hat{\boldsymbol{\theta}}} = \frac{1}{N} \sum_{\mu=1}^N \hat{\boldsymbol{\theta}}_{\mu} , \quad (2.61)$$

$$\hat{\mathbf{K}} = \frac{1}{N-1} \sum_{\mu=1}^N (\hat{\boldsymbol{\theta}}_{\mu} - \bar{\hat{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}}_{\mu} - \bar{\hat{\boldsymbol{\theta}}})^t . \quad (2.62)$$

The estimated bias $\mathbf{b}$ of the estimator is written as

$$\mathbf{b} = \bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}_{\text{true}} . \quad (2.63)$$

The equations above apply whether the data are simulated or real, although an advantage of simulated experiments is that the true parameters underlying the data are exactly known.

Computer simulated experiments would allow us to examine the effects of a *model-mismatch*, that is, we can quantify how much estimation error results from deficiencies in the forward model, by purposely using a different model in the estimation procedure than was used to generate the data. In doing so, we can determine the robustness of the estimator (Bard, 1974).

2.7 Nuisance parameters

A *nuisance parameter* is one that influences the data, but is of no immediate interest to the estimation problem. However, it must be factored into the likelihood function in

order to completely specify the PDF on the data, otherwise the results can confound the estimation problem.

Suppose we let

$$\theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix},$$

where α represents the parameters of interest and β represents the nuisance parameters.

Barrett *et al.* (2007) propose the following options for handling β :

1. Disregard the problem and let $\text{pr}(\mathbf{g}|\theta) \approx \text{pr}(\mathbf{g}|\alpha)$.
2. Replace β with a suitable value β_0 and let $\text{pr}(\mathbf{g}|\alpha, \beta) \approx \text{pr}(\mathbf{g}|\alpha, \beta_0)$.
3. Estimate β independently from an auxiliary data set, then apply method (2).
4. Assume or measure a prior distribution $\text{pr}(\beta)$, then marginalize over β .
5. Estimate α and β simultaneously and ignore the estimate of β .

Each of these options leads to considerable practical issues. The first approach is essentially equivalent to ignoring modeling errors, which would undoubtedly lead to errors in the estimates of α . The second option is no different from assuming a prior distribution $\text{pr}(\beta)$ and treating it as a delta function, which would clearly be a strong and unrealistic prior on β . The third approach may lead to better estimates of α , but it requires an additional estimation problem altogether.

Barrett and Myers (2004) show that the optimal strategy is to marginalize over the nuisance parameters rather than estimate them, as in the fourth approach. However, this assumes a meaningful prior distribution $\text{pr}(\boldsymbol{\beta})$, not one that is simply based on belief or selected for mathematical ease.

Now we will discuss the fifth approach, which is to simultaneously estimate both the parameters of interest and the nuisance parameters, and we shall examine the consequences on the variance on the estimates. We showed in (2.15) that the minimum variance on an estimate of the p^{th} parameter θ_p is given by

$$\text{Var}\{\hat{\boldsymbol{\theta}}_p\} \geq [\mathbf{F}^{-1}]_{pp}, \quad (2.64)$$

where $\boldsymbol{\theta}$ is a $P \times 1$ vector parameter and $\mathbf{F}$ is the $P \times P$ Fisher information matrix for $\boldsymbol{\theta}$.

Suppose only θ_p is unknown and needs to be estimated, for sake of this discussion. Then the minimum variance on the p^{th} parameter becomes

$$\text{Var}\{\hat{\boldsymbol{\theta}}_p\} \geq [\mathbf{F}^{-1}]_{pp}, \quad (2.65)$$

For real vectors $\mathbf{a}$ and $\mathbf{b}$, the extended Cauchy-Schwarz inequality in matrix form is given by

$$|\mathbf{a}^t \mathbf{b}|^2 \leq (\mathbf{a}^t \mathbf{K} \mathbf{a})(\mathbf{b}^t \mathbf{K}^{-1} \mathbf{b}), \quad (2.66)$$

where $\mathbf{K}$ is a positive-definite matrix and $\mathbf{a}^t$ is the transpose of the column vector $\mathbf{a}$. The equality is true if and only if $\mathbf{b} = c\mathbf{Ka}$, where c is a scalar. Since $\mathbf{F}$ is positive-definite, we can write

$$|\mathbf{a}^t \mathbf{b}|^2 \leq (\mathbf{a}^t \mathbf{F} \mathbf{a})(\mathbf{b}^t \mathbf{F}^{-1} \mathbf{b}). \quad (2.67)$$

Now consider $\mathbf{e}_p$, the column vector $\{0, \dots, 1, \dots, 0\}^t$ with the p^{th} component equal to 1 and all other components equal to zero. For $\mathbf{a} = \mathbf{b} = \mathbf{e}_p$, we have

$$|\mathbf{e}_p^t \mathbf{e}_p|^2 \leq (\mathbf{F}_{pp})(\mathbf{F}^{-1})_{pp}. \quad (2.68)$$

Since the left side is equal to one, (2.68) gives

$$\frac{1}{(\mathbf{F}_{pp})} \leq (\mathbf{F}^{-1})_{pp}. \quad (2.69)$$

Therefore, the minimum variance on the estimated parameter when nuisance parameters are absent is less than the minimum variance when nuisance parameters are present.

2.8 Gaussian distributions and electronic noise

In any practical electronic system, a substantial number of electrons contribute to overall fluctuations in nearly an independent fashion. Therefore, Gaussian statistics provides an accurate representation of electronic noise by virtue of the central-limit theorem. We first

propose a general form for the PDF, making no simplifying assumptions regarding noise characteristics. Then we introduce assumptions that simplify the probability model and suggest a tractable, yet realistic, PDF on the noise.

If we assume a discrete array of detector elements with electronic coupling between elements, then the noise is correlated. The PDF for electronic noise in the absence of other noise sources is a multivariate Gaussian, given by

$$\text{pr}(\mathbf{g} | \boldsymbol{\theta}) = \frac{1}{(2\pi)^{M/2} \{\det[\mathbf{K}(\boldsymbol{\theta})]\}^{1/2}} \exp\left\{-\frac{1}{2}[\mathbf{g} - \bar{\mathbf{g}}(\boldsymbol{\theta})]^t [\mathbf{K}(\boldsymbol{\theta})]^{-1} [\mathbf{g} - \bar{\mathbf{g}}(\boldsymbol{\theta})]\right\}, \quad (2.70)$$

where $\mathbf{K}(\boldsymbol{\theta})$ is the covariance matrix conditioned on $\boldsymbol{\theta}$, $\det[\mathbf{K}(\boldsymbol{\theta})]$ is its determinant, and both the mean and covariance are functions of $\boldsymbol{\theta}$.

If we assume that the detector elements are uncorrelated, then the covariance matrix has diagonal components equal to the variances in the detector elements

$$[\mathbf{K}(\boldsymbol{\theta})]_{mm} = \sigma_m^2, \quad (2.71)$$

with all other components equal to zero, where σ_m^2 is the variance on the noise at the m^{th} detector element. Moreover, the determinant of a matrix is the product of its eigenvalues, and for a diagonal matrix, the eigenvalues are the diagonal entries:

$$\det[\mathbf{K}(\boldsymbol{\theta})] = \prod_{m=1}^M \sigma_m^2. \quad (2.72)$$

The signal generated by the optical illumination is not zero-mean, but assuming the noise is independent of the illumination, the illumination simply shifts the PDF on the noise. Therefore, only the mean data vector in the PDF depends on the underlying parameters. Finally, the PDF reduces to a product of univariate PDFs, written as

$$\text{pr}(\mathbf{g} | \boldsymbol{\theta}) = \prod_{m=1}^M \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left\{-\frac{[g_m - \bar{g}_m(\boldsymbol{\theta})]^2}{2\sigma_m^2}\right\}, \quad (2.73)$$

where g_m is the measured signal at the m^{th} element. Note that the uncorrelated components are also statistically independent, which is true only for a normal random vector, although the reverse is always true (Barrett & Myers, 2004).

Barrett *et al.* (2007) suggest that the PDF given in (2.73) provides a more accurate representation of the data, since the pixels in commercial CCD detectors may have significant variation in dark current and responsivity. These effects can be corrected on average with digital post-processing by measuring and subtracting a dark current map, then dividing by a gain map, but this process does not result in a uniform variance across the pixels. For instance, a pixel with low response would be divided by a small gain factor, which would actually enhance the variance non-uniformity. The PDF could be utilized by measuring the variance in each element after corrections are made on the data.

If there is good variance uniformity, we can treat the detector elements as identical with constant variance. Thus, the noise is modeled as i.i.d. zero-mean Gaussian:

$$\text{pr}(\mathbf{g} | \boldsymbol{\theta}) = \prod_{m=1}^M \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[g_m - \bar{g}_m(\boldsymbol{\theta})]^2}{2\sigma^2}\right\}. \quad (2.74)$$

To perform ML estimation, it is especially convenient in this case to take the logarithm of the PDF, given by

$$\ln \text{pr}(\mathbf{g} | \boldsymbol{\theta}) = -\frac{1}{2}M \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{m=1}^M [g_m - \bar{g}_m(\boldsymbol{\theta})]^2. \quad (2.75)$$

We immediately observe that the first term on the right is a constant, and the second term is a sum of squares preceded by a minus sign. So, maximizing the likelihood for Gaussian i.i.d. data reduces to nonlinear least-squares fitting between the measured data and the average data vector:

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{m=1}^M [g_m - \bar{g}_m(\boldsymbol{\theta})]^2. \quad (2.76)$$

For Gaussian i.i.d. data, we will derive the FIM components using (2.14) and (2.75):

$$F_{jk} = \left\langle \frac{\partial^2}{\partial \theta_j \partial \theta_k} \left\{ \frac{1}{2}M \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{m=1}^M [g_m - \bar{g}_m(\boldsymbol{\theta})]^2 \right\} \right\rangle_{\mathbf{g}|\boldsymbol{\theta}}. \quad (2.77)$$

Since the first term in the curly brackets is a constant, we have

$$F_{jk} = \frac{1}{2\sigma^2} \left\langle \sum_{m=1}^M \frac{\partial^2}{\partial \theta_j \partial \theta_k} [g_m - \bar{g}_m(\boldsymbol{\theta})]^2 \right\rangle_{\mathbf{g}|\boldsymbol{\theta}}. \quad (2.78)$$

Carrying out the differentiation gives

$$\begin{aligned} F_{jk} &= -\frac{1}{\sigma^2} \left\langle \sum_{m=1}^M \frac{\partial}{\partial \theta_j} \left\{ [g_m - \bar{g}_m(\boldsymbol{\theta})] \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_k} \right\} \right\rangle_{\mathbf{g}|\boldsymbol{\theta}} \\ &= -\frac{1}{\sigma^2} \left\langle \sum_{m=1}^M \left\{ -\frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_k} + [g_m - \bar{g}_m(\boldsymbol{\theta})] \frac{\partial^2}{\partial \theta_j \partial \theta_k} \bar{g}_m(\boldsymbol{\theta}) \right\} \right\rangle_{\mathbf{g}|\boldsymbol{\theta}}. \end{aligned} \quad (2.79)$$

Since the data are zero-mean, the second term vanishes when averaging over the data,

which leads to

$$F_{jk} = \frac{1}{\sigma^2} \sum_{m=1}^M \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_k}. \quad (2.80)$$

Similarly, for uncorrelated but non-identical pixels as described by (2.71), we have

$$F_{jk} = \sum_{m=1}^M \frac{1}{\sigma_m^2} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_k}. \quad (2.81)$$

The FIM components in (2.80) and (2.81) depend solely on the average data vector evaluated at the underlying parameter and are inversely proportional to the variance. Therefore, lower noise levels yield greater Fisher information.

2.9 Practical challenges

In practice, measurement procedures have limited accuracy, repeated measurements of the same quantity produce different values, and the idealized conditions under which the model was derived are never perfectly achievable. We also know that unpredicted randomness, not taken into account in deterministic models, always occurs. The randomness is a part of physical reality as much as the values of parameters underlying the data, so the model would be incomplete without an accurate description of the random phenomena. One major challenge in ML estimation is that an accurate probability model must be used that includes all sources of randomness. For instance, one might incorporate a Poisson distribution of errors to emulate photon noise, Gaussian statistics for electronic noise, or a combination thereof. Since the systems discussed in this research are not light-starved as in astronomical applications, we will adhere to a Gaussian noise model.

Careful attention must also be made to avoid misalignments of the test optics, since such misalignments would manifest as errors in the estimates. One solution is to treat uncertain alignments, distances, magnifications, and so on, in the system as nuisance parameters, then utilize one of the methods in Section 2.6 for their treatment.

We mentioned in Chapter 1 that a significant limitation in inverse optical design is that the ML estimation step is very computationally intensive, particularly for complicated probability surfaces. Making the method practical requires rapid processing techniques and dedicated computer hardware. Several hardware platforms are available for implementing parallel algorithms, including the graphics processing unit (GPU) in video cards, capable of massively parallel high-performance computing in scientific and engineering fields. We will elaborate on GPU technology in Section 6.2.

CHAPTER 3

OPTIMIZATION METHODS

As we discussed in Chapter 2, algorithms that implement the ML approach to estimating parameters θ perform a search through the space defined by all possible values to find a point that maximizes the probability of generating the observed data $\mathbf{g}$:

$$(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3 \dots) = \arg \max_{\theta_1, \theta_2, \theta_3 \dots} [\log \text{pr}(\mathbf{g} \mid \theta_1, \theta_2, \theta_3 \dots)] . \quad (3.1)$$

Practically all search (i.e., optimization) algorithms locate an extremum through iterative methods that execute in a variable number of steps depending on the starting location, the complexity of the probability surface, and values selected for convergence factors.

In Section 3.1, we describe various considerations in selecting a suitable search algorithm, while distinguishing between local and global algorithms. Section 3.2 provides a qualitative overview of global optimization algorithms and sets stochastic algorithms apart from deterministic ones.

Much of this chapter is dedicated to the simulated annealing (SA) algorithm for global optimization, covered in Section 3.3. We begin with a general overview of SA, including its many desirable properties and ability to handle extremely complicated functions. Embedded in each phase of the SA algorithm is the Metropolis algorithm, which was originally developed in the context of statistical mechanics. We discuss the

fundamental concepts of statistical mechanics that are vital to understanding the origins of SA, as well as the applicability to general optimization problems based on an interesting analogy to thermodynamics. Lastly, we describe in detail the specific SA algorithm used in this research for continuous minimization of multimodal functions. We explain different factors in choosing an appropriate annealing schedule.

We saw in Chapter 2 that ML estimation reduces to nonlinear least-squares fitting (i.e., minimizing the squares) for an i.i.d. Gaussian noise model, so we will regard the optimization problem as minimization for sake of this discussion. Nothing is sacrificed by restricting our attention to minimization, since maximizing a function is equivalent to minimizing its negative.

3.1 Selecting a search algorithm

If the objective (or cost) function of interest is well-behaved and unimodal within a specified domain, there are many search algorithms to choose from to solve the optimization problem. Direct search algorithms, which only accept downhill moves along the surface to be minimized, rely exclusively on values of the objective function, therefore they are relatively easy to implement. Classical direct-search methods, at least in the realm of unconstrained minimization, include pattern search methods (Hooke & Jeeves, 1961; Polak, 1971; Davidon, 1991; Torczon, 1997), simplex methods (Spendley, Hext, & Himsworth 1962; Nelder & Mead, 1965), and methods with adaptive sets of search directions (Rosenbrock, 1960; Powell, 1964, 1965). Fletcher (1965) provides an excellent review of direct search algorithms.

Other available search algorithms incorporate the derivatives of the objective function, such as the gradient descent or conjugate gradient method (Fletcher & Reeves, 1963, 1964; Hestenes & Stiefel, 1952; Hestenes, 1969) and Newton's method or quasi-Newton methods (Greenstadt, 1967; Spang, 1962). Although derivative-based methods are often faster and more reliable than direct-search algorithms, they are liable to terminate far from the true solution if the objective function is ill-conditioned (Corana, Marchesi, Martini, & Ridella, 1987).

Conversely, if the objective function contains multiple minima, straightforward minimization will terminate at the quickest nearby solution, even if the global optimum exists some distance away. A multimodal function can occur for a number of reasons, for instance, if the mean data vector is a nonlinear functional of the parameters (Barrett & Myers, 2004). In inverse optical design, strong coupling occurs between various pairs of prescription parameters, resulting in a complicated probability surface with many local minima. The effects of parametric coupling are also evidenced in the Fisher information matrix, many of which will be presented in Chapters 5 – 7. Since the amount of local minima in the domain of interest can increase exponentially with the number of parameters to be estimated, an exhaustive search for the global minimum would require an impractical number of function evaluations. Thus the search must be limited somehow, either deterministically or stochastically.

Fig. 3.1: A multimodal test function in two dimensions, exhibiting a high degree of nonlinearity and various local minima.

3.2 Global optimization algorithms

Global optimization algorithms commonly have two phases. An exhaustive global search in parameter space is performed, iteratively identifying a promising starting point to be used in a local search. A local minimum is then determined from each starting point, usually through a deterministic local descent algorithm that is executed by the global phase. This is performed as a black-box procedure without deeper insight into the local structure of the objective function, which makes the process amenable to broad classes of

problems. The global phase is either stochastic or deterministic. In contrast to deterministic global phases, stochastic global phases are often heuristic by nature (Liberti & Maculan, 2006).

Deterministic algorithms usually operate with a divide and conquer scheme; the search space is recursively partitioned into smaller subspaces. Each subspace is then solved for globally, and upper and lower bounds of the objective function in the subspace are computed to test for optimality of the local minimum. If the difference between the bounds is less than a specified threshold, the local optimum is regarded as the global solution. In every step of a deterministic algorithm, there exists at most one way to proceed; if no way to proceed exists, the algorithm terminates. Since deterministic algorithms do not use random numbers in their instructions, a given input always produces the same output. Although deterministic algorithms can guarantee optimality and precision in their solutions, they are less efficient in problems with high dimensionality or complicated features, and perform at their best on small- to medium-scale problems with more obvious algebraic formulations.

Stochastic global phases find the starting points for local searches through random sampling, or by trying to escape from the local basin of the current local minimum, or by employing a combination of the two techniques. Stochastic algorithms include at least one instruction that incorporates random numbers, thus violating the condition of determinism. Although stochastic global phases do not guarantee a specified amount of optimality in their solutions, they do guarantee asymptotic convergence to the local

minimum as the number of evaluated points in the search space increases. They also tend to be very efficient, although the level of efficiency strongly depends on tuning parameters such as the sampling intensity, escaping capacity, and termination criteria (Liberti & Maculan, 2006). These parameters are usually found empirically and will vary depending on the behavior of the objective function.

Examples of deterministic global phases are Branch-and-Select (Tuy, 1998), spatial Branch-and-Bound (Falk & Soland, 1969), and state space search. Some of the many examples of stochastic global optimization algorithms are multistart, genetic algorithms (Goldberg & Richardson, 1987), differential evolution (Storn & Price, 1997), adaptive Lagrange multiplier methods (Wah & Wang, 1999), dynamic tunneling methods (Levy & Montalvo, 1985; RoyChowdury, Singh, & Chansarkar, 2000), and variable neighborhood search (Hansen & Mladenović, 2001, 2002). Another example is simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983, 1984), which was investigated for inverse optical design.

3.3 Simulated annealing

3.3.1 Overview

Simulated annealing is a feasible candidate for the task of global optimization, since it provides a good approximation to the global optimum of a function in a high-dimensional search space with many local extrema. In fact, it has vast utility in large-scale problems with up to tens of thousands of variables (Kirkpatrick *et al.*, 1983, 1984; Romeo,

Vincentelli, & Sechen 1984; Smith, Barrett, & Paxman, 1983; Smith, Paxman, & Barrett, 1985; White, 1984). Simulated annealing is a stochastic algorithm that allows transitions out of a local optimum during the search based on a probability criterion. It is able to process objective functions with high degrees of nonlinearities, discontinuities, and randomness due to noise sources. It can also distinguish between gross features and finer “wrinkles” in the function. Macroscopic features of the eventual state of the system appear earlier in the search process, and the system explores high-cost configurations, irrespective of small local minima; finer details develop later in the search, when the system is less likely to escape from the current local basin. Since simulated annealing performs constrained optimization, the search space can contain arbitrary boundary conditions and constraints, allowing one to incorporate *a priori* information about the system.

Although simulated annealing guarantees the true solution under very stringent conditions, satisfying these requirements would lead to the global optimum much too slowly for practical use; the conditions are instead relaxed to trade-off computational time and optimality of the solution (Barrett & Myers, 2004). Even with this compromise, if simulated annealing does not find the true solution, it will find a near-optimal one (Corana *et al.*, 1987). While simulated annealing is very promising in many aspects, its major criticism is high computational demand compared to straightforward optimization methods.

Simulated annealing was originally developed by Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and Teller (1953) from the perspective of statistical mechanics, an application of probability theory in thermal physics, which includes a mathematical framework for dealing with large populations of atoms or molecules. Metropolis created a simple algorithm to simulate the thermodynamic equation of state for a complex system of atoms at a given temperature in thermal equilibrium. Since the number of atoms is on the order of 10^{23} per cubic centimeter, ensemble averages were replaced by sample averages through Monte Carlo sampling, more specifically, the technique known as *Markov-chain Monte Carlo* (MCMC).

It was later discovered by Kirkpatrick *et al.* (1983, 1984) that the Metropolis algorithm could be applied to general optimization problems by using its underlying concepts to simulate the physical process of *annealing* in materials science. Annealing is the heating of a substance, such as a crystal, to a molten state, followed by a gradual reduction in temperature until the crystalline structure is frozen in. Each temperature should be held long enough for the substance to reach equilibrium, and more time must be spent near the freezing point. If done properly, the material will reach a crystalline state of high order and translational symmetry, with the ground state being the configuration of perfect order and minimum energy. However, *quenching* will occur if the temperature is lowered too rapidly; the substance will depart from equilibrium and the crystal will lock in many irregularities and defects, with an energy level higher than that of a perfect crystal.

Kirkpatrick was particularly interested in the optimal design of integrated circuits on computer chips, densely wired with an elaborate network of interconnections and electronic gates. The design variables involved the placement of the gates and partitioning of electrical components, while the objective function was a measure of system performance. In the analogy to thermodynamics, the configuration of gates and components correspond to the atomic positions in a gas or liquid, while the objective function corresponds to energy. Thus the state of lowest cost represents the ground state, or the state of lowest energy. Kirkpatrick essentially wanted to minimize the length of connections, given that wire lengths were proportionate to time delays in signal propagation. However, the configuration with the shortest possible wires did not necessarily give the best solution, because this would likely lead to congestion and noise such as interference between nearby wires. Since the objective function was defined in a discrete domain, or *configuration space*, this was a problem in combinatorial optimization. Nevertheless, the method was later modified to perform continuous optimization of objective functions of continuous variables, which will be described in Section 3.5

The classic example of combinatorial optimization is the traveling salesman problem, described by Kirkpatrick *et al.* (1983): “Given a list of N Cities and a means of calculating the cost of traveling between any two cities, one must plan the salesman’s route, which will pass through each city once and return finally to the starting point, minimizing the total cost.” Many problems involving scheduling and design, such as in computer science and engineering are akin to the traveling salesman problem. Two

secondary problems are to predict the expected cost of the salesman's route, and to estimate the computing effort required to determine the route.

Fig. 3.2: Illustration of the travelling salesman problem and its solution.

3.3.2 Basic concepts in statistical mechanics

In an ensemble of N identical particles, such as in a gas or liquid, the state of the n^{th} particle is defined by its position $\mathbf{r}_n$ and velocity $\mathbf{v}_n$ in 3D space. Each possible system configuration is described by a set of $6N$ coordinates, which can be regarded as a point in $6N$ -dimensional phase space (Barrett & Myers, 2004). The energy of the ensemble is denoted by $\mathcal{E}(\{\mathbf{r}_n, \mathbf{v}_n\})$, where the brackets comprise the set of N particles; in the j^{th} state

of the system, $\mathcal{E}_j = \mathcal{E}(\{\mathbf{r}_{nj}, \mathbf{v}_{nj}\})$. Due to the random behavior of the system, the energy fluctuates about some average value, where the average is taken over the ensemble of identical systems. Assuming the system is in thermal contact with a heat bath, both the temperature and mean energy remain constant.

One of the most useful functions in statistical mechanics is the *partition function*, defined as

$$Z \equiv \sum_j \exp\left(-\frac{\mathcal{E}_j}{k_B T}\right) = \sum_j \exp\left(-\frac{\mathcal{E}_j}{\tau}\right). \quad (3.2)$$

The summation is over all possible states of the system, where each state is weighted by its Boltzmann factor, given by $\exp(-\mathcal{E}_j / k_B T)$. Here, $\mathcal{E}_j$ is the energy of the system in state j , k_B is Boltzmann's constant with units of energy per kelvin, and T is the absolute temperature in kelvin. Note that the fundamental temperature $\tau = k_B T$ differs from the absolute temperature by a scale factor k_B and has units of energy. The partition function is the normalizing factor between the probability of the system in state j at thermal equilibrium $\text{Pr}(j)$ and the respective Boltzmann factor:

$$\text{Pr}(j) = \frac{1}{Z} \exp\left(-\frac{\mathcal{E}_j}{\tau}\right). \quad (3.3)$$

Derivatives, as well as logarithms, of the partition function lead to a series of important thermodynamic quantities. For instance, the mean energy of a system is obtained by averaging over all states of the ensemble:

$$\langle \varepsilon(\tau) \rangle \equiv \bar{\varepsilon}(\tau) = \sum_j \varepsilon_j \text{Pr}(j) . \quad (3.4)$$

This ensemble average energy represents the states of a system that can exchange energy with a reservoir. Using the relationship

$$\frac{\partial \ln Z}{\partial Z} = \frac{1}{Z} , \quad (3.5)$$

we have

$$\begin{aligned} \bar{\varepsilon}(\tau) &= \frac{1}{Z} \sum_j \varepsilon_j \exp\left(-\frac{\varepsilon_j}{\tau}\right) \\ &= -\frac{1}{Z} \frac{\partial Z}{\partial(1/\tau)} = -\frac{\partial \ln Z}{\partial Z} \frac{\partial Z}{\partial(1/\tau)} = -\frac{\partial \ln Z}{\partial(1/\tau)} . \end{aligned} \quad (3.6)$$

Rearranging further gives

$$\bar{\varepsilon}(\tau) = -\frac{\partial \ln Z}{\partial \tau} \frac{\partial \tau}{\partial(1/\tau)} = \tau^2 \frac{\partial \ln Z}{\partial \tau} , \quad (3.7)$$

The *entropy* of the system, denoted as $\sigma(\tau)$, is essentially the logarithm of the “number of ways that the state can be constructed from indistinguishable molecules” (Barrett & Myers, 2004), or the number of possible configurations in the ensemble of particles. The fundamental entropy is given by

$$\sigma(\tau) = -\sum_j \text{Pr}(j) \ln \text{Pr}(j) = \ln(Z) + \frac{\bar{\varepsilon}(\tau)}{\tau}. \quad (3.8)$$

Entropy is central to the second law of thermodynamics, also called the *law of increase in entropy*, which states that the entropy of a thermally isolated closed system cannot decrease; if a constraint internal to the system is removed, then the entropy tends to increase (Kittel & Kroemer, 1980). The probability distribution that maximizes the statistical entropy is the *Boltzmann distribution* (Bonomi & Lutton, 1984). Another important quantity that utilizes the partition function is the *Helmholtz free energy*, which is related to the logarithm of Z and carries information regarding the mean energy and the entropy:

$$-\tau \ln(Z) = F(\tau) = \bar{\varepsilon}(\tau) - \tau \sigma(\tau), \quad (3.9)$$

The free energy conveys how to “balance the conflicting demands of a system for minimum energy and maximum entropy” and is at a minimum when the system is coupled to a reservoir, provided that the volume is constant (Kittel & Kroemer, 1980).

In fundamental units, the *heat capacity* of a system (at constant volume) is defined as the rate of change of the energy with temperature, which can be shown mathematically to be proportional to the variance in energy:

$$C_V(\tau) \equiv \frac{\partial \bar{\varepsilon}(\tau)}{\partial \tau} = \frac{\langle \varepsilon(\tau)^2 \rangle - \langle \varepsilon(\tau) \rangle^2}{\tau^2}. \quad (3.10)$$

Observe that the heat capacity according to (3.10) is a dimensionless quantity. In conventional units (i.e., energy per kelvin), however, it is written as

$$C_V(T) \equiv \frac{\partial \bar{\varepsilon}(T)}{\partial T} = \frac{\langle \varepsilon(T)^2 \rangle - \langle \varepsilon(T) \rangle^2}{k_B T^2}. \quad (3.11)$$

We can prove the relationship in (3.10) by first expanding the mean energy in terms of the Boltzmann factors, as shown in (3.06):

$$C_V(\tau) \equiv \frac{\partial \bar{\varepsilon}(\tau)}{\partial \tau} = \frac{\partial}{\partial \tau} \left[\frac{1}{Z(\tau)} \sum_j \varepsilon_j \exp(-\varepsilon_j / \tau) \right]. \quad (3.12)$$

Applying the derivative leads to

$$C_V(\tau) = \frac{Z(\tau) \frac{\partial}{\partial \tau} \sum_j \varepsilon_j \exp(-\varepsilon_j / \tau) - \frac{\partial Z(\tau)}{\partial \tau} \sum_j \varepsilon_j \exp(-\varepsilon_j / \tau)}{Z(\tau)^2}$$

$$\begin{aligned}
&= \frac{\frac{Z(\tau)}{\tau^2} \sum_j \varepsilon_j^2 \exp(-\varepsilon_j / \tau) - \frac{1}{\tau^2} \left[\sum_j \varepsilon_j \exp(-\varepsilon_j / \tau) \right]^2}{Z(\tau)^2} \\
&= \frac{1}{\tau^2} \left\{ \frac{1}{Z(\tau)} \sum_j \varepsilon_j^2 \exp(-\varepsilon_j / \tau) - \left[\frac{1}{Z(\tau)} \sum_j \varepsilon_j \exp(-\varepsilon_j / \tau) \right]^2 \right\}, \quad (3.13)
\end{aligned}$$

and using the Boltzmann probability in (3.3), we obtain

$$C_V(\tau) = \frac{1}{\tau^2} \left\{ \sum_j \varepsilon_j^2 \Pr(j) - \left[\sum_j \varepsilon_j \Pr(j) \right]^2 \right\}. \quad (3.14)$$

The first and second sums are simply the expected values of ε_j^2 and ε_j , respectively, so that (3.14) is identical (3.10).

The heat capacity per unit mass is called the *specific heat*. An abrupt change in either the heat capacity (or specific heat) with a small change in temperature indicates a phase transition of a thermodynamic system. Accordingly, a large value of the heat capacity signifies a change in state of a system, which can be used during optimization to indicate that freezing has begun, and that gradual cooling is required to avoid quenching (Kirkpatrick *et al.*, 1983).

3.3.3 The Metropolis algorithm

For complex thermodynamic systems, the partition function for the entire ensemble of particles can be extremely difficult to calculate. In 1953, Metropolis *et al.* developed a simple algorithm to simulate the equation of state for a many-body system in thermal equilibrium at a given temperature. In this approach, many random samples of various molecular (or atomic) configurations $\{\mathbf{r}_n, n = 1, \dots, N\}$ are generated through Monte Carlo sampling; thus properties of the system can be estimated by using sample averages in lieu of ensemble averages.

The algorithm proceeds iteratively: At every iteration, a particle undergoes a small random displacement, resulting in a change of energy $\Delta\mathcal{E}$ in the configuration of particles. If $\Delta\mathcal{E} \leq 0$, the new configuration is automatically accepted, and the system moves to a state of lower energy. If $\Delta\mathcal{E} > 0$, the configuration is accepted with probability $\text{Pr}(\Delta\mathcal{E}) = \exp(-\Delta\mathcal{E}/\tau)$. To implement this, a random number ξ uniformly distributed in the range (0, 1) is drawn; if $\xi < \text{Pr}(\Delta\mathcal{E})$, the new configuration is accepted and the particle moves to its new position, but if $\xi > \text{Pr}(\Delta\mathcal{E})$, the old configuration is retained. In either case, the current configuration contributes to the calculation of sample averages:

$$\bar{F} = \frac{1}{M} \sum_{i=1}^M F_i, \quad (3.15)$$

where F represents a property of the system, and i is the iteration number, and M is the total number of iterations so far. Through repetitive execution of the basic step described above, one can simulate the thermal motion of particles in equilibrium at a fixed temperature. The choice of $\text{Pr}(\Delta\mathcal{E})$ means that the system approaches a Boltzmann distribution, as long as all possible states can eventually be reached (Barrett & Myers, 2004). The method is said to be *ergodic* if this condition is satisfied, that is, if all states can be reached from any other in a finite number of steps.

3.3.4 Continuous minimization by simulated annealing

An adaptive form of simulated annealing was developed by Corana *et al.* (1987) for minimizing multimodal functions in a continuous domain, in other words, to determine the position in N -dimensional space of the minimum of a given function of N variables. This algorithm considers unique physical constraints (e.g., different finite ranges) for each parameter, as well as different sensitivities of the cost function along different parametric axes. It also optimizes computational time by attempting to maintain a one-to-one ratio between the number of accepted and rejected moves, thereby searching the parameter-space more efficiently. The precise method and description of the algorithm as implemented by Corana are outlined in the remainder of this section.

Method

Suppose $Q(\boldsymbol{\theta})$ is the objective function to minimize, where $\boldsymbol{\theta} = \{\theta_n, n = 1, \dots, N\}$ is a vector in R^N . The N variables each range over a continuous, but finite, interval:

$a_1 < \theta_1 < b_1, \dots, a_N < \theta_N < b_N$. Though Q must be bounded, it may be discontinuous. To be clear, θ_{ni} denotes the n^{th} component of $\boldsymbol{\theta}$ at the i^{th} iteration.

The algorithm has an iterative scheme, which is outlined in Fig. 3.3. Starting with a given point $\boldsymbol{\theta}_0$ called the *initial guess*, it generates a succession of points, $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_i, \dots$, in search of the minimum of the function. The initial guess can be made by the use of prior information, such as estimates computed from previous studies, known values determined from other systems, or values calculated through theoretical means.

Fig. 3.3: The simulated annealing algorithm implemented by Corana *et al.*(1987).

The set of points that are accessible in a single step are said to be in the *neighborhood* of the current point. A new candidate point is generated in the neighborhood of the current point θ_i by making a random move along a single coordinate direction; this step is repeated for the remaining directions, yielding a cycle through all variables. The step vector $\mathbf{v}$, also a vector in R^N , represents the size of the neighborhood around θ_i . New coordinate values are uniformly distributed in a bracket centered on the respective coordinate of θ_i , and half the size of this bracket for each coordinate is recorded in $\mathbf{v}$. A new point is discarded if it falls outside the definition domain of Q , and another point is generated until one within the definition domain is obtained.

The Metropolis criterion determines whether a candidate point θ' is accepted or rejected (Metropolis *et al.*, 1953):

If $\Delta Q \leq 0$, then accept the new point: $\theta_{i+1} = \theta'$,

else accept the new point with probability: $\Pr(\Delta Q) = \exp(-\Delta Q/\tau)$,

where $\Delta Q = Q(\theta') - Q(\theta_i)$ and τ is an effective temperature. Therefore, downhill moves are always accepted, while uphill moves are made probabilistically.

At a given temperature, the succession of points $\theta_0, \theta_1, \dots, \theta_i, \dots$ is not purely downhill, except at $\tau = 0$, when uphill moves are no longer possible. For large values of $\tau \gg \Delta Q_{\text{avg}}$, where the average is taken over random pairs of points inside the definition domain of Q , nearly all new points are accepted, resulting in a random sampling of Q over the entire definition domain.

The search begins at a high temperature τ_0 declared by the user. A succession of points $\theta_0, \theta_1, \dots, \theta_i, \dots$ is generated until the system equilibrates, that is when the average value of the objective function Q_{avg} converges as the number of iterations increases. In the thermodynamics analogy, Q_{avg} corresponds to the internal energy of the system. The best point obtained is recorded as θ_{opt} , which is used as a starting point for the next temperature phase. The temperature τ is reduced after each phase according to an *annealing schedule*, which will be discussed later. At each temperature, the step vector $\mathbf{v}_m$ is periodically adjusted to adapt to the function behavior, attempting to equalize the number of accepted and rejected moves; the index m describes successive step vector adjustments during a single temperature phase. The process is terminated at a very low temperature when there is no more improvement is possible, according to a terminating criterion.

In the optimization context, an iterative search accepting only configurations that lower the objective function is like rapidly quenching a thermodynamic system, so that it is most likely to become trapped in a metastable, local minimum. Conversely, the effective temperature τ acts as a control parameter that allows the system to make uphill moves and explore high-cost configurations. At higher temperatures, the system is more sensitive to the large-scale features of the function. As the temperature decreases, finer details in the surface emerge, while the system is less likely to escape a local basin. Although simulated annealing does not provide a certificate of optimality for the final point, the method searches for a better solution in the presence of many local minima.

Annealing schedule

The simulated annealing algorithm guarantees the true minimum as long as the temperature is proportional to the reciprocal of the logarithm of the iteration number:

$$\tau_k = \tau_0 \frac{\ln k_0}{\ln k}, \quad (3.16)$$

where τ is the temperature, k is the iteration number, and k_0 is some starting index. For large k , (3.16) can be rewritten as

$$\Delta \tau = -\tau_0 \frac{\ln k_0 \Delta k}{k(\ln k)^2}, \quad k \gg 1, \quad (3.17)$$

or using recursive indices,

$$\tau_{k+1} = \tau_k - \tau_0 \frac{\ln k_0}{k(\ln k)^2}. \quad (3.18)$$

Such a logarithmic temperature schedule is consistent with the Boltzmann algorithm, but it converges to the true minimum much too slowly for practical use (Ingber, 1993). Thus in practice, the optimality of the solution is traded for better computational time. To expedite the search process, some researches employ an exponential annealing schedule:

$$\tau_{k+1} = r \tau_k, \quad 0 < r < 1, \quad (3.19)$$

where r is the reduction coefficient. When the iteration number k is large,

$$\frac{\Delta\tau}{\tau_k} = (r-1)\Delta k, \quad k \gg 1, \quad (3.20)$$

or equivalently,

$$\tau_k = \tau_0 \exp[k(r-1)]. \quad (3.21)$$

The tail of the exponential function enforces a gradual decline in temperature toward the end of the search when freezing takes place. Algorithms that use a rapid cooling schedule are referred to as *simulated quenching* (Gillet & Sheng, 1999; Ingber, 1996; Sato, 1997).

One of the many questions that arise when choosing an efficient annealing schedule concerns the choice of starting temperature τ_0 . In combinatorial problems, a recommended value is on the order of magnitude of the standard deviation of the objective function in the domain of interest (White, 1984). In the case of continuous optimization, however, this leads to starting temperatures that are higher than necessary, thereby wasting computational time (Corana *et al.*, 1987).

The success or failure of the simulated annealing algorithm often depends on the choice of annealing schedule, involving important parameters such as the starting temperature, the reduction rate between successive temperature phases, the number of function evaluations at each temperature, and the time at which to terminate the search

process. An effective combination of control parameters is best determined empirically for a given problem, or by carefully monitoring the optimization process. (This constitutes yet another optimization problem!) Nonetheless, physical intuition of the problem also plays an instrumental role in simulated annealing.

The algorithm

The following is a description of the algorithm used in this research, as outlined by Corana *et al.* (1987), apart from notational changes:

Step 0 (Initialization)

Specify:

- A starting point in the parameter space θ_0 .
- An initial temperature τ_0 .
- An initial step vector $\mathbf{v}_0$.
- A terminating criterion δ and a number of consecutive temperature reductions to test for termination N_δ .
- A test for change in the step vector N_s and a varying criterion c .
- A test for temperature reduction N_T and a reduction coefficient r_T .

Set i, j, m, k to 0, where i is the index for successive points, j denotes successive cycles along every direction, m describes successive step adjustments, and k is for successive temperature reductions.

Set h to 1, where h is the index denoting the direction along which the trial point is generated, starting from the last accepted point.

Compute $Q_0 = Q(\theta_0)$.

Set $\theta_{\text{opt}} = \theta_0$.

Set $Q_{\text{opt}} = Q_0$.

Set $n_u = 0, \quad u = 1, \dots, N$.

Set $Q_u^* = Q_0, \quad u = 0, -1, \dots, -N_\delta + 1$.

Step 1

Starting from the point θ_i , generate a random point θ' along the direction h :

$$\theta' = \theta_i + rv_{m,h}\mathbf{e}_h, \quad (3.22)$$

where r is a random number generated in the range $(-1, 1)$ by a pseudorandom number generator, $\mathbf{e}_h$ is the vector of the h^{th} coordinate direction, and $v_{m,h}$ is the component of the step vector $\mathbf{v}_m$ along the same direction.

Step 2

If the h^{th} coordinate of $\boldsymbol{\theta}'$ lies outside the definition domain of Q (i.e., if $x_h' < a_h$ or $x_h' > b_h$), then return to step 1.

Step 3

Compute $Q' = Q(\boldsymbol{\theta}')$.

If $Q' \leq Q_i$, then accept the new point:

Set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}'$,

set $Q_{i+1} = Q'$,

add 1 to i ,

add 1 to n_h ,

if $Q' < Q_{\text{opt}}$, then set

$\boldsymbol{\theta}_{\text{opt}} = \boldsymbol{\theta}'$,

$Q_{\text{opt}} = Q'$.

endif;

else ($Q' > Q_i$) accept or reject the point with probability p (Metropolis criterion):

$$p = \exp\left(\frac{Q_i - Q'}{\tau_k}\right). \quad (3.23)$$

Generate a random number p' uniformly distributed over the range (0, 1) and compare to p : If $p' < p$, the point is accepted, otherwise it is rejected.

In the case of acceptance:

Set $\theta_{i+1} = \theta'$,

set $Q_{i+1} = Q'$,

add 1 to i ,

add 1 to n_h .

Step 4

Add 1 to h .

If $h \leq n$, then go to step 1;

else set h to 1 and add 1 to j .

Step 5

If $j < N_S$, then go to step 1;

else update the step vector $\mathbf{v}_m$:

For each direction u the new step vector component $\mathbf{v}_u'$ is

$$v'_u = v_{m,u} \left(1 + c_u \frac{n_u/N_S - 0.6}{0.4} \right) \quad \text{if } n_u > 0.6N_S, \quad (3.24)$$

$$v'_u = \frac{v_{m,u}}{1 + c_u \frac{0.4 - n_u/N_S}{0.4}} \quad \text{if } n_u < 0.4N_S, \quad (3.25)$$

$$v'_u = v_{m,u} \quad \text{otherwise.} \quad (3.26)$$

Set $\mathbf{v}_{m+1} = \mathbf{v}'$,

set j to 0,

set n_u to 0, $u = 1, \dots, N$,

add 1 to m .

The control parameter for the step variation along the u^{th} direction is denoted c_u . The modifications in step length attempt to equalize the number of accepted and rejected moves for maximal efficiency. A high number of accepted moves indicates that the trial points are too close to the initial ones and the system evolves too slowly, whereas a high number of rejected moves means that the candidate points are too distant. Either of these cases results in lower computational efficiency and wasted effort.

Step 6

If $m < N_T$, then go to step 1;

else, reduce the temperature τ_k :

Set $\tau_{k+1} = r_\tau \cdot \tau_k$,

set $Q_k^* = Q_i$,

add 1 to k ,

set m to 0.

Notice that an exponential annealing schedule is used. Furthermore, a temperature reduction occurs after N_T step adjustments, or $N_S \cdot N_T$ cycles of moves, where one cycle is over N coordinate directions.

Step 7 (Terminating criterion)

If:

$$|Q_k^* - Q_{k-u}^*| \leq \delta, \quad u = 1, \dots, N_\delta$$

$$Q_k^* - Q_{\text{opt}} \leq \delta,$$

then stop the search;

else:

Add 1 to i ,

set $\theta_i = \theta_{\text{opt}}$,

set $Q_i = Q_{\text{opt}}$.

Go to step 1.

Corana *et al.* (1987) suggests the following reasonable values for the control parameters in the algorithm to be used for initial test runs:

$$N_S = 20,$$

$$N_T = \max(100, 5 \cdot N),$$

$$c_i = 2, \quad i = 1, \dots, N,$$

$$N_\delta = 4,$$

$$r_T = 0.85.$$

Of course, suitable control parameters depend on the particular objective function being optimized. They are determined heuristically through carefully observing the system over successive test runs.

CHAPTER 4

PROPAGATION OF LIGHT

Light is an electromagnetic wave phenomenon governed by the same principles ascribed to the complete spectrum of electromagnetic radiation. Although electromagnetic waves are *vectorial* in nature, propagating as a mutually coupled electric-field wave and magnetic-field wave, a *scalar* wave theory in which light is treated as a scalar wavefunction often suffices in describing many optical phenomena. The effects of the wave nature of light are most appreciable when light propagates around objects whose size compares to a wavelength, but when the objects are much larger, the wave behavior can be approximated by rectilinear light rays that travel through different media according to a set of geometrical laws. The term *diffraction* refers to the deviation from rectilinear propagation.

The branch of optics that treats light as electromagnetic waves is called *electromagnetic optics*, while the scalar wave theory and the ray theory are referred to as *wave optics* and *ray optics* or *geometrical optics*, respectively. In principle, geometrical optics is the limit of wave optics when the wavelength becomes infinitesimally small, and wave optics is an approximation of electromagnetic optics. These theories are encompassed by *classical optics*, but certain optical phenomena are quantum mechanical in nature, requiring a quantum theory of light (*quantum optics*).

This chapter imparts fundamental results in the classical regime, beginning with the basic equations of electromagnetism, called *Maxwell's equations*, in Section 4.1. From these equations, we derive the general time-dependent and time-independent wave equations, which decouple the electric-field and magnetic-field wave components. We discuss two important solutions to these equations, plane waves and spherical waves, in Section 4.2.

In Section 4.3, we use Maxwell's equations once again to derive the basic equation of geometrical optics, namely, the *eikonal equation*. We then develop several theoretical principles governing the propagation of light rays through dielectric media, as well as reflection and refraction at a planar dielectric boundary.

Section 4.4 is dedicated to a facet of diffraction theory. We deal with the specific case where a wave propagates through a planar aperture and travels some distance in free space, and we formulate several laws for calculating the diffraction pattern in different regions of space.

4.1 The electromagnetic field

The electromagnetic field is a real physical entity that occupies the space surrounding electric charges. It is described by four vector fields, each a function of a 3D position vector, denoted as $\mathbf{r} = (x, y, z)$, and time t . These vector fields are the electric field $\mathbf{E}(\mathbf{r}, t)$, the magnetic field $\mathbf{H}(\mathbf{r}, t)$, the electric flux density or displacement $\mathbf{D}(\mathbf{r}, t)$, and the magnetic flux density or induction $\mathbf{B}(\mathbf{r}, t)$. While $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$ are regarded as

the basic field vectors, $\mathbf{D}(\mathbf{r}, t)$ and $\mathbf{B}(\mathbf{r}, t)$ represent the influence of matter, which is described by an electric current density $\mathbf{j}(\mathbf{r}, t)$ (current per unit area) and a charge density $q(\mathbf{r}, t)$ (charge per unit volume). The standard notation for charge density is ρ , however, we are reserving it to denote spatial frequency later in this chapter.

4.1.1 Maxwell's equations

All classical electromagnetic phenomena are governed by Maxwell's equations, which describe the dynamics of charged particles interacting with electromagnetic fields.

Maxwell's equations are a set of four coupled first-order partial differential equations, written in the International System of Units, abbreviated SI, as

$$\nabla \cdot \mathbf{D}(\mathbf{r}, t) = q(\mathbf{r}, t), \quad (4.1a)$$

$$\nabla \cdot \mathbf{B}(\mathbf{r}, t) = 0, \quad (4.1b)$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\frac{\partial}{\partial t} \mathbf{B}(\mathbf{r}, t), \quad (4.1c)$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \mathbf{j}(\mathbf{r}, t) + \frac{\partial}{\partial t} \mathbf{D}(\mathbf{r}, t). \quad (4.1d)$$

These equations incorporate the divergence and curl operators, denoted as $\nabla \cdot$ and $\nabla \times$, respectively, where the vector *del* operator in Cartesian coordinates is given by

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right). \quad (4.2)$$

The divergence of a vector field, such as in (4.1a) and (4.1b), becomes a scalar field.

Thus, the charge density in (4.1a) is scalar-valued.

Since the divergence of a curl is *always* zero, it follows from (4.1d) that

$$\nabla \cdot \mathbf{j}(\mathbf{r}, t) = -\nabla \cdot \frac{\partial \mathbf{D}(\mathbf{r}, t)}{\partial t} = -\frac{\partial}{\partial t} \nabla \cdot \mathbf{D}(\mathbf{r}, t). \quad (4.3)$$

After invoking (4.1a), we have

$$\nabla \cdot \mathbf{j}(\mathbf{r}, t) + \frac{\partial q(\mathbf{r}, t)}{\partial t} = 0, \quad (4.4)$$

which is referred to as the *continuity equation* and expresses local conservation of charge.

The divergence theorem states that

$$\int_V (\nabla \cdot \mathbf{v}) d^3r = \int_S \mathbf{v} \cdot \hat{\mathbf{n}} da, \quad (4.5)$$

where $\mathbf{v}$ is an arbitrary vector field, V is a volume enclosed by a closed surface S with area element da , and $\hat{\mathbf{n}}$ is a unit outward normal on S . Applying (4.5) to (4.4) leads to

$$\int_S \mathbf{j}(\mathbf{r}, t) \cdot \hat{\mathbf{n}} da + \frac{\partial}{\partial t} \int_V q(\mathbf{r}, t) d^3r = 0, \quad (4.6)$$

where the first integral represents the total current flowing through the surface

$$J = \int_S \mathbf{j}(\mathbf{r}, t) \cdot \hat{\mathbf{n}} \, da, \quad (4.7)$$

and the second integral is the total charge in the volume

$$Q(t) = \int_V q(\mathbf{r}, t) d^3r. \quad (4.8)$$

Thus, the total charge contained in the volume can only change with the flow of electric current.

4.1.2 Constitutive relations

In order to uniquely determine the field vectors that appear in Maxwell's equations for a given distribution of charge and current, we need to describe the behavior of material media when influenced by a field. These relationships are called the *constitutive relations* (or *material equations*). In isotropic linear media these relations are:

$$\mathbf{P}(\mathbf{r}, t) = \varepsilon_0 \chi_e \mathbf{E}(\mathbf{r}, t), \quad (4.9)$$

$$\mathbf{M}(\mathbf{r}, t) = \chi_m \mathbf{H}(\mathbf{r}, t), \quad (4.10)$$

where $\mathbf{P}(\mathbf{r}, t)$ is the *polarization*, $\mathbf{M}(\mathbf{r}, t)$ is the *magnetization*, χ_e is the *electric susceptibility*, χ_m is the *magnetic susceptibility*, and ε_0 is the *permittivity of free space*.

In a dielectric medium, the polarization is defined as the macroscopic average of the electric dipole moment per unit volume. The magnetization is similarly defined for magnetic dipole moments.

Polarization results from distortion of the charge distribution in a medium in the presence of an external electric field. For many materials, the polarization is proportional to the electric field so that (4.9) holds, provided the field is not too strong. If so, we can express the electric flux density as

$$\mathbf{D}(\mathbf{r}, t) = \varepsilon_0 \mathbf{E}(\mathbf{r}, t) + \mathbf{P}(\mathbf{r}, t) \equiv \varepsilon \mathbf{E}(\mathbf{r}, t), \quad (4.11)$$

where $\varepsilon \equiv \varepsilon_0(1 + \chi_e)$ is the *dielectric constant* (or *permittivity*). Apart from *ferroelectric* media, the polarization of a material is impermanent and is only induced by an instantaneous external field. In ferroelectrics the electric displacement is determined by the past history of the field, instead of its instantaneous value. This effect is referred to as *hysteresis*.

Magnetization occurs when an applied magnetic field creates a net alignment of magnetic dipoles in a magnetic substance. When the magnetization is linearly proportional to the field as in (4.10), we can write

$$\mathbf{B}(\mathbf{r}, t) = \mu_0 [\mathbf{H}(\mathbf{r}, t) + \mathbf{M}(\mathbf{r}, t)] \equiv \mu \mathbf{H}(\mathbf{r}, t), \quad (4.12)$$

where $\mu \equiv \mu_0(1 + \chi_m)$ is the *magnetic permeability* and μ_0 is the *permeability of free space*. The magnetic susceptibility χ_m is very close to zero ($\mu \approx \mu_0$) for most materials, but for magnetic media, χ_m is substantially far from zero. Unlike electric polarization, which is usually in the same direction as $\mathbf{E}(\mathbf{r}, t)$, certain materials experience a magnetization parallel to $\mathbf{H}(\mathbf{r}, t)$ (e.g., oxygen, aluminum, platinum) and others opposite to $\mathbf{H}(\mathbf{r}, t)$ (e.g., water, copper, silver). These materials are said to be *paramagnetic* ($\chi_m > 0$) and *diamagnetic* ($\chi_m < 0$), respectively. Except for *ferromagnetic* materials (e.g., iron, cobalt, nickel), the magnetization is not retained after the external field is removed. Nonetheless, optical frequencies do not have an effect on magnetic dipoles as they do on electric ones (Barrett & Myers, 2004), so we can safely assume that $\chi_m = 0$ as in free-space.

For exceedingly strong fields, which can be obtained, for instance, by focusing a laser beam, the constitutive relations can be nonlinear functions of $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$. Throughout this chapter, we will deal with light propagation in linear media. We will also consider transparent, non-conducting media, in which light traverses without considerable weakening.

In free space where there are no charges or currents, the constitutive relations become

$$\mathbf{D}(\mathbf{r}, t) = \varepsilon_0 \mathbf{E}(\mathbf{r}, t), \quad (4.13)$$

$$\mathbf{B}(\mathbf{r}, t) = \mu_0 \mathbf{H}(\mathbf{r}, t). \quad (4.14)$$

and Maxwell's equations can be expressed solely in terms of $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$:

$$\nabla \cdot \mathbf{E}(\mathbf{r}, t) = 0, \quad (4.15a)$$

$$\nabla \cdot \mathbf{H}(\mathbf{r}, t) = 0, \quad (4.15b)$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\mu_0 \frac{\partial}{\partial t} \mathbf{H}(\mathbf{r}, t), \quad (4.15c)$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \varepsilon_0 \frac{\partial}{\partial t} \mathbf{E}(\mathbf{r}, t). \quad (4.15d)$$

4.1.3 Time-dependent wave equation

Consider a region without material media, although it may contain charges and currents to generate an electromagnetic field. Using the constitutive relations (4.13) and (4.14), Maxwell's equations become

$$\nabla \cdot \mathbf{E}(\mathbf{r}, t) = \frac{1}{\varepsilon_0} q(\mathbf{r}, t), \quad (4.16a)$$

$$\nabla \cdot \mathbf{H}(\mathbf{r}, t) = 0, \quad (4.16b)$$

$$\nabla \times \mathbf{E}(\mathbf{r}, t) = -\mu_0 \frac{\partial}{\partial t} \mathbf{H}(\mathbf{r}, t), \quad (4.16c)$$

$$\nabla \times \mathbf{H}(\mathbf{r}, t) = \mathbf{j}(\mathbf{r}, t) + \varepsilon_0 \frac{\partial}{\partial t} \mathbf{E}(\mathbf{r}, t) . \quad (4.16d)$$

To derive the wave equation for the electric field, we will need the following identity from vector calculus:

$$\nabla \times (\nabla \times \mathbf{v}) = \nabla (\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v} , \quad (4.17)$$

where the Laplacian operator in Cartesian coordinates is

$$\nabla^2 = \left(\frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}, \frac{\partial^2}{\partial z^2} \right) . \quad (4.18)$$

Taking the curl of (4.16c) leads to

$$\nabla \times [\nabla \times \mathbf{E}(\mathbf{r}, t)] = -\mu_0 \frac{\partial}{\partial t} \nabla \times \mathbf{H}(\mathbf{r}, t) . \quad (4.19)$$

Applying (4.16d) and the vector identity (4.17) gives

$$\nabla [\nabla \cdot \mathbf{E}(\mathbf{r}, t)] - \nabla^2 \mathbf{E}(\mathbf{r}, t) = -\mu_0 \frac{\partial}{\partial t} \left[\mathbf{j}(\mathbf{r}, t) + \varepsilon_0 \frac{\partial}{\partial t} \mathbf{E}(\mathbf{r}, t) \right] . \quad (4.20)$$

Finally, using (4.16a) leads to the wave equation for $\mathbf{E}(\mathbf{r}, t)$:

$$\left(\nabla^2 - \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \right) \mathbf{E}(\mathbf{r}, t) = \mu_0 \frac{\partial}{\partial t} \mathbf{j}(\mathbf{r}, t) + \frac{1}{\varepsilon_0} \nabla q(\mathbf{r}, t). \quad (4.21)$$

A similar approach leads to the wave equation for $\mathbf{H}(\mathbf{r}, t)$:

$$\left(\nabla^2 - \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \right) \mathbf{H}(\mathbf{r}, t) = -\nabla \times \mathbf{j}(\mathbf{r}, t). \quad (4.22)$$

We have now reduced the set of four Maxwell's equations to two second-order equations in which $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$ are uncoupled. Thus, we can directly solve for the field vectors if $q(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ are given. Note, however, that (4.21) and (4.22) each contain three equations, since $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$ each consist of three Cartesian components, while these six equations are also uncoupled from one another.

The wave equations in (4.21) and (4.22) each have the basic form,

$$\left(\nabla^2 - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) u(\mathbf{r}, t) = s(\mathbf{r}, t), \quad (4.23)$$

called the *time-dependent inhomogeneous scalar wave equation*. Here, $u(\mathbf{r}, t)$ represents a scalar field and $s(\mathbf{r}, t)$ is a corresponding scalar source distribution of charges and (or) currents, which can be derived from Maxwell's equations (Jackson, 1975; Barrett &

Myers, 2004). When there are no sources, $s(\mathbf{r}, t) = 0$, so that (4.23) becomes *homogeneous*.

The constant v in (4.23) symbolizes the speed of wave propagation in the medium, while the choice of v depends on the type of media and the type of wave. For the material media described in Section 4.1.2,

$$v^2 = 1/\mu_0 \varepsilon \equiv c^2/n^2, \quad (4.24)$$

where c is the speed of light in vacuum and n is the refractive index of the medium. If the material is homogeneous, then both c and n are constants. In general, however, they can have a spatial or temporal dependence.

By introducing Fourier transformations, we can convert (4.23) from a partial differential equation to an algebraic one. Assuming that $u(\mathbf{r}, t)$ and $s(\mathbf{r}, t)$ have the Fourier integral representations,

$$u(\mathbf{r}, t) = \int_{-\infty}^{\infty} d^3\sigma \int_{-\infty}^{\infty} d\nu U(\sigma, \nu) \exp[2\pi i(\sigma \cdot \mathbf{r} - \nu t)], \quad (4.25a)$$

$$s(\mathbf{r}, t) = \int_{-\infty}^{\infty} d^3\sigma \int_{-\infty}^{\infty} d\nu S(\sigma, \nu) \exp[2\pi i(\sigma \cdot \mathbf{r} - \nu t)], \quad (4.25b)$$

we can write the inverse relations as

$$U(\boldsymbol{\sigma}, \nu) = \int_{-\infty}^{\infty} d^3\mathbf{r} \int_{-\infty}^{\infty} dt u(\mathbf{r}, t) \exp[-2\pi i(\boldsymbol{\sigma} \cdot \mathbf{r} - \nu t)], \quad (4.26a)$$

$$S(\boldsymbol{\sigma}, \nu) = \int_{-\infty}^{\infty} d^3\mathbf{r} \int_{-\infty}^{\infty} dt s(\mathbf{r}, t) \exp[-2\pi i(\boldsymbol{\sigma} \cdot \mathbf{r} - \nu t)], \quad (4.26b)$$

where $\boldsymbol{\sigma}$ denotes the 3D spatial-frequency. Note that according to the sign convention in (4.25) and (4.26), a wave component travelling in the $+\boldsymbol{\sigma}$ direction has positive temporal frequency ν . Using these transformations, the time-dependent scalar wave equation in the Fourier domain is given by

$$-4\pi^2 \left(\boldsymbol{\sigma}^2 - \frac{\nu^2}{v^2} \right) U(\boldsymbol{\sigma}, \nu) = S(\boldsymbol{\sigma}, \nu). \quad (4.27)$$

4.1.4 Time-independent wave equation

The time dependence of the scalar wave equation can be removed in the case when the source oscillates at a single frequency, ν_0 , such that

$$s(\mathbf{r}, t) = s(\mathbf{r}) \exp(-2\pi i \nu_0 t). \quad (4.28)$$

Since $s(\mathbf{r}, t)$ must be real, we take the real part of the complex exponential. We use complex notation for the scalar source and field for mathematical ease in calculations,

then we can take the real part of the final expression to represent the physical quantity of interest. This process is valid due to linearity of the wave equation. Note that $s(\mathbf{r})$ can be complex, so the magnitude and phase of the oscillation may change with position.

In the Fourier domain, the source is given by

$$S(\boldsymbol{\sigma}, \nu) = S(\boldsymbol{\sigma})\delta(\nu - \nu_0). \quad (4.29)$$

A *monochromatic* source is one that oscillates at a single frequency, satisfying both (4.28) and (4.29). Combining (4.27) and (4.29) gives

$$-4\pi^2 \left(\boldsymbol{\sigma}^2 - \frac{\nu^2}{\nu_0^2} \right) U(\boldsymbol{\sigma}, \nu) = S(\boldsymbol{\sigma})\delta(\nu - \nu_0). \quad (4.30)$$

The only solution is to have

$$U(\boldsymbol{\sigma}, \nu) = U(\boldsymbol{\sigma})\delta(\nu - \nu_0), \quad (4.31)$$

which is equivalent to

$$u(\mathbf{r}, t) = u(\mathbf{r})\exp(-2\pi i \nu_0 t). \quad (4.32)$$

Here we observe that the field has the same monochromatic time dependence as the source, although there may be a phase shift, since $u(\mathbf{r})$ can be complex. The reason is that the wave equation represents a temporal linear shift-invariant system, of which

$\exp(-2\pi i \nu_0 t)$ is an eigenfunction. Thus any other choice of time dependence for the source, besides the complex exponential, would not lead to the same time dependence of the field (Barrett & Myers, 2004).

For a monochromatic source, the Fourier transform of the wave amplitude satisfies

$$-4\pi^2 \left(\sigma^2 - \frac{\nu_0^2}{\nu^2} \right) U(\sigma, \nu) = S(\sigma), \quad (4.33)$$

or in the space domain,

$$\left(\nabla^2 + k^2 \right) u(\mathbf{r}) = s(\mathbf{r}), \quad (4.34)$$

where $k = 2\pi \nu_0 / \nu$. Each Fourier component of the scalar field satisfies (4.33) and (4.34).

This equation is called the *time-independent scalar wave equation*, more commonly referred to as the *Helmholtz equation*. It is called the *homogeneous Helmholtz equation* when $s(\mathbf{r}) = 0$. Notice that there is no implicit time dependence in (4.34).

4.2 Plane waves and spherical waves

An important feature of the time-dependent and time-independent wave equations is the existence of travelling wave solutions, representing the transport of electromagnetic energy. In this section, we will examine two fundamental solutions, plane waves and

spherical waves. Only simple media, free of sources and characterized by spatially constant permeability and susceptibility, will be considered here.

4.2.1 Plane waves

The simplest solution of the homogeneous wave equation is the monochromatic plane wave, which has the form

$$u(\mathbf{r}, t) = \exp(i\mathbf{k} \cdot \mathbf{r} - 2\pi i \nu_0 t), \quad (4.35)$$

where the frequency ν_0 is related to the magnitude of the wave vector $k = |\mathbf{k}|$ by

$$k = \frac{2\pi \nu_0}{v}. \quad (4.36)$$

Parallel to the direction of $\mathbf{k}$, the function (4.35) is periodic with a wavelength of $\lambda = 2\pi/k$.

It is particularly convenient to replace the wave vector with the 3D spatial frequency $\boldsymbol{\sigma}$, defined in Cartesian components as

$$\mathbf{k} = 2\pi\boldsymbol{\sigma} = (2\pi\xi, 2\pi\eta, 2\pi\zeta). \quad (4.37)$$

Inserting (4.35) and (4.37) into the homogeneous wave equation (4.23) leads to

$$\begin{aligned}
& \left(\nabla^2 - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) \exp[2\pi i(\boldsymbol{\sigma} \cdot \mathbf{r} - v_0 t)] \\
& = -4\pi^2 \left(\xi^2 + \eta^2 + \zeta^2 - \frac{v_0^2}{v^2} \right) \exp[2\pi i(\boldsymbol{\sigma} \cdot \mathbf{r} - v_0 t)] = 0, \tag{4.38}
\end{aligned}$$

which has solution

$$\sigma^2 = \xi^2 + \eta^2 + \zeta^2 = \frac{v_0^2}{v^2}. \tag{4.39}$$

This is equivalent to

$$\xi^2 + \eta^2 + \zeta^2 = \frac{1}{\lambda^2}, \tag{4.40}$$

since $\sigma = 2\pi/k = 1/\lambda$. From (4.40), we see that the components of $\boldsymbol{\sigma}$ depend on each other through the wavelength λ , so that if we know two of the components, say ξ and η , ζ can be determined by

$$\zeta = \pm \sqrt{\frac{1}{\lambda^2} - \xi^2 - \eta^2}. \tag{4.41}$$

Although there is sign ambiguity, a wave that propagates in the $+z$ direction requires the positive sign (Barrett & Myers, 2004).

4.2.2 Spherical waves

A monochromatic spherical wave is written as

$$u(\mathbf{r}, t) = \frac{1}{|\mathbf{r} - \mathbf{r}_0|} \exp(ik|\mathbf{r} - \mathbf{r}_0| - 2\pi i \nu_0 t), \quad (4.42)$$

which has spherical symmetry about the point $\mathbf{r}_0$. Showing that (4.42) is a solution to the wave equation requires us to take its Laplacian, but we first make a useful change of variables,

$$\mathbf{R} \equiv \mathbf{r} - \mathbf{r}_0. \quad (4.43)$$

In spherical coordinates centered on $\mathbf{r}_0$, $\mathbf{R}$ has components (R, θ_R, ϕ_R) , but because $u(\mathbf{r}, t)$ has no angular dependency, the Laplacian is simply

$$\nabla^2 \left[\frac{\exp(ikR)}{R} \right] = \frac{1}{R^2} \frac{\partial}{\partial R} \left[R^2 \frac{\partial}{\partial R} \frac{\exp(ikR)}{R} \right] = -k^2 \frac{\exp(ikR)}{R}, \quad (4.44)$$

given that $R \neq 0$. The behavior at $R = 0$ involves a discussion of Green's functions, but we will not elaborate on that here. Using (4.44) in the homogeneous wave equation (4.23) gives

$$\left(\nabla^2 - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) \frac{\exp(ikR - 2\pi i \nu_0 t)}{R} = - \left(k^2 - \frac{4\pi^2 \nu_0^2}{v^2} \right) \frac{\exp(ikR - 2\pi i \nu_0 t)}{R} = 0. \quad (4.45)$$

We observe from (4.45) that the spherical wave in (4.42) satisfies the homogeneous wave equation as long as $k = 2\pi\nu_0/\nu$, as in the case for plane waves. The significance of these results is that we can “decompose an arbitrary solution of the homogeneous wave equation into monochromatic plane waves or spherical waves, and for each component we can define a wavelength λ and an associated $k = 2\pi/\lambda$ ” (Barrett & Myers, 2004).

4.3 Geometrical optics

The fundamental equation of geometrical optics, called the eikonal equation, is a direct implication of Maxwell’s equations in the short-wavelength limit. This section begins with a derivation of the eikonal equation, which we use to define geometrical wavefronts and geometrical light rays. We then develop a set of mathematical laws governing the propagation of light rays through a dielectric medium, as well as the refraction and reflection of rays at the interface between two dielectric media.

4.3.1. The eikonal equation

A general time-harmonic field in a non-conducting isotropic medium can be written as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0(\mathbf{r}) \exp(-i\omega_0 t), \quad (4.46a)$$

$$\mathbf{H}(\mathbf{r}, t) = \mathbf{H}_0(\mathbf{r}) \exp(-i\omega_0 t), \quad (4.46b)$$

where $\mathbf{E}_0(\mathbf{r})$ and $\mathbf{H}_0(\mathbf{r})$ are complex vector functions of position. It is understood that the real parts of the expressions on the right-hand side represent the fields.

On substituting (4.46) into (4.1), we find that $\mathbf{E}_0(\mathbf{r})$ and $\mathbf{H}_0(\mathbf{r})$ satisfy the time-free Maxwell's equations. In source-free media that satisfy the assumptions in Section 4.1.2, these equations are given by

$$\nabla \cdot \varepsilon(\mathbf{r})\mathbf{E}_0(\mathbf{r}) = 0, \quad (4.47a)$$

$$\nabla \cdot \mu(\mathbf{r})\mathbf{H}_0(\mathbf{r}) = 0, \quad (4.47b)$$

$$\nabla \times \mathbf{E}_0(\mathbf{r}) = i\omega_0\mu(\mathbf{r})\mathbf{H}_0(\mathbf{r}), \quad (4.47c)$$

$$\nabla \times \mathbf{H}_0(\mathbf{r}) = -i\omega_0\varepsilon(\mathbf{r})\mathbf{E}_0(\mathbf{r}), \quad (4.47d)$$

where the constitutive relations (4.11) and (4.12) have been applied. Here we let the permittivity ε and magnetic permeability μ , and therefore the refractive index

$n = c(\mu\varepsilon)^{1/2}$, vary with position. We also have

$$\omega_0 = 2\pi\nu_0 = ck_0 = \frac{2\pi c}{\lambda_0}, \quad (4.48)$$

with λ_0 denoting the vacuum wavelength.

Deriving the eikonal equation begins by representing the complex vectors $\mathbf{E}_0(\mathbf{r})$ and $\mathbf{H}_0(\mathbf{r})$ in the form

$$\mathbf{E}_0(\mathbf{r}) = \mathbf{e}(\mathbf{r})\exp[ik_0S(\mathbf{r})], \quad (4.49a)$$

$$\mathbf{H}_0(\mathbf{r}) = \mathbf{h}(\mathbf{r})\exp[ik_0S(\mathbf{r})], \quad (4.49b)$$

where $S(\mathbf{r})$ is the *eikonal*, a real scalar function of position. Conversely, $\mathbf{e}(\mathbf{r})$ and $\mathbf{h}(\mathbf{r})$ are complex vector functions of position; letting these functions be complex allows all possible polarization states to be included (Born & Wolf, 1999).

Applying the familiar vector identities,

$$\nabla \cdot (f\mathbf{v}) = f(\nabla \cdot \mathbf{v}) + \mathbf{v} \cdot (\nabla f), \quad (4.50)$$

$$\nabla \times (f\mathbf{v}) = f(\nabla \times \mathbf{v}) - \mathbf{v} \times (\nabla f), \quad (4.51)$$

to (4.49) results in

$$\nabla \times \mathbf{H}_0(\mathbf{r}) = [\nabla \times \mathbf{h}(\mathbf{r}) + ik_0 \nabla S(\mathbf{r}) \times \mathbf{h}(\mathbf{r})]\exp[ik_0S(\mathbf{r})], \quad (4.52a)$$

$$\nabla \cdot \mu(\mathbf{r})\mathbf{H}_0(\mathbf{r}) = [\mu(\mathbf{r})\nabla \cdot \mathbf{h}(\mathbf{r}) + \mathbf{h}(\mathbf{r}) \cdot \nabla \mu(\mathbf{r}) + ik_0\mu(\mathbf{r})\mathbf{h}(\mathbf{r}) \cdot \nabla S(\mathbf{r})]\exp[ik_0S(\mathbf{r})], \quad (4.52b)$$

$$\nabla \times \mathbf{E}_0(\mathbf{r}) = [\nabla \times \mathbf{e}(\mathbf{r}) + ik_0 \nabla S(\mathbf{r}) \times \mathbf{e}(\mathbf{r})]\exp[ik_0S(\mathbf{r})], \quad (4.52c)$$

$$\nabla \cdot \varepsilon(\mathbf{r})\mathbf{E}_0(\mathbf{r}) = [\varepsilon(\mathbf{r})\nabla \cdot \mathbf{e}(\mathbf{r}) + \mathbf{e}(\mathbf{r}) \cdot \nabla \varepsilon(\mathbf{r}) + ik_0 \varepsilon(\mathbf{r})\mathbf{e}(\mathbf{r}) \cdot \nabla S(\mathbf{r})] \exp[ik_0 S(\mathbf{r})]. \quad (4.52d)$$

Now combining (4.47) and (4.52) gives

$$\mathbf{e}(\mathbf{r}) \cdot \nabla S(\mathbf{r}) = -\frac{1}{ik_0} [\nabla \cdot \mathbf{e}(\mathbf{r}) + \mathbf{e}(\mathbf{r}) \cdot \nabla \ln \varepsilon(\mathbf{r})], \quad (4.53a)$$

$$\mathbf{h}(\mathbf{r}) \cdot \nabla S(\mathbf{r}) = -\frac{1}{ik_0} [\nabla \cdot \mathbf{h}(\mathbf{r}) + \mathbf{h}(\mathbf{r}) \cdot \nabla \ln \mu(\mathbf{r})], \quad (4.53b)$$

$$\nabla S(\mathbf{r}) \times \mathbf{e}(\mathbf{r}) - c\mu(\mathbf{r})\mathbf{h}(\mathbf{r}) = -\frac{1}{ik_0} \nabla \times \mathbf{e}(\mathbf{r}), \quad (4.53c)$$

$$\nabla S(\mathbf{r}) \times \mathbf{h}(\mathbf{r}) + c\varepsilon(\mathbf{r})\mathbf{e}(\mathbf{r}) = -\frac{1}{ik_0} \nabla \times \mathbf{h}(\mathbf{r}). \quad (4.53d)$$

In the short-wavelength limit (large k_0), we can neglect the right-hand sides of (4.53), provided that the multiplicative factors of $1/ik_0$ are not extremely large (Born & Wolf, 1999). Thus, we have

$$\mathbf{e}(\mathbf{r}) \cdot \nabla S(\mathbf{r}) = 0, \quad (4.54a)$$

$$\mathbf{h}(\mathbf{r}) \cdot \nabla S(\mathbf{r}) = 0, \quad (4.54b)$$

$$\nabla S(\mathbf{r}) \times \mathbf{e}(\mathbf{r}) - c\mu(\mathbf{r})\mathbf{h}(\mathbf{r}) = 0, \quad (4.54c)$$

$$\nabla S(\mathbf{r}) \times \mathbf{h}(\mathbf{r}) + c \varepsilon(\mathbf{r}) \mathbf{e}(\mathbf{r}) = 0. \quad (4.54d)$$

Solving for $\mathbf{h}(\mathbf{r})$ in (4.54c) and substituting into (4.55d) leads to

$$\nabla S(\mathbf{r}) \times [\nabla S(\mathbf{r}) \times \mathbf{e}(\mathbf{r})] + c^2 \mu(\mathbf{r}) \varepsilon(\mathbf{r}) \mathbf{e}(\mathbf{r}) = 0, \quad (4.55)$$

and applying the vector identity,

$$\mathbf{v}_1 \times (\mathbf{v}_1 \times \mathbf{v}_2) = \mathbf{v}_1 (\mathbf{v}_1 \cdot \mathbf{v}_2) - \mathbf{v}_2 (\mathbf{v}_1 \cdot \mathbf{v}_1), \quad (4.56)$$

results in

$$\nabla S(\mathbf{r}) [\mathbf{e}(\mathbf{r}) \cdot \nabla S(\mathbf{r})] - \mathbf{e}(\mathbf{r}) |\nabla S(\mathbf{r})|^2 + \mu(\mathbf{r}) \varepsilon(\mathbf{r}) c^2 \mathbf{e}(\mathbf{r}) = 0. \quad (4.57)$$

The first term vanishes due to (4.54a), so we are finally left with the eikonal equation,

$$|\nabla S(\mathbf{r})|^2 = [n(\mathbf{r})]^2, \quad (4.58)$$

or equivalently,

$$\left(\frac{\partial S}{\partial x}\right)^2 + \left(\frac{\partial S}{\partial y}\right)^2 + \left(\frac{\partial S}{\partial z}\right)^2 = n^2(x, y, z). \quad (4.59)$$

The surfaces over which $S(\mathbf{r})$ is constant are called *geometrical wavefronts* or *geometrical wave surfaces*, therefore, the eikonal equation relates these surfaces to only the refractive index function of the medium.

4.3.2 Differential equation of light rays

Orthogonal trajectories to geometric wavefronts are referred to as *geometrical light rays*.

If $\mathbf{r}(s)$ is the position vector of an arbitrary point on a ray and s is the line element along a ray path s , then $d\mathbf{r}/ds = \mathbf{s}$, so that the equation of the ray is given by

$$n(\mathbf{r}) \frac{d\mathbf{r}}{ds} = \nabla S(\mathbf{r}) . \quad (4.60)$$

Although (4.60) describes the ray in terms of the function $S(\mathbf{r})$, we can obtain a differential equation of the ray in terms of just the refractive index function. We do this by differentiating (4.60) with respect to s , giving

$$\begin{aligned} \frac{d}{ds} \left[n(\mathbf{r}) \frac{d\mathbf{r}}{ds} \right] &= \frac{d}{ds} [\nabla S(\mathbf{r})] \\ &= \frac{d\mathbf{r}}{ds} \cdot \nabla [\nabla S(\mathbf{r})] \\ &= \frac{1}{n(\mathbf{r})} \nabla S(\mathbf{r}) \cdot \nabla [\nabla S(\mathbf{r})] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2n(\mathbf{r})} \nabla [\nabla S(\mathbf{r})]^2 \\
&= \frac{1}{2n(\mathbf{r})} \nabla [n(\mathbf{r})^2]
\end{aligned}$$

so that

$$\frac{d}{ds} \left[n(\mathbf{r}) \frac{d\mathbf{r}}{ds} \right] = \nabla n(\mathbf{r}). \quad (4.61)$$

In a homogeneous medium, $n = \text{constant}$ and (4.61) reduces to

$$\frac{d^2 \mathbf{r}}{ds^2} = 0, \quad (4.62)$$

which has the solution

$$\mathbf{r} = s\mathbf{a} + \mathbf{b}, \quad (4.62)$$

where $\mathbf{a}$ and $\mathbf{b}$ are constant vectors. Therefore, light rays in a homogeneous medium propagate in straight lines.

4.3.3 Refraction and reflection

So far, we have considered the behavior of light rays in media with a continuously varying refractive index function $n(\mathbf{r})$. Now we introduce a surface discontinuity in $n(\mathbf{r})$, that is, a planar interface that separates two media with different refractive indices.

Let $\hat{\mathbf{n}}$ be the unit vector normal to the interface and $\hat{\mathbf{k}}_{inc}$ and $\hat{\mathbf{k}}_{tr}$ be unit vectors parallel to the incident and transmitted wavevectors, respectively. If the incident ray propagates through a medium with refractive index n_1 , and the transmitted ray through index n_2 , then the law of refraction is written as (Barrett & Myers, 2004; Born & Wolf, 1999; Stavroudis, 1972)

$$n_1(\hat{\mathbf{k}}_{inc} \times \hat{\mathbf{n}}) = n_2(\hat{\mathbf{k}}_{tr} \times \hat{\mathbf{n}}). \quad (4.63)$$

This is Snell's law in vector form; unlike the scalar version of this law, (4.63) is not specific to a coordinate system, rendering it very useful in optical design programs.

We can make two important observations from Snell's law. Firstly, the tangential component of the ray vector $n\hat{\mathbf{k}}$ is continuous across the interface, or equivalently, the vector $\mathbf{N}_{12} = n_2\hat{\mathbf{k}}_{tr} - n_1\hat{\mathbf{k}}_{inc}$ is normal to the interface. Secondly, the refracted ray lies in the same plane as both the incident ray and the normal to the surface, called the *plane of incidence* (Born & Wolf, 1999).

Suppose we know n_1 , n_2 , and $\hat{\mathbf{n}}$, and would like to determine $\hat{\mathbf{k}}_{tr}$ for a given $\hat{\mathbf{k}}_{inc}$.

We can do this with an alternative version of Snell's law, particularly, by forming an

orthonormal basis for the plane of incidence. The vectors $\hat{\mathbf{n}}$ and $\hat{\mathbf{\kappa}}_{inc}$ are linearly independent, except at normal incidence, and constitute a basis for the plane of incidence; however, they are not orthonormal. Using Gram-Schmidt orthogonalization (Arfken & Weber, 2001), we can construct an orthonormal basis containing $\hat{\mathbf{n}}$ and an orthonormal vector,

$$\hat{\mathbf{n}}_{\perp} = \frac{\hat{\mathbf{\kappa}}_{inc} - (\hat{\mathbf{\kappa}}_{inc} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}}{\sqrt{1 - (\hat{\mathbf{\kappa}}_{inc} \cdot \hat{\mathbf{n}})^2}}, \quad (4.64)$$

such that $\hat{\mathbf{n}} \cdot \hat{\mathbf{n}}_{\perp} = 0$ and $|\hat{\mathbf{n}}_{\perp}|^2 = 1$. Since $\hat{\mathbf{n}}_{\perp}$ is normal to $\hat{\mathbf{n}}$, which is normal to the interface, we see that $\hat{\mathbf{n}}_{\perp}$ lies in the interface plane. More specific, $\hat{\mathbf{n}}_{\perp}$ coincides with the intersection between the interface plane and the plane of incidence (Barrett & Myers, 2004).

In terms of the orthonormal vectors $\hat{\mathbf{n}}$ and $\hat{\mathbf{n}}_{\perp}$, the constraint that both $\hat{\mathbf{\kappa}}_{inc}$ and $\hat{\mathbf{\kappa}}_{tr}$ lie in the plane of incidence is expressed as

$$(\hat{\mathbf{n}}_{\perp} \times \hat{\mathbf{n}}) \cdot \hat{\mathbf{\kappa}}_{inc} = (\hat{\mathbf{n}}_{\perp} \times \hat{\mathbf{n}}) \cdot \hat{\mathbf{\kappa}}_{tr} = 0. \quad (4.65)$$

Snell's law can be rewritten using $\hat{\mathbf{n}}_{\perp}$:

$$n_1 (\hat{\mathbf{\kappa}}_{inc} \cdot \hat{\mathbf{n}}_{\perp}) = n_2 (\hat{\mathbf{\kappa}}_{tr} \cdot \hat{\mathbf{n}}_{\perp}). \quad (4.66)$$

Finally, we have an expression for $\hat{\mathbf{k}}_{tr}$ that satisfies both (4.65) and (4.66):

$$\begin{aligned}\hat{\mathbf{k}}_{tr} &= \frac{n_1}{n_2} (\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}}_{\perp}) \hat{\mathbf{n}}_{\perp} + \left[\sqrt{1 - \left(\frac{n_1}{n_2} \right)^2 (\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}}_{\perp})^2} \right] \hat{\mathbf{n}} \\ &= \frac{n_1}{n_2} [\hat{\mathbf{k}}_{inc} - (\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}}) \hat{\mathbf{n}}] + \left[\sqrt{1 + \left(\frac{n_1}{n_2} \right)^2 [(\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}})^2 - 1]} \right] \hat{\mathbf{n}}.\end{aligned}\quad (4.67)$$

The corresponding equation for the reflected wavevector is given by

$$\begin{aligned}\hat{\mathbf{k}}_{refl} &= (\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}}_{\perp}) \hat{\mathbf{n}}_{\perp} - \left[\sqrt{1 - (\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}}_{\perp})^2} \right] \hat{\mathbf{n}} \\ &= (\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}}_{\perp}) \hat{\mathbf{n}}_{\perp} - (\hat{\mathbf{k}}_{inc} \cdot \hat{\mathbf{n}}) \hat{\mathbf{n}},\end{aligned}\quad (4.68)$$

which lies in the plane of incidence, along with $\hat{\mathbf{k}}_{inc}$, $\hat{\mathbf{k}}_{tr}$, and $\hat{\mathbf{n}}$ (Barrett & Myers, 2004). From (4.68), we see that the angle of reflection equals the angle of incidence.

These last two results summarize the law of reflection.

4.4 Diffraction by a planar aperture

4.4.1 A brief history of diffraction theory

Sommerfeld (1954) defined the term *diffraction* as “any deviation of light rays from rectilinear paths which cannot be interpreted as reflection or refraction.” Diffraction

results from the lateral confinement of a wave, and the effect is greatest when the size of the confinement is comparable to the wavelength of radiation involved. There is a rich history regarding the discovery and evolution of diffraction theory, which will be described briefly here.

The first advocate of the wave theory of light was Christian Huygens (1678), who expressed intuitively that if each point on a wavefront gave rise to a *secondary* diverging spherical wave, then the wavefront at a later instant would be the envelope of the secondary *wavelets*. It was not until 140 years later in 1818 that Augustin Jean Fresnel made various assumptions about the amplitudes and phases of the secondary wavelets, and was able to accurately predict the distribution of light in diffraction patterns by allowing the wavelets to interfere with each other. The principle of interference was established by Thomas Young (1802).

A significant step in the evolution of the wave theory of light occurred in 1860 when Maxwell identified light as an electromagnetic wave phenomenon. In 1882, Gustav Kirchhoff formed a stronger mathematical framework from the combined ideas of Huygens and Fresnel, but he based his theory on two assumptions regarding the boundary values of the wave impinging on an obstacle placed in the path of propagation. However, inconsistencies in these assumptions were later demonstrated by Henri Poincaré (1892) and Arnold Sommerfeld (1896). Sommerfeld then modified the Kirchhoff theory by abandoning one of Kirchhoff's assumptions dealing with the light amplitude at the boundary; he accomplished this through the utilization of Green's functions. The results

became known as the *Rayleigh-Sommerfeld diffraction theory*, which is well accepted for dealing with particular problems in optics.

One should be aware, however, that the Kirchhoff and Rayleigh-Sommerfeld theories involve broad simplifications and approximations. The most consequential simplification is that the vectorial nature of electromagnetic waves is ignored, so that light is treated simply as a scalar phenomenon (Goodman, 2005). Nonetheless, the scalar theory still offers great accuracy provided that two conditions are satisfied: the size of the diffracting aperture must be much greater than a wavelength, and the fields must be observed sufficiently far from the diffracting aperture (Silver, 1962). Both of these conditions will be met for the problems described in this chapter.

For a broad overview of diffraction theory, see Baker and Copson (1949) and Bouwkamp (1954).

4.4.2 *Geometry of the problem*

In optics and imaging, we are often interested in diffraction of light by an open planar aperture in an otherwise opaque screen. As illustrated in Fig. 4.1, a wave is assumed to impinge on the aperture from the left and the field is calculated at the arbitrary point $\mathbf{r}$ in the observation plane, which lies in the z plane. An arbitrary point in the aperture plane is denoted $\mathbf{r}_0$. For convenience, the aperture is placed in the $z = 0$ plane. Since we know which planes are the input and output planes, it suffices to use the 2D vectors $\mathbf{r}_0$ and $\mathbf{r}$ instead of their 3D counterparts, $\mathbf{r}_0$ and $\mathbf{r}$, respectively.

Fig. 4.1: Geometry for diffraction by a planar aperture (Barrett & Myers, 2004).

In this geometry, the distance between the points $\mathbf{r}_0$ and $\mathbf{r}$, denoted R , is given by

$$R = \sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2}, \quad (4.69)$$

where we have the expansion,

$$|\mathbf{r} - \mathbf{r}_0|^2 = (x - x_0)^2 + (y - y_0)^2. \quad (4.70)$$

We define θ as the angle between $\mathbf{r} - \mathbf{r}_0$ and the z axis, so that

$$\cos \theta = \frac{z}{\sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2}} = \frac{z}{\sqrt{(x - x_0)^2 + (y - y_0)^2 + z^2}}. \quad (4.71)$$

4.4.3 Huygens-Fresnel principle

According to Rayleigh-Sommerfeld diffraction theory, the Huygens-Fresnel principle can be stated mathematically as

$$u_z(\mathbf{r}) = \frac{1}{i\lambda} \int_{\infty} d^2r_0 u_{inc}(\mathbf{r}_0) t_{ap}(\mathbf{r}_0) \cos \theta \frac{\exp(ikR)}{R}, \quad (4.72)$$

where $u_z(\mathbf{r})$ is the field evaluated on a plane of fixed z for all (x,y) and $t_{ap}(\mathbf{r})$ is the *amplitude transmittance* of the aperture:

$$t_{ap}(\mathbf{r}) \equiv \begin{cases} 1 & \text{if } \mathbf{r} \text{ lies in the clear aperture} \\ 0 & \text{if } \mathbf{r} \text{ is behind the opaque screen.} \end{cases} \quad (4.73)$$

Built into (4.72) are two basic approximations, namely, the approximation inherent in scalar diffraction theory and the *radiation approximation*, which states that the observation distance is many wavelengths from the aperture, $z \gg \lambda$ (Goodman, 2005).

The Huygens-Fresnel principle states that the observed field $u_z(\mathbf{r})$ is a superposition of secondary diverging spherical waves $\exp(ikR)/R$ called *Huygens' wavelets* emanating from every point $\mathbf{r}_0$ within the aperture. Each wavelet has a 90° phase shift relative to the incident wave, as expressed in the factor $1/i$, as well as a directivity pattern, or obliquity factor, $\cos \theta$. The amplitude of each wavelet is proportional to the amplitude of the excitation $u_{inc}(\mathbf{r}_0)$ at the respective point in the aperture.

Note that (4.72) is readily seen as a convolution integral, since R and $\cos \theta$ are both functions of $\mathbf{r} - \mathbf{r}_0$, written symbolically as

$$u_z(\mathbf{r}) = [u_{inc}(\mathbf{r})t_{ap}(\mathbf{r})] * p_z(\mathbf{r}), \quad (4.74)$$

where $p_z(\mathbf{r})$ is the 2D point spread function (PSF) for propagation,

$$p_z(\mathbf{r}) = \frac{1}{i\lambda} \frac{z}{\sqrt{r^2 + z^2}} \frac{\exp\left(ik\sqrt{r^2 + z^2}\right)}{\sqrt{r^2 + z^2}}, \quad (4.75)$$

and $r = |\mathbf{r}|$. The input to the convolution is simply the incident field after being modified by the aperture,

$$u_0(\mathbf{r}) \equiv u_{inc}(\mathbf{r})t_{ap}(\mathbf{r}), \quad (4.76)$$

so that (4.72) becomes

$$u_z(\mathbf{r}) = \frac{1}{i\lambda} \int_{\infty} d^2 r_0 u_0(\mathbf{r}_0) \cos \theta \frac{\exp(ikR)}{R}. \quad (4.77)$$

The ability to express the Huygens-Fresnel principle as a convolution integral is a direct consequence of the linearity and shift-invariance of the diffraction operation, at

least as a 2D mapping from the aperture plane to a parallel plane some distance away.

Thus, if $u_0(\mathbf{r})$ and the observation point are shifted together, the result is the same.

If we are only interested in points close to the z -axis, we can apply the paraxial approximation, according to which $\cos\theta \approx 1$. Using this approximation, (4.75) results in

$$p_z(\mathbf{r}) \approx \frac{1}{i\lambda z} \exp\left(ik\sqrt{r^2 + z^2}\right), \quad (4.78)$$

and (4.77) becomes

$$u_z(\mathbf{r}) \approx \frac{1}{i\lambda z} \int_{\infty} d^2r_0 u_0(\mathbf{r}_0) \exp\left(ik\sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2}\right). \quad (4.79)$$

To be clear, the exponential factor in (4.79) still represents a spherical wave originating from $\mathbf{r}_0$ in the $z = 0$ plane, observed at $\mathbf{r}$ in the z plane. The next subsection deals with approximating this factor, which requires much carefulness, as k is often a very large number (on the order of 10^5 cm^{-1} at optical wavelengths).

4.4.4 Fresnel diffraction

To reduce (4.79) to a more practical form, we now introduce an approximation to the distance R through a binomial expansion for $z > |\mathbf{r} - \mathbf{r}_0|$, so that

$$R = \sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2} = z + \frac{|\mathbf{r} - \mathbf{r}_0|^2}{2z} - \frac{|\mathbf{r} - \mathbf{r}_0|^4}{8z^3} + \dots \quad (4.80)$$

Therefore, we can rewrite the exponential factor in (4.79) as

$$\exp\left(ik\sqrt{|\mathbf{r}-\mathbf{r}_0|^2+z^2}\right)=\exp(ikz)\exp\left(ik\frac{|\mathbf{r}-\mathbf{r}_0|^2}{2z}\right)\exp\left(-ik\frac{|\mathbf{r}-\mathbf{r}_0|^4}{8z^3}\right)\dots \quad (4.81)$$

We can disregard the quartic term and higher in (4.80) if

$$\frac{k|\mathbf{r}-\mathbf{r}_0|^4}{8z^3} \ll \frac{\pi}{4}, \quad (4.82)$$

or equivalently,

$$|\mathbf{r}-\mathbf{r}_0|^4 \ll \lambda z^3. \quad (4.83)$$

If this condition is satisfied, (4.79) becomes

$$u_z(\mathbf{r}) \approx \frac{\exp(ikz)}{i\lambda z} \int_{\infty} d^2r_0 u_0(\mathbf{r}_0) \exp\left(i\pi\frac{|\mathbf{r}-\mathbf{r}_0|^2}{\lambda z}\right) = u_0(\mathbf{r}) * p_z(\mathbf{r}), \quad (4.84)$$

and the 2D PSF is reduced to

$$p_z(\mathbf{r}) \approx \frac{\exp(ikz)}{i\lambda z} \exp\left(i\pi\frac{r^2}{\lambda z}\right). \quad (4.85)$$

Equations (4.84) and (4.85) reflect the *Fresnel approximation*, in which the PSF is a constant (i.e., independent of $\mathbf{r}$) multiplied by a quadratic phase exponential (Barrett & Myers, 2004). Thus, the spherical wavefronts observed on a plane are now approximated by paraboloids. The region where this approximation holds is called the *near field* of the aperture.

The Fresnel diffraction integral (4.84) is readily converted into a Fourier transform by substituting the term,

$$|\mathbf{r} - \mathbf{r}_0|^2 = r^2 + r_0^2 + 2\mathbf{r} \cdot \mathbf{r}_0, \quad (4.86)$$

which yields

$$u_z(\mathbf{r}) \approx \frac{\exp(ikz)}{i\lambda z} \exp\left(i\pi \frac{r^2}{\lambda z}\right) \int_{\infty} d^2 r_0 u_0(\mathbf{r}_0) \exp\left(i\pi \frac{r_0^2}{\lambda z}\right) \exp\left(-2\pi i \frac{\mathbf{r} \cdot \mathbf{r}_0}{\lambda z}\right). \quad (4.87)$$

Therefore, (4.87) is immediately seen as the 2D Fourier transform of the product

$u_0(\mathbf{r}_0) \exp(i\pi r_0^2 / \lambda z)$, with the spatial frequency given by $\boldsymbol{\rho} = \mathbf{r} / \lambda z$:

$$u_z(\mathbf{r}) \approx \frac{\exp(ikz)}{i\lambda z} \exp\left(i\pi \frac{r^2}{\lambda z}\right) \mathbf{F}_2 \left\{ u_0(\mathbf{r}_0) \exp\left(i\pi \frac{r_0^2}{\lambda z}\right) \right\}_{\boldsymbol{\rho}=\mathbf{r}/\lambda z}. \quad (4.88)$$

Despite the transform, the output $u_z(\mathbf{r})$ is related to the input $u_0(\mathbf{r}_0)$ in the space domain; substituting $\boldsymbol{\rho} = \mathbf{r} / \lambda z$ converts things from the frequency domain back into the

space domain (Barrett & Myers, 2004). The Fourier transform is simply more convenient than performing a spatial convolution with the quadratic phase factor.

4.4.5 Fraunhofer diffraction

We will consider a more stringent approximation, which is more difficult to satisfy, but will further simplify calculations when valid. Suppose the clear aperture fits into a circle of radius a , that is, $\mathbf{r}_0 < a$ for all $\mathbf{r}_0$ in the range of integration in (4.88). Now, if

$z \gg a^2 / \lambda$, then we can approximate $\exp(i\pi r_0^2 / \lambda z)$ by unity, so that (4.88) results in

$$u_z(\mathbf{r}) \approx \frac{\exp(ikz)}{i\lambda z} \exp\left(i\pi \frac{r^2}{\lambda z}\right) \mathbf{F}_2\{u_0(\mathbf{r}_0)\}_{\rho=\mathbf{r}/\lambda z}, \quad z \gg a^2 / \lambda. \quad (4.89)$$

When the approximation $z \gg a^2 / \lambda$ is valid, we are said to be in the *Fraunhofer zone* or *far field*. At optical frequencies, the condition of validity can be quite demanding. For instance, at a wavelength of 0.6 μm and aperture diameter of 2.5 cm, the observation distance must satisfy $z \gg 260$ meters. Of course, the greater the distance, the better accuracy in the approximation.

The *irradiance* of the diffraction pattern, denoted $I(\mathbf{r})$, is defined as the optical power per unit area incident on a surface. It can be shown that the irradiance is proportional to $|u(\mathbf{r})|^2$, under certain assumptions, with a proportionality constant that

relates to the physical interpretation of $u(\mathbf{r})$ (Barrett & Myers, 2004). Here we will

disregard the constant and let $I(\mathbf{r}) = |u(\mathbf{r})|^2$, so that

$$I(\mathbf{r}) = |u_z(\mathbf{r})|^2 \approx \frac{1}{\lambda^2 z^2} \left| \int_{-\infty}^{\infty} d^2 r_0 u_0(\mathbf{r}_0) \exp(-2\pi i \mathbf{r} \cdot \mathbf{r}_0 / \lambda z) \right|^2 = \frac{1}{\lambda^2 z^2} \left| U_0\left(\frac{\mathbf{r}}{\lambda z}\right) \right|^2, \quad (4.90)$$

where $U_0(\boldsymbol{\rho})$ is the 2D Fourier transform of $u_0(\mathbf{r})$. We see from (4.90) that the

irradiance is proportional to the squared modulus of the Fourier transform of the input field.

CHAPTER 5

INVERSE OPTICAL DESIGN OF THE HUMAN EYE USING LIKELIHOOD METHODS AND WAVEFRONT SENSING

In the preceding chapters, we provided the theoretical building blocks of this research, including concepts in estimation theory, global optimization methods with an emphasis on simulated annealing, and relevant topics in geometrical optics and diffraction theory. This chapter presents our results for the first of three applications that utilize the mathematical framework developed in the previous chapters. Specifically, it deals with the original motivation of our research, which is to develop a new approach to studying the human eye by estimating the complete set of ocular parameters for a given patient.

We begin in Section 5.1 by describing the basic optical components of the eye that are integral to the remainder of the chapter. Section 5.2 provides an overview on schematic eye models and their usefulness in evaluating the optical properties of the eye, followed by an algebraic method for ray-tracing through an arbitrary eye model.

Section 5.3 discusses the fundamental concepts in wavefront sensor technology and why we chose not to perform traditional wavefront sensing. In Section 5.4, we provide the details on our data acquisition system and optical-design program. We then present several results from the program, including the final WFS data to be used as input to inverse optical design.

In Section 5.5, we explore the Fisher information matrices and Cramér-Rao lower bounds for different configurations of the imaging system to demonstrate the impact of changes in the system parameters. In Section 5.6, we get a feel for the behavior of the probability surface, or the objective function to be optimized, as we vary the parameters of the eye. The final ML estimation results from a series of trials using simulated annealing are provided in Section 5.7.

We conclude in Section 5.8 with a set of ideas for future work related to practical application. In particular, we consider various enhancements to our optical design program, primarily to increase model robustness, since performing ML estimation on real data requires a vastly accurate probability model.

5.1 Basic anatomy of the human eye

The human eye is an extraordinary and highly complex organ with many integrated structures and dynamic, working parts. It is responsible for receiving light and converting it into an electrical signal, which follows the visual pathway to the brain where visual perception takes place.

Light first enters the *cornea*, a transparent layer forming the front of the eye, followed by a cavity filled with a clear fluid, called the *aqueous humor*, which occupies the *anterior chamber* and provides necessary nutrients to the cornea (Fig. 5.1). It then passes through the *pupil*, an opening in the opaque *iris*, with variable size to regulate the amount of light that eventually reaches the *retina*. Behind the iris is the crystalline *lens*, a transparent biconvex structure, whose shape is controlled by ciliary muscles at its edge.

Light then travels through the *vitreous humor*, a clear gelatinous substance that fills the central chamber of the eye. The final destination is the retina, a curved surface in back of the eye that is densely covered with nearly 130 million light-sensitive *photoreceptors*. These photoreceptors convert photons to an electrochemical neural signal, which leaves the eye via the *optic nerve* to the visual centers of the brain, where the optical information is processed (Palmer, 1999). This process is referred to as *visual phototransduction*.

Fig. 5.1: Basic anatomy of the human eye, as seen through a cross-sectional view.

The lens plays a chief role in proper image formation due to its variable focusing ability, which is achieved by changing the shape of the lens, a process called *accommodation*. Light from distant objects is brought into focus on the retina when the

ciliary muscles are relaxed, resulting in a thin lens. To focus on nearby objects, the ciliary muscles are contracted, so that the lens is thicker and provides more optical power. The total optical power of a relaxed eye, for which the focal length is longest, is about 60 diopters. Roughly two-thirds of this power is provided by the air-cornea interface, and the remaining one-third by the crystalline lens. As the ciliary muscles contract, the *lenticular* power, or power of the lens, increases and the total focal length of the eye decreases.

It has been well known since 1909 that the crystalline lens has a graded-index (GRIN) distribution, which increases not only the refractive power of the lens, but also the degree of accommodation (Gullstrand, 1962). There has been much interest in recent years for more accurate measurements and mathematical models to have greater understanding of lens functionality, for instance, regarding the distribution changes with accommodation (short-term) and age (long-term). Certain models approximate the lens with a shell structure, involving concentric *iso-indical* surfaces of constant index with the maximum value at the center (Atchison & Smith, 1995; Goncharov & Dainty, 2007; Navarro *et al.*, 2007, 2007a). Due to limited *in vivo* experimental data, however, there is plenty of debate surrounding how to best model the index changes with radial and axial position. The goal of the research by Navarro *et al.* (2007, 2007a) is to develop an adaptive model with adjustable parameters, so that individual data can be fitted for a range of ages and accommodation levels.

There are two distinct classes of photoreceptors in the retinal layer, rods and cones, which differ in shape and functionality. Rods are usually longer, narrower, and

have straight, rod-like ends, while cones are shorter, wider, and have tapered, cone-like ends. Rods are much more abundant, with about 120 million cells distributed all throughout the retina except for the very center, called the *fovea*. They are extremely sensitive to light and used only for vision under *scotopic* conditions, that is, at very low light levels. On the other hand, there are about 8 million cones scattered throughout the retina, but with a heavy concentration in the fovea. Cones are much less sensitive to light and are designed for most normal lighting, or *photopic*, conditions. They are solely responsible for color vision (Palmer, 1999).

Interestingly, cone photoreceptors in the human eye exhibit a directional sensitivity to light, in that marginal rays passing through the periphery of the pupil (off-axis light) are perceived as less intense than rays passing through the center of the pupil (axial light), a phenomenon called the *Stiles-Crawford effect* (SCE) (Stiles & Crawford, 1933). Cones essentially act like microscopic waveguides, funneling light from one end to another, with an associated effective acceptance angle of approximately 5 degrees. The SCE is presumably advantageous to visual performance by ameliorating the effects of defocus and aberrations for large pupils, although there have not been many theoretical or experimental studies to verify this (Atchison, Scott, Joblin, & Smith, 2000). If this evolutionary strategy is true, however, it would only pay off under photopic conditions; but this is precisely when cones are more sensitive. Conversely, rod photoreceptors are not as directionally sensitive as cones and have larger acceptance angles, as they are designed for dim light and cannot afford to waste photons.

The theoretical analysis of the physical properties of photoreceptors, using electromagnetic principles, and their influence on the SCE is a very complicated problem, even with many simplifying assumptions. It combines the directional characteristics and relative orientation of individual receptors, as well as the light leakage or cross-talk between cones (He, Marcos, & Burns, 1999). The simplest mathematical representation of this phenomenon incorporates an apodization in the pupil plane, as conceived by Westheimer (1959) and developed by Metcalf (1965). For the greatest simplicity, however, the retina is often treated as a perfect Lambertian reflector.

5.2 Ray-tracing through a schematic eye

Ray-tracing through schematic eye models has vast utility in ophthalmology and vision science for evaluating the optical properties of normal and pathologic eyes. We adopted an algebraic method for non-paraxial ray-tracing through an optical system containing aspheric surfaces to second-order, in which a surface is represented by a 4×4 matrix. (Langenbucher, Viestenz, Viestenz, Brnner, & Seitz, 2006). The advantages of using second-order, or quadric, surfaces is that the ray-surface intersection, surface normal vector, and direction of the refracted ray can be determined analytically. We applied this matrix-based approach to the Navarro wide-angle schematic eye and incorporated it into our inverse design system for estimating ocular parameters

Schematic eye models can vary greatly in terms of their complexity. The earliest eye models were developed over a century ago by Gullstrand (1962) and Von Helmholtz (1910). These models integrated spherical surfaces for the cornea and lens determined

from clinical data in order to predict first-order optical properties. More recent models have been used to simulate optical functions including retinal illumination (Kooijman, 1983), chromatic aberration (Thibos & Bradley, 1999), and retinal image formation (Camp, Maguire, Cameron, & Rob, 1990a). Kooijman (1983) utilized aspheric surfaces for the cornea, lens, and retina into a model to study retinal illumination, while the Indiana eye was developed to model chromatic aberration of the eye (Thibos & Bradley, 1999). Camp *et al.* (1990a, 1990b) integrated corneal topography data into a single refracting surface to model the optical imaging properties of the anterior corneal surface.

Eye models that emphasize anatomical accuracy incorporate a GRIN distribution for the crystalline lens, or even non-axially-symmetrical features, such as decentered lenses or pupils, which can have a strong impact optical performance. However, the large number of parameters involved in these advanced models can render them impractical, so there remains much interest in simplified, reduced schematic eyes. Reduced models are often rotationally symmetric and utilize an effective refractive index for the lens, making them more amenable to ray-tracing. Reduced models are generally able to reproduce certain ocular aberrations, such as axial spherical aberration (El Hage & Berny, 1973; Lotmar, 1971; Thibos, Ye, Zhang, & Bradley, 1997) or chromatic aberration (Thibos, Ye, Zhang, & Bradley, 1992).

The Navarro wide-angle schematic eye is based on anatomical data from clinical measurements and contains four centered quadric refracting surfaces with rotational symmetry, plus a spherical image surface representing the retina (Navarro *et al.*, 1985). Therefore, we are currently utilizing an effective refractive index for the crystalline lens,

but will later incorporate a GRIN distribution according to models suggested in the latest vision science studies (Goncharov & Dainty, 2007; Navarro, Palos, & González, 2007, 2007a). This model has minimal complexity, but on average, it can accurately predict optical performance across the visual field, including longitudinal and transverse chromatic aberration (Escudero-Sanz & Navarro, 1999). Table 5.1 provides a complete parametric description of the eye model, including radii, thicknesses, conic constants, and refractive indices. To calculate refractive indices at 780 nm, we applied the chromatic dispersion model developed by Atchison and Smith (2005) to the reference values provided by Navarro *et al.* (2007, 2007a).

Table 5.1: Navarro wide-angle schematic eye model at $\lambda = 780$ nm.

Surface	Radius [mm]	Thickness [mm]	Conic Constant	Refractive Index	Optical Medium
Anterior cornea	7.72	0.55	-0.26	1.3729	Cornea
Posterior cornea	6.50	3.05	0	1.3329	Aqueous
Pupil	Infinity	0	0	N/A	Aqueous
Anterior lens	10.20	4.00	-3.1316	1.4138	Lens
Posterior lens	-6.00	16.3203	-1.0	1.3311	Vitreous
Retina	-12.00	N/A	0	N/A	N/A

We implemented the Navarro wide-angle model for this study, but varied the values of the parameters to emulate a realistic eye (Table 5.2). In addition, we decentered

the lens by 0.20 mm in the horizontal direction and -0.10 mm in the vertical direction, which is consistent with the experimental range of lens misalignments (Rosales & Marcos, 2006).

Table 5.2: Geometry of eye model used to generate WFS data.

Surface	Radius [mm]	Thickness [mm]	Conic Constant	Refractive Index	Decentration (X, Y) [mm]
Anterior cornea	7.46	0.554	-0.24	1.3729	(0, 0)
Posterior cornea	6.38	3.37	0	1.3329	(0, 0)
Pupil	Infinity	0	0	N/A	(0, 0)
Anterior lens	10.85	4.09	-3.1304	1.4138	(0.2, -0.1)
Posterior lens	-5.92	16.40	-0.97	1.3317	(0.2, -0.1)
Retina	-12.00	N/A	0	N/A	(0, 0)

Description of refracting quadric surfaces

A quadric surface $S(x,y,z)$ is implicitly defined by

$$\begin{aligned}
 S(x, y, z) = & Ax^2 + By^2 + Cz^2 + 2Dxy + 2Eyz + 2Fxz \\
 & + 2Gx + 2Hy + 2Iz + K = 0,
 \end{aligned}
 \tag{5.1}$$

where x , y , and z are the Cartesian coordinates and $A, B, C, D, E, F, G, H, I, K$ are coefficients. In the matrix method developed by Langenbucher *et al.* (2006), (5.1) is written as

$$\mathbf{x}^t \mathbf{S} \mathbf{x} = 0, \quad (5.2)$$

where $\mathbf{x}$ is the generalized coordinate vector,

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \quad (5.3)$$

and $\mathbf{S}$ describes the quadric surface in matrix form,

$$\mathbf{S} = \begin{bmatrix} A & D & F & G \\ D & B & E & H \\ F & E & C & I \\ G & H & I & K \end{bmatrix}. \quad (5.4)$$

For instance, suppose we have a convex conic surface whose apex is at the origin and axis of rotation coincides with the z -axis. The surface sag is given by

$$z = \frac{\frac{x^2 + y^2}{R}}{1 + \sqrt{1 - (1 + \kappa) \frac{x^2 + y^2}{R^2}}}, \quad (5.5)$$

with conic constant κ and apical radius R . Generally, a paraboloid is described by $\kappa = -1$, an ellipsoid by $-1 < \kappa < 0$, a sphere by $\kappa = 0$, and a hyperboloid by $\kappa > 0$. After rearranging (5.5), we have

$$S_{\text{conic}}(x, y, z) = x^2 + y^2 + (1 + \kappa)z^2 - 2Rz = 0, \quad (5.6)$$

so that the matrix form of the surface is written as

$$\mathbf{S}_{\text{conic}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 + \kappa & -R \\ 0 & 0 & -R & 0 \end{bmatrix}. \quad (5.7)$$

Translating a quadric surface

When the quadric surface is shifted by a translation vector $\mathbf{x}_T = (x_T, y_T, z_T)$, (5.1) becomes

$$\begin{aligned} S(x, y, z) = & A(x - x_T)^2 + B(y - y_T)^2 + C(z - z_T)^2 + 2D(x - x_T)(y - y_T) \\ & + 2E(y - y_T)(z - z_T) + 2F(x - x_T)(z - z_T) + 2G(x - x_T) \\ & + 2H(y - y_T) + 2I(z - z_T) + K = 0, \end{aligned} \quad (5.8)$$

or equivalently,

$$(\mathbf{x} - \mathbf{x}_T)^t \mathbf{S} (\mathbf{x} - \mathbf{x}_T) = \mathbf{x}^t \mathbf{S}_T \mathbf{x} = 0. \quad (5.9)$$

The elements of the translated surface matrix $\mathbf{S}_T$ in (5.9) are given by

$$A_T = A, \quad B_T = B, \quad C_T = C,$$

$$D_T = D, \quad E_T = E, \quad F_T = F,$$

$$G_T = G - Ax_T - Dy_T - Fz_T,$$

$$H_T = H - Dx_T - By_T - Ez_T,$$

$$I_T = I - Fx_T - Ey_T - Cz_T,$$

$$K_T = K + Ax_T^2 + By_T^2 + Cz_T^2 + 2Ey_T z_T + 2Fx_T z_T - 2Gx_T - 2Hy_T - 2Iz_T. \quad (5.10)$$

We can now use (5.9) to generate the surface matrices for the quadric surfaces described in Table 5.2, including the decentration of the lens. If the axes of rotation are parallel to the z -axis and the apex of the anterior corneal surface is at the origin, we have

$$\mathbf{S}_{\text{cornea,anterior}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.76 & -7.46 \\ 0 & 0 & -7.46 & 0 \end{bmatrix}, \quad (5.11a)$$

$$\mathbf{S}_{\text{cornea,posterior}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -6.93 \\ 0 & 0 & -6.93 & 7.38 \end{bmatrix}, \quad (5.11b)$$

$$\mathbf{S}_{\text{lens,anterior}} = \begin{bmatrix} 1 & 0 & 0 & -0.20 \\ 0 & 1 & 0 & 0.10 \\ 0 & 0 & -2.1304 & -2.49 \\ -0.20 & 0.10 & -2.49 & 52.40 \end{bmatrix}, \quad (5.11c)$$

$$\mathbf{S}_{\text{lens,posterior}} = \begin{bmatrix} 1 & 0 & 0 & -0.20 \\ 0 & 1 & 0 & 0.10 \\ 0 & 0 & 0.03 & 5.68 \\ -0.20 & 0.10 & 5.68 & -92.91 \end{bmatrix}, \quad (5.11d)$$

$$\mathbf{S}_{\text{retina}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -12.41 \\ 0 & 0 & -12.41 & 10.11 \end{bmatrix}. \quad (5.11e)$$

Similar matrices are determined for the second pass through the ocular, after light is reflected from the retina and advances toward the detector:

$$\mathbf{S}_{\text{retina}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -12.00 \\ 0 & 0 & -12.00 & 0 \end{bmatrix}, \quad (5.12a)$$

$$\mathbf{S}_{\text{lens,posterior}} = \begin{bmatrix} 1 & 0 & 0 & 0.20 \\ 0 & 1 & 0 & 0.10 \\ 0 & 0 & 0.03 & -6.41 \\ 0.20 & 0.10 & -6.41 & 202.29 \end{bmatrix}, \quad (5.12b)$$

$$\mathbf{S}_{\text{lens,anterior}} = \begin{bmatrix} 1 & 0 & 0 & 0.20 \\ 0 & 1 & 0 & 0.10 \\ 0 & 0 & -2.1304 & 54.50 \\ 0.20 & 0.10 & 54.50 & -1339.01 \end{bmatrix}, \quad (5.12c)$$

$$\mathbf{S}_{\text{cornea,posterior}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -17.48 \\ 0 & 0 & -17.48 & 264.85 \end{bmatrix}, \quad (5.12d)$$

$$\mathbf{S}_{\text{cornea,anterior}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.76 & -11.09 \\ 0 & 0 & -11.09 & -345.70 \end{bmatrix}, \quad (5.12e)$$

Determining the surface normal vector

The normal vector to the quadric surface described in (5.1), denoted $\mathbf{n}$, can be determined analytically by taking the gradient of $S(x,y,z)$,

$$\mathbf{n} = \nabla S(x, y, z), \quad (5.13)$$

so that the components are given by

$$\frac{\partial S(x, y, z)}{\partial x} = 2Ax + 2Dy + 2Fz + 2G, \quad (5.14a)$$

$$\frac{\partial S(x, y, z)}{\partial y} = 2Dx + 2By + 2Ez + 2H, \quad (5.14b)$$

$$\frac{\partial S(x, y, z)}{\partial z} = 2Fx + 2Ey + 2Cz + 2I. \quad (5.14c)$$

We can rewrite (5.14) in matrix notation,

$$\mathbf{n} = \begin{bmatrix} A & D & F & G \\ D & B & E & H \\ F & E & C & I \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = 2\mathbf{S}(1:3, 1:4)\mathbf{x}, \quad (5.15)$$

where $\mathbf{S}(1:3, 1:4)$ is the upper 3×4 submatrix of $\mathbf{S}$. Finally, the unit normal vector is determined in the usual way,

$$\hat{\mathbf{n}} = \frac{\mathbf{n}}{\sqrt{\mathbf{n} \cdot \mathbf{n}}}. \quad (5.16)$$

Determining the ray-surface intersection

A ray is simply characterized by the coordinates of its origin $\mathbf{x}_0 = (x_0, y_0, z_0)$ and a direction vector $\mathbf{x}_d = (x_d, y_d, z_d)$,

$$\mathbf{x} = \mathbf{x}_0 + k\mathbf{x}_d = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + k \begin{bmatrix} x_d \\ y_d \\ z_d \end{bmatrix}, \quad (5.17)$$

where k is the scalar propagation constant. The intersection of an arbitrary ray with a quadric surface can be determined analytically with the quadratic equation. Substituting (5.17) into (5.1) leads to

$$k = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = k_1, k_2, \quad (5.18)$$

with the coefficients a , b , and c given by

$$a = Ax_d^2 + By_d^2 + Cz_d^2 + 2Dx_dy_d + 2Ey_dz_d + 2Fx_dz_d, \quad (5.19a)$$

$$b = 2 \left[\begin{array}{c} Ax_0x_d + By_0y_d + Cz_0z_d + Gx_d + Hy_d + Iz_d \\ + D(x_0y_d + y_0x_d) + E(y_0z_d + z_0y_d) + F(x_0z_d + z_0x_d) \end{array} \right], \quad (5.19b)$$

$$\begin{aligned} c = & Ax_0^2 + By_0^2 + Cz_0^2 + 2Dx_0y_0 + 2Ey_0z_0 \\ & + 2Fx_0z_0 + 2Gx_0 + 2Hy_0 + 2Iz_0 + K. \end{aligned} \quad (5.19c)$$

Note the following special cases and how to address them:

$$a = 0 \rightarrow k = -\frac{c}{b}, \quad (5.20a)$$

$$b^2 - 4ac < 0 \rightarrow \text{ray - surface intersection is imaginary,} \quad (5.20b)$$

$$a \neq 0 \rightarrow k \text{ is smallest positive value of } k_1 \text{ and } k_2. \quad (5.20c)$$

Direction of the transmitted ray

We saw in Section 4.3 that Snell's law in vector form yields an expression for the transmitted wavevector (ray) at a refracting interface,

$$\hat{\mathbf{x}}'_d = \hat{\mathbf{x}}_d \frac{n}{n'} [\hat{\mathbf{x}}_d - (\hat{\mathbf{x}}_d \cdot \hat{\mathbf{n}}) \hat{\mathbf{n}}] + \left[\sqrt{1 + \left(\frac{n}{n'} \right)^2 [(\hat{\mathbf{x}}_d \cdot \hat{\mathbf{n}})^2 - 1]} \right] \hat{\mathbf{n}}, \quad (5.21)$$

where $\hat{\mathbf{x}}_d$ and $\hat{\mathbf{x}}'_d$ are unit vectors parallel to the incident and transmitted rays, respectively, and n and n' are the corresponding refractive indices on both sides of the interface.

When using (5.21) in an optical-design program, it is imperative to verify that the transmitted ray propagates in the $+z$ direction, provided that the rays travel from left to right. If not, we must negate the unit normal vector $\hat{\mathbf{n}}$ in this equation, resulting in

$$\hat{\mathbf{x}}'_d = \hat{\mathbf{x}}_d \frac{n}{n'} [\hat{\mathbf{x}}_d - (\hat{\mathbf{x}}_d \cdot \hat{\mathbf{n}}) \hat{\mathbf{n}}] - \left[\sqrt{1 + \left(\frac{n}{n'} \right)^2 [(\hat{\mathbf{x}}_d \cdot \hat{\mathbf{n}})^2 - 1]} \right] \hat{\mathbf{n}}. \quad (5.22)$$

Figure 5.2 illustrates the eye model corresponding to the parameters in Table 5.2, with an 8-mm pupil and rays from a collimated, on-axis source. A close-up of the focal region emphasizes the spherical aberration in this schematic eye.

Fig. 5.2: Geometrical eye model corresponding to parameters in Table 5.2, with an on-axis source and 8-mm pupil to demonstrate spherical aberration.

5.3 Shack-Hartmann wavefront sensors

The Shack-Hartmann wavefront sensor (SHWFS) is a simple optical instrument used to characterize aberrations in an imaging system. It is a technological advancement of the Hartmann screen test, which was originally developed for optical testing by German

astrophysicist Johannes Hartmann at the turn of the 20th century (Schwiegerling & Neal, 2005). The SHWFS was invented in the 1960s for applications in astronomy to improve the resolution in images from ground-based telescopes, since the resolution is compromised by atmospheric turbulence (Platt & Shack, 2001). As the technology reached greater sophistication, alternative applications using these sensors received considerable impetus from the expertise developed by astronomers, including applications in ophthalmology, quality laser beam measurement, and optical system alignment (Neal, Copland, & Neal, 2002).

The SHWFS contains a two-dimensional lenslet array for measuring distortions in a wavefront, providing valuable information about aberrations in an optical system. The lenslet array is typically conjugated to the pupil plane of the system. Upon sampling an incoming wavefront, the lenslets produce a grid of focused spots on a CCD placed some distance behind the array, normally in the focal plane. Note that this assumes a locally uniform wavefront over each lenslet, which we will discuss in detail in Section 5.3.1. We know from basic Fourier theory that the amount of displacement of each spot from its ideal, on-axis location is proportional to the average local wavefront slope at the respective lenslet. Thus, if the wavefront is perfectly uniform (and normally incident), there is no shift in spots and we observe the focal plane image shown in Figure 5.3, while Figure 5.4 depicts the same image for an aberrated wavefront.

Fig. 5.3: Shack-Hartmann WFS measuring a perfect incoming wavefront.

Fig. 5.4: Shack-Hartmann WFS measuring an aberrated incoming wavefront.

In classical wavefront sensing, an algorithm processes the detected image and attempts to estimate the centroids (focal-spot positions) produced by each lenslet in the detector plane. The local wavefront slopes are computed from the centroids by

comparing to a reference, then the wavefront is reconstructed from the array of wavefront slopes.

For a measured $M \times 1$ irradiance distribution $\mathbf{g}(\boldsymbol{\theta})$, the centroid positions are computed from the first moments:

$$\hat{\tau}_{xj}(\boldsymbol{\theta}) = \frac{\sum_{m \in AOI, j} g_m(\boldsymbol{\theta}) x_m}{\sum_{m \in AOI, j} g_m(\boldsymbol{\theta})} \approx \frac{\sum_{m \in AOI, j} g_m(\boldsymbol{\theta}) x_m}{\bar{g}_{\text{tot}}}, \quad (5.23a)$$

$$\hat{\tau}_{yj}(\boldsymbol{\theta}) = \frac{\sum_{m \in AOI, j} g_m(\boldsymbol{\theta}) y_m}{\sum_{m \in AOI, j} g_m(\boldsymbol{\theta})} \approx \frac{\sum_{m \in AOI, j} g_m(\boldsymbol{\theta}) y_m}{\bar{g}_{\text{tot}}}, \quad (5.23b)$$

where the index j corresponds to the lenslet number and the index m is over only those detector elements that receive a signal from that lenslet, in the Area-of-Interest AOI, j .

The sum of irradiance values is denoted as g_{tot} , and the average sum as $\bar{g}_{\text{tot}}$.

The wavefront slope distribution is computed by comparing the measured centroids $\hat{\boldsymbol{\tau}}_j = (\hat{\tau}_{xj}, \hat{\tau}_{yj})$ to those determined by a reference wavefront

$\boldsymbol{\tau}_{j, \text{ref}} = (\tau_{xj, \text{ref}}, \tau_{yj, \text{ref}})$, given by

$$\begin{pmatrix} \langle \partial W / \partial x \rangle \\ \langle \partial W / \partial y \rangle \end{pmatrix}_j = \begin{pmatrix} \beta_x \\ \beta_y \end{pmatrix}_j \approx \frac{1}{d} \begin{pmatrix} \hat{\tau}_x - \tau_{x, \text{ref}} \\ \hat{\tau}_y - \tau_{y, \text{ref}} \end{pmatrix}_j, \quad (5.24)$$

where $W = W(x,y)$ is the two-dimensional wavefront and d is the distance between the lenslet array and detector, normally equal to the lenslet focal length f . The angled brackets denote the average, as the spot displacements are proportional to the average local wavefront slope.

Wavefront reconstruction is accomplished by relating the set of slopes to the wavefront gradient:

$$\nabla W = \frac{\partial W}{\partial x} \hat{x} + \frac{\partial W}{\partial y} \hat{y}. \quad (5.25)$$

Since the local derivatives are approximated by the average over the respective lenslet, this can introduce significant errors in the reconstructed wavefront (fitting errors), particularly for larger lenslet areas (Neal, Topa, & Copland, 2001). Different techniques are used to reconstruct the wavefront from the slope measurements, such as direct numerical integration (zonal) or polynomial fitting (modal). Southwell (1980) provides a useful description of these methods.

5.3.1 Centroid estimation and Fisher information

Blurred spots produced by the lenslets signify a departure from the assumption of a locally uniform wavefront. In reality there may be wavefront features that are smaller than the lenslet diameter, so that these finer details manifest in the spot profiles (Fig. 5.5). Therefore, both the spot positions and profiles provide indispensable information about aberrations in the optical system, or in our case, the human eye. In an effort to preserve

all information, our method does not involve centroid estimation or wavefront reconstruction; the data consist of all detector irradiance values in the focal plane of the WFS, which we refer to as the raw detector outputs.

Fig. 5.5: Blurred spot profiles in the focal plane of a Shack-Hartmann WFS.

To illustrate the last point, we examined the Fisher information matrix when the data consisted of centroid positions and were used to estimate ocular parameters. If the index i denotes the Cartesian direction in the detector plane and the index j represents the lenslet number, we can write

$$\hat{\tau}_{ij} = \begin{cases} \hat{\tau}_{xj}, & \text{if } i = 1 \\ \hat{\tau}_{yj}, & \text{if } i = 2, \end{cases} \quad (5.26)$$

where $i = \{1, 2\}$, $j = \{1, \dots, L\}$, and L is the total number of lenslets. By combining (5.23) and (5.26), we can formulate the average spot position for the j^{th} lenslet:

$$\bar{\hat{\tau}}_{xj}(\boldsymbol{\theta}) \approx \frac{\sum_{m \in AOI, j} \bar{g}_m(\boldsymbol{\theta}) x_m}{\bar{g}_{\text{tot}}}, \quad (5.27a)$$

$$\bar{\hat{\tau}}_{yj}(\boldsymbol{\theta}) \approx \frac{\sum_{m \in AOI, j} \bar{g}_m(\boldsymbol{\theta}) y_m}{\bar{g}_{\text{tot}}}. \quad (5.27b)$$

The FIM components are then expressed as

$$\begin{aligned} F_{kl} &= \frac{1}{\sigma_\tau^2} \sum_{i=1}^2 \sum_{j=1}^L \frac{\partial \bar{\hat{\tau}}_{ij}(\boldsymbol{\theta})}{\partial \theta_k} \frac{\partial \bar{\hat{\tau}}_{ij}(\boldsymbol{\theta})}{\partial \theta_l} \\ &= \frac{1}{\sigma_\tau^2} \sum_{j=1}^L \left(\frac{\partial \bar{\hat{\tau}}_{xj}(\boldsymbol{\theta})}{\partial \theta_k} \frac{\partial \bar{\hat{\tau}}_{xj}(\boldsymbol{\theta})}{\partial \theta_l} + \frac{\partial \bar{\hat{\tau}}_{yj}(\boldsymbol{\theta})}{\partial \theta_k} \frac{\partial \bar{\hat{\tau}}_{yj}(\boldsymbol{\theta})}{\partial \theta_l} \right), \end{aligned} \quad (5.28)$$

where σ_τ^2 is the variance in the centroid estimates, and θ_k and θ_l denote the k^{th} and l^{th} parameters, respectively.

The variance σ_τ^2 is attributed to centroid estimation error and is a direct consequence of the electronic noise in detector systems. For a set of N successive measurements, we have

$$\sigma_\tau^2 = \frac{1}{N} \sum_{n=1}^N \left\{ \frac{1}{L} \sum_{j=1}^L [(\hat{\tau}_{xj} - \bar{\hat{\tau}}_{xj})^2 + (\hat{\tau}_{yj} - \bar{\hat{\tau}}_{yj})^2] \right\}_n. \quad (5.29)$$

This simple formula of course assumes that the x and y centroids, as well as the detector elements and individual measurements, are statistically independent (Neal *et al.*, 2002). The centroid estimation error can be measured by performing consecutive measurements of the same true wavefront, then analyzing the centroid positions.

Computation of centroids can possibly lead to correlations in the estimates, which would require the variance in (5.29) to be replaced with a suitable covariance matrix. Barrett *et al.* (2007) suggests a full treatment of the statistical properties of the estimates $\hat{\boldsymbol{\tau}}_j = (\hat{\tau}_{xj}, \hat{\tau}_{yj})$, by initially declaring a conditional PDF $\text{pr}(\hat{\boldsymbol{\tau}}_j | \theta)$. They suggest that a more realistic PDF on the data may be a correlated multivariate normal distribution.

The information loss when replacing raw detector outputs with centroid positions as data in the FIM can be demonstrated by observing the increase in the Cramér-Rao lower bounds.

5.4 Data-acquisition system

5.4.1 System configuration

The data acquisition system for performing inverse optical design was modeled after a clinical Shack-Hartmann aberrometer for measurement of aberrations in human eyes, developed by Straub, Schwiegerling, and Gupta (2001), as shown in Figure 1.1. In the clinical aberrometer, a narrow collimated beam from a laser diode produces a spot on the retina, which acts as a laser beacon and fills the dilated 6-mm pupil upon reflection (Fig. 1.1). The 30-nm bandwidth of the source, centered at 780 nm, reduces speckle

noise in the real system, but we used a single wavelength of 780 nm in our computerized model to minimize computation time. We considered an ideal beamsplitter and illuminated the eye with a Gaussian beam (1 mm at FWHM) at multiple angles of 0, 6, and 12 degrees in the vertical direction to assess both on-axis and off-axis aberrations, thereby increasing the amount of Fisher information in the system. The center of the beam was coincident with the intersection of the optical axis and anterior corneal surface. We treated the aberrated retinal spot as a perfect diffuse scatterer and did not account for scattering within the ocular media or internal reflections. While the clinical configuration uses relay optics to conjugate the exit pupil of the eye to the SHWFS, we simply placed the lenslet array 10 mm away from the corneal apex. Global wavefront tip and tilt were discarded by rotating the sensor for off-axis angles. The lenslets were 0.6 mm in diameter and 24 mm in focal length, and the detector pitch was 8.0 μm . Our configuration is shown in Figure 5.6.

Fig. 5.6: Data acquisition system for estimating ocular parameters.

5.4.2 Optical-design program

We developed an optical-design program in C that performs non-paraxial ray-tracing through quadric surfaces for ocular parameters comparable to those in the Navarro wide-angle schematic eye model, given in Table 5.2. Each detector data set resulted from tracing a 256×256 bundle of rays through the double-pass system of the eye, where approximately 60% of the rays survived the double-pass after the diffuse retinal reflection and vignetting at the pupil. The WFS lenslets were treated as ideal thin lenses in our model.

Next, we assumed that the system is not photon-starved and used Gaussian statistics to represent electronic noise. We presented in Section 2.8 that for i.i.d. detector

elements and noise that is independent of the illumination level, the probability density function (PDF) from which the data are drawn is given by

$$\text{pr}(\mathbf{g} | \boldsymbol{\theta}) = \prod_{m=1}^M \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[g_m - \bar{g}_m(\boldsymbol{\theta})]^2}{2\sigma^2}\right\}, \quad (5.30)$$

where $\mathbf{g}$ is the $M \times 1$ vector set of random data, $\boldsymbol{\theta}$ is the set of estimable parameters, and σ^2 is the variance in each detector element. Since the noise is zero-mean Gaussian, $\bar{g}_m(\boldsymbol{\theta})$ is simply the output of the optical-design program. We fixed the variance to obtain a modest peak SNR of 10^3 and applied noise in the data using a Gaussian random number generator.

The geometrical eye model corresponding to the ocular parameters provided in Table 5.2 is illustrated in Figure 5.7, which was generated using our optical design program. Sample rays for the multiple beam angles used in this study $\alpha = \{0^\circ, 6^\circ, 12^\circ\}$ are also plotted.

Fig. 5.7: Geometrical eye model used to generate WFS data, corresponding to ocular parameters in Table 5.2.

The trial set of WFS data for the different beam angles, used as input to inverse optical design, is provided in Figures 5.8 – 5.10. In each image, central focal spots are sharp and focused, while spots near the periphery are blurred and smeared due to the worsening of aberrations in this region. The peripheral spots become increasingly smeared for larger off-axis angles.

Fig. 5.8: WFS data used as input to inverse optical design, for beam angle $\alpha = 0^\circ$.

Fig. 5.9: WFS data used as input to inverse optical design, for beam angle $\alpha = 6^\circ$.

Fig. 5.10: WFS data used as input to inverse optical design, for beam angle $\alpha = 12^\circ$.

The corresponding focal spots on the retina are shown in Figures 5.11 – 5.13 with the coordinate system centered on the optical axis, so that the position on the retina can

be read from the axes. Notice that the spot for beam angle $\alpha = 0$ is centered at (0.027, 0.056) mm, due to the decentration of the lens (Fig. 5.11).

Fig. 5.11: Focal spot on the retina for a source beam angle of $\alpha = 0^\circ$.

As the beam angle increases, off-axis aberrations such as coma and astigmatism become more apparent in the retinal image (Figs. 5.12 & 5.13).

Fig. 5.12: Focal spot on the retina for a source beam angle of $\alpha = 6^\circ$.

Fig. 5.13: Focal spot on the retina for a source beam angle of $\alpha = 12^\circ$.

5.5 Fisher information and Cramér-Rao lower bounds

Adjustable system parameters to increase Fisher information include the beam size, beam angle, lenslet array geometry, detector element spacing, and variance in the detector elements. Multiple output planes can be combined to form a larger, diversified data set to decouple pairs of parameters in the FIM. Here we present the Fisher information matrices and Cramér-Rao lower bounds for various system configurations. In each case, the changes will be compared to the original system configuration as described in Section 5.4.

In this proof-of-principle study, we estimated a reduced set of ocular parameters, including the posterior radius, thickness, and refractive index of the cornea, the thickness and index of the anterior chamber, the anterior and posterior radius, thickness, and equivalent index of the crystalline lens, and the thickness of the vitreous body, for a total of 11 parameters. Thus, the FIM is an 11×11 symmetric matrix for each system configuration. This matrix is order-specific, and the indices of the estimated parameters are listed in Table 5.3. Note that the jk^{th} entry has units of

$$\text{units of } F_{jk} = \frac{1}{(\text{units of } \theta_j)(\text{units of } \theta_k)} . \quad (5.23)$$

In Section 2.7, we derived the FIM components for i.i.d. Gaussian data, given by

$$F_{jk} = \frac{1}{\sigma^2} \sum_{m=1}^M \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_k}. \quad (5.24)$$

We computed the FIM according to (5.24) for the system configuration, provided in Figure 5.14 on a base-10 logarithmic scale, with values ranging from $10^{9.32}$ to $10^{16.29}$. The large values in the FIM indicate that the data are very sensitive to changes in the parameters, however, the magnitudes of off-diagonal entries reveal a high degree of detrimental coupling between pairs of parameters. This is intuitive since first-order geometrical and optical parameters combine to form various optical quantities; for example, refractive index and thickness in optical path length, and curvatures, index, and thickness combine in optical power. Interestingly, the lenticular thickness ($p = 6$) is much less coupled to the lenticular anterior radius ($p = 2$) than the posterior radius ($p = 3$), with $F_{2,6} = 10^{9.50}$ and $F_{3,6} = 10^{11.38}$. The same goes for the lenticular refractive index ($p = 10$), since $F_{2,10} = 10^{11.40}$ and $F_{3,10} = 10^{13.36}$. As another example, the structure of the FIM for the four refractive indices, $p = \{8, 9, 10, 11\}$, show that the coupling is stronger for pairs (8,11) and (9,10), but weaker for other pairs.

Fig. 5.14: FIM for the chosen system configuration (log scale).

The inverse of the FIM is shown in Figure 5.15, also displayed on a logarithmic scale. We read off the diagonal entries to determine the CRB for each parameter estimate; the corresponding standard deviations, denoted γ_1 , are given in Table 5.3. These diminutive values permit the estimation of parameters to high precisions even under pessimistic noise levels, since we implemented a peak SNR of only 10^3 , while the CRB can immediately be improved by decreasing the variance in the detector pixels.

Fig. 5.15: Inverse of the FIM for the chosen system configuration (log scale).

Recall from Chapter 2 that for an efficient estimator, the inverse of the FIM is the covariance matrix of the estimates. Despite the feasibility of the CRB, there is also a considerable amount of off-axis structure in the inverse of the FIM, suggesting a high level of coupling between the parameter estimates.

One straightforward solution to eliminate the coupling among the parameters is to diagonalize the FIM, which is possible due to the properties of *Hermitian* matrices. A Hermitian matrix is one that is equal to its own conjugate transpose; the FIM fulfills this requirement, since it is real and symmetric. Barrett and Myers (2004) and Strang (1980) show that a Hermitian matrix can be diagonalized by an appropriate unitary transformation, possibly invoking Gram-Schmidt orthogonalization. The only question,

however, is whether the estimable parameters after the transformation would be useful from an ophthalmological standpoint.

Table 5.3: Square-root of the CRB (standard deviation) for various system configurations.

No.	Parameter	True Value	γ_1	γ_2	γ_3	γ_4
1	Cornea posterior radius [mm]	6.38	6.2×10^{-6}	4.1×10^{-5}	2.0×10^{-6}	2.0×10^{-5}
2	Lens anterior radius [mm]	10.85	5.4×10^{-6}	3.6×10^{-5}	2.1×10^{-6}	1.0×10^{-5}
3	Lens posterior radius [mm]	-5.92	7.7×10^{-6}	4.9×10^{-5}	1.9×10^{-6}	3.5×10^{-5}
4	Cornea thickness [mm]	0.554	5.6×10^{-6}	3.9×10^{-5}	1.8×10^{-6}	2.5×10^{-5}
5	Anterior chamber thickness [mm]	3.37	7.8×10^{-6}	4.6×10^{-5}	2.7×10^{-6}	3.3×10^{-5}
6	Lens thickness [mm]	4.09	7.1×10^{-6}	4.9×10^{-5}	2.4×10^{-6}	3.0×10^{-5}
7	Vitreous thickness [mm]	16.40	5.0×10^{-7}	3.5×10^{-6}	1.6×10^{-7}	8.0×10^{-6}
8	Cornea refractive index	1.3729	9.0×10^{-8}	6.3×10^{-7}	2.5×10^{-8}	1.1×10^{-7}
9	Anterior chamber refractive index	1.3329	5.6×10^{-8}	3.7×10^{-7}	1.6×10^{-8}	4.7×10^{-7}
10	Lens refractive index	1.4138	8.4×10^{-8}	4.9×10^{-7}	2.7×10^{-8}	1.3×10^{-7}
11	Vitreous refractive index	1.3317	9.9×10^{-9}	6.8×10^{-8}	2.9×10^{-9}	9.4×10^{-8}

Adjusting the detector element size from 8 μm to 25 μm , but leaving all other system parameters unchanged, reduces the total number of detector elements that receive a signal from the source. Since only these elements contribute to the FIM, the pixel enlargement causes an increase in the CRB; the square-root of the CRB is denoted as γ_2 in Table 5.3. Although the overall FIM is smaller in magnitude, there is a minimal effect on the structure of the FIM (Fig. 5.16) and inverse FIM (Fig. 5.17), so that the relative degree of coupling between parameters is roughly the same.

Fig. 5.16: FIM for the system after increasing the detector element size.

Fig. 5.17: Inverse of the FIM after increasing the detector element size.

A very similar effect is observed after increasing the beam diameter (from 1 mm to 2 mm) and the pupil diameter (from 6 mm to 8 mm), with no other changes. Note that we are again comparing to the original system configuration with 8- μm detector elements. The larger beam and pupil sizes produce a bigger focal spot on the retina, which in turn leads to bigger spots in the focal plane of the WFS. Figure 5.18 provides the new detector data for $\alpha = 0^\circ$, showing larger focal spots throughout the image. More pixels now receive a non-zero signal, which increases the Fisher information in the system, therefore the CRB becomes smaller. The corresponding standard deviations are labeled γ_3 in Table 5.3. Once again, however, there is a negligible difference in the structure of the FIM (Fig. 5.19) and inverse FIM (5.20).

Fig. 5.18: Detector data for $\alpha = 0^\circ$ after increasing the beam and pupil diameters.

Fig. 5.19: FIM for the system after increasing the beam and pupil diameters.

Fig. 5.20: Inverse of the FIM after increasing the beam and pupil diameters.

Something more interesting happens when we reduce the number of beam angles to a single angle of $\alpha = 0^\circ$. In this case the system loses its sensitivity to off-axis aberrations, which has a considerable impact on the structure of the FIM (Fig. 5.21). Compared to Figure 5.14, we immediately observe greater entanglement among the refractive indices, $p = \{8, 9, 10, 11\}$, based on the off-axis structure in this region. Upon closer inspection, we can also see relatively higher coupling between the indices and other parameters, aside from an overall decrease in the magnitude of the FIM.

Fig. 5.21: FIM for the system after reducing the number of beam angles to one.

Inversion of the FIM (Fig. 5.22) leads to a matrix that dramatically differs from the original inverse matrix, shown in (5.15). There is a moderate increase in the CRB,

whose square-root is denoted as γ_4 in Table 5.3, but this would probably be less troublesome during the estimation step, compared to the greater parametric coupling in the system.

Fig. 5.22: Inverse of the FIM after reducing the number of beam angles.

5.6 Likelihood surfaces

When solving optimization problems, it helps to have a strong sense of the objective function to be optimized. This is particularly useful when fitting nonlinear, multivariate functions that may be plagued with numerous local extrema.

In the context of ML estimation, a plot of $\text{pr}(\mathbf{g}|\boldsymbol{\theta})$ versus $\boldsymbol{\theta}$ for a particular $\mathbf{g}$ is called the *likelihood surface* for that data vector. We demonstrated in Section 2.7 that for a purely Gaussian noise model, ML estimation reduces to nonlinear least-squares fitting between the data and the output of the optical-design program:

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{m=1}^M [g_m - \bar{g}_m(\boldsymbol{\theta})]^2. \quad (5.25)$$

Although we are minimizing the sum-of-squares (i.e., negative likelihood), we will refer to this objective function as the likelihood surface.

For a vector set of N parameters, the likelihood surface exists in an N -dimensional hyperspace. However, we are currently restricted to plotting the likelihood surface while varying up to two parameters at a time. We selected a handful of pairs of parameters for the following figures and applied the ranges given in Table 5.4. In each plot, the fixed parameters were set to the true values underlying the data.

Figures 5.23 – 5.32 illustrate the likelihood surface along the axes of the posterior corneal radius, $R_{\text{cornea, posterior}}$, and each of the other 10 parameters. For every pair of parameters, including those not shown in the following plots, there exists a groove along which the likelihood is nearly constant and that runs through the global minimum. We speculate that these ridges are caused by the same coupling of parameters that is evident in the Fisher information matrices.

Of the aforementioned plots, Figures 5.24, 5.26 – 5.28, 5.31, and 5.32 feature a high barrier followed by a local minimum at the boundary. The absence of a local minimum in the other figures does not guarantee an absence of local minima at other points in the search space. Due to the strong entanglement between parameters, a 2-D likelihood plot is very likely to change in shape or scale, if the 9 fixed parameters are altered to some extent, which will likely shift the position of the local minimum .

The $\times$ signs correspond to the final ML estimates, which we will save for discussion in the next section.

Fig. 5.23: Likelihood surface along $R_{\text{cornea,posterior}}$ and $R_{\text{lens,anterior}}$ axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.24: Likelihood surface along $R_{\text{cornea,posterior}}$ and $R_{\text{lens,posterior}}$ axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.25: Likelihood surface along $R_{\text{cornea,posterior}}$ and Δt_{cornea} axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.26: Likelihood surface along $R_{\text{cornea,posterior}}$ and $\Delta t_{\text{ant.chamber}}$ axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.27: Likelihood surface along $R_{\text{cornea,posterior}}$ and Δt_{lens} axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.28: Likelihood surface along $R_{\text{cornea,posterior}}$ and $\Delta t_{\text{vitreous}}$ axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.29: Likelihood surface along $R_{\text{cornea,posterior}}$ and n_{cornea} axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.30: Likelihood surface along $R_{\text{cornea,posterior}}$ and $n_{\text{ant.chamber}}$ axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.31: Likelihood surface along $R_{\text{cornea,posterior}}$ and n_{lens} axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.32: Likelihood surface along $R_{\text{cornea, posterior}}$ and n_{vitreous} axes. Final ML estimates indicated by $\times$ sign.

Figures 5.33 – 5.36 illustrate how the likelihood surface varies with both the thickness Δt and refractive index n of each optical medium in the eye, which closely resemble the previous plots. At first glance, Figure 5.33 appears different than the others, but bear in mind that the range in corneal thickness Δt_{cornea} is only 0.54 – 0.56 mm, based on the low variation across populations. Interestingly, both the likelihood and optical path length, $\text{OPL} = n\Delta t$, are roughly constant along each groove, since the grooves are straight.

Fig. 5.33: Likelihood surface along Δt_{cornea} and n_{cornea} axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.34: Likelihood surface along $\Delta t_{ant.chamber}$ and $n_{ant.chamber}$ axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.35: Likelihood surface along Δt_{lens} and n_{lens} axes. Final ML estimates indicated by $\times$ sign.

Fig. 5.36: Likelihood surface along $\Delta t_{\text{vitreous}}$ and n_{vitreous} axes. Final ML estimates indicated by $\times$ sign.

The likelihood plots comparing pairs of thicknesses or pairs of indices, as well as the remaining pairs, are markedly similar to the plots we have shown so far. As mentioned, the common features are a low groove containing the global minimum and a high ridge that is roughly parallel to the groove. To understand this general behavior, we picked a plot exhibiting these features, Figure 5.28, and ran our ray-trace program for several different points in parameter space (Fig. 5.37). P1 represents the true values of the parameters underlying the data, corresponding to a nearsighted (*myopic*) eye, which focuses before the retina when accommodation is relaxed (Fig. 5.38). We formed a line perpendicular to the ridge and groove, running through P1 and two additional points, P2 and P3. As we follow this line, we pass through paraxial focus, represented by P3 and a peak cost function value (Fig. 5.39). The cost function then decreases as we approach P2, corresponding to a farsighted (*hyperopic*) eye (Fig. 5.40). Other selected points on the ridge, P4 and P5, lead to paraxial focus as well (Figs. 5.41 & 5.42). This illustrates that the likelihood is largely a function of defocus and that points in parameter space having the same likelihood correspond to comparable levels of defocus.

Fig. 5.37: Understanding the likelihood as a function of defocus. P1 corresponds to the true minimum and a myopic eye (focuses before retina); P3, P4, and P5 are high points and correspond to zero defocus; P2 corresponds to a hyperopic eye (focuses behind retina).

Fig. 5.38: Level of defocus at P1 ($R_{\text{cornea,posterior}} = 6.381$ mm, $n_{\text{vitreous}} = 16.40$ mm).

5.39: Level of defocus at P3 ($R_{\text{cornea,posterior}} = 6.188$ mm, $n_{\text{vitreous}} = 15.97$ mm).

5.40: Level of defocus at P2 ($R_{\text{cornea,posterior}} = 6.000 \text{ mm}$, $n_{\text{vitreous}} = 15.50 \text{ mm}$).

5.41: Level of defocus at P4 ($R_{\text{cornea,posterior}} = 6.512 \text{ mm}$, $n_{\text{vitreous}} = 15.85 \text{ mm}$).

5.42: Level of defocus at P5 ($R_{\text{cornea,posterior}} = 5.871$ mm, $n_{\text{vitreous}} = 16.09$ mm).

5.7 Maximum-likelihood estimation of ocular parameters

After generating the data, we pretended not to know the values of selected parameters and estimated them by maximizing the likelihood, or minimizing the sum-of-squares as in (5.25), since we used a Gaussian noise model. We performed the optimization with the simulated annealing algorithm presented in Section 3.3.

After testing a series of tuning parameter combinations, we chose the following values during initialization (see Sec. 3.3.4):

$$\tau_0 = 10^5,$$

$$\delta = 0.1, \quad N_\delta = 5,$$

$$N_S = 80, \quad c = 2.0,$$

$$N_T = 60, \quad r_T = 0.9,$$

$$\mathbf{v}_0 = 0.5 (\boldsymbol{\theta}_{\text{upper}} - \boldsymbol{\theta}_{\text{lower}}),$$

where the initial temperature τ_0 has the same units as the cost function, and $\boldsymbol{\theta}_{\text{upper}}$ and $\boldsymbol{\theta}_{\text{lower}}$ are the upper and lower limits in the parameter space during the search process.

We verified that τ_0 was large enough for the system to perform a random search in the parameter space (i.e., nearly all proposed configurations are accepted), so that the local minima may be sufficiently sampled. We selected limits that compared well with the normal ranges in the parameters, based on clinical population studies, and chose the center of the search space as the starting point. The number of iterations per temperature phase is $N_I = N_S \times N_T = 4800$, with an iteration defined as one cycle through all parameters.

Figure 5.42 shows the results from 16 optimization trials, plotting the optimal cost function versus the iteration number on a log-log scale. Each trial represented a different noise realization for the same parameter values, so that the variance in the estimates could be determined, and the same estimation procedure was operated on each trial. The plots begin at the end of the first temperature phase, or at 4800 iterations. After normalization by the detector variance and total number of elements, the average sum-of-squares described in (5.25) was 462.25 at the starting point in the optimization, 1.3136 at the termination point, and 1.0011 at the true global minimum. As the number of detector elements increases, the normalized true minimum should approach unity.

Fig. 5.43: 16 simulated annealing trials for the estimation of ocular parameters.

On a 64-bit AMD246 Opteron CPU with 3 GFLOPS of peak computing power, the average computation time per ray was $0.230 \mu\text{s}$. There were three output planes per forward propagation, corresponding to 45.2 ms of computation time for the bundle of rays. Thus, each temperature phase took 39.8 min to compute. There was an average of 80.1 phases for the optimization trials, so the average computation time per trial was 53 hrs .

We discussed in Section 6.2 that the maximum floating-point computing speed of a single NVIDIA Tesla C2075 GPU is 1030 GFLOPS . There are many other factors that affect performance in computers, such as memory speed and architecture, storage

technology, cache coherence, internal bus speeds, and software (i.e., operating system and application), so the clock rate alone does not accurately gauge relative performance, except when comparing it to other processors in the same line. However, assuming the computational time can be rescaled with a quoted computing speed, the total time would be reduced from 53 hours to 9.3 min with a single Tesla C2075.

The parameter estimates are listed in Table 5.4, along with the true values underlying the data, the starting point in the ML search, the upper and lower bounds of the search space, the final estimates, and the standard deviation in each estimate. All of the parameters were estimated to within one standard deviation, and each estimate has a very small bias and variance. The accuracy in the estimates were down to three to four decimal places for radii and thickness and four to five decimal places for refractive indices.

Table 5.4: Estimated ocular parameters, including the true values, starting point in the search, upper and lower limits in the search space, and estimated values with standard deviations.

No.	Parameter	True Value	Lower Limit	Starting Point	Upper Limit	ML Estimate
1	Cornea posterior radius [mm]	6.38	5.75	6.25	6.75	6.3805 ± 0.0008
2	Lens anterior radius [mm]	10.85	10.50	11.00	11.50	10.846 ± 0.007
3	Lens posterior radius [mm]	-5.92	-7.00	-6.00	-5.00	-5.922 ± 0.006
4	Cornea thickness [mm]	0.554	0.54	0.55	0.56	0.553 ± 0.002
5	Anterior chamber thickness [mm]	3.37	3.00	3.75	4.50	3.368 ± 0.006
6	Lens thickness [mm]	4.09	3.25	4.00	4.75	4.08 ± 0.01
7	Vitreous thickness [mm]	16.40	15.50	16.50	17.50	16.397 ± 0.005
8	Cornea refractive index	1.3729	1.3700	1.3750	1.3800	1.3728 ± 0.0002
9	Anterior chamber refractive index	1.3329	1.3300	1.3350	1.3400	1.33287 ± 0.00007
10	Lens refractive index	1.4138	1.4100	1.4150	1.4200	1.4139 ± 0.0001
11	Vitreous refractive index	1.3317	1.3300	1.3350	1.3400	1.33175 ± 0.00008

Figure 5.41 illustrates the eye model reconstructed from the estimates, showing precise overlap with the true model. Though the corneal anterior radius was not

estimated, we plotted it in the reconstruction to indicate the estimated corneal thickness. Another way to visualize the ML estimates is to represent them as points on the likelihood surface; we used $\times$ marks rather than actual points in the likelihood plots in Figures 5.23 – 5.41. Notice that all of these marks fall on the ridges that run through the true minimum.

Fig. 5.43: Reconstructed eye model of the estimated parameters, superimposed with the true values underlying the data.

5.8 Summary of Chapter 5

We estimated patient-specific ocular parameters using simulated WFS data with Gaussian noise, ML estimation, and an adaptive simulated annealing algorithm tailored to our

probability surface. Our optical-design program performed non-paraxial ray-tracing through quadric surfaces that may contain surface misalignments.

An important result of this study is that the centroiding process in traditional wavefront sensing results in severe information loss, which we demonstrated by examining the FIM. In our method, we do not perform centroid estimation or wavefront reconstruction, but instead use the raw detector outputs of the WFS as input to IOD. We saw that another way of increasing information is to enlarge the beam and pupil diameters, resulting in larger spots on the retina and in the focal plane of the WFS. This is also opposite to classical wavefront sensing, in which smaller focal spots are preferred for the centroid estimation step. The key point here is that IOD actually prefers poor imaging for great information yield and smaller CRBs.

After investigating the Fisher information in various system configurations, we implemented multiple input angles of the source beam to assess both on- and off-axis aberrations, which produced a feasible Cramér-Rao lower bound and reduced parametric coupling of ocular parameters.

Although we obtained excellent results in this proof-of-principle study, we are still far from working with real patient data. The greatest obstacle in performing inverse optical design of the human eye is certainly the requirement of an accurate forward model of the eye that includes all sources of randomness. Fluctuations of ocular aberrations associated with live imaging of the eye, such as the optical tear film effect, would have to be taken into account. Future studies must consider more complexities in the model, such as the coherence properties of the source, the GRIN distribution of the crystalline lens,

irregularities in the corneal surface, scattering in the ocular media, Fresnel reflections, and the Stiles-Crawford effect. It may even be necessary to consider other sources of noise, since an i.i.d. Gaussian model is very idealized, as well a more realistic model for the WFS lenslets. We also anticipate practical issues such as stray light and misalignments, or other effects that may contribute to estimation errors. These issues are beyond the scope of this stage of the research, but may be dealt with in subsequent stages.

To enhance the inverse optical design algorithm, prior information obtained from statistical studies or by other reliable modalities, such as corneal topography, can be incorporated. This information can be used to select a starting point or to narrow the search space. An approximate method based on optimization with reverse ray-tracing (Goncharov, Nowakowski, Sheehan, & Dainty, 2008) could provide a promising starting point in our likelihood approach.

In order to reach clinical application, rapid processing techniques must be explored to improve computation time, which would allow an increase in model robustness. This can be accomplished with dedicated computer hardware and the parallelization of the optical-design program. Global search algorithms such as simulated annealing are very time consuming, though if we could get to a point of negligible parameter coupling and a unimodal likelihood surface, perhaps straightforward optimization could be used instead. This would not only reduce the computation time, but may also result in more reliable estimates. Ultimately, the goal is to make the system

practical in the clinical setting and to obtain accurate estimates of the full set of ocular parameters for a given patient.

CHAPTER 6

MAXIMUM-LIKELIHOOD ESTIMATION OF PARAMETERIZED WAVEFRONTS USING MULTIFOCAL DATA

In this chapter, we present the second of three applications of likelihood methods in optics as it applies to high-precision optical testing. Section 6.1 introduces the general approach to acquiring irradiance data from which wavefront parameters are estimated. Section 6.2 is dedicated to modeling the wave propagation and emphasizes the complexity and accuracy of the computations needed for finding ML estimates. We conclude with a discussion on the graphics processing unit (GPU) for rapid processing, since the estimation procedure requires substantial computation time.

In Section 6.3 and Section 6.4, we present proof-of-principle results obtained in simulation and in experiment, respectively. We discuss the added challenge of large-aberration wavefronts in the numerical study. On the experimental side, we discuss the accuracy of the propagation algorithm and the handling of nuisance parameters. In both cases, we examine Fisher information matrices, Cramér-Rao bounds, and likelihood surfaces, after which we provide ML estimates obtained by simulated annealing.

6.1 Formulation of the problem

Phase retrieval (PR) is a useful method for recovering the phase distribution in the pupil of an optical system from the irradiance distribution in the focal plane. However, the

usual PR problem is ill posed, since both distributions are unrestricted 2D functions (Stefanescu, 1985) and a single irradiance measurement does not ensure that the recovered phase is unique (Seldin & Fienup, 1990; Teague, 1983). A straightforward solution to avoiding the ambiguities is to make multiple irradiance measurements near the focal plane.

Another approach to avoiding the phase ambiguities is to estimate the parameters describing the phase function, instead of the function itself (Brady & Fienup, 2004, 2005; Brady, Guizar-Sicairos, Fienup, 2009). Parameterization of the phase is achieved with a set of expansion functions. Obvious choices for circular pupils are Zernike polynomials, although there are many other options.

In our method, the data consist of irradiance measurements in multiple planes near the focus of an aberrated optical element (Fig. 6.1). From the multifocal data, we estimate the coefficients of phase polynomials in a wavefront expansion, as proposed by Brady and Fienup, but with various extensions. As mentioned briefly in Section 1.2, we optimize the data-acquisition system by analyzing the Fisher information matrix and Cramér-Rao lower bound, as well as the likelihood surface. Additionally, we have developed our method to handle very large wavefront aberrations. To deal with the resulting high computational demand, we employ rapid-processing techniques using dedicated computer hardware.

Fig. 6.1: Data-acquisition system for collecting multiple irradiance patterns near the focus of an optical element.

6.2 Propagation algorithm

6.2.1 Diffraction propagation vs. ray-tracing

An imaging system whose aberrations are well corrected will convert a diverging spherical wave emitted by a point source into a converging spherical wave centered on the geometrical image point, through which all geometrical rays or wavefront normals will pass. Since there are an infinite number of rays and each ray contributes a finite amount of energy, the geometrical treatment predicts an infinite irradiance at the focus and zero irradiance everywhere else in the image plane. We know that this is nonphysical, therefore geometrical optics is invalid in the focal region (Stamnes, 1986).

For an imaging system that is not well corrected, the converging wave leaving the system will no longer be spherical and all rays will not meet at the geometrical focus. If

the aberrations are sufficiently large, the ray-density may provide a crude approximation of the irradiance distribution in the focal region, though the imaging is poor. Even under this circumstance, the prediction by geometrical optics is still invalid in the vicinity of the caustic. Thus, in all cases, we must consider diffraction theory when determining the irradiance in the focal region (Stamnes, 1986). While our method of estimating wavefront parameters from irradiance data does not make demands of the image quality, it does require an accurate forward model of the system when working with real data.

The diffraction integral without the Fresnel approximation is given by (4.79),

$$u_z(\mathbf{r}) = \frac{1}{i\lambda z} \int_{\infty} d^2r_0 u_0(\mathbf{r}_0) \exp\left(ik\sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2}\right), \quad (6.1)$$

which is essentially a mathematical refinement of the Huygens wavelet formulation. The field at each observation point in the image plane is the sum of an infinite number of secondary waves emanating from the aperture. Since the integrand of the double integral in (6.1) can vary rapidly over the integration domain, especially when the aberrations are large, a significant amount of sampling in the aperture is required to accurately predict the irradiance distribution in the image plane. Thus, numerical evaluations of the Huygens diffraction integral are prohibitively time-consuming and an approximation to this integral is necessary to reduce the computing problem.

Note that both the Huygens wavelet formation and the Fresnel approximation assume scalar diffraction theory, which we assume to be adequate for our method of wavefront measurement with multifocal data, even when large aberrations are present.

6.2.2 Diffraction equation for a converging spherical wave

Suppose a converging wave has wavefront error $W(\mathbf{r}_0)$ in the exit pupil of an optical system. Then the field in the exit pupil is given by

$$u_0(\mathbf{r}_0) = \exp[ikW(\mathbf{r}_0)] \frac{\exp[-ikR_f(\mathbf{r}_0)]}{R_f(\mathbf{r}_0)}, \quad (6.2)$$

where $R_f(\mathbf{r}_0) = \sqrt{|\mathbf{r}_0|^2 + f^2}$ and f is defined as the radius of curvature of the unaberrated wave in the exit pupil (i.e., the distance between the exit pupil and paraxial focus). Inserting (6.2) into (6.1) leads to

$$u_z(\mathbf{r}) = \frac{1}{i\lambda z} \int_{xp} d^2r_0 \exp[ikW(\mathbf{r}_0)] \frac{\exp[-ikR_f(\mathbf{r}_0)]}{R_f(\mathbf{r}_0)} \exp\left(ik\sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2}\right), \quad (6.3)$$

or equivalently,

$$u_z(\mathbf{r}) = \frac{1}{i\lambda z} \int_{xp} d^2r_0 \frac{\exp[ikW(\mathbf{r}_0)]}{R_f(\mathbf{r}_0)} \exp\left(-ik \frac{\mathbf{r} \cdot \mathbf{r}_0}{z}\right)$$

$$\times \exp \left[ik \left(\sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2} - \sqrt{|\mathbf{r}_0|^2 + f^2} + \frac{\mathbf{r} \cdot \mathbf{r}_0}{z} \right) \right]. \quad (6.4)$$

Note that the last exponential term is part of the integrand. A binomial expansion for

$z > |\mathbf{r} - \mathbf{r}_0|$ and $f > r_0$ leads to

$$\sqrt{|\mathbf{r} - \mathbf{r}_0|^2 + z^2} - \sqrt{|\mathbf{r}_0|^2 + f^2} + \frac{\mathbf{r} \cdot \mathbf{r}_0}{z} = z - f + \frac{r^2}{2z} + \frac{r_0^2}{2} \left(\frac{1}{z} - \frac{1}{f} \right) + \text{HOT}, \quad (6.5)$$

where HOT denotes the higher-order terms in $\mathbf{r}_0$:

$$\text{HOT} = -\frac{|\mathbf{r} - \mathbf{r}_0|^4}{8z^3} + \frac{|\mathbf{r}_0|^4}{8f^3} + \frac{|\mathbf{r} - \mathbf{r}_0|^6}{16z^5} - \frac{|\mathbf{r}_0|^6}{16f^5} + \dots \quad (6.6)$$

Combining (6.4) and (6.5) results in

$$u_z(\mathbf{r}) = A(r) \int_{xp} d^2 r_0 \frac{\exp[ikW(\mathbf{r}_0)]}{R_f(\mathbf{r}_0)} \exp \left(-ik \frac{\mathbf{r} \cdot \mathbf{r}_0}{z} \right) \exp \left\{ ik \left[\frac{r_0^2}{2} \left(\frac{1}{z} - \frac{1}{f} \right) + \text{HOT} \right] \right\}. \quad (6.7)$$

where the function $A(r)$ is given by

$$A(r) = \frac{1}{i\lambda z} \exp \left[ik \left(z - f + \frac{r^2}{2z} \right) \right]. \quad (6.8)$$

If the higher-order terms in (6.7) can be ignored, that equation can be represented by a 2D Fourier transform,

$$u_z(\mathbf{r}) \approx A(r) \mathbf{F}_2 \left\{ \frac{1}{R_f(\mathbf{r}_0)} \exp[ikW(\mathbf{r}_0)] \exp \left[ik \frac{r_0^2}{2} \left(\frac{1}{z} - \frac{1}{f} \right) \right] \right\}_{\boldsymbol{\rho}=\mathbf{r}/\lambda z}, \quad (6.9)$$

where the spatial frequency is $\boldsymbol{\rho} = \mathbf{r} / \lambda z$, as we saw in Section 4.4.4. The corresponding irradiance under this approximation is given by

$$I(\mathbf{r}) = |u_z(\mathbf{r})|^2 \approx \frac{1}{\lambda^2 z^2} \left| \mathbf{F}_2 \left\{ \frac{1}{R_f(\mathbf{r}_0)} \exp[ikW(\mathbf{r}_0)] \exp \left[ik \frac{r_0^2}{2} \left(\frac{1}{z} - \frac{1}{f} \right) \right] \right\}_{\boldsymbol{\rho}=\mathbf{r}/\lambda z} \right|^2. \quad (6.10)$$

The approximation in (6.9) breaks down as the numerical aperture of the optical system increases. Under circumstances when the higher-order terms are not negligible, we might consider including, say, the fourth-order terms in (6.6). The caveat is that $\mathbf{r}$ and $\mathbf{r}_0$ are inseparable in these terms, so including them in the integral means that we cannot take advantage of the FFT. Since a brute-force computation of the diffraction integral is too computationally expensive and impractical, as we will show in Section 6.4.5, the best we can do is minimize the terms in (6.6) by considering planes sufficiently close to nominal focus, so that $z \approx f$ and $\mathbf{r}$ is small.

6.2.3 *Parameterized wavefront description*

The unknown function of interest in these equations is the wavefront error $W(\mathbf{r}_0)$, which we represent in a parameterized form in the exit pupil of the optical system. Our approach is based on the fundamental assumption that the continuous wavefront can be approximated to sufficient accuracy by a finite set of expansion functions. We choose to represent this function by expanding it in some number of Zernike polynomials,

$$W(\mathbf{r}_0) \approx \sum_{n=1}^N \alpha_n Z_n(\mathbf{r}_0), \quad (6.11)$$

where $Z_n(\mathbf{r}_0)$ is the n^{th} Zernike polynomial with coefficient α_n and $\mathbf{r}_0$ is a 2D position vector in the pupil.

The parameters to be estimated are the Zernike coefficients $\{\alpha_n, n = 1, \dots, N\}$, but an important determination is the number of coefficients necessary for an accurate representation of the wavefront. Even if a small number of terms is used in the expansion, this does not imply that the wavefront aberration is small; representing large wavefront errors simply requires large coefficients. In our approach, we assume that the wavefront is smoothly varying, so that sufficient accuracy can be achieved with a relatively small value of N . We choose $N = 37$, the maximum number of coefficients calculated in ZEMAX.

Specifically, we choose to use the Fringe Zernike polynomials, provided in Appendix A. These are identical to the original Zernike polynomials, except for the manner and order in which they are listed.

6.2.4 Sampling considerations

Let D_{xp} denote the diameter of the exit pupil, $P \times P$ the array size in the pupil plane, and $F \times F$ the array size for the FFT propagation including zero-padding, so that $F > P$.

Then we define the following parameters,

$$\Delta x_p \equiv \frac{D_{xp}}{P}, \quad (6.12a)$$

$$\Delta \nu \equiv \frac{1}{F \Delta x_p} = \frac{P}{F} \frac{1}{D_{xp}}, \quad F > P, \quad (6.12b)$$

$$\Delta x_d \equiv \lambda z \Delta \nu = \frac{P}{F} \frac{\lambda z}{D_{xp}}, \quad (6.12c)$$

where Δx_p is the pupil element spacing, $\Delta \nu$ is the spatial frequency spacing, and Δx_d is the detector element spacing in each transverse direction. Most of the $F \times F$ detector elements do not receive a signal, which only creates unnecessary computation time when computing the objective function during optimization. For this reason, we automatically extract the $M_{ID} \times M_{ID}$ innermost elements from the $F \times F$ output of each FFT operation

and discard the remainder. Note that we use M_{1D} rather than M , since the latter is reserved for $M \times 1$ data vector $\mathbf{g}$.

6.2.5 *Parallel processing with the graphics processing unit*

Among the earliest applications driving the development of the graphics processing unit were computer-aided design and flight simulation in the 1960s. GPU technology has come a long way since then, and modern uses now include web browser graphics, complex mechanical CAD, DVD video playback, and 3D video games approaching cinematic realism. The same fundamental technology used in dedicated systems such as entertainment consoles and medical imaging stations are now being used for massively parallel high-performance computing in the scientific and engineering fields.

General-purpose computing with the GPU became possible at the turn of the century when pioneer programmers realized that the pixel shaders on graphics chips can be treated as *stream processors* or *thread processors* with their own registers and local memory. However, the programming model was extremely awkward and clumsy, requiring graphics application interfaces (APIs), such as OpenGL and Cg. By 2006, the GPU was modified with the added support of C/C++, which allowed significant simplification of the programming model and accessibility to a larger community of application programmers.

Modern GPU programming is accomplished using an extension to the C programming language called CUDA (“Compute Unified Device Architecture”), which provides software development tools and allows functions in C to be implemented on a

GPU's multiple stream processors. The programming model is *heterogeneous*; the sequential part runs on a *host* CPU and the computationally-intensive part on one or more compute *devices*, which are massively parallel coprocessors. CUDA devices support the Single-Instruction Multiple-Data (SIMD) model, in which all concurrent threads are based on the same code, though the path of execution may differ between threads. An important aspect of CUDA is that it features a hardware abstraction mechanism, by which the runtime transparently compiles the data-parallel computation to shader programs (Ryoo *et al.*, 2008).

CUDA programming is achieved with a minimal set of keywords and extensions of the standard ANSI C language, which assign *kernels*, or data-parallel functions, and their associated data structures to the compute devices. The kernels provide instructions to single threads, usually calling upon thousands of threads at a time. Threads are organized by the developer into thread bundles, or *thread blocks*, in which they can exchange data in their own *shared memory* and synchronize actions, while they also have access to *global memory* and read-only access to *constant memory* and *texture memory*. Through a language integration programming interface, the CUDA runtime supports the execution of standard C functions on the device, including library functions for managing device memory and transferring data between the host and device (Ryoo *et al.*, 2008).

There are many advantages of using GPU hardware as opposed to other hardware platforms, such as the field-programmable gate array (FPGA) or the cell broadband engine architecture (CBEA) found in the Sony PlayStation3. New hardware becomes available on a continual basis, providing much hardware flexibility based on

programming needs, plus a lot of memory exists on the device and host machine. Also, CUDA is comparatively straightforward to use with many useful library routines (e.g., FFTW, BLAS), resulting in high productivity of the programmer.

The latest, most advanced GPUs currently on the market include the NVIDIA Tesla models, such as the C1060 and C2075, whose specifications are provided in Table 6.1. For instance, the C2075 contains 448 processing cores and offers an unprecedented 515 GFLOPS of peak double-precision floating-point performance, where FLOPS stands for “floating-point operations per second”.

Table 6.1: Product specifications for NVIDIA® Tesla™ C1060 and C2075 models.

	Tesla C1060	Tesla C2075
Peak double-precision floating-point performance	78 GFLOPS	515 GFLOPS
Peak single-precision floating-point performance	933 GFLOPS	1030 GFLOPS
CUDA cores	240	448
Memory size	4 GB	6 GB
Memory bandwidth	102 GB/sec	144 GB/sec

6.3 Numerical studies

For the numerical proof-of-principle system, we chose a rather large lens with substantial aberrations. We examined the ideal amount of pupil sampling for a good representation of the irradiance data, but without wasting computational effort. We thoroughly investigated Fisher information matrices, Cramér-Rao bounds, and likelihood surfaces. Due to the multitude of local minima in the cost function, we used simulated annealing to obtain ML estimates of wavefront parameters.

6.3.1 Test lens description

Much of the work presented in this chapter was completed under a contract with a corporation that funded our research. Due to the confidentiality agreement between our institution and this corporation, we cannot disclose any design parameters of the lens system referenced in this numerical study. For the purpose of estimating wavefront parameters, however, the design parameters are not pertinent. We are primarily interested in the wavefront emerging from the system and the region of space between the exit pupil and the image planes.

We chose to operate the lens, which is rotationally-symmetric, at finite conjugates in our optical design program by placing an on-axis point source 113.0 mm from the entrance pupil of the lens. Table 6.2 summarizes the relevant system data calculated by ZEMAX using a wavelength of $\lambda = 0.6328 \mu\text{m}$, including an exit pupil diameter of $D_{xp} = 88.41 \text{ mm}$ and a working f-number of $f/\#_w = 1.796$. Note that the position of the paraxial

Table 6.2: System data provided by ZEMAX™ for the highly aberrated test lens at $\lambda = 0.6328 \mu\text{m}$.

Effective Focal Length [mm]	66.71521
Back Focal Length [mm]	40.80047
Image Space f/#	0.9309758
Paraxial Working f/#	1.785467
Working f/#	1.796028
Image Space NA	0.2696646
Object Space NA	0.3022556
Entrance Pupil Diameter [mm]	71.6616
Exit Pupil Diameter [mm]	88.40518
Paraxial Focal Plane Position [mm]	157.8443

focal plane at $z = z_f = 157.8 \text{ mm}$ is measured from the exit pupil, which lies in the $z = 0$ plane. An illustration of the focal region of the lens is provided in Figure 6.2, showing a distance of 5.0 mm between marginal and paraxial focus.

Fig. 6.2: Focal region of the highly aberrated test lens at $\lambda = 0.6328 \mu\text{m}$. Paraxial focal plane is at $z = z_f = 157.8 \text{ mm}$.

The Fringe Zernike coefficients describing the wavefront error $W(\mathbf{r}_0)$ in the exit pupil, according to ZEMAX, are provided for $N = 37$ in Table 6.3. Since the system has rotational symmetry, the coefficients for all non-rotationally symmetric Zernike terms are equal to zero, while non-zero coefficients correspond to piston, defocus, and various orders of spherical aberration. Figure 6.3 shows the wavefront error map computed with these coefficients as a function of normalized radius, indicating a peak-to-valley measurement of 149.1λ .

Table 6.3: Fringe Zernike coefficients $\{\alpha_n, n = 1, \dots, 37\}$, peak-to-valley, RMS, and variance, provided by ZEMAX for the highly aberrated test lens at $\lambda = 0.6328 \mu\text{m}$.

Unlisted coefficients are zero.

Index	Aberration Type [λ]	Design Value [λ]
1	Piston	50.73020573
4	Defocus	75.17884184
9	Spherical Aberration, Primary	23.81690475
16	Spherical Aberration, Secondary	-0.60555615
25	Spherical Aberration, Tertiary	0.04113341
36	Spherical Aberration, Quaternary	-0.00287742
37	Spherical Aberration, 12th-order Term	-0.01310320
Peak-to-valley [λ]	N/A	149.14956942
RMS [λ]	N/A	44.18240916
Variance [λ^2]	N/A	1952.08527937

Fig. 6.3: Wavefront error in the exit pupil of the highly aberrated test lens at $\lambda = 0.6328$ μm , as a function of normalized radius. Units are in waves.

6.3.2 Pupil sampling

Before simulating the irradiance data, we determined the optimal amount of pupil sampling to accurately represent the irradiance data without unnecessary increases in computation time.

A critical design option is the number and location of planes at which to measure the irradiance. It is well known that two-plane measurements are sufficient to determine the pupil phase, so to minimize computation time, we selected two output planes just before paraxial focus,

$$z = z_1 = z_f - 0.25 \text{ mm},$$

$$z = z_2 = z_f - 0.43 \text{ mm}.$$

For each plane, we computed the irradiance using (6.10) with pupil sampling levels of $P = 256$, $P = 512$, and $P = 1024$ (Figs. 6.4 & 6.5). In each case, we held the ratio P/F in (6.12) constant at $1/2$, thereby fixing the size of the detector elements in the output plane and allowing us to isolate the pupil sampling effect on the data. The data computed with $P = 256$ and $F = 512$ (Figs. 6.4c & 6.5c) exhibit severe signs of undersampling, such as discontinuities and D4 symmetry (i.e., symmetry of a square) from the rectilinear grid of the input function, and the irradiance patterns are clearly unphysical. While the data with $P = 512$ and $F = 1024$ (Figs. 6.4b & Fig. 6.5b) show much improvement, the sampling artifacts are still evident. Since these artifacts are diminished for $P = 1024$ and $F = 2048$ (Figs. 6.4a and 6.5a), we chose this pupil sampling for the study.

Fig. 6.4: Detector data at $z = z_1$ for the highly aberrated test lens using a pupil sampling of: (a) $P = 1024$, (b) $P = 512$, and (c) $P = 256$.

Fig. 6.5: Detector data at $z = z_2$ for the highly aberrated test lens using a pupil sampling of: (a) $P = 1024$, (b) $P = 512$, and (c) $P = 256$.

The final data set, containing electronic noise with a peak SNR of 10^4 , is shown in Figure 6.6. According to (5.12c), the detector element size is roughly $0.56\mu\text{m}$.

Fig. 6.6: Detector data for the highly aberrated test lens using a pupil sampling of $P = 1024$ at image plane: (a) $z = z_1$ and (b) $z = z_2$.

We saw that for a given system and image plane, the FFT method requires a minimum value of P to avoid sampling artifacts. Likewise, there is a finite range in the

image position, for which a given P is sufficient. For the system configuration described in Section 6.3.1 and with $P = 1024$, this range is limited to approximately

$$z_f - 1.8 \text{ mm} < z < z_f - 0.2 \text{ mm},$$

which is roughly one-third of the range between marginal and paraxial focus,

$$z_f - 5.0 \text{ mm} < z < z_f.$$

Beyond this range, the irradiance distribution degrades very rapidly.

6.3.3 Fisher information and Cramér-Rao lower bounds

We computed the Fisher information matrix using (2.76), according to an i.i.d. Gaussian noise model, for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 37\}$ (Fig. 6.7). Since α_1 corresponds to piston and does not influence the irradiance data, we had no interest in estimating it and disregarded it in the FIM. We determined the variance in the detector elements σ^2 with a peak SNR of 10^4 in the data and evaluated the FIM components at θ , based on the values in Table 6.3. (See Appendix A for a list of Fringe Zernike polynomials.)

The high diagonal values in the FIM indicate that there is abundant information in the data with respect to each estimable parameter. However, the off-diagonal structure is also very pronounced, indicating strong parametric coupling that may confound the estimation problem. Recall that the degree of coupling between two parameters is proportional to the magnitude of the respective FIM component.

Fig. 6.7: FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 37\}$, in the exit pupil of the highly aberrated test lens (log scale).

When an optical system has rotational symmetry, there is immediate interest in coefficients associated with defocus and spherical aberration, which are $\{\alpha_n, n = 4, 9, 16, 25, 36, 37\}$. Unsurprisingly, the FIM indicates strong coupling between these parameters, which only intensifies as the pupil sampling is reduced.

There is also significant coupling among terms with bilateral symmetry about the x-axis, including $\{\alpha_n, n = 2, 7, 10, 14, 19, 23, 26, 30\}$, which are x-axis tilt and all orders of x-axis coma and trefoil. The same is true for the y-axis counterparts of these terms, that is, for $\{\alpha_n, n = 3, 8, 11, 15, 20, 24, 27, 31\}$. However, the coupling is minimal for any pair that includes a parameter from each set, such as (α_2, α_3) or (α_7, α_{11}) . After all, it only makes sense that the system does not confuse changes in the x-direction with those in the y-direction, but it is reassuring that the FIM verifies this.

A similar observation can be made about terms with two axes of bilateral symmetry. Forming one group, we have the terms $\{\alpha_n, n = 5, 12, 21, 32\}$, whose axes of symmetry are the x- and y-axes. These correspond to all orders of astigmatism at 0° . Another group is formed with $\{\alpha_n, n = 6, 13, 22, 33\}$, the orders of astigmatism 45° . In other words, the system can distinguish any two astigmatic terms with relative ease, as long as one is at 0° and the other at 45° , but it has difficulty otherwise.

We computed the inverse of the FIM (Fig. 6.8) and read off its diagonal components to determine the Cramér-Rao bound for the parameters, whose square-root is provided in Table 6.3. The diminutive values in $(\text{CRB})^{1/2}$, on the order of 10^{-9} to 10^{-8} , permit the estimation of the wavefront parameters to very high precision, provided that the forward model is exact.

Fig. 6.8: Inverse of the FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 37\}$ in the exit pupil of the highly aberrated test lens (log scale).

Table 6.4: Square-root of the CRB for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 37\}$ in the exit pupil of the highly aberrated test lens at $\lambda = 0.6328 \mu\text{m}$. Units are in waves λ .

Index	True Value [λ]	$(\text{CRB})^{1/2}$ [λ]	Index	True Value [λ]	$(\text{CRB})^{1/2}$ [λ]
2	0	1.4×10^{-8}	20	0	1.2×10^{-8}
3	0	1.4×10^{-8}	21	0	4.8×10^{-9}
4	75.17884184	1.3×10^{-8}	22	0	4.5×10^{-8}
5	0	6.6×10^{-9}	23	0	7.7×10^{-9}
6	0	7.4×10^{-8}	24	0	7.7×10^{-9}
7	0	6.9×10^{-8}	25	0.04113341	5.5×10^{-9}
8	0	6.9×10^{-8}	26	0	7.6×10^{-9}
9	23.81690475	2.4×10^{-8}	27	0	7.7×10^{-9}
10	0	2.0×10^{-8}	28	0	2.8×10^{-9}
11	0	2.0×10^{-8}	29	0	2.9×10^{-9}
12	0	2.9×10^{-9}	30	0	7.8×10^{-9}
13	0	1.2×10^{-8}	31	0	7.8×10^{-9}
14	0	1.1×10^{-8}	32	0	1.9×10^{-9}
15	0	1.1×10^{-8}	33	0	1.7×10^{-8}
16	-0.60555615	2.4×10^{-8}	34	0	3.6×10^{-9}
17	0	4.1×10^{-9}	35	0	3.6×10^{-9}
18	0	5.6×10^{-9}	36	-0.00287742	1.2×10^{-8}
19	0	1.2×10^{-8}	37	-0.01310320	4.3×10^{-9}

In this proof-of-principle study, we chose to estimate the Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$, which we will discuss in Section 6.3.5. In a real physical system, the

coefficients related to tilt and off-axis aberrations (i.e., coma and astigmatism) would be useful in determining misalignments in the optical system, while the remaining coefficients represent defocus and spherical aberration, relating to optical power and the curvatures and asphericities of the refractive surfaces.

The reduced FIM for the selected coefficients is shown in Figure 6.9. After the inversion of this matrix (Fig. 6.10), we actually see an improvement in $(\text{CRB})^{1/2}$ by roughly one to two orders of magnitude for each parameter. These values are provided in Table 6.5.

Fig. 6.9: FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$, in the exit pupil of the highly aberrated test lens (log scale).

Fig. 6.10: Inverse of the FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$, in the exit pupil of the highly aberrated test lens (log scale).

Table 6.5: Square-root of the CRB for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$, in the exit pupil of the highly aberrated test lens at $\lambda = 0.6328 \mu\text{m}$. Units are in waves λ .

Index	True Value [λ]	$(\text{CRB})^{1/2}$ [λ]
2	0	6.6×10^{-10}
3	0	6.6×10^{-10}
4	75.17884184	1.2×10^{-10}
5	0	6.8×10^{-10}
6	0	1.5×10^{-9}
7	0	3.6×10^{-10}
8	0	3.6×10^{-10}
9	23.81690475	2.2×10^{-9}
16	-0.60555615	1.5×10^{-9}

6.3.4 Likelihood surfaces

Reiterating from Chapters 2 and 5, the cost function to be minimized for our i.i.d. Gaussian noise model is the sum-of-squares between the data and the output of our optical-design program, expressed in (2.72). As before, we will refer to the following plots of the cost function as likelihood surfaces. Using the data in Figure 6.6, we computed the surface along two parametric axes at a time, where the parameters were selected from $\{\alpha_n, n = 2, \dots, 9, 16, 25, 36, 37\}$. Each plot is centered about the true minimum, and the range for each parameter is given in Table 6.6.

Table 6.6: Range in likelihood surface plots for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16, 25, 36, 37\}$ in the exit pupil of the highly aberrated test lens. Units are in waves λ .

Index	True value [λ]	Range
2	0	$\pm 4\lambda$
3	0	$\pm 4\lambda$
4	75.17884184	$\pm 10\lambda$
5	0	$\pm 4\lambda$
6	0	$\pm 4\lambda$
7	0	$\pm 2\lambda$
8	0	$\pm 2\lambda$
9	23.81690475	$\pm 4\lambda$
16	-0.60555615	$\pm \lambda/2$
25	0.04113341	$\pm \lambda/2$
36	-0.00287742	$\pm \lambda/4$
37	-0.01310320	$\pm \lambda/4$

We first examined pairs of the rotationally-symmetric terms, $\{\alpha_n, n = 4, 9, 16, 25, 36, 37\}$. Several examples are provided in Figures 6.11 – 6.14. These likelihood plots are very reminiscent of those shown in Chapter 5, in which we estimated ocular parameters, and we concluded that the likelihood shape was primarily a function of the defocus level in the eye. Although the plots here incorporate the various spherical aberration terms, there is intense entanglement between these terms and defocus, as evidenced in the FIM in Section 6.3.2. A key difference, though, is that these plots contain fine wrinkles, specifically along the α_4 and α_9 axes (Figs. 6.11 – 6.13), whereas the plots in Chapter 5 are smoothly varying. These wrinkles are likely to cause difficulty in the later stages of optimization using simulated annealing, when finer features in the cost function begin to emerge.

Fig. 6.11: Likelihood surface along the α_4 (defocus) and α_9 (primary spherical aberration) axes for the highly aberrated test lens.

Fig. 6.12: Likelihood surface along the α_4 (defocus) and α_{25} (tertiary spherical aberration) axes for the highly aberrated test lens.

Fig. 6.13: Likelihood surface along the α_9 (primary spherical aberration) and α_{16} (secondary spherical aberration) axes for the highly aberrated test lens.

Fig. 6.14: Likelihood surface along the α_{16} (secondary spherical aberration) and α_{25} (tertiary spherical aberration) axes for the highly aberrated test lens.

Likelihood plots comparing a rotationally-symmetric term with an off-axis term, such as tilt, α_2 and α_3 , astigmatism, α_5 and α_6 , or coma, α_7 and α_8 , tend to resemble one of the plots in Figures 6.15 and 6.16. These are clearly marked by a number of local basins and a predictable axis of symmetry. Conversely, combining the x- and y-axis versions of an off-axis term (e.g., x- and y-axis tilt) results in a plot with full rotational symmetry, as in Figure 6.17. Any other combination between these off-axis terms results in a plot similar to those in Figures 6.18 and 6.19, once again characterized by a large number of local extrema.

Fig. 6.15: Likelihood surface along the α_4 (defocus) and α_5 (primary astigmatism at 0°) axes for the highly aberrated test lens.

Fig. 6.16: Likelihood surface along the α_{25} (tertiary spherical aberration) and α_7 (primary coma, x-axis) axes for the highly aberrated test lens.

Fig. 6.17: Likelihood surface along the α_2 (tilt, y-axis) and α_3 (tilt, x-axis) axes for the highly aberrated test lens.

Fig. 6.18: Likelihood surface along the α_5 (primary astigmatism at 0°) and α_7 (primary coma, x-axis) axes for the highly aberrated test lens.

Fig. 6.19: Likelihood surface along the α_3 (tilt, y-axis) and α_8 (primary coma, y-axis) axes for the highly aberrated test lens.

6.3.5 Maximum-likelihood estimates

After generating the data (Figure 6.6), we disregarded our knowledge of Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$, then estimated them using ML estimation according to our Gaussian noise model.

We tried a variety of tuning-parameter combinations for the simulated annealing algorithm described in Section 3.3, observing the cost function and the trajectory taken during the search process. Ultimately, we decided on the following values during initialization:

$$\tau_0 = 10^6,$$

$$\delta = 1.0, N_\delta = 5,$$

$$N_S = 10, \quad c = 2.0,$$

$$N_T = 20, \quad r_T = 0.90,$$

$$\mathbf{v}_0 = 0.5 (\boldsymbol{\theta}_{\text{upper}} - \boldsymbol{\theta}_{\text{lower}}),$$

where $\boldsymbol{\theta}_{\text{upper}}$ and $\boldsymbol{\theta}_{\text{lower}}$ are respectively the upper and lower limits in the parameter space, based on the ranges in Table 6.6. To give local minima throughout the entire search space an equal chance to be sampled, we imposed a high initial temperature τ_0 that would enable the system to accept virtually all proposed configurations. We produced a random starting point in the search, as listed in Table 6.7.

We ran 12 optimization trials, with each trial representing a distinct noise realization for the same wavefront parameters, and implemented the same estimation procedure on each. Figure 6.20 illustrates the optimal cost function versus iteration number for the trials, starting at the end of the first temperature phase. The average final cost and number of temperature phases were 147.7 and 87.3, respectively. This number of phases is equivalent to a final temperature of 116.1.

Fig. 6.20: 12 simulated annealing trials for the estimation of wavefront parameters in the exit pupil of the highly aberrated test lens (log-log scale).

As we discussed in Chapter 2, bias and variance are used to specify estimator performance, where bias is defined as the deviation of the average parameter estimate from the true value, and variance is the mean-square fluctuation of the estimate about its mean. In essence, bias is related to accuracy and variance to precision. Inherent bias in the estimator and systematic errors due to miscalibration or inaccurate modeling of the system both factor into the overall bias. Variance provides a measure of random errors that fluctuate from one measurement to another for a given wavefront.

Final average estimates for Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ are given in Table 6.7, along with the respective standard deviation. Both the bias and variance are very small for every parameter, while all parameters were estimated to within one standard deviation. Based on the low biases, it is not surprising that the true and estimated irradiance patterns are virtually indistinguishable (Fig. 6.21).

Table 6.7: ML estimates of wavefront parameters for the highly aberrated test lens at $\lambda = 0.6328 \mu\text{m}$, including their standard deviations and the starting point in the search. Units are in waves λ .

Index	Aberration Type	True Value [λ]	Start Point [λ]	Estimate [λ]
2	Tilt (x-axis)	0	2.33765863	0.11 ± 0.34
3	Tilt (y-axis)	0	3.67593941	-0.09 ± 0.30
4	Defocus	75.17884184	81.29481398	75.5 ± 1.0
5	Astigmatism, Primary (0° or 90°)	0	-3.71430657	-0.09 ± 0.28
6	Astigmatism, Primary ($\pm 45^\circ$)	0	2.79303444	0.17 ± 0.23
7	Coma, Primary (x-axis)	0	1.73597299	0.08 ± 0.33
8	Coma, Primary (y-axis)	0	-1.81531443	-0.10 ± 0.36
9	Spherical Aberration, Primary	23.81690475	20.60294575	23.58 ± 0.59
16	Spherical Aberration, Secondary	-0.60555615	-0.29345783	-0.604 ± 0.047

Throughout this dissertation, we have emphasized the need for an accurate probability model when performing ML estimation, particularly when working with physical data. In this simulation study, we assumed the validity of the Fresnel approximation despite the low f-number of the test lens. This did not pose any problems, however, since any modeling errors were mapped in both the forward and inverse problems. Thus, the negligible bias probably resulted from the lack of systematic errors in the two-way mapping.

Regarding the variance, the large deviations from the CRB must have risen from the host of local extrema in the likelihood surface. If the magnitude of the fluctuations are unacceptable to the specific application, then one must either devote more computation time to navigating the search space or find a system configuration with less extrema.

Fig. 6.21: Comparison between the true and estimated irradiance patterns for the highly aberrated test lens.

An iteration is defined as one cycle through every parameter, so there were $N_I = N_S \times N_T = 200$ iterations, or 1.8×10^3 forward propagations, per temperature phase. For the pupil sampling of $P = 1024$ and FFT grid size of $F = 2048$, the computation time was 790 ms per forward propagation, including both output planes. So, the average of 87.3 temperature phases per trial corresponds to 34.5 hours of computation time. We

carried out this study using an NVIDIA Tesla C1060 GPU, whose peak double-precision (DP) performance is 78 GFLOPS. Had we used the Tesla C2075 model, offering 515 GFLOPS of DP power, the computation time per trial would have been roughly 5.2 hours. Even further improvement can be achieved with a cluster of GPUs. For instance, the VSC455 V8 GPU workstation by Velocity Micro combines 8 Tesla C2075s for over 4 TFLOPS of DP power, which would turn the 5.2 hours into just 40 minutes.

6.4 Experimental results

In the experimental proof-of-principle study, we selected a relatively benign lens with less aberrations. In contrast to the numerical study, we dealt with nuisance parameters in the system and evaluated the accuracy of the Fresnel approximation. As usual, we determined the optimal pupil sampling and investigated the FIM, CRB, and likelihood surface. Despite the lower aberrations, the cost function still contained numerous local minima, so we performed ML estimation by simulated annealing.

6.4.1 System configuration

We obtained multifocal irradiance data in the focal region of a spherical test lens, described in detail in Section 6.4.2, again at finite conjugates and with a HeNe source ($\lambda = 0.6328 \mu\text{m}$) (Fig. 6.22). The distance from the on-axis point source, created with a 40X microscope objective and 10- μm pinhole, to the test lens was 457 mm. To generate an increase in information by illuminating substantially more detector elements, we magnified the intermediate image with an imaging lens placed just before the CCD. This

had the added benefit of eliminating the CCD saturation without the use of neutral density filters. The imaging lens was a Nikon 40X microscope objective ($NA = 0.95$), configured at the design conjugates. When used at the proper conjugates, the manufacturer claims that the objective lens is corrected for spherical aberration. This particular objective was infinity-corrected with an optical tube length of 160 mm. To maintain these conjugates, we placed the imaging lens and CCD on a translation stage, allowing us to scan through the focal region of the test lens. We used a CCD with a fine pixel size of $4.4\ \mu\text{m}$ for high information yield.

Fig. 6.22: Data-acquisition system for collecting multiple irradiance patterns near the focus of a spherical test lens, including a movable imaging lens.

6.4.2 Test lens description

The test lens for this study was a double-convex spherical lens by Edmund Optics (part no. NT45-891), with a diameter of $D_{\text{lens}} = 25$ mm and radius of curvature of $R_1 = -R_2 = 76.66$ mm. According to ZEMAX, the exit pupil diameter and working f-number was $D_{xp} = 25.39$ mm and $f/\#_w = 3.471$, respectively (Table 6.8). It was imperative that the image-space NA of the test lens ($\text{NA} = 0.14$) was less than that of the imaging lens to avoid information loss. The position of the paraxial focal plane was $z = z_f = 90.83$ mm, where the $z = 0$ plane contained the exit pupil, and the distance between marginal and paraxial focus was 4.0 mm (Fig. 6.23).

Table 6.8: System data provided by ZEMAX™ for the spherical test lens at $\lambda = 0.6328$ μm .

Effective Focal Length [mm]	74.99634
Back Focal Length [mm]	73.83225
Image Space f/#	2.999853
Paraxial Working f/#	3.576688
Working f/#	3.471054
Image Space NA	0.1384479
Object Space NA	0.02729433
Entrance Pupil Diameter [mm]	25.0
Exit Pupil Diameter [mm]	25.39417
Paraxial Focal Plane Position [mm]	90.82701

Fig. 6.23: Focal region of the spherical test lens. Paraxial focal plane is at $z = z_f = 90.83$ mm.

The ZEMAX calculations of the Fringe Zernike coefficients ($N = 37$) in the exit pupil of the lens are provided in Table 6.9. As before, the coefficients for non-rotationally-symmetric terms are zero. With a smaller peak-to-valley wavefront error of 30.6λ , we anticipated less stringent requirements on the pupil sampling in our propagation algorithm. A map of the wavefront error according to the design parameters is shown in Figure 6.24.

Table 6.9: Fringe Zernike coefficients $\{\alpha_n, n = 1, \dots, 37\}$, peak-to-valley, RMS, and variance, provided by ZEMAX for the spherical test lens at $\lambda = 0.6328 \mu\text{m}$. Unlisted coefficients are zero. Units are in waves λ .

Index	Aberration Type [λ]	Design Value [λ]
1	Piston	10.15431626
4	Defocus	15.26099035
9	Spherical Aberration, Primary	5.12576917
16	Spherical Aberration, Secondary	0.01863364
25	Spherical Aberration, Tertiary	-0.00048193
36	Spherical Aberration, Quaternary	-0.00002109
37	Spherical Aberration, 12th-order term	-0.00000061
Peak-to-valley [λ]	N/A	30.55920575
RMS [λ]	N/A	8.99563403
Variance [λ^2]	N/A	80.92143161

Fig. 6.24: Theoretical wavefront error in the exit pupil of the spherical lens as a function of normalized radius. Units are in waves.

6.4.3 Experimental data

Figure 6.25 displays the physical data that we used as input to our optimization algorithm, consisting of two image planes just before paraxial focus, where the scale bar corresponds to the intermediate image plane between the test lens and objective lens. In Section 6.4.8, we discuss the estimation of nuisance parameters in the system, namely, the image plane locations and the true magnification of the objective lens. Since the computer simulations in Sections 6.4.4 – 6.4.7 were rerun after obtaining these estimates, we will simply quote the results here:

$$z_1 = z_f + \Delta z_1 = z_f - 0.6745 \text{ mm}, \quad M_1 = 39.96, \quad (6.13a)$$

$$z_2 = z_f + \Delta z_2 = z_f - 0.7945 \text{ mm}, \quad M_2 = 40.09, \quad (6.13b)$$

where Δz is the distance from paraxial focus and M is the magnification.

Fig. 6.25: Experimental data for the spherical test lens for image planes: (a) $z = z_1$ and (b) $z = z_2$. Scale bar corresponds to the intermediate image plane just before the imaging lens.

6.4.4 Pupil sampling

As previously mentioned, the required amount of pupil sampling is generally proportional to the degree of aberrations in the exit pupil, while other factors to consider are the curvature of the unaberrated reference wave and the distance of the output plane from optimal focus. Simulated detector data for various sampling levels, $P = 128$, $P = 256$, and $P = 1024$, are provided for both output planes in Figures 6.26 and 6.27. The ratio P/F was fixed at $1/4$ for all plots, resulting in a detector pitch of roughly $0.56 \mu\text{m}$. Sampling artifacts are quite evident in the data for $P = 128$, $F = 512$ (Figs. 6.26c & 6.27c), as the

irradiance pattern has a cogwheel appearance. There are no apparent signs of undersampling for $P = 256$, $F = 1024$ (Figs. 6.26b & 6.27b), and the irradiance is smooth and seemingly physical. Since the irradiance is unaffected by further increase in sampling to $P = 512$, $F = 2048$, we decided on $P = 256$ for this particular study.

Fig. 6.26: Detector data at $z = z_1$ for spherical lens using pupil sampling of: (a) $P = 512$, (b) $P = 256$, and (c) $P = 128$.

Fig. 6.27: Detector data at $z = z_2$ for spherical lens using pupil sampling of: (a) $P = 512$, (b) $P = 256$, and (c) $P = 128$.

6.4.5 Huygens' method vs. Fresnel propagation

Since this estimation task dealt with physical data, accurate forward modeling of the system was imperative. Here we compare the Fresnel approximation, or FFT method, with a brute-force evaluation of the Huygens wavelet formula in (6.1), which we will treat as the gold standard. We computed the detector data in both output planes for the spherical lens using the two methods (Figs. 6.28 & 6.29). Each irradiance pattern was normalized to unity area, consistent with conservation of power. Displayed in Figures 6.28c and 6.29c are the differential images, indicating modest peak discrepancies of 1.5% for $z = z_1$ and 1.7% for $z = z_2$.

Despite the slight variation in irradiance between the methods, there is a tremendous difference in computation time. With an NVIDIA Tesla C1060, the computation time per output plane with the FFT method for a pupil sampling of $P = 256$ and FFT grid size of $F = 1024$ was $T_{\text{FFT}} = 154$ ms, independent of the number of useful elements we choose to extract from the 1024×1024 FFT output. Table 7.10 provides the computation time using the Huygens integral (T_{int}) for the same P , but various detector grid sizes M_{1D} . The ratio $T_{\text{int}}/T_{\text{FFT}}$ is 377 for $M_{\text{1D}} = 128$ to an astounding 2.42×10^4 for $M_{\text{1D}} = 1024$. Although $M_{\text{1D}} = 128$ is large enough to include the irradiance pattern for the wavefront parameters in Table 6.9, it may not be large enough as the parameter space is explored. Regardless, all of the computation times listed below are prohibitively long for any extensive optimization routine.

Table 6.10: Computation time using Huygens' method for a pupil sampling of 256×256 and various detector grid sizes.

P	M_{ID}	T_{int} [sec]	T_{int} [min]	T_{int}/T_{FFT}
256	1024	3728	62.1	2.42×10^4
256	512	932	15.5	6.05×10^3
256	256	233	3.88	1.51×10^3
256	128	58.1	0.968	3.77×10^2

Fig. 6.28: Irradiance data at $z = z_1$ for the spherical lens: (a) Fresnel approximation, (b) Huygens integral, (c) difference.

Fig. 6.29: Irradiance data at $z = z_2$ for the spherical lens: (a) Fresnel approximation, (b) Huygens integral, (c) difference.

6.4.6 Fisher information and Cramér-Rao lower bounds

We computed the FIM for the Fringe Zernike coefficients to be estimated, $\{\alpha_n, n = 2, \dots, 9, 16\}$, evaluating it at the point in parameter space corresponding to the design parameters in Table 6.9. As in Section 6.3.2, we based the FIM on an i.i.d. Gaussian noise model and a peak SNR of 10^4 . The FIM (Fig. 6.30) structure is comparable to that for the highly aberrated lens, including strong coupling among the rotationally-symmetric terms, α_4 , α_9 , and α_{16} . From the inverse matrix (Fig. 6.31), we determined the CRB for the set of parameters (Table 6.11), which again takes on diminutive values. Thus, the variance from detector noise is unlikely to preclude high-precision estimates.

Fig. 6.30: FIM for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the spherical test lens (log scale).

Fig. 6.31: Inverse of the FIM for Fringe Zernike coefficients α_2 – α_9 and α_{16} in the exit pupil of the spherical test lens (log scale).

Table 6.11: Square-root of the CRB for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the spherical test lens at $\lambda = 0.6328 \mu\text{m}$. Units are in waves λ .

Index	Design Value [λ]	$(\text{CRB})^{1/2}$ [λ]
2	0	8.3×10^{-9}
3	0	8.3×10^{-9}
4	15.26099035	4.2×10^{-9}
5	0	3.9×10^{-9}
6	0	3.9×10^{-9}
7	0	5.7×10^{-9}
8	0	5.7×10^{-9}
9	5.12576917	3.5×10^{-9}
16	0.01863364	2.4×10^{-9}

6.4.7 Likelihood surfaces

We greatly expanded the search space in this study, compared to the numerical study, which revealed many new and interesting characteristics of the likelihood surface. The adjusted ranges are listed in Table 6.12 and several examples are shown in Figures 6.32 – 6.36.

Table 6.12: Range in likelihood surface plots for Fringe Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ in the exit pupil of the spherical test lens. Units are in waves λ .

Index	Design value [λ]	Range
2	0	$\pm 10\lambda$
3	0	$\pm 10\lambda$
4	15.26099035	$\pm 20\lambda$
5	0	$\pm 5\lambda$
6	0	$\pm 5\lambda$
7	0	$\pm 5\lambda$
8	0	$\pm 5\lambda$
9	5.12576917	$\pm 10\lambda$
16	0.01863364	$\pm \lambda$

Fig. 6.32: Likelihood surface along α_4 (defocus) and α_{16} (secondary spherical aberration) axes for the spherical test lens.

Fig. 6.33: Likelihood surface along α_7 (primary coma, x-axis) and α_9 (primary spherical aberration) axes for the spherical test lens.

Fig. 6.34: Likelihood surface along α_2 (tilt, x-axis) and α_4 (defocus) axes for the spherical test lens.

Fig. 6.35: Likelihood surface along α_5 (primary astigmatism at 0°) and α_{16} (secondary spherical aberration) axes for the spherical test lens.

Fig. 6.36: Likelihood surface along α_2 (tilt, x-axis) and α_7 (primary coma, x-axis) axes for the spherical test lens.

6.4.8 Nuisance parameters

In Section 2.7, we discussed various methods for dealing with nuisance parameters, which are defined as parameters that affect the data and are therefore fundamental to the probability model, but are not useful to the estimation task.

Suppose α denotes the wavefront parameters of interest, β denotes the wavefront parameters of no immediate interest, and χ denotes all other nuisance parameters, so that the entire vector parameter can be written as $\theta = (\alpha, \beta, \chi)^t$. Therefore, α is comprised

of Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$, β is comprised of all remaining coefficients, and χ consists of unknown system parameters to be discussed next. We dealt with β by replacing it (up to the ZEMAX limit of $N = 37$) with the respective design coefficients in Table 6.9, denoted as β_0 , then separately estimated χ prior to the primary estimation task. Finally, we set $\text{pr}(\mathbf{g} | \alpha, \beta, \chi) \approx \text{pr}(\mathbf{g} | \alpha, \beta_0, \hat{\chi})$ and proceeded to estimate α .

Two of the critical nuisance parameters in the system were the absolute image plane positions, $z_1 = z_f + \Delta z_1$ and $z_2 = z_f + \Delta z_2$. During acquisition of the detector data, it was difficult to accurately pinpoint the paraxial focal plane, $z = z_f$, which served as the reference plane in our studies. Another nuisance parameter was the true magnification of the microscope objective just before the detector, denoted as M_1 and M_2 for the first and second image, respectively. To achieve the design magnification of 40X, the detector must be placed 160 mm from the rear principal plane of the objective lens, however, we had no knowledge of the exact location of this plane.

For the i^{th} output plane, we estimated z_i and M_i from the respective detector data in Figure 7.25 by performing a straightforward 2D grid search. During this step, we fixed all wavefront parameters, α and β , to the design parameters, α_0 and β_0 , and computed the cost function for both image planes assuming a Gaussian noise model (Figs. 6.37 & 6.38). Prior to generating these plots, we first did a broader search for local minima in the region of interest, but did not find any. The final nuisance parameter estimates are given by

$$\hat{z}_1 = 90.153 \text{ mm}, \quad \hat{M}_1 = 39.96, \quad (5.14a)$$

$$\hat{z}_2 = 90.033 \text{ mm}, \quad \hat{M}_2 = 40.09, \quad (5.14b)$$

Since the detector element size is $4.4 \text{ } \mu\text{m}$, the effective pixel size after magnification is

$\Delta x_d = 0.1101 \text{ } \mu\text{m}$ and $\Delta x_d = 0.1098 \text{ } \mu\text{m}$.

Fig. 6.37: Determining the nuisance parameters in the system for image plane $z = z_1$ via a 2D grid search prior to the estimation of wavefront parameters.

Fig. 6.38: Determining the nuisance parameters in the system for image plane $z = z_2$ via a 2D grid search prior to the estimation of wavefront parameters.

6.4.9 Maximum-likelihood estimates

To be able to compare the physical data with the output of the optical design program during optimization, we first interpolated the data, so that its coordinates matched that of the FFT output. Then we normalized the irradiance pattern in both the data and FFT output, which is analogous to normalizing the power of the source beam.

We minimized the cost function for an i.i.d. Gaussian noise model, given by (2.72), using the broad parametric ranges in Table 6.12 and the following tuning parameters:

$$\tau_0 = 10^3,$$

$$\delta = 1.0, N_\delta = 5,$$

$$N_S = 10, \quad c = 2.0,$$

$$N_T = 25, \quad r_T = 0.90$$

$$\mathbf{v}_0 = 0.5 (\boldsymbol{\theta}_{\text{upper}} - \boldsymbol{\theta}_{\text{lower}}),$$

The starting point $\boldsymbol{\theta}_0$ in the search space was based on the design values for the Zernike coefficients, since this offered the best possible initial guess.

Figure 6.39 illustrates the optimal cost function versus number of iterations for 12 trials, each with a different noise realization in the data. The average final cost function was 0.150, while the average number of temperature phases was 105, corresponding to a temperature of 0.0174.

Fig. 6.39: 12 simulated annealing trials for the estimation of wavefront parameters in the exit pupil of the spherical test lens.

In each temperature phase, there were $N_I = N_S \times N_T = 250$ iterations, equaling 2.25×10^3 forward propagations for 9 parameters. With $P = 256$ and $F = 1024$, the computation time was 308 ms per forward propagation. There was an average of 105 temperature phases per trial, corresponding to 20.2 hours of computation time with a single NVIDIA Tesla C1060. On a Tesla C2075 model, the computation time per trial would be 3.1 hours, and clusters of GPUs could be used as well.

Final average estimates and standard deviations for Zernike coefficients $\{\alpha_n, n = 2, \dots, 9, 16\}$ are given in Table 6.13. Without knowledge of the true values or a gold standard, we have no basis for evaluating biases in the estimates. In theory, estimator bias should be insignificant as long as the global minimum of the cost function can be found, while systematic bias is likely to arise from errors in the measurement plane positions or the source position. Still, the estimated values are within one standard deviation from the design values. Once again, large departures from the CRB must have resulted from the numerous local basins in the cost function.

Table 6.13: ML estimates of wavefront parameters for the spherical test lens at $\lambda = 0.6328 \mu\text{m}$, including their standard deviations. Design values were used as a starting point in the search. Units are in waves λ .

Index	Aberration Type	Design Value [λ]	ML Estimate [λ]
2	Tilt (x-axis)	0	0.10 ± 0.30
3	Tilt (y-axis)	0	-0.13 ± 0.28
4	Defocus	15.26099035	15.57 ± 0.64
5	Astigmatism, Primary (0° or 90°)	0	-0.09 ± 0.19
6	Astigmatism, Primary ($\pm 45^\circ$)	0	0.13 ± 0.26
7	Coma, Primary (x-axis)	0	0.08 ± 0.18
8	Coma, Primary (y-axis)	0	-0.11 ± 0.23
9	Spherical Aberration, Primary	5.12576917	4.92 ± 0.32
16	Spherical Aberration, Secondary	0.01863364	0.0196 ± 0.0014

Fig. 6.40: Comparison between the true and estimated irradiance patterns for the spherical test lens.

6.5 Summary of Chapter 6

In both the numerical and experimental studies, the data-acquisition method involved multiple irradiance patterns collected near the focus of an optical system for the purpose of estimating the pupil phase distribution. We considered various approaches for a suitable propagation algorithm to accurately model the wave propagation and developed

the mathematical framework for an aberrated wave emerging from the exit pupil, where we parameterized the wavefront using expansion functions, particularly, Fringe Zernike polynomials. To substantially reduce the computation time, we implemented parallel processing with a state-of-the-art GPU.

We obtained proof-of-principle results in both simulation and experiment. In each study, we evaluated the sampling requirements and verified that significant pupil sampling is needed for large wavefront errors. Fisher information matrices featured prominent coupling among specific groups of parameters, such as the group containing only rotationally-symmetric Zernike terms. The associated Cramér-Rao bounds were incredibly small, thereby permitting high-precision estimates, although this generally requires an accurate forward model and a search algorithm that can reliably locate the global minimum of the cost function.

After discovering numerous local extrema in the likelihood surfaces, we chose simulated annealing for the estimation of selected Zernike coefficients. In the numerical study with the highly aberrated lens, the estimate biases were negligible, probably because of the lack of systematic errors in the estimation procedure. Although the variances were fairly small, they were far from the CRB, certainly due to entrapment in local basins in the cost function.

In the experimental study, we used a benign test lens with significantly less aberrations and a larger working f-number of $f/\#_w = 3.47$. Since the imaging (i.e., relay) lens was much faster with $f/\# \approx 0.53$ ($\text{NA} = 0.95$), it should not have resulted in information loss from the suppression of high spatial frequencies. However, this may

become an issue if the f-number of the test lens is low enough. To avoid having to include the imaging lens in the forward model here, one solution is to use the translation stage to place a ground glass in any plane where an irradiance measurement is desired, then relay the incoherent irradiance, instead of the field, to the detector (Fig. 8.1). Note that it is necessary to have a rotating ground glass to decorrelate the resulting speckle noise. An alternative to a rotating diffuser is a liquid-crystal diffuser operated in a dynamic scattering mode.

Fig. 6.41: Data-acquisition system for collecting multiple irradiance patterns near the focus of an optical element, including a movable diffuser and imaging lens.

Experimentally, we were challenged with the absolute requirement of an accurate probability model, as well as the presence of nuisance parameters in the system. We verified the accuracy of the Fresnel approximation in our forward model compared to the Huygens wavelet formula. Nuisance parameters included image plane positions and the

true magnification of the imaging lens. We dealt with them by means of a 2D grid search and located a single extremum for each image plane. The estimate variances were comparable to those in the simulation study, and each estimate did not exceed one standard deviation of the design values.

CHAPTER 7

INVERSE OPTICAL DESIGN FOR OPTICAL TESTING

Inverse optical design provides a unique approach to testing graded-index and aspheric lenses to ensure that they have been fabricated to specification. In our method of optical testing via parametric modeling, the parameters to be estimated may include coefficients in the refractive index distribution of GRIN lenses or coefficients describing the high-order surfaces of precision aspheres.

We present results from numerical studies for both types of lenses. Section 7.1 is devoted to aspheric lenses and Section 7.2 to GRIN lenses. In Section 7.1, we discuss our rapid ray-tracing algorithm that was developed in CUDA on the GPU platform. In Section 7.2, we outline the theoretical framework for tracing rays through GRIN-rod lenses, which involves analytic solutions to the eikonal equation. In both cases, we provide Fisher information matrices, Cramér-Rao bounds, and likelihood surfaces. As usual, we provide maximum-likelihood estimates obtained with simulated annealing.

7.1 Inverse optical design of aspheric lenses

The primary objective in this particular application is to process high-order aspheric surfaces by means of ray-tracing. One problem, however, is that the ray-surface intercepts cannot be determined analytically for surfaces beyond fourth-order, so that

iterative techniques must be implemented (Blinn, 2006, 2006a). We begin this section with a detailed description of an iterative algorithm used in our optical design program.

7.1.1 Optical-design program

We developed a rapid ray-trace algorithm in CUDA that performs non-paraxial ray-tracing through high-order aspheric surfaces. The ray-surface intersection is determined iteratively through a marching-points algorithm for root-isolation and repeated bisections for root-refinement.

Description of refracting high-order aspheric surfaces

We assume the following expression to describe a high-order even asphere:

$$z = \frac{r^2/R}{1 + \sqrt{1 - (1 + \kappa)r^2/R^2}} + \alpha_4 r^4 + \alpha_6 r^6 + \alpha_8 r^8 + \alpha_{10} r^{10} + \dots, \quad (7.1)$$

where z is the sag of the surface, $r^2 = x^2 + y^2$.

$$\begin{aligned} S(r, z) = & (1 + \kappa)[z - (\alpha_4 r^4 + \alpha_6 r^6 + \alpha_8 r^8 + \alpha_{10} r^{10})]^2 \\ & - 2R[z - (\alpha_4 r^4 + \alpha_6 r^6 + \alpha_8 r^8 + \alpha_{10} r^{10})] + r^2, \end{aligned} \quad (7.2)$$

Determining the ray-surface intersection

Singh and Narayanan (2007) developed a simple ray-tracing algorithm for ray-tracing general implicit surfaces that is well suited to the SIMD architecture of the GPU. We applied this method to aspheric surfaces described by (7.2) and found that it delivers high performance through robust root-finding.

While analytical solutions are possible for polynomials of fourth-order or lower, roots of higher-order polynomials must be determined iteratively. There are many root-finding techniques that are popular for ray-tracing, such as the Newton-Raphson, Newton-Bisection, and Laguerre methods (Kajiya, 82; Press, Teukolsky, Vetterling, & Flannery, 1992; Wyvill & Trotman, 1990), or extensions of these methods that integrate interval arithmetic (Duff, 1992; Mitchell, 1990). Other techniques for finding roots of polynomials incorporate auxiliary polynomials (Sederberg & Change, 1993) or Sturm sequences (Nister, 2004). However, many of these methods are difficult to implement with the SIMD model and can be quite complicated for higher-order surfaces; quick, simple computations perform best on the GPU (Singh & Narayanan, 2007).

In the method that we adopted, a simple *marching-points* scheme is used to isolate the smallest positive root for a given ray, whereby a ray is sampled at consecutive points and the first bracket containing a root is returned. A straightforward test for root-containment is applied. Root-refinement is achieved with repeated bisections, which recursively splits the bracket into sub-intervals and keeps the first one containing a root.

Prior to root-finding, the surface $S(x, y, z) = 0$ is cast to the form

$$F_f(k) = 0 \quad (7.3)$$

by substituting the ray-equation (5.17) for an arbitrary surface-fragment f , where k is the ray-parameter. Recall that a ray is defined as

$$\mathbf{x}(k) = \mathbf{x}_0 + k\mathbf{x}_d. \quad (7.4)$$

An alternative to computing the univariate polynomial (7.3) for a given k is to evaluate the points according to (7.4), then substitute the values into the surface equation, which now becomes

$$S(x, y, z) = S(\mathbf{x}(k)) = 0. \quad (7.5)$$

The computational implications can vary greatly between the two approaches, and for higher-order polynomials, the expression $F_f(k)$ usually has a large number of terms. Choice of which expression to evaluate must be determined for a particular surface.

In the root-isolation stage, bounds on the total search range $[k_s, k_e]$ for the rays are initially specified and then the range is divided into N equal intervals. The interval endpoints for a given ray are evaluated one-by-one until a root is found, that is, until $F_f(k)$ crosses zero between two neighboring points. If this occurs at the i^{th} iteration, then the algorithm returns $[k_i, k_{i+1}]$ as containing a root.

A sign test is used to check for root-containment, where a root exists in the i^{th} interval if the function changes sign between the endpoints:

$$S(p(k_i)) * S(p(k_{i+1})) < 0, \quad \text{root exists,} \quad (7.6a)$$

$$S(p(k_i)) * S(p(k_{i+1})) > 0, \quad \text{root does not exist,} \quad (7.6b)$$

Although this test does not produce false roots, it may miss roots if an interval contains an even number of roots (Singh & Narayanan, 2007).

In the bisection method for root-refinement, the bracket $[k_i, k_{i+1}]$ is divided into two sub-intervals, $[k_i, k_m]$ and $[k_m, k_{i+1}]$, using the midpoint k_m . The first sub-interval containing a root is identified, then the process is repeated until the maximum number of bisections is reached or a tolerance condition is met. This method is robust and never fails, as long as the bracketing is correct.

Since the function $F_f(k)$ is crudely evaluated only at the sample points, the computation complexity is very low. Provided that the total number of point evaluations is roughly the same for all rays, the algorithm fits the parallel architecture of the GPU very well. When this is not the case, Singh and Narayanan (2007) propose an *adaptive marching-points* scheme that samples each ray non-uniformly based on the algebraic distance to the surface, as well as the angle relative to the surface normal. This optimizes the performance of the GPU by balancing the computation load across threads, but adds to the computational burden and involves a derivative computation. For the precision asphere described in Table 7.1, the expense outweighs the benefit, so we used the basic marching-points approach.

Direction of the transmitted ray

As discussed in Section 5.2, the direction of the refracted ray $\hat{\mathbf{x}}'_d$ is determined by (5.21):

$$\hat{\mathbf{x}}'_d = \hat{\mathbf{x}}_d \frac{n}{n'} [\hat{\mathbf{x}}_d - (\hat{\mathbf{x}}_d \cdot \hat{\mathbf{n}}) \hat{\mathbf{n}}] + \left[\sqrt{1 + \left(\frac{n}{n'} \right)^2 [(\hat{\mathbf{x}}_d \cdot \hat{\mathbf{n}})^2 - 1]} \right] \hat{\mathbf{n}}. \quad (7.7)$$

The unit vector $\hat{\mathbf{n}}$, generally given by (5.16), for the 10th-order precision asphere has components

$$\frac{\partial S(x, y, z)}{\partial x} = 2x - 4Bx[(1 + \kappa)(z - A) - R], \quad (7.8a)$$

$$\frac{\partial S(x, y, z)}{\partial y} = 2y - 4By[(1 + \kappa)(z - A) - R], \quad (7.8b)$$

$$\frac{\partial S(x, y, z)}{\partial z} = 2[(1 + \kappa)(z - A) - R], \quad (7.8c)$$

where

$$A(r) = \alpha_4 r^4 + \alpha_6 r^6 + \alpha_8 r^8 + \alpha_{10} r^{10}, \quad (7.9)$$

$$B(r) = 2\alpha_4 r^2 + 3\alpha_6 r^4 + 4\alpha_8 r^6 + 5\alpha_{10} r^8. \quad (7.10)$$

7.1.2 Test lens description and system configuration

We simulated detector data for a precision plano-convex aspheric lens with parameters based on Edmund Optics Precision Asphere NT47-731 (Table 7.1). The parameters we chose to estimate in this study, R_C , κ , α_4 , and α_6 , contain small deviations from the design values of this lens. Table 7.2 provides the system data computed by ZEMAX, where the positions of the entrance and exit pupils are measured from the first surface of the lens.

Table 7.1: True values of parameters underlying the irradiance data for the precision asphere, and design values of Edmund Optics Precision Asphere NT47-731.

Parameter	Units	Design value	True value
Radius of curvature, R_C	mm	18.41	18.55
Conic constant, κ	N/A	-1.607913	-1.737913
4th-order aspheric coefficient, α_4	mm^{-3}	2.0634554×10^{-5}	2.413455×10^{-5}
6th-order aspheric coefficient, α_6	mm^{-5}	$-7.6489765 \times 10^{-9}$	-7.438977×10^{-9}
8th-order aspheric coefficient, α_8	mm^{-7}	1.117573×10^{-11}	1.117573×10^{-11}
10th-order aspheric coefficient, α_{10}	mm^{-9}	$-1.010058 \times 10^{-14}$	$-1.010058 \times 10^{-14}$
Refractive index at $\lambda = 632.8 \text{ nm}$, n	N/A	1.58708982	1.58708982
Center thickness, t	mm	6.50	6.50

Table 7.2: System data provided by ZEMAX™ for the precision asphere at $\lambda = 0.6328$ μm .

Effective Focal Length [mm]	31.59653
Back Focal Length [mm]	31.59653
Image Space f/#	1.579826
Paraxial Working f/#	4.432388
Working f/#	19.32893
Image Space NA	0.112095
Object Space NA	0.1995864
Entrance Pupil Diameter [mm]	20
Entrance Pupil Position [mm]	4.095546
Exit Pupil Diameter [mm]	20
Exit Pupil Position [mm]	6.5

Our model incorporated two point sources, one on-axis and the other displaced by $y = 10$ mm from the optical axis (z -axis), while both were placed 45 mm before the test lens along the optical axis (Fig. 7.1). A 20-mm-diameter iris was positioned immediately before the lens. Since larger spot sizes produce greater information yield, we intentionally oriented the lens opposite to the manufacturer-intended orientation, so that the flat surface faced the source. We traced rays through the system for two on-axis image plane positions at $z = \{95, 100\}$ mm from the lens, plus one off-axis position at $z = 90$ mm. Our ray-trace results are practically identical to those of ZEMAX (Fig. 7.2).

Fig. 7.1: Ray-trace data from our CUDA algorithm for the precision asphere.

Fig. 7.2: Ray-trace data computed by ZEMAX™ for the precision asphere.

Using our Gaussian noise model, we emulated electronic noise in the detector data with a modest peak signal-to-noise ratio (SNR) of 10^3 and a pixel size $12\text{ }\mu\text{m}$ (Fig. 7.3). The irradiance patterns compare very well with the output of ZEMAX (Fig. 7.4).

Fig. 7.3: Irradiance data computed at: (a) $z = 95$ mm after lens for on-axis source, (b) $z = 100$ mm for same on-axis source, and (c) $z = 90$ mm for off-axis source.

Fig. 7.4: Irradiance data computed with ZEMAXTM at: (a) $z = 95$ mm after lens for on-axis source, (b) $z = 100$ mm for same on-axis source, and (c) $z = 90$ mm for off-axis source.

7.1.3 Fisher information and Cramér-Rao bounds

We originally computed the FIM for the parameters appearing in (7.2), including R_C , κ , α_4 , α_6 , α_8 , and α_{10} . However, the presence of α_8 and α_{10} results in a singular, non-invertible matrix, since the data are not sensitive to changes in these higher-order coefficients. Although there are methods for dealing with singular information matrices that often involve their pseudoinverses (Hero, Fessler, & Usman, 1996; Rao, 1973), there are no unbiased estimators for the confounding parameters (i.e., α_8 and α_{10}) with finite variance (Stoica, 2001). The simplest solution is to simply exclude them from both the FIM and the general estimation procedure.

We computed the FIM according to (2.76), as well as its inverse, based our standard Gaussian noise model and a peak SNR of 10^3 (Fig. 7.5). While the resulting CRBs (Table 7.3) are impressively small, thus indicating substantial information in the system for the selected parameters, it is meaningful only if we can locate the global maximum in the likelihood function.

Fig. 7.5: (a) FIM and (b) inverse of the FIM for prescription parameters describing the precision asphere. (logarithmic scale)

Table 7.3: Square-root of the CRB for prescription parameters describing the precision asphere.

Index	Parameter	Units	True value	$(\text{CRB})^{1/2}$
1	Radius of curvature, R_C	mm	18.55	1.5×10^{-7}
2	Conic constant, κ	N/A	-1.737913	1.7×10^{-7}
3	4 th -order aspheric coefficient, α_4	mm^{-3}	2.413455×10^{-5}	2.6×10^{-12}
4	6 th -order aspheric coefficient, α_6	mm^{-5}	-7.438977×10^{-9}	2.8×10^{-15}

7.1.4 Likelihood surfaces

Likelihood surfaces for all six pairs of parameters are provided in Figures 7.6 – 7.11, based on the parametric ranges in Table 7.4. Since the ranges are centered on the parameters underlying the data, the global minimum occurs at the center of each plot.

Most of the irregularities and local minima occur along the R_C axis, as seen in Figures 7.6 – 7.8. (Note that Figure 7.7 is shown on a logarithmic scale to bring out features otherwise suppressed by a very strong and narrow peak.) In contrast, there is barely any variation along the α_6 axis (Figs. 7.8, 7.10, & 7.11), which could hinder the estimation process. Finally, there is an interesting dynamic in the pairs (κ, α_4) (Fig. 7.9) and (α_4, α_6) (Fig. 7.11) that leads to the same likelihood plots that we saw in the estimation of ocular parameters, as well as wavefront coefficients, in Chapters 5 and 6.

Table 7.4: Range in likelihood surfaces for parameters describing the precision asphere, relative to the true values.

Parameter	Units	True value	Range
Radius of curvature, R_C	mm	18.55	± 3.0
Conic constant, κ	N/A	-1.737913	± 5.0
4 th -order aspheric coefficient, α_4	mm^{-3}	2.413455×10^{-5}	$\pm 5.0 \times 10^{-5}$
6 th -order aspheric coefficient, α_6	mm^{-5}	-7.438977×10^{-9}	$\pm 5.0 \times 10^{-9}$

Fig. 7.6: Likelihood surface along R_C and κ axes. Global minimum is located at center of plot.

Fig. 7.7: Likelihood surface along R_C and α_4 axes. Global minimum is located at center of plot. (logarithmic scale)

Fig. 7.8: Likelihood surface along R_C and α_6 axes. Global minimum is located at center of plot.

Fig. 7.9: Likelihood surface along κ and α_4 axes. Global minimum is located at center of plot.

Fig. 7.10: Likelihood surface along κ and α_6 axes. Global minimum is located at center of plot.

Fig. 7.11: Likelihood surface along α_4 and α_6 axes. Global minimum is located at center of plot.

7.1.5 Maximum-likelihood estimates

We chose to estimate a subset of the parameters, including the radius of curvature R_C , conic constant κ , and the 4th- and 6th-order aspheric coefficients, α_4 and α_6 . Pretending to know nothing of the true values of these parameters, we estimated them according to (2.72) with the following tuning parameters in our simulated annealing algorithm:

$$\begin{aligned}\tau_0 &= 10^3, \\ \delta &= 1.0, \quad N_\delta = 5, \\ N_S &= 10, \quad c = 2.0, \\ N_T &= 20, \quad r_T = 0.90, \\ \mathbf{v}_0 &= 0.5 (\boldsymbol{\theta}_{\text{upper}} - \boldsymbol{\theta}_{\text{lower}}).\end{aligned}$$

Bounds on the parameters during the search process are based on the ranges in Table 7.4.

To assess the variance in our estimates, we computed 20 data sets for the given set of parameters in Table 7.1, where each data set represented a different noise realization (Fig. 7.12). We rescaled the cost function to equal unity at the true minimum. The average final cost and number of temperature phases were 1.0908 and 65.3, respectively, where the latter corresponds to a final temperature of 1.1790.

Fig. 7.12 : 20 simulated annealing trials for the estimation of prescription parameters describing the precision asphere.

Both the estimate bias and variance are very small for each parameter (Table 7.5), except for the broad variance in $\hat{\alpha}_6$. This is not unexpected, since the likelihood surface hardly varies along the α_6 axis. Conversely, the radius of curvature R_C demonstrated the best performance, with accuracy to $0.1 \mu\text{m}$, despite the complicated features in the likelihood surface along this axis.

Table 7.5: ML estimates of prescription parameters describing the precision asphere, including standard deviations. Design values were used as a starting point in the search.

Parameter	Units	True value	ML estimate
Radius of curvature, R_C	mm	18.55	18.5500 ± 0.0004
Conic constant, κ	N/A	-1.737913	-1.738 ± 0.008
4th-order aspheric coefficient, α_4	mm^{-3}	2.413455×10^{-5}	$(2.41 \pm 0.02) \times 10^{-5}$
6th-order aspheric coefficient, α_6	mm^{-5}	-7.438977×10^{-9}	$(-7.46 \pm 0.41) \times 10^{-9}$

When a bundle of rays is launched for a particular source location, the rays are propagated through the system in parallel on the GPU's architecture. Using an NVIDIA Tesla C1060 graphics card, it takes 0.65 sec to compute detector data for a 1024×1024 bundle of rays, equivalent to approximately $0.6 \mu\text{sec}$ per ray. Thus, the computation time for one forward propagation (or two source positions) is roughly 1.3 sec. There were $N_I = N_S \times N_T = 200$ iterations or 800 forward propagations per temperature phase, which takes 17.3 min. So, the total computation time for 65 phases is 18.7 hrs.

In applications where ray optics does not provide an accurate representation of the irradiance data, the ray-trace code can be modified to keep track of the optical path length (OPL) along a ray and then construct a wavefront perpendicular to all of the rays in a specified reference plane, by some algorithm. The wavefront can then be propagated to the final image plane by FFT propagation. Such a hybridized approach would relieve some of the computational burden in using solely diffraction propagation. One caveat in

this process is that the ray bundle is no longer uniformly distributed in the reference plane, so an interpolation scheme must be applied prior to the FFT. This is non-trivial, as standard interpolators convert a regular grid to an irregular one, but not the other way around.

7.2 Inverse optical design of GRIN-rod lenses

Ray-tracing through GRIN lenses requires solutions of the eikonal equation, discussed in Section 4.3.1. Finding such solutions is generally complicated, except for a few simple textbook problems, while exact solutions do not exist for GRIN lenses. We begin by presenting approximate analytic solutions to the equation, which we used in our optical design program.

7.2.1 Ray-tracing through a GRIN-rod lens

Toyokazu Sakamoto (1993, 1995) worked out analytic solutions of the eikonal equation for both meridional and skew rays using the perturbation method of Streifer and Paxton (1971). Since we are interested in systems with rotational symmetry, we will outline the theoretical framework for meridional rays, as presented by Sakamoto (1993).

A basic assumption in the analytic solutions is that the refractive index distribution $n(r)$ of a GRIN-rod lens can be represented as

$$n^2(r) = n_0^2[1 - (gr)^2 + h_4(gr)^4 + h_6(gr)^6 + h_8(gr)^8 + \dots], \quad (7.11)$$

where r is the distance from the optical axis z , n_0 is the refractive index along the axis, g is the focusing parameter, and h_4 , h_6 , and h_8 are the fourth-, sixth-, and eighth-order refractive index solutions, respectively. Equation (7.11) assumes perfect rotational symmetry and an index profile independent of z , which may not be realistic in optical testing, but is a start for this proof-of-principle study.

Ray equation

The ray equation (4.61) for a medium with refractive index $n(r)$ is given by

$$\frac{d}{ds} \left[n(r) \frac{d\mathbf{r}}{ds} \right] = \nabla n(r), \quad (7.12)$$

where $r^2 = (x^2 + y^2)$, ds is the line element along a ray path s , and $\mathbf{r} = (x, y, z)$ is the 3D position vector for an arbitrary point on the ray. Since $n(r)$ is independent of z , the z component of (7.12) can be written as

$$n(r) \frac{dz}{ds} = n_i \cos \gamma_i, \quad (7.13)$$

where n_i and $\cos \gamma_i$ are the refractive index and z component of the directional cosine, respectively, at the initial ray position $(x_i, y_i, 0)$. Combining (7.13) and (7.12) leads to the ray equation for a meridional ray in a GRIN-rod lens,

$$\frac{d^2x}{dz^2} = \frac{1}{2n_i^2 \cos^2 \gamma_i} \frac{\partial n^2(r)}{\partial x}. \quad (7.14)$$

Eikonal equation

The eikonal equation (4.58) for a medium with refractive index $n(r)$ can be rewritten as

$$|dS/ds|^2 = [n(r)]^2, \quad (7.15)$$

where $S = S(\mathbf{r})$ is the eikonal or optical path length. Inserting (7.13) into (7.15) gives

$$\frac{dS}{dz} = \frac{n^2(r)}{n_i \cos \gamma_i}. \quad (7.16)$$

Variable transformations

The following variable transformations are made for mathematical ease:

$$n_i \cos \gamma_i \zeta = n_0 g z, \quad (7.17a)$$

$$n_0 g^2 W = g \omega - n_0 \zeta. \quad (7.17b)$$

Equations (7.11) and (7.17) can be used to express the ray equation as a second-order differential equation in x alone,

$$\frac{d^2x}{d\zeta^2} + x = 2h_4g^2x^3 + 3h_6g^4x^5 + 4h_8g^6x^7 + \dots, \quad (7.18)$$

while an alternative expression for the eikonal equation is

$$\frac{dW}{d\zeta} + x^2 = h_4g^2x^4 + h_6g^4x^6 + h_8g^6x^8 + \dots. \quad (7.19)$$

Perturbation method

For many GRIN-rod lenses, $(gx)^2 \ll 1$, so that g^2 can be treated as a perturbation parameter. To begin, the ray path $x(z)$ and the parameter $W(z)$ are respectively modeled as

$$x(z) = x_0 + g^2x_1 + g^4x_2 + g^6x_3 + \dots, \quad (7.20a)$$

$$W(z) = \Delta + W_0 + g^2W_1 + g^4W_2 + g^6W_3 + \dots, \quad (7.20b)$$

so the zeroth-order perturbation solutions, x_0 and W_0 , are given by

$$x_0 = a \cos \psi, \quad (7.21a)$$

$$W_0 = -\frac{a^2}{2}\psi - \frac{a^2}{4}\sin 2\psi, \quad (7.21b)$$

where

$$\psi = \omega\zeta + \psi_i, \quad (7.21c)$$

$$\omega = 1 + g^2\omega_1 + g^4\omega_2 + g^6\omega_3 + \dots \quad (7.21d)$$

Initial conditions specify the constants a , ψ_i , and Δ , and the removal of a *secular* term determines the frequency correction terms, ω_1 , ω_2 , and ω_3 .

By substituting (7.20) and (7.21) into (7.18) and (7.19), then matching terms of the same order in g^2 , the first-order approximation of g^2 leads to the following coupled differential equations:

$$\frac{d^2x_1}{d\psi^2} + x_1 = 2\omega_1x_0 + 2h_4x_0^3, \quad (7.22a)$$

$$\frac{dW_1}{d\psi} + \omega_1 \frac{dW_0}{d\psi} = -2x_0x_1 + h_4x_0^4. \quad (7.22b)$$

Inserting (7.21a) into (7.22a) and applying a Fourier-series expansion to the right-hand side gives

$$\frac{d^2x_1}{d\psi^2} + x_1 = \left(2\omega_1 + \frac{3}{2}a^2h_4\right)a\cos\psi + \frac{1}{2}a^3h_4\cos 3\psi. \quad (7.23)$$

The secular term in (7.23) (i.e., the $\cos \psi$ term) is the one responsible for boundless growth as ψ increases, and eliminating it allows the approximation to hold for all ψ .

Removing this term requires that

$$\omega_1 = -\frac{3}{4}a^2h_4, \quad (7.24)$$

which can be substituted into (7.22a), along with x_0 , to obtain

$$x_1 = -\frac{1}{2^4}a^2h_4 \cos 3\psi. \quad (7.25a)$$

W_1 is determined by substituting the zeroth- and first-order perturbation solutions in (7.22b):

$$W_1 = \frac{1}{2^6}a^4(6h_4 \sin 2\psi + 3h_4 \sin 4\psi). \quad (7.25b)$$

The procedure for finding the third-order solutions is too lengthy and complicated to reproduce here, but can be found in Appendix A of Sakamoto (1993).

Ray path

Based on the perturbation solutions up to third-order, the ray path $x(z)$ for meridional rays is expressed as

$$gx(z) = ga \cos \psi + \sum_{j=1}^3 A_{2j+1} \cos(2j+1)\psi, \quad (7.26)$$

where the coefficients, A_3 , A_5 , and A_7 , are given by

$$A_3 = -\frac{(ga)^3}{2^4} h_4 - \frac{(ga)^5}{2^8} (21h_4^2 + 30h_6) - \frac{(ga)^7}{2^{12}} (417h_4^3 + 984h_4h_6 + 672h_8), \quad (7.27a)$$

$$A_5 = \frac{(ga)^5}{2^8} (h_4^2 - 2h_6) + \frac{(ga)^7}{2^{12}} (43h_4^3 + 24h_4h_6 - \frac{224}{3}h_8), \quad (7.27b)$$

$$A_7 = -\frac{(ga)^7}{2^{12}} (h_4^3 - 6h_4h_6 - \frac{16}{3}h_8). \quad (7.27c)$$

A similar equation can be determined for the optical path length $S(z)$, however, we are only interested in $x(z)$ for now.

Initial conditions

As previously mentioned, the initial conditions of the ray and eikonal equations specify the constants a , ψ_i , and Δ . Since Δ is primarily related to $S(z)$, it will not be discussed here. The necessary initial conditions are the initial position and slope:

$$x(0) = x_i, \quad (7.28a)$$

$$\frac{dx(0)}{dz} = \tan \gamma_i. \quad (7.28b)$$

Note that the period of oscillation in $x(z)$ is governed by the initial ray angle. Inserting (7.28) into (7.26) leads to these coupled equations:

$$gx_i = ga \cos \psi_i + \sum_{j=1}^3 A_{2j+1} \cos(2j+1)\psi_i, \quad (7.29a)$$

$$\frac{n_i}{n_0} \sin \gamma_i = -\omega \left[ga \sin \psi_i + \sum_{j=1}^3 (2j+1) A_{2j+1} \sin(2j+1)\psi_i \right]. \quad (7.29b)$$

To express a and ψ_i as functions of x_i and γ_i , the following expressions are assumed:

$$a = a_0 + g^2 a_1 + g^4 a_2 + g^6 a_3 + \dots, \quad (7.30a)$$

$$\psi_i = \psi_0 + g^2 \psi_1 + g^4 \psi_2 + g^6 \psi_3 + \dots \quad (7.30b)$$

Substituting (7.30) into (7.29) and again matching terms of the same order in g^2 leads to the following coupled equations:

(zeroth order)

$$a_0 \cos \psi_0 = x_i, \quad (7.31a)$$

$$a_0 \sin \psi_0 = -\frac{n_i}{n_0 g} \sin \gamma_i, \quad (7.31b)$$

(first order)

$$a_1 = \frac{a_0^3}{2^4} (6h_4 - 4h_4 \cos 2\psi_0 - h_4 \cos 4\psi_0), \quad (7.32a)$$

$$\psi_1 = \frac{a_0^2}{2^4} (8h_4 \sin 2\psi_0 + h_4 \sin 4\psi_0). \quad (7.32b)$$

7.2.2 Test lens description

Prescription parameters of the GRIN-rod test lens used in this numerical study are summarized in Table 7.6, where the refractive index distribution $n(r)$ is specified up to fourth-order by n_0 , g , and h_4 . Figure 7.13 provides a 2D plot of $n(r)$, indicating a 0.20% change in index between the center and the periphery.

Table 7.6: Design parameters of the GRIN-rod test lens at an arbitrary design wavelength. Included are the distances in the optical system used in the simulations.

Parameter	Value
Index at center, n_0	1.4140
Index at edge, $n(R)$	1.4112
Focusing parameter, g	0.018 mm^{-1}
Fourth-order coefficient, h_4	-120
Radius of lens, R	3 mm
Length of lens, L	160 mm
Distance from source to entrance face	200 mm
Distance from exit face to CCD	100 mm

Fig. 7.13: Refractive index distribution of the GRIN-rod lens.

When the front face of the lens is placed 200 mm from an on-axis point source, it subtends a half-angle of 0.86° . Using (7.26), we traced a 1200×1200 bundle of rays through the lens system for this source position and a detector positioned at 100 mm from the exit face. Sample rays are illustrated in Figure 7.14. The irradiance pattern, computed for a pixel size of $25 \mu\text{m}$ and a peak SNR of 10^3 , features a pronounced central peak and a relatively faint outer ring (Fig. 7.15).

Fig. 7.14: Real eikonal rays traced through the GRIN-rod lens. Plot is expanded in the transverse direction to show detail.

Fig. 7.15: (a) Irradiance distribution in the detector plane and (b) irradiance profile for the GRIN-rod test lens.

7.2.3 Fisher information and Cramér-Rao bounds

We computed the FIM and its inverse for the parameters describing the index distribution (i.e., n_0 , g , and h_4), which indicate that the optical system, as simple as it is, contains a wealth of information (Fig. 7.16). The Cramér-Rao bound on the parameters are once again incredibly small, enabling very precise estimates.

Fig. 7.16: (a) FIM and (b) inverse of the FIM for the parameters describing the refractive index distribution of the GRIN-rod lens. (logarithmic scale)

Table 7.7: Square-root of the CRB for the parameters describing the refractive index distribution of the GRIN-rod lens.

No.	Parameter	Units	True value	$(\text{CRB})^{1/2}$
1	Index at center, n_0	N/A	1.4140	9.5×10^{-7}
2	Focusing parameter, g	mm^{-1}	0.018	1.7×10^{-8}
3	Fourth-order coefficient, h_4	N/A	-120	6.3×10^{-4}

7.2.4 Likelihood surfaces

Figures 7.17 – 7.19 provide likelihood surfaces for the three pairs of parameters. Each plot is centered on the true minimum. While the surfaces are slowly varying along the n_0 axis, the behavior along the g axis is quite complicated. However, the SA algorithm should be able to process this cost function with relative ease, based on its prior performance with convoluted, high-dimensional surfaces.

Table 7.8: : Range in likelihood surfaces for parameters describing the GRIN-rod lens, relative to the true values.

No.	Parameter	Units	True value	Range
1	Index at center, n_0	N/A	1.4140	± 0.06
2	Focusing parameter, g	mm^{-1}	0.018	± 0.006
3	Fourth-order coefficient, h_4	N/A	-120	± 40

Fig. 7.17: Likelihood surface along n_0 and g axes. Global minimum is located at center of plot.

Fig. 7.18: Likelihood surface along n_0 and h_4 axes. Global minimum is located at center of plot.

Fig. 7.19: Likelihood surface along g and h_4 axes. Global minimum is located at center of plot.

7.3 Summary of Chapter 7

Through numerical analysis, we demonstrated how inverse optical design can be used in optical testing to profile high-order aspheric surfaces or to measure the refractive index distribution of GRIN lenses.

Although the propagation algorithm consisted of ray-tracing for both cases, the two types of lenses required vastly different approaches. An iterative method was employed to find the precise ray-surface intersection for an aspheric lens, involving a marching-points and repeated-bisections scheme for root-isolation and root-refinement, respectively. We showed that the output of our rapid ray-tracing program compared extremely well with ZEMAX. Conversely, rays were traced through a GRIN-rod lens

using analytic but approximate solutions to the eikonal equation based on the perturbation method.

In each application, the FIMs indicated high sensitivity of the data to changes in the parameters, thereby resulting in diminutive CRBs, even with a modest peak SNR. The exceptions were the higher-order aspheric coefficients, α_8 and α_{10} , which caused a singular, non-invertible information matrix, so we excluded them from the FIM, as well as the estimation task.

Since the likelihood surfaces again revealed many local extrema, we implemented global optimization with simulated annealing for the precision asphere. The estimate biases and variances were very small, except for the variance in the sixth-order aspheric coefficient, α_6 . Averaged over an increasing number of trials, however, the bias for this coefficient approached zero.

The next stage in this research is to repeat the estimation procedure on real data. An important objective is to identify deficiencies in the forward model, including additional sources of noise. In cases where ray-tracing is inadequate for representing the irradiance data, we suggested a combined approach of ray-tracing and diffraction propagation to reduce the computational intensiveness.

CHAPTER 8

CONCLUSION AND FUTURE WORK

In this dissertation, we presented the theoretical framework and several applications of our basic method of inverse optical design. The results for these applications will be summarized here.

In Chapter 5, we presented results for the original application of IOD, which is to estimate the complete set of patient-specific ocular parameters from Shack-Hartmann WFS data. We developed an optical design program that performs non-paraxial ray-tracing through the quadric surfaces of the eye, which may incorporate surface misalignments. The system configuration involved multiple beam angles to detect on- and off-axis aberrations, resulting in reduced parametric coupling, greater Fisher information, and smaller CRBs. One of the key points in our approach is that we do not perform centroid estimation as in classical wavefront sensing, since this results in severe information loss; instead, the raw detector outputs of the WFS are used as input to IOD. Due to the multitude of local extrema in the likelihood surface, we implemented SA search algorithm.

The bias and variance for each estimate were very small, giving much hope for success in a real experiment. However, this does not take into account any modeling errors, since the same program that generated the data was also used during the optimization. For this method to succeed with real patient data, there must be very

accurate modeling of the extremely complex optical system of the eye. As mentioned in Chapter 5, we did not consider factors such as the optical tear film effect, the GRIN distribution of the crystalline lens, irregularities in the corneal surface, scattering in the ocular media, the Stiles-Crawford effect, and so on.

One way to test estimator robustness is to intentionally use a different model during optimization than was used to generate the data, so that we can determine the estimation error due to model deficiencies. For instance, the data may include the GRIN distribution of the lens, while the estimation procedure may involve an equivalent refractive index.

Making our method practical in the clinical setting requires rapid processing techniques, especially if the computational time increases due to enhancements in the forward model. The studies in Chapter 5 were performed in the early stages of this project, prior to the use of GPU technology within our research group, so there is still much to explore here.

Although the original motivation of our work relates to vision science and ophthalmology, we are also leveraging the basic method in optical shop testing. In Chapter 6, we provided both numerical and experimental results for parameterized wavefront estimation. Here we estimated the pupil phase distribution of an optical system from multiple irradiance measurements near focus, by first parameterizing the wavefront with a set of expansion functions, the Zernike polynomials. We developed a parallel algorithm in CUDA to simulate the wave propagation that takes advantage of the FFT in the Fresnel approximation. The required amount pupil sampling was carefully examined for lenses of different f-numbers and aberration levels.

In the numerical study, our method was successful for a test lens with a large peak-to-valley wavefront error of 150λ . Both the estimate biases and variances were negligible, although the diminutive CRBs were not achieved. We declared that the only way to attain the CRBs is to locate the global minimum of the cost function, which is plagued with numerous local minima for this application. If the biases or variances are unsatisfactory for a particular purpose, either more time must be spent on searching for the global basin, or a new configuration with less local extrema must be found.

In the experimental wavefront estimation study, we used a benign test lens with a larger f-number and much lower aberrations, therefore, the pupil sampling requirement was relatively relaxed and the Fresnel approximation more valid. However, the use of real data created additional challenges, including the presence of nuisance parameters and the stringent requirement of an accurate forward model. The nuisance parameters were the actual magnification of the imaging lens (a microscope objective) and the exact image plane locations. We performed a 2D grid search of the unimodal likelihood surface for these parameters, assuming the design Zernike coefficients for the lens.

We examined the accuracy of the Fresnel approximation by comparing the output irradiance patterns to those of the Huygens diffraction integral, and the peak discrepancies were marginally 1.5–1.7% for the output planes of interest. Without knowledge of the true values of the parameters, we were unable to evaluate estimate biases, however, the estimates were within one standard deviation from the design values. Interestingly, the variances were comparable to those in the simulation study, perhaps because the variance is primarily influenced by stagnation in local basins, rather than

potential systematic errors or model deficiencies. Model-mismatch effects can be determined by using the Huygens formula to generate the data and the Fresnel FFT method in the estimation procedure.

In Chapter 7, we presented numerical results for additional applications in optical testing, which were parametric surface profilometry of precision aspheres and GRIN lens testing. For the former application, we produced a rapid ray-trace algorithm in CUDA that iteratively finds ray-surface intercepts for high-order aspheric surfaces, resulting in a computational time of 0.6 μ sec per ray with a single GPU. It might be the case that ray-tracing does not generate accurate irradiance data, and that diffraction propagation might be necessary when working with real data. A useful model-mismatch study would be to use wave propagation, having greater accuracy, for data generation and ray-tracing during the optimization. Aspheric coefficients beyond sixth-order and larger were excluded from the estimation step, since they did not influence the data and resulted in a singular FIM. Each ML estimate contained a very small bias and variance, apart from the variance of the sixth-order coefficient.

For the testing of GRIN-rod lenses, the estimated parameters were the coefficients of the index distribution. We developed a ray-trace program using analytic solutions of the eikonal equation. As usual, we discovered substantial Fisher information and multiple minima in the cost function.

A common characteristic of the likelihood surfaces across all applications is the presence of multiple local extrema, so we chose simulated annealing for each estimation procedure. Although SA was very successful in every case, a major drawback of this

search algorithm is its slowness. One way to mitigate this is to switch to a local descent algorithm once the system hones in on the final basin. When the temperature is low enough during the search process, the system essentially behaves like straightforward optimization anyway, since it is very unlikely to accept uphill moves. However, SA makes an inefficient local descent algorithm, and it would save time to use a suitable algorithm at this point.

A concerted effort was put into the computational aspects of this project, including the development of various optical design programs for implementation on both the CPU and GPU platforms. Rapid processing techniques were fully necessary to address the high computational demands of the maximum-likelihood approach, particularly due to the complicated objective functions encountered in IOD. We also performed extensive statistical analysis with the Fisher information matrix and visualization of the likelihood surfaces, investigating parametric coupling and information content in numerous system configurations through simulation. Furthermore, proof-of-principle studies for the various applications have been primarily computational. Although more work needs to be done in the way of physical experiments, we believe that the computational work and theory have been sufficiently developed, such that the next researcher can readily perform IOD on real data.

APPENDIX A

FRINGE ZERNIKE POLYNOMIALS

Optical imaging systems typically have a circular or annular pupil, as well as an axis of rotational symmetry. There are many applications in which it is useful to expand the wave aberration function of these systems in a power series or a complete set of orthogonal polynomials. Zernike polynomials (Zernike, 1934) are excellent candidates for this task, since they are orthogonal over a circular pupil and represent balanced aberrations with minimum variance.

A Zernike polynomial of a particular order in pupil coordinates achieves balanced aberrations by including terms of equal or lower order in a power series expansion, such that the variance is minimized (Born & Wolf, 1999; Mahajan, 2001, 2004). Note that this is different from balanced aberrations that yield minimum variance with respect to ray aberrations (Malacara, 2007).

Consider an optical system whose optical axis coincides with the z -axis. Let $\mathbf{r}$ be a position vector in the exit pupil, which is orthogonal to the optical axis. Using the standard convention for the polar angle θ , defined as the angle of $\mathbf{r}$ with the x -axis, we have

$$x = r \cos \theta, \quad y = r \sin \theta,$$

where $r = |\mathbf{r}|$.

For the purpose of wavefront estimation, discussed in Chapter 6, we are interested in systems with a circular pupil that are not necessarily rotationally symmetric. The wave aberration function for such a system will consist of terms in both $\cos m\theta$ and $\sin m\theta$, where $m \geq 0$, and can be expanded in terms of orthogonal Zernike polynomials $Z_n(\rho, \theta)$:

$$W(\rho, \theta) = \sum_n \alpha_n Z_n(\rho, \theta), \quad \text{A.1}$$

where α_n are the expansion coefficients. Without going into detail on the many mathematical properties of Zernike polynomials, we will simply quote the Fringe Zernike polynomials, developed at the University of Arizona. These are identical to the standard polynomials, except for the indexing format and the order in which they are listed. The expressions are provided in Table A.1 for $\{n = 1, \dots, 37\}$, with the corresponding plots in Figure A.1. Note that these are the orthogonal, not orthonormal, versions of the Fringe Zernike polynomials.

Table A.1: Fringe Zernike Polynomials $\{Z_n, n = 1, \dots, 37\}$.

n	Fringe Zernike Polynomial	Aberration Type
1	1	Piston
2	$\rho \cos \theta$	Distortion - Tilt (x-axis)
3	$\rho \sin \theta$	Distortion - Tilt (y-axis)
4	$2\rho^2 - 1$	Defocus - Field Curvature
5	$\rho^2 \cos 2\theta$	Astigmatism, Primary (0° or 90°)
6	$\rho^2 \sin 2\theta$	Astigmatism, Primary ($\pm 45^\circ$)
7	$(3\rho^3 - 2\rho) \cos \theta$	Coma, Primary (x-axis)
8	$(3\rho^3 - 2\rho) \sin \theta$	Coma, Primary (y-axis)
9	$6\rho^4 - 6\rho^2 + 1$	Spherical Aberration, Primary
10	$\rho^3 \cos 3\theta$	Trefoil, Primary (x-axis)
11	$\rho^3 \sin 3\theta$	Trefoil, Primary (y-axis)
12	$(4\rho^4 - 3\rho^2) \cos 2\theta$	Astigmatism, Secondary (0° or 90°)
13	$(4\rho^4 - 3\rho^2) \sin 2\theta$	Astigmatism, Secondary ($\pm 45^\circ$)
14	$(10\rho^5 - 12\rho^3 + 3\rho) \cos \theta$	Coma, Secondary (x-axis)
15	$(10\rho^5 - 12\rho^3 + 3\rho) \sin \theta$	Coma, Secondary (y-axis)
16	$20\rho^6 - 30\rho^4 + 12\rho^2 - 1$	Spherical Aberration, Secondary
17	$\rho^4 \cos 4\theta$	Tetrafoil, Primary (x-axis)
18	$\rho^4 \sin 4\theta$	Tetrafoil, Primary (y-axis)
19	$(5\rho^5 - 4\rho^3) \cos 3\theta$	Trefoil, Secondary (x-axis)
20	$(5\rho^5 - 4\rho^3) \sin 3\theta$	Trefoil, Secondary (y-axis)
21	$(15\rho^6 - 20\rho^4 + 6\rho^2) \cos 2\theta$	Astigmatism, Tertiary (0° or 90°)
22	$(15\rho^6 - 20\rho^4 + 6\rho^2) \sin 2\theta$	Astigmatism, Tertiary ($\pm 45^\circ$)
23	$(35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho) \cos \theta$	Coma, Tertiary (x-axis)
24	$(35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho) \sin \theta$	Coma, Tertiary (y-axis)

25	$70\rho^8 - 140\rho^6 + 90\rho^4 - 20\rho^2 + 1$	Spherical Aberration, Tertiary
26	$\rho^5 \cos 5\theta$	Pentafoil, Primary (x-axis)
27	$\rho^5 \sin 5\theta$	Pentafoil, Primary (y-axis)
28	$(6\rho^6 - 5\rho^4) \cos 4\theta$	Tetrafoil, Secondary (x-axis)
29	$(6\rho^6 - 5\rho^4) \sin 4\theta$	Tetrafoil, Secondary (y-axis)
30	$(21\rho^7 - 30\rho^5 + 10\rho^3) \cos 3\theta$	Trefoil, Tertiary (x-axis)
31	$(21\rho^7 - 30\rho^5 + 10\rho^3) \sin 3\theta$	Trefoil, Tertiary (y-axis)
32	$(56\rho^8 - 105\rho^6 + 60\rho^4 - 10\rho^2) \cos 2\theta$	Astigmatism, Quaternary (0° or 90°)
33	$(56\rho^8 - 105\rho^6 + 60\rho^4 - 10\rho^2) \sin 2\theta$	Astigmatism, Quaternary ($\pm 45^\circ$)
34	$(126\rho^9 - 280\rho^7 + 210\rho^5 - 60\rho^3 + 5\rho) \cos \theta$	Coma, Quaternary (x-axis)
35	$(126\rho^9 - 280\rho^7 + 210\rho^5 - 60\rho^3 + 5\rho) \sin \theta$	Coma, Quaternary (y-axis)
36	$252\rho^{10} - 630\rho^8 + 560\rho^6 - 210\rho^4 + 30\rho^2 - 1$	Spherical Aberration, Quaternary
37	$924\rho^{12} - 2772\rho^{10} + 3150\rho^8 - 1680\rho^6 + 420\rho^4 - 42\rho^2 + 1$	Spherical Aberration, 12th order

Fig. A.1: Fringe Zernike Polynomials 2-37.

APPENDIX B

LIST OF ACRONYMS

AO	: adaptive-optics
API	: application programming interface
CBEA	: cell broadband engine architecture
CRB	: Cramér-Rao lower bound
CUDATM	: Compute Unified Device Architecture
DP	: double-precision
FIM	: Fisher information matrix
FOV	: field-of-view
FPGA	: field-programmable gate array
FWHM	: full width at half maximum
GPU	: graphics processing unit
FLOPS	: floating-point operations per second
GRIN	: graded-index
i.i.d.	: independent and identically distributed
IOD	: inverse optical design
IOL	: intraocular lens
MAP	: maximum a posteriori
MCAO	: multi-conjugate adaptive optics

MCMC	: Markov-chain Monte Carlo
ML	: maximum-likelihood
MSE	: mean-square error
OPL	: optical path length
PDF	: probability density function
PR	: phase retrieval
PSF	: point-spread function
SA	: simulated annealing
SCE	: Stiles-Crawford effect
SHWFS	: Shack-Hartmann wavefront sensor
SI	: International System of Units
SIMD	: single-instruction multiple-data
SNR	: signal-to-noise ratio
WFS	: wavefront sensor

REFERENCES

- Aldrich, J. (1997). "R. A. Fisher and the making of maximum likelihood 1912 – 1922." *Stat. Sci.*, **12**, 162-176.
- Arfken, G. B. and Weber, H. J. (2001). *Mathematical Methods for Physicists*, Fifth Edition. Academic Press, San Diego.
- Artal, P. and Guirao, A. (1998). "Contributions of the cornea and lens to the aberrations of the human eye." *Opt. Lett.*, **23**, 1713-1715.
- Atchison, D. A. and Smith, G. (1995). "Continuous gradient index and shell models of the human lens." *Vision Res.*, **35**, 2529-2538.
- Atchison, D. A., Scott, D. H., Joblin, A., and Smith, G. (2000). "Influence of Stiles-Crawford effect apodization on spatial visual performance with decentered pupils." *J. Opt. Soc. Am. A*, **18**, 1201-1211.
- Atchison, D. A. and Smith, G. (2005). "Chromatic dispersions of the ocular media of human eyes." *J. Opt. Soc. Am. A*, **22**, 29-37.
- Audet, C. and Dennis, Jr., J. E. (2000). "Pattern search algorithms for mixed variable programming." *SIAM J. Optimiz.*, **11**, 573-594.
- Audet, C. and Dennis, Jr., J. E. (2003). "Analysis of generalized pattern searches." *SIAM J. Optimiz.*, **13**, 889-903.
- Baker, B. B. and Copson, E. T. (1949). *The Mathematical Theory of Huygens' Principle*, Second Edition. Clarendon Press, Oxford.
- Bará, S. and Navarro, R. (2003). "Wide-field compensation of monochromatic eye aberrations: expected performance and design trade-offs." *J. Opt. Soc. Am. A*, **20**, 1-10.
- Barankin, E. W. (1949). "Locally best unbiased estimates." *Ann. Math. Statist.*, **20**, 477-501.
- Bard, Y. (1974). *Nonlinear Parameter Estimation*. Academic Press, New York and London.

- Barrett, H. H. and Myers, K. J. (2004). *Foundations of Image Science*. Wiley, New Jersey.
- Barrett, H. H., Dainty, C., and Lara, D. (2007). "Maximum-likelihood methods in wavefront sensing: stochastic models and likelihood functions." *J. Opt. Soc. Am. A*, **24**, 391-414.
- Barrett, H. H., Sakamoto, J. A., and Goncharov, A., "Inverse optical design", U. S. Patent 7,832,864. Issued on 11/16/2010.
- Bhattacharyya, A. (1946). "On some analogues of the amount of information and their use in statistical estimation. Part 1." *Sankhya*, **8**, 1-14.
- Bhattacharyya, A. (1947). "On some analogues of the amount of information and their use in statistical estimation. Part 2." *Sankhya*, **8**, 201-218.
- Bhattacharyya, A. (1948). "On some analogues of the amount of information and their use in statistical estimation. Part 3." *Sankhya*, **8**, 315-328.
- Blinn, J. F. (2006). "How to solve a cubic equation, part 1: The shape of the discriminant." *IEEE Comput. Graph. Appl.*, **26**(3), 84-93.
- Blinn, J. F. (2006a). "How to solve a cubic equation, part 3: General depression and a new covariant." *IEEE Comput. Graph. Appl.*, **26**(6), 92-102.
- Bonomi, E. and Lutton, J. (1984). "The N -city travelling salesman problem: Statistical mechanics and the Metropolis algorithm." *SIAM Rev.*, **26**, 551-568.
- Booth, G. W. and Peterson, T. I. (1958). "Nonlinear Estimation." IBM SHARE Program Pa. No. 687 WLNLI.
- Born, M. and Wolf, E. (1999). *Principles of Optics*, 7th edition. Cambridge University Press, Cambridge.
- Bouwkamp, C. J. (1954). "Diffraction Theory." In A. C. Strickland, editor, *Reports on Progress in Physics*, Vol. XVII. The Physical Society, London.
- Box, G. E. P. and Muller, M. E. (1958). "A note on the generation of random normal deviates." *Ann. Math. Statist.* **29**, 610-611.
- Box, G. E. P. and Lucas, H. L. (1959). "Design of experiments in nonlinear situations." *Biometrika*. **46**, 77-90.

- Box, G. E. P. and Hunter, W. G. (1962). "A useful method for model-building." *Technometrics*, **4**, 301-318.
- Brady, G. R. and Fienup, J. R. (2004). "Improved optical metrology using phase retrieval." *2004 Optical Fabrication & Testing Topical Meeting*, OSA, Rochester, NY, paper OTuB3.
- Brady, G. R. and Fienup, J. (2005). "Phase retrieval as an optical metrology tool." *Optifab: Technical digest, SPIE Technical Digest TD03*, pp. 139-141.
- Brady, G. R., Guizar-Sicairos, M., and Fienup, J. (2009). "Optical wavefront measurement using phase retrieval with transverse translation diversity." *Opt. Express*, **17**, 624-639.
- Camp, J. J., Maguire, L. J., Cameron, B. M., and Robb, R. A. (1990a). "A computer model for the evaluation of the effect of corneal topography on optical performance." *Am. J. Ophthalmol.*, **109**(4), 379-386.
- Camp, J. J., Maguire, L. J., Cameron, B. M., and Robb, R. A. (1990b). "An efficient ray tracing algorithm for modeling visual performance from corneal topography." In *Proc. First Conf. on Visualization in Biomedical Computing*, Atlanta, GA, May 22-25. Piscataway, NJ, IEEE.
- Corana, A., Marchesi, M., Martini, C., and Ridella, S. (1987). "Minimizing multimodal functions of continuous variable with the 'simulated annealing' algorithm." *ACM T. Math. Software*, **13**, No. 3, 262-280.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.
- Daniels, H. E. (1961). "The asymptotic efficiency of a maximum likelihood estimator." *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, **1**, 151.
- Davidon, W. C. (1991). "Variable metric method for minimization." *SIAM J. Optim.*, **1**, 1-17.
- Dubbelman, M. and Van der Heijde, G. L. (2001). "The shape of the aging human lens: curvature, equivalent refractive index and the lens paradox." *Vision Res.*, **41**, 1867-1877.
- Dubbelman, M., Weeber, H. A., Van der Heijde, G. L., and Völker-Dieben, H. J. (2002). "Radius and asphericity of the posterior corneal surface determined by corrected Scheimpflug photography." *Acta. Ophthalmol. Scand.*, **80**, 379-383.

- Duff, T. (1992). "Interval arithmetic recursive subdivision for implicit functions and constructive solid geometry." *SIGGRAPH Comput. Graph.*, **26(2)**, 131-138.
- Dugué, D. (1937). "Application des propriétés de la limite au sens du calcul des probabilités a l'étude de diverse questions d'estimation." *Ecol. Poly.*, **3(4)**, 305-372.
- Eisenpress, H., Bomberault, A., and Greenstadt, J. (1966a). "Nonlinear Regression Equations and Systems, Estimation and Prediction (IBM) 7090." Computer program 7090-G2 IBM0035 G2, IBM, Hawthorne, New York.
- Eisenpress, H. and Greenstadt, J. (1966b). "The estimation of nonlinear econometric systems." *Econometrica*, **34**, 851-861.
- El Hage, S. G. and Berny, F. (1973). "Contribution of the crystalline lens to the spherical aberration of the eye." *J. Opt. Soc. Am.*, **63**, 205-211.
- Escudero-Sanz, I. and Navarro, R. (1999). "Off-axis aberrations of a wide-angle schematic eye model." *J. Opt. Soc. Am. A*, **16**, 1881-1891.
- Falk, J. E. and Soland, R. M. (1969). "An algorithm for separable nonconvex programming problems." *Manage. Sci.*, **15**, 550-569.
- Ferguson, T. S. (1967). *Mathematical Statistics, A Decision Theoretic Approach*. Academic Press, New York.
- Fisher, R. A. (1912). "On an absolute criterion for fitting frequency curves." *Messenger Math.*, **41**, 155-160.
- Fisher, R. A. (1922). "On the mathematical foundations of theoretical statistics." *Philos. Trans. Roy. Soc. London Ser. A*, **222**, 309-368.
- Fisher, R. A. (1925). "Theory of statistical estimation." *Proc. Cambridge Philos. Soc.*, **22**, 700-725.
- Fisher, R. A. (1934). "Two new properties of mathematical likelihood." *Proc. Roy. Soc. Ser. A*, **144**, 285-307.
- Fisher, R. A. (1935). "The logic of inductive inference." *J. Roy. Statist. Soc.*, **98(1)**, 39-54.
- Fletcher, R. and Powell, M. J. D. (1963). "A rapidly convergent descent method for minimization." *Comput. J.*, **6**, 163-168.

- Fletcher, R. and Reeves, C. M. (1964). "Function minimization by conjugate gradients." *Comput. J.*, **7**, 149-154.
- Fletcher, R. (1965). "Function minimization without evaluating derivatives—a review." *Comput. J.*, **8**, 33-41.
- Gauss, K. F. (1809). "Theoria Motus Corporum Coelestium." *Werke*, **7**, 240-254.
- Gillet, J. and Sheng, Y. (1999). "Simulated quenching with temperature rescaling for designing diffractive optical elements." In *Proc. 18th Congress Int. Commission for Optics*, volume 3749 of *Proc. SPIE*, pp. 683-684.
- Goldberg, D. E. and Richardson, J. (1987). "Genetic algorithms with sharing for multimodal function optimization." In Grefenstette, J. J., editor, *Proc. Second Int. Conf. on Genetic Algorithms*, 41-49. Lawrence Erlbaum.
- Goncharov, A. V. and Dainty, C. (2007). "Wide-field schematic eye models with gradient-index lens." *J. Opt. Soc. Am. A*, **24**, 2157-2174.
- Goncharov, A. V., Nowakowski, M., Sheehan, M. T., and Dainty, C. (2008). "Reconstruction of the optical system of the human eye with reverse ray-tracing." *Opt. Express*, **16**, 1692-1703.
- Goodman, J. W. (2005). *Introduction to Fourier Optics*, 3rd edition. Roberts & Company Publishers, Englewood.
- Gray, H. L. and Schucany, W. R. (1972). *The Generalized Jackknife Statistic*. Dekker, New York.
- Greenstadt, J. (1967). "On the relative efficiencies of gradient methods." *Math. Comp.*, **21**, 360-366.
- Guirao, A. and Artal, P. (2000). "Corneal wave aberration from videokeratography: accuracy and limitations of the procedure." *J. Opt. Soc. Am. A*, **17**, 955-965.
- Gullstrand, A. (1962). *Helmholtz's Handbuch der Physiologischen Optik*, 3rd Edition. English translation edited by J. P. Southall (Optical Society of America) Vol. 1, 351-352.
- Halmos, P. R. and Savage, L. J. (1949). "Application of the Radon-Nikodym theorem to the theory of sufficient statistics." *Ann. Math. Statist.*, **20**, 225-241.
- Hansen, P. and Mladenović, N. (2001). "Variable neighborhood search: Principles and applications." *Eur. J. Oper. Res.*, **130**, 449-467.

- Hansen, P. and Mladenović, N. (2002). "Variable neighborhood search." In P. Pardalos and M. Resende, editors, *Handbook of Applied Optimization*, Oxford, 221-234.
- He, J. C., Marcos, S., and Burns, S. A. (1999). "Comparison of cone directionality determined by psychophysical and reflectometric techniques." *J. Opt. Soc. Am. A*, **16**, 2363-2369.
- Hemenger, R. P., Garner, L. F., and Ooi, C. S. (1995). "Change with age of the refractive index gradient of the human ocular lens." *Invest. Ophthalm. Vis. Sci.*, **36**, 703-707.
- Hero, III, A. O., Fessler, J. A., and Usman, M. (1996). "Exploring estimator bias-variance tradeoffs using the uniform CR bound." *IEEE Trans. Signal Processing*, **44**, 2026-2041.
- Hestenes, M. R. and Stiefel, E. (1952). "Methods of conjugate gradients for solving linear systems." *J. Res. N. B. S.*, **49**, 409-436.
- Hestenes, M. R. (1969). "Multiplier and gradient methods." *J. Opt. Theo. Applns.*, **4**, 303-320, and in *Computing Methods in Optimization Problems*, 2 (Eds. L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan), Academic Press, New York, 1969.
- Hofer, H., Chen, L., Yoon, G., Singer, B., Yamauchi, Y., and Williams, D. R. (2001). "Improvement in retinal image quality with dynamic correction of the eye's aberrations." *Opt. Express*, **8**, 631-643.
- Hood, W. C. and Koopmans, T. C., eds. (1953). *Studies in Econometric Method*. Wiley, New York.
- Hooke, R. and Jeeves, T. A. (1961). "Direct search solution of numerical and statistical problems." *J. Assoc. Comput. Mach.*, **8**, 212-229.
- Huber, P. J. (1972). "Robust statistics: a review." *Ann. Math. Statist. and Prob.*, **1**, 221.
- Huygens, C. (1690). *Traité de la Lumière*. Leyden; Engl. Transl. Thompson, S. P. (1912). *Treatise on Light*. Macmillan, London.
- Ingber, L. (1993). "Simulated annealing: Practice versus theory ." *Math. Comp. Model.*, **18**, 29-57.
- Ingber, L. (1996). "Adaptive simulated annealing (asa): Lessons learned." *Control Cybern.*, **25**, 33-54.

- Jackson, J. D. (1975). *Classical Electrodynamics*, 2nd edition. John Wiley & Sons, New York.
- Jones, C. E., Atchison, D. A., Meder, R., and Pope, J. M. (2005). "Refractive index distribution and optical properties of the isolated human lens measured using magnetic resonance imaging (MRI)." *Vision Res.*, **45**, 2352-2366.
- Kajiya, J. T. (1982). "Ray tracing parametric patches." In *SIGGRAPH '82*, pp. 245-254.
- Kendall, M. and Stuart, A. (1979). *The Advanced Theory of Statistics, Vol. 2: Inference and Relationship*, 4th edition. Charles Griffin & Co. Ltd., London.
- Kirkpatrick, S., Gelatt, Jr., C. D., Vecchi, M. P. (1983). "Optimization by simulated annealing." *Science*, **220**, No. 4598, 671-680.
- Kirkpatrick, S., Gelatt, Jr., C. D., Vecchi, M. P. (1984). "Optimization by simulated annealing: Quantitative study." *J. Stat. Phys.*, **34**, 975.
- Kittel, C. and Kroemer, H. (1980). *Thermal Physics*. W.H. Freeman and Company, New York.
- Kooijman, A. C. (1983). "Light distribution on the retina of a wide-angle theoretical eye." *J. Opt. Soc. Am.*, **73**, 1544-1550.
- Koretz, J. F., Strenk, S. A., Strenk, L. M., and Semmlow, J. L. (2004). "Scheimpflug and high-resolution magnetic resonance imaging of the anterior segment: a comparative study." *J. Opt. Soc. Am. A*, **21**, 346-354.
- Langenbucher, A., Viestenz, A., Viestenz, A., Brunner, H., and Seitz, B. (2006). "Ray tracing through a schematic eye containing second-order (quadric) surfaces using 4×4 matrix notation." *Ophthalm. Physiol. Opt.*, **26**, 180-188.
- Lecam, L. (1970). "On the assumptions used to prove asymptotic normality of maximum likelihood estimates." *Ann Math. Statist.*, **41**, 802.
- Legendre, A. M. (1805). *Nouvelles Méthodes pour la Détermination des Orbites de Comètes*. Paris.
- Levy, A. V. and Montalvo, A. (1985). "The Tunneling Algorithm for the Global Minimization of Functions." *SIAM J. Sci. Stat. Comp.*, **6**, 15-29.
- Liang, J., Williams, D. R., and Miller, D. (1997). "Supernormal vision and high-resolution retinal imaging through adaptive optics." *J. Opt. Soc. Am. A*, **14**, 2884-2892.

- Liberti, L. and Maculan, N. (2006). *Global Optimization: From Theory to Implementation*. Springer, New York.
- Lotmar, W. (1971). "Theoretical eye model with aspherics." *J. Opt. Soc. Am.*, **61**, 1522-1529.
- Mahajan, V. N. (2001). *Optical Imaging and Aberrations, Part I: Ray Geometrical Optics*. SPIE Press, Bellingham, Washington, second printing.
- Mahajan, V. N. (2004). *Optical Imaging and Aberrations, Part I: Wave Diffraction Optics*. SPIE Press, Bellingham, Washington, second printing.
- Malacara, D. (2007). *Optical Shop Testing*, Third edition. Wiley, Hoboken.
- Mallen, E. and Kashyap, P. (2007). "Technical note: measurement of retinal contour and supine axial length using the Zeiss IOLMaster." *Ophthalm. Physiol. Opt.*, **27**, 404-411.
- McAulay, R. J. & Hofstetter, E. M. (1971). "Barankin bounds on parameter estimation." *IEEE T. Inform. Theory*, **17**, 669-676.
- Melsa, J. L. and Cohn, D. L. (1978). *Decision and Estimation Theory*. McGraw-Hill, New York.
- Metcalf, H. J. (1965). "Stiles-Crawford apodization." *J. Opt. Soc. Am.*, **55**, 72-74.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). "Equation of state calculations by fast computing machines." *J. Chem. Phys.*, **21**, No. 6, 1087-1092.
- Miller, L. H. (1964). "A trustworthy jackknife." *Ann. Math. Statist.*, **35**, 1549-1605.
- Mitchell, D. P. (1990). "Robust ray intersection with interval arithmetic." In *Proc. Graphics Interface '90*, pp. 68-74.
- Moffat, B. A., Atchison, D. A., and Pope, J. M. (2002). "Age-related changes in refractive index distribution and power of the human lens as measured by magnetic resonance micro-imaging in vitro." *Vision Res.*, **42**, 1683-1693.
- Navarro, R., Santamaría, J., and Bescós, J. (1985). "Accommodation-dependent model of the human eye with aspherics." *J. Opt. Soc. Am. A*, **2**, 1273-1280.

Navarro, R., Moreno, E., and Dorronsoro, C. (1998). "Monochromatic aberrations and point-spread functions of the human eye across the visual field." *J. Opt. Soc. Am. A*, **15**, 2522-2529.

Navarro, R., González, L., and Hernández, J. L. (2006). "Optics of the average normal cornea from general and canonical representations of its surface topography." *J. Opt. Soc. Am. A*, **23**, 219-232.

Navarro, R., Palos, F., and González, L. M. (2007). "Adaptive model of the gradient index of the human lens. I. optics formulation and model of aging ex vivo lenses." *J. Opt. Soc. Am. A*, **24**, 2175-2185.

Navarro, R., Palos, F., and González, L. M. (2007a). "Adaptive model of the gradient index of the human lens. II. optics of the accommodating aging lens." *J. Opt. Soc. Am. A*, **24**, 2911-2920.

Neal, D. R., Topa, D. M., and Copland, J. (2001). "The effect of lenslet resolution on the accuracy of ocular wavefront measurements." In *Proc. SPIE*, **4245**, 78-91.

Neal, D. R., Copland, J., and Neal, D. (2002). "Shack-Hartmann wavefront sensor precision and accuracy." In *Proc. SPIE*, **4779**, 148-160.

Nelder, J. A. and Mead, R. (1965). "A simplex method for function minimization." *Comput. J.*, **20**, 308-313.

Neyman, J. (1935). "Sur un teorema concernente le cosiddette statistiche sufficienti." *Inst. Ital. Atti. Giorn.*, **6**, 320-334.

Neyman, J. (1937). "Outline of a theory of statistical estimation based on the classical theory of probability." *Philos. Trans. Roy. Soc. London Ser. A*, **236**, 333-380.

Neyman, J. and Pearson, E. S. (1936). "Contributions to the theory of testing statistical hypotheses. I. Unbiased critical regions of type A and type A_1 ." *Stat. Res. Mem.*, **1**, 1-37.

Nister, D. (2004). "An efficient solution to the five-point relative pose problem." *IEEE Trans. Pattern Anal. Mach. Intell.*, **26(6)**, 756-777.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge.

Pearson, K. (1894). "Contributions to the mathematical theory of evolution." *Philos. Trans. Roy. Soc. London Ser. A*, **185**, 71-110.

- Pearson, K. (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling." *Philos. Mag. Fifth Series*, **50**, 157-175.
- Pearson, K. (1901). "Of lines and planes of closest fit to systems of points in space." *Philos. Mag. Sixth Series*, **2**, 559-572.
- Pearson, K. (1936). "Method of moments and method of maximum likelihood." *Biometrika*, **28**, 34-59.
- Plackett, R. L. (1972). "The discovery of the method of least squares." *Biometrika*, **67**, 239-251.
- Platt, B. and Shack, R. (2001). "History and principles of Shack-Hartmann wavefront sensing." *J. Refract. Surg.*, **17**, S573 – S577.
- Poincaré, H. (1892). *Théorie Mathématique de la Lumière vol. II*. Georg Carre, Paris.
- Polak, E. (1971). *Computational Methods in Optimization: A Unified Approach*. Academic Press, New York.
- Powell, M. J. D. (1964). "An efficient method for finding the minimum of a function of several variables without calculating derivatives." *Comput. J.*, **7**, 155-162.
- Powell, M. J. D. (1965). "A method for minimizing a sum of squares of nonlinear functions without calculating derivatives." *Comput. J.*, **7**, 303-307.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*, 2nd edition. Cambridge, New York.
- Quenouille, M. H. (1956). "Notes on bias in estimation." *Biometrika*, **43**, 353-360.
- Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Graduate School of Business Administration, Harvard Univ., Boston.
- Rao, C. R. (1945). "Information and accuracy attainable in the estimation of statistical parameters." *Bull. Calcutta Math. Soc.* **37**, 81-91.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Second Edition. Wiley, New York.

- Redding, D., Dumont, P., and Yu, J. (1993). "Hubble Space Telescope prescription retrieval." *Appl. Opt.*, **32**, 1728-1736.
- Robson, D. S. and Whitlock, J. H. (1964). "Estimation of a truncation point." *Biometrika*, **51**, 33-39.
- Romeo, F., Sangiovanni Vincentelli, A., and Sechen, C. (1984). "Research on simulated annealing at Berkeley." In *Proc. IEEE Int. Conf. on Computer Design, ICCD 84*, IEEE New York, 652-657.
- Roorda, A., Romero-Borja, F., Donnelly, III, W., Queener, H., Hebert, T., and Campbell, M. (2002). "Adaptive optics scanning laser ophthalmoscopy." *Opt. Express*, **10**, 405-412.
- Rosales, P. and Marcos, S. (2006). "Phakometry and lens tilt and decentration using a custom-developed Purkinje imaging apparatus: Validation and measurements." *J. Opt. Soc. Am. A*, **23**, 509-520.
- Rosales, P., Dubbelman, M., Marcos, S., and Van der Heijde, G. L. (2006a). "Crystalline lens radii of curvature from Purkinje and Scheimpflug imaging." *J. Vis.*, **6**, 1057-1067.
- Rosenbrock, H. H. (1960). "An automatic method for finding the greatest or least value of a function." *Comput. J.*, **3**, 175-184.
- RoyChowdury, P., Singh, Y. P., and Chansarkar, R. A. (2000). "Hybridization of gradient descent algorithms with dynamic tunneling methods for global optimization." *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, **30**, 384-390.
- Rynders, M., Lidkea, B., Chisholm, W., and Thibos, L. N. (1995). "Statistical distribution of foveal transverse chromatic aberration, pupil centration, and angle psi in a population of young adult eyes." *J. Opt. Soc. Am. A*, **12**, 2348-2357.
- Ryoo, S., Rodrigues, C. I., Bagsorkhi, S. S., Stone, S. S., Kirk, D. B., and Hwu, W. W. (2008). "Optimization principles and application performance evaluation of a multithreaded GPU using CUDA." In *Proc. 13th ACM SIGPLAN Symp. Principles and Practice of Parallel Programming*, ACM Press, 73-82.
- Sakamoto, J. A., Barrett, H. H., and Goncharov, A. V. (2008). "Inverse optical design of the human eye using likelihood methods and wavefront sensing." *Opt. Express*, **16**, 304-314.

- Sakamoto, T. (1993). "Analytic solutions of the eikonal equation for a GRIN-rod lens 1. Meridional rays." *J. Mod. Optics*, **40**, 503-516.
- Sakamoto, T. (1995). "Analytic solutions of the eikonal equation for a GRIN-rod lens 2. Skew rays." *J. Mod. Optics*, **42**, 1575-1592.
- Sato, S. (1997). "Simulated quenching: A new placement method for module generation." In *ICCAD'97: Proc. 1997 IEEE/ACM Int. Conf. on Computer-aided Design*, 538-541, San Jose, California, United States.
- Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.
- Schwiegerling, J., Greivenkamp, J. E., and Miller, J. M. (1995). "Representation of videokeratographic height data with Zernike polynomials." *J. Opt. Soc. Am. A*, **12**, 2105-2113.
- Schwiegerling, J. (2004). *Field Guide to Visual and Ophthalmic Optics*. SPIE Press, Bellingham, Washington.
- Schwiegerling, J. and Neal, D. "Historical development of the Shack-Hartmann wavefront sensor." In Harvey, J. E. and Hooker, R. B. (2005) *Robert Shannon and Roland Shack: Legends in Applied Optics*. SPIE Press Monograph Vol. PM148, pp. 132-139.
- Seal, H. L. (1967). "The historical development of the Gauss linear model." *Biometrika*, **54**, 1-24.
- Sederberg, T. W. and Chang, G. (1993). "Isolating the real roots of polynomials using isolator polynomials. In *Algebraic Geometry and Applications*, Springer Verlag.
- Seldin, J. H. and Fienup, J. R. (1990). "Numerical investigation of the uniqueness of phase retrieval." *J. Opt. Soc. Am. A*, **7**, 412-427.
- Sheehan, M. T., Goncharov, A. V., O'Dwyer, V. M., Toal, V., and Dainty, C. (2007). "Population study of the variation in monochromatic aberrations of the normal human eye over the central visual field." *Opt. Express*, **15**, 7367-7380.
- Silver, S. (1962). "Microwave aperture antennas and diffraction theory." *J. Opt. Soc. Am.*, **52**, 131.
- Singh, J. M. and Narayanan, P. J. (2007). "Real-time ray tracing of implicit surfaces on the GPU." Technical Report IIIT/TR/2007/72, International Institute of Information Technology, Hyderabad, India.

- Smith, G., Atchison, D. A., Pierscione, B. K. (1992). "Modeling the power of the aging human eye." *J. Opt. Soc. Am. A*, **9**, 2111-2117.
- Smith, W. E., Barrett, H. H., and Paxman, R. G. (1983). "Reconstruction of objects from coded images by simulated annealing." *Opt. Lett.*, **8**, 199-201.
- Smith, W.E., Paxman, R. G., and Barrett, H. H. (1985). "Application of simulated annealing to coded-aperture design and tomographic reconstruction." *IEEE Trans. Nuc. Sci.*, NS-32, 758-761.
- Sommerfeld, A. (1896). Mathematische theorie der diffraction. *Math. Ann.*, **47**, 317.
- Sommerfeld, A. (1954). *Optics*, volume IV of *Lectures on Theoretical Physics*. Academic Press, New York.
- Southwell, W. (1980). "Wave-front estimation from wave-front slope measurements." *J. Opt. Soc. Am.*, **70**, 998-1006.
- Spang, III, H. A. (1962). "A review of minimization techniques for nonlinear functions." *SIAM Rev.*, **4**, 343-365.
- Spendley, W., Hext, G. R., Himsforth, F. R. (1962). "Sequential application of simplex designs in optimization and evolutionary operation." *Technometrics*, **4**, 441-461.
- Stamnes, J. J. (1986). *Waves in Focal Regions*. Taylor & Francis Group, New York.
- Stavroudis, O. N. (1972). *The Optics of Rays, Wavefronts, and Caustics*. Academic Press, New York and London.
- Stefanescu, I. S. (1985). "On the phase retrieval problem in two dimensions." *J. Math. Phys.*, **26**, 2141-2160.
- Stiles, W. S. and Crawford, B. H. (1933). "The luminous efficiency of rays entering the eye pupil at different points." *P. Roy. Soc. Lond. B*, **112**, 428-450.
- Stoica, P. (2001). "Parameter estimation problems with singular information matrices." *IEEE T. Signal Proces.*, **49**, 87-90.
- Storn, R. and Price, K. (1997). "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces." *J. Global Optim.*, **11**, 341-359.

- Strang, G. (1980). *Linear Algebra and Its Applications*, 2nd edition. Academic Press, Orlando, FL.
- Straub, J., Schwiegerling, J., and Gupta, A. (2001). "Design of a compact Shack-Hartmann aberrometer for real-time measurement of aberrations in human eyes." *Vision Science and Its Applications, OSA Technical Digest* (Optical Society of America, Washington DC), 110-113.
- Streifer, W. and Paxton, K. B. (1971). "Analytic solution of ray equations in cylindrically inhomogeneous guiding media. 1: meridional rays." *Appl. Optics*, **10**, 769-775.
- Strenk, S. A., Semmlow, J. L., Strenk, L. M., Munoz, P., Gronlund-Jacob, J., and DeMarco, J. K. (1999). "Age-related changes in human ciliary muscle and lens: a magnetic resonance imaging study." *Invest. Ophthalmol. Visual Sci.*, **40**, 1162-1169.
- Tabernero, J., Benito, A., Nourrit, V., and Artal, P. (2006). "Instrument for measuring the misalignments of ocular surfaces." *Opt. Express*, **14**, 10945-10956.
- Tan, P. and Drossos, C. (1975). "Invariance properties of maximum likelihood estimators." *Math. Mag.*, **48**, 37-41.
- Teague, M. R. (1983). "Deterministic phase retrieval." *J. Opt. Soc. Am.*, **73**, 1434-1441.
- Thibos, L. N., Ye, M., Zhang, X., and Bradley, A. (1992). "The chromatic eye: a new reduced-eye model of ocular chromatic aberration in humans." *Appl. Optics*, **31**, 3594-3600.
- Thibos, L. N., Ye, M., Zhang, X., and Bradley, A. (1997). "Spherical aberration of the reduced schematic eye with elliptical refracting surface." *Optom. Vision Sci.*, **74**, 548-556.
- Thibos, L. N., Bradley, A. (1999). "Modeling the refractive and neurosensor systems of the eye. Published as Chapter 4 (pp. 101-159) in *Visual Instrumentation: Optical Design and Engineering Principles*. Pantazis Mouroulis, ed. McGraw-Hill, New York.
- Torczon, V. (1997). "On the convergence of pattern search algorithms." *SIAM J. Optim.*, **7**, 1-25.
- Trotter, H. F. (1957). "Gauss's work (1803-1826) on the theory of least squares." Technical Report 5, Statistical Techniques Research Group, Princeton Univ.
- Tuy, H. (1998). *Convex Analysis and Global Optimization*. Kluwer, Dordrecht.

- Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory*. Wiley, New York.
- von Helmholtz, H. (1910). *Handbuch der physiologischen Optik*. Translation by Southall, J. P. C. (1925). *Helmholtz's treatise on physiologic optics*. Dover, New York, Vol III.
- Wah, B. W. and Wang, T. (1999). "Efficient and adaptive Lagrange-multiplier methods for nonlinear continuous global optimization." *J. Global Optim.*, **14**, 1-25.
- Wald, A. (1939). "Contributions to the theory of statistical estimation and testing hypotheses." *Ann. Math. Stat.*, **10**, 299-326.
- Wald, A. (1945). "Statistical decision functions which minimize the maximum risk." *Ann. Math. Second Series*, **46**, 265-280.
- Wald, A. (1950). *Statistical decision functions*. Wiley, New York.
- Westheimer, G. (1965). "Retinal light distribution for circular apertures in Maxwellian view." *J. Opt. Soc. Am.*, **49**, 41-44.
- White, S. R. (1984). "Concepts of scale in simulated annealing." In *Proceedings of the IEEE International Conference on Computer Design, ICCD 84*, New York, 646-651.
- Wyvill, G. and Trotman, A. (1990). "Ray-tracing soft objects." In *CG International '90*, pp. 469-476, New York, NY. Springer-Verlag New York, Inc.
- Young, T. (1802). "On the theory of light and colours." *Phil. Trans. R. Soc.*, **92**, 12-48.
- Zernike, F. (1934). "Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode." *Physica*, **1**, 689-704.
- Zhou, F., Hong, X., Miller, D. T., Thibos, L. N., and Bradley, A. (2004). "Validation of a combined corneal topographer and aberrometer based on Shack-Hartmann wave-front sensing." *J. Opt. Soc. Am. A*, **21**, 683-696.