McGRAW-HILL PROFESSIONAL ENGINEERING




# Modern Lens Design

Second Edition

- ✔ Detailed procedures for designing any major lens
- ✔ Lens CAD with OSLO – managing the design software
- ✔ Lens data tables – all new or updated

Warren J. Smith

Petitioner has added romanettes to pages i - v. Otherwise, it leaves the original page numbering.

MOTOROLA MOBILITY
EXHIBIT 1010

# Modern Lens Design

**Warren J. Smith**
*Chief Scientist*
*Kaiser Electro-Optics, Inc., Carisbad, California*
*and Consultant in Optics and Design*

**Second Edition**

**McGraw-Hill**
New York Chicago San Francisco Lisbon London Madrid
Mexico City Milan New Delhi San Juan Seoul
Singapore Sydney Toronto

The McGraw-Hill Companies

Copyright © 2005, 1992 by The McGraw-Hill Companies, Inc. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

ISBN: 978-0-07-177726-1

MHID: 0-07-177726-1

The material in this eBook also appears in the print version of this title: ISBN: 978-0-07-143830-8,

MHID: 0-07-143830-0.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. To contact a representative please e-mail us at bulksales@mcgraw-hill.com.

Information contained in this work has been obtained by The McGraw-Hill Companies, Inc. ("McGraw-Hill") from sources believed to be reliable. However, neither McGraw-Hill nor its authors guarantee the accuracy or completeness of any information published herein, and neither McGraw-Hill nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that McGraw-Hill and its authors are supplying information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. ("McGrawHill") and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill's prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED "AS IS." McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

*To Rudolf Kingslake (1903–2003)—professor, mentor, colleague, gentleman, and friend*

# Contents

Preface xiii

Chapter 1. Introduction 1

1.1 Lens Design Books 1
1.2 Reference Material 2
1.3 Specifications 2
1.4 Lens Design 4
1.5 Lens Design Program Features 7
1.6 About This Book 28

Chapter 2. Automatic Lens Design: Managing the Lens Design Program 11

2.1 Optimization 11
2.2 The Merit Function 13
2.3 Local Minima 19
2.4 The Landscape Lens 21
2.5 Types of Merit Functions 28
2.6 Stagnation 29
2.7 Generalized Simulated Annealing 30
2.8 Considerations about Variables for Optimization 31
2.9 How to Increase the Speed or Field of a System and Avoid Ray Failure Problems 36
2.10 Test Plate Fits, Melt Fits, Thickness Fits, and Reverse Aberration Fits 37
2.11 Spectral Weighting 40
2.12 How to Get Started 41

Chapter 3. Improving a Design 47

3.1 Lens Design Tip Sheet: Standard Improvement Techniques 47
3.2 Glass Changes: Index and *V*-value 51
3.3 Splitting Elements 52
3.4 Separating a Cemented Doublet 55

3.5 Compounding an Element 55
3.6 Vignetting and Its Uses 58
3.7 Eliminating a Weak Element—the Concentric Problem 60
3.8 Balancing Aberrations 60
3.9 The Symmetrical Principle 67
3.10 Aspheric Surfaces 68

Chapter 4. Evaluation: How Good Is This Design? 71

4.1 The Uses of a Preliminary Evaluation 71
4.2 OPD versus Measures of Performance 71
4.3 Geometric Blur Spot Size versus Certain Aberrations 80
4.4 Interpreting MTF—The Modulation Transfer Function 82
4.5 Fabrication Considerations 83

Chapter 5. Lens Design Data 85

5.1 About the Sample Lens Designs 85
5.2 Lens Prescriptions, Drawings, and Aberration Plots 87
5.3 Estimating the Potential of a Redesign 92
5.4 Scaling a Design, Its Aberrations, and Its Modulation Transfer Function 96
5.5 Notes on the Interpretation of Ray Intercept Plots 98
5.6 Various Evaluation Plots 103

Chapter 6. Telescope Objectives 109

6.1 The Thin Airspaced Doublet 109
6.2 Merit Function for a Telescope Objective 110
6.3 The Design of an *f*/7 Cemented Doublet Telescope Objective 115
6.4 Spherochromatism 118
6.5 Zonal Spherical Aberration 123
6.6 Induced Aberrations 124
6.7 Three-Element Objectives 125
6.8 Secondary Spectrum (Apochromatic Systems) 125
6.9 The Design of an *f*/7 Apochromatic Triplet 133
6.10 The Diffractive Surface in Lens Design 145
6.11 A Final Note 150

Chapter 7. Eyepieces and Magnifiers 151

7.1 Eyepieces 151
7.2 A Pair of Magnifier Designs 155
7.3 The Simple, Classical Eyepieces 155
7.4 Design Story of an Eyepiece for a 6 × 30 Binocular 160
7.5 Four-Element Eyepieces 176
7.6 Five-Element Eyepieces 187
7.7 Very High Index Eyepiece/Magnifier 187
7.8 Six- and Seven-Element Eyepieces 200

Chapter 8. Cooke Triplet Anastigmats 201

8.1 Airspaced Triplet Anastigmats 201
8.2 Glass Choice 205
8.3 Vertex Length and Residual Aberrations 206
8.4 Other Design Considerations 209
8.5 A Plastic, Aspheric Triplet Camera Lens 215
8.6 Camera Lens Anastigmat Design "from Scratch"—The Cooke Triplet 223
8.7 Possible Improvements to Our "Basic" Triplet 234
8.8 The Rare Earth (Lanthanum) Glasses 236
8.9 Aspherizing the Surfaces 237
8.10 Increasing the Element Thickness 246

Chapter 9. Split Triplets 247

Chapter 10. The Tessar, Heliar, and Other Compounded Triplets 259

10.1 The Classic Tessar 259
10.2 The Heliar/Pentac 266
10.3 The Portrait Lens and the Enlarger Lens 266
10.4 Other Compounded Triplets 272
10.5 Camera Lens Anastigmat Design "from Scratch"—The Tessar and Heliar 272

Chapter 11. Double-Meniscus Anastigmats 297

11.1 Meniscus Components 297
11.2 The Hypergon, Topogon, and Metrogon 297
11.3 A Two Element Aspheric Thick Meniscus Camera Lens 299
11.4 Protar, Dagor, and Convertible Lenses 302
11.5 The Split Dagor 305
11.6 The Dogmar 305
11.7 Camera Lens Anastigmat Design "from Scratch"—The Dogmar Lens 305

Chapter 12. The Biotar or Double-Gauss Lens 319

12.1 The Basic Six-Element Version 319
12.2 Twenty-Eight Things That Every Lens Designer Should Know About the Double-Gauss/Biotar Lens 329
12.3 The Seven-Element Biotar—Split-Rear Crown 334
12.4 The Seven-Element Biotar—Broken Contact Front Doublet 340
12.5 The Seven-Element Biotar—One Compounded Outer Element 340
12.6 The Eight-Element Biotar 340
12.7 A "Doubled Double-Gauss" Relay 350

Chapter 13. Telephoto Lenses 355

13.1 The Basic Telephoto 355
13.2 Close-up or Macro Lenses 356

13.3 Telephoto Designs 358
13.4 Design of a 200-mm *f*/4 Telephoto for a 35-mm Camera "from Scratch" 367

Chapter 14. Reversed Telephoto (Retrofocus and Fish-Eye) Lenses 395

14.1 The Reversed Telephoto Principle 395
14.2 The Basic Retrofocus Lens 397
14.3 Fish-Eye, or Extreme Wide-Angle Reversed Telephoto, Lenses 402

Chapter 15. Wide-Angle Lenses with Negative Outer Elements 415

Chapter 16. The Petzval Lens; Head-up Display Lenses 423

16.1 The Petzval Portrait Lens 423
16.2 The Petzval Projection Lens 423
16.3 The Petzval with a Field Flattener 426
16.4 Very High Speed Petzval Lenses 429
16.5 Head-up Display (HUD) Lenses, Biocular Lenses, and Head/Helmet Mounted Display (HMD) Systems 437

Chapter 17. Microscope Objectives 441

17.1 General Considerations 441
17.2 Classical Objective Design Forms: The Aplanatic Front 442
17.3 Flat-Field Objectives 446
17.4 Reflecting Objectives 446
17.5 The Microscope Objective Designs 447

Chapter 18. Mirror and Catadioptric Systems 455

18.1 The Good and the Bad Points of Mirrors 455
18.2 The Classical Two-Mirror Systems 456
18.3 Catadioptric Systems 469
18.4 Aspheric Correctors and Schmidt Systems 473
18.5 Confocal Paraboloids 476
18.6 Unobscured Systems 476
18.7 Design of a Schmidt-Cassegrain "from Scratch" 482

Chapter 19. Infrared and Ultraviolet Systems 503

19.1 Infrared Optics 503
19.2 IR Objective Lenses 504
19.3 IR Telescopes 507
19.4 Laser Beam Expanders 511
19.5 Ultraviolet Systems 514
19.6 Microlithographic Lenses 514

Chapter 20. Zoom Lenses 521

20.1 Zoom Lenses 521
20.2 Zoom Lenses for Point and Shoot Cameras 526

20.3 A 20x Video Zoom Lens 539
20.4 A Zoom Scanner Lens 541
20.5 A Possible Zoom Lens Design Procedure 542

Chapter 21. Projection TV Lenses and Macro Lenses 551

21.1 Projection TV Lenses 551
21.2 Macro Lenses 553

Chapter 22. Scanner/*f*-$\theta$, Laser Disk and Collimator Lenses 561

22.1 Monochromatic Systems 561
22.2 Scanner Lenses 561
22.3 Laser Disk, Focussing, and Collimator Lenses 571

Chapter 23. Tolerance Budgeting 573

23.1 The Tolerance Budget 573
23.2 Additive Tolerances 578
23.3 Establishing the Tolerance Budget 583

Chapter 24. Formulary 587

24.1 Sign Conventions, Symbols, and Definitions 587
24.2 The Cardinal Points 588
24.3 Image Equations 590
24.4 Paraxial Ray Tracing (Surface by Surface) 592
24.5 Invariants 594
24.6 Paraxial Ray Tracing (Component by Component) 594
24.7 Two-Component Relationships 595
24.8 Third-Order Aberrations—Surface Contributions 596
24.9 Third-Order Aberrations—Thin Lens Contributions: The G-Sum Equations 598
24.10 Stop Shift Equations 600
24.11 Third-Order Aberrations—Contributions from Aspheric Surfaces 601
24.12 Conversion of Aberrations to Wavefront Deformation (Optical Path Difference) 601

Glossary 605
References 621
Index 623

# Preface

My personal optical design experience has spanned more than five decades. They have been exciting, fascinating, and delightful decades; I have enjoyed each one. During that half century, lens design has changed radically. In the mid-twentieth century, lens design was still a semi-intuitive art, practiced by a few dedicated individuals of great perseverance, knowledge, and skill. And by mid-century most of the classic lens design forms had already been created. To this day, these designs are still the basis of many excellent modern optical systems.

Of course the practice of lens design today is radically different from what it was in the 1940s and 50s. Then, most optical design was done with an electromechanical desk calculator (e.g., Marchant, Frieden, and Monroe), and the raytracing rate, measured in terms of the number of surfaces through which one could trace the path of a ray in a given amount of time, was to the order of one ray surface in about 250s (if one were to work at it continuously through the day). Thus, using the current dimensions for raytracing speed, one did about 0.004 ray-surfaces per second. And these were only meridional two-dimensional rays, not the three-dimensional general rays ordinarily traced today. A great deal of ingenuity (and elegant theory) went into finding ways to avoid tracing any more rays than were absolutely necessary.

Thanks to the modern personal computer or PC, the computing rate has increased almost unbelievably. Today a run-of-the-mill PC is capable of calculating several million ray-surfaces per second; this is about nine or ten orders of magnitude faster. Needless to say the techniques of lens design today differ mightily from those of fifty or sixty years ago. Then, the designer might calculate the derivatives of a few aberrations with respect to a limited number of constructional parameters and solve a small set of simultaneous, linear equations in the course of correcting his lens. These limited calculations were all carefully selected on the basis of theory, experience, and intuition. (Interestingly, one of the very real problems facing designers today is that the computer spews out

numbers so rapidly that it takes strong self-discipline just to make oneself stop and think.)

In modern lens design work, a computer program almost instantaneously calculates and solves equations which are far more than an order of magnitude more complex and extensive than those cited above. It is not atypical for the computer program to control about 50 lens performance characteristics by adjusting the values of some 20 or 30 construction parameters of the optical system. These latter numbers imply a design space with 20 or 30 dimensions, a complex space indeed.

There are, however, some real limitations on the power of a so-called *automatic lens design* program. The typical program proceeds from a given starting design and drives the design to the nearest *local optimum*, a form at which any small structural changes will degrade the system performance. System performance is judged by a set of calculated characteristics defined in a *merit function*, which would be better termed a *defect* or *error* function, since the characteristics in it represent departures from desired values.

Obviously then, the final automatic design solution is completely and uniquely determined by (a) the merit function, (b) the starting design form, and (c) the algorithm by which the computer solves the problem of locating an optimum design form with the minimum value of the merit function.

When the first edition of *Modern Lens Design* (MLD) was published, there was a great need for a collection of suitable design forms at which to start the design process, and MLD provided almost 300 lens designs for this purpose. These designs were selected not only as starting points, but also as illustrations of important design principles. At the present time the need for sample designs, while still real, is significantly less, largely because most optical design programs now include libraries of lens designs. (These programs also include random search design capabilities which permit large changes in lens forms.) For example, all of the lens designs in the first edition of MLD (plus many others) are included in the lens libraries of the optical design program OSLO (a product of Lambda Research Corp.). Another program, LensVIEW by Brian Caldwell, is a compilation of over 30,000 lens designs and patents.

That said, it is (at least it is for me) far more easy and convenient to scan and compare a series of printed design pages than it is to do the same thing on a computer screen (even with the multiwindow capabilities of many programs). For this and other reasons this second edition of MLD has retained about half of the original designs and has added some new ones. The reader may also find some additional designs in the works referenced at the end of the book.

The practice of lens design is now essentially an engineering discipline. While this book is intended to be self-contained, we deliberately

do not include a lot of derivations, or even the mechanics of exact ray tracing. And as valuable and cherished as they may be in academia, we happily omit any derivations from first principles, Maxwell's equations, or Fermat's principle. These are simply not necessary for a book on lens design. We make one exception to the no "ray tracing" rule, namely for the tracing of paraxial rays, which a lens designer often carries out by hand, or with a programmed pocket calculator. This topic is covered in the Formulary of Chapter 24, along with other valuable and frequently used geometrical optics relationships.

However, there is currently a growing need for a more detailed exposition of basic lens design and theory in a single volume. The first edition of MLD was a "companion" volume to the author's *Modern Optical Engineering*. Several very basic lens design books have recently appeared; some are almost extended user manuals written for a specific design program. This edition of MLD is definitely not intended as a user manual, or as a guide to any specific program. It is an attempt to go well beyond this level by presenting both the basics of, and a more advanced approach to, lens design. The intent is to advise the reader how to get the most from any computer lens design program. To this end, about half of the lens designs in the first edition of MLD have been eliminated to make room for quite a bit of new material.

The text is, as far as possible, completely program neutral. I have tried to make the material regarding design programs as generic as I could, discussing features that are available in almost all commercial software. I have used OSLO for the design work demonstrated in the text, and for preparing the new figures. (The lens analysis figures in the first edition of MLD were prepared with a customized version of the program GENII, using a new and unique presentation style which is now widely available; for an example, see the OSLO aberration plots herein.)

Most neophyte lens designers very quickly get past the basics and learn to use their computer programs with a high level of proficiency. At this point, what they need most is an answer to the question, "What do I do *now*?" Much of the new material in this edition is designed to this end and takes the form of actual design projects carried out from scratch, warts and all. (In other words, I have not papered over the blunders I made in the design process.) These designs include a cemented doublet, a triplet anastigmat, a Tessar, a Heliar, a Dogmar, a telephoto, a Schmidt cassegrain, a binocular eyepiece, an apochromatic triplet, and a landscape lens. Many of these design stories are carried out to some length to illustrate all of the possible steps that can be taken to improve a design. Every initial assumption is explained and justified. These design descriptions not only show the basic design approach, but continue on with advanced steps and the rationale for them.

I have surveyed the literature at some length for any design techniques which might have a general applicability (as well as the reported specific use for the writer's specific problem). Some were found in the references listed at the end of the book. For the most part, the design techniques described here are those which I have found to be useful in working with an optimization program. Many of the techniques have been developed or refined during more than two decades of teaching courses in lens design; indeed some of these ideas were suggested or inspired by my students. Other valuable sources were the many informal discussions that I have been fortunate to have with my colleagues.

For better or worse, one can never seem to squeeze all the material that you want into a book. At the manuscript deadline date there is always at least one more feature that you wished there was enough time to develop, write, and include. But I suppose that if there were time, no book would ever be finished.

Surprisingly, there are only a modest number of well-understood and widely utilized principles of optical design. If one can master a thorough understanding of these principles, their effects, and their mechanisms, it is easy to recognize them in existing designs and also easy to apply them to one's own design work. It is the intent here to promote such understanding by presenting both expositions and annotated design examples of these principles.

Readers are free to use the designs in this book as starting points for their own design efforts, or in any other way they see fit. The reader must accept full responsibility for meeting whatever limitations are imposed on the use of these designs by any patent, copyright, or other (whether indicated herein or not).

And finally, I want to express my gratitude to Kaiser Electro-Optics and Jerry Carollo for accomodating me, and to Lambda Research Corp. and Leo Gardner for providing the lens design program OSLO and for helping me use it to its full potential. PBGT.

WARREN J. SMITH
*Vista, California*

# Modern Lens Design

Chapter

# 1

# Introduction

## 1.1 Lens Design Books

*Modern Lens Design* has several primary aims. The text contains extensive discussions of the principles and design techniques, which are appropriate to design work with a modern automatic lens design computer program. It is effectively a compendium of the many design techniques available to lens designers today. It is also a prescription source book for a variety of already designed lens types; these serve the dual purpose of providing starting points for the lens designer's efforts and also providing a basis for discussions of the application of many of the classic lens design principles. In the formulary of Chap. 24 we have provided a compact collection of most of the equations and relationships, which find frequent day-to-day use in lens design.

*Modern Lens Design* is intended as an aid to lens designers who work with the commercially available lens design computer programs. The approach used here assumes that the reader understands basic optical principles and may, in fact, have a command of the fundamentals of classical optical design methods. For those who want or need information in these areas, the following books should prove helpful. This author's *Modern Optical Engineering: The Design of Optical Systems*, 3rd ed., McGraw-Hill, 2000, is a comprehensive coverage of optical system design; it includes several chapters dealing specifically with lens design in considerable detail. Rudolf Kingslake's *Optical System Design* (1983), *Fundamentals of Lens Design* (1978), and *A History of the Photographic Objective* (1989), all by Academic Press, are complete, authoritative, very well written, and a pleasure to read.

## 1.2 Reference Material

There have been many books written about the subject of lens design. Some go deeply into the mathematical aspects of the subject; some are effectively expanded "user manuals" for a particular lens design program. (And then there are the program manuals themselves, many of which could benefit from an expanded and cross-referenced index, at the least.) Without further editorial comment I refer the reader to the list of references that follows Chap. 24. These are books, chapters, and articles whose subject is primarily *lens design, not optical engineering* or *optical system* design. Many of these primary references contain additional references that may be of interest to the reader who wishes to explore the edges and corners of the field of lens design. Many of the references are worthy of a thorough reading and study, in and of themselves. In the text we have cited these publications by their reference numbers as superscripts.

Adjacent to the reference list is a very complete glossary of the terms common to the practice of lens design. It is as complete, current, and authoritative as I could make it without becoming encyclopedic. I strongly recommend it to the reader. For the experienced designer it should serve to establish a common language. For the less experienced, a perusal of the glossary can serve almost as an introductory text.

## 1.3 Specifications

Before beginning the actual lens design work on a project, it is extremely wise to firmly nail down as many specifications as possible. Note that by "nail down" we do not mean cast in concrete; we mean that they should be at least considered and preliminarily agreed upon. An important part of defining a specification is determining which specifications are minimums, which are really just goals on someone's wish list, and which have some wiggle room, and if so, how much. It *always* pays to question the specifications. Are they theoretically possible and practically feasible in light of the diffraction limit and the conservation of radiance? Can they be met within the constraints of materials, spatial requirements, and the expected state of the fabrication technology? Which ones are those that define what the customer really wants and needs?

Note well that specifications are often prepared by someone who is not knowledgible in optics or optical design. The designer must convert the customer specifications into terms meaningful in optics, and should also be sure that the reality of fabrication tolerances is taken into account.

The following is a check list of possible specifications, which ought to prevent overlooking some requirement that might bite you later on.

### 1.3.1 Possible specification list

A checklist of specifications and considerations, which should be established before the lens design begins. Note that these items may be redundant or contradictory.

Effective focal length

Field angle or object size

F-number (infinity versus working)

Numerical aperture

Clear aperture

Size, weight, and packaging constraints

Wavelength and bandwidth

Spectral distribution and weighting

Transmission

Object and image distance

Object to image distance (*track length*)

Internal image required

Erect or inverted image

Magnification (and range)

Flat or curved field

Back focal length

Working distance (clearance)

Front focal length

Pupil location(s) and size; eye relief

Stops: Aperture, field, and cold

Space for iris diaphragm

Telecentric

Vignetting/illumination uniformity

Distortion

Zoom range: Continuous or stepped

Focus range

Focusing: Shift part or all of system

Anamorphic ratio

Performance
- versus diffraction limit
- versus field
- versus F-stop
- Resolution
- *Modulation transfer function* (MTF)
- Encircled/ensquared energy
- *Root mean square* (rms) spot or blur size
- Point spread function
- Wavefront aberration, *optical path difference* (OPD) (rms, P-V)
- Specific aberrations

Depth of focus/field

Lens type

Refractive versus mirror versus catadioptric

Plastic elements

Aspheric surfaces

Diffractive surfaces

Cost, quantity, delivery schedule

Materials

Environment, ambient conditions

Athermalization

Sensor, spectral response

Visual, photographic, projection

Patent considerations

Ergonomics

## 1.4 Lens Design

There are many commonly asked questions about lens design. Many of them are variations of the following:

1. Where do the aberrations come from?
2. Do we get spherical aberration because the surfaces are spherical?
3. Are aspheric surfaces free of aberration?
4. Are the aberrations related to each other?

The *complete* answer to no. 1 is that aberrations arise from Snell's law, $n \sin I = n' \sin I'$, which governs the refraction of rays at a surface. Because this is a trigonometric expression, it is nonlinear and the effect

on a ray's direction varies across the aperture of the lens surface. And Snell's law is especially nonlinear as the angles $I$ or $I'$ approach 90°.

The answer to no. 2 is that spherical surfaces usually do have spherical aberration, but when used in a certain way (the aplanatic case) they are free of spherical aberration, coma, and astigmatism.

For question no. 3 the answer is similar to that for no. 2. The conic aspherics (e.g., paraboloid, ellipsoid, hyperboloid, and the sphere) have no spherical aberration if the object point is at one focal point and the image is at the other. But if we depart from this condition we definitely do get spherical aberration. A paraboloid, noted for having no spherical aberration when the object is at infinity, has spherical aberration when the object is at any other distance.

Question no. 4: The aberrations are in fact related, as is illustrated by Fig. 1.1, where a single spherical refracting surface is shown with an aperture stop to its left. Consider an axial object point A that has its paraxial image at point A′. If we pivot the axis about the center of curvature of the surface, we generate a curved object surface, A-C, and a curved image surface, A′-C′. The departure of these surfaces from the ideal flat object and image surfaces, A-B and A′-B′, is the source of the *Petzval field curvature*. If we were to move object point C back to point B, putting it on the flat object surface, then the image point C′ will also shift to the left, increasing the field curvature at the image. Note that this is a basic field curvature and does not depend on astigmatism.

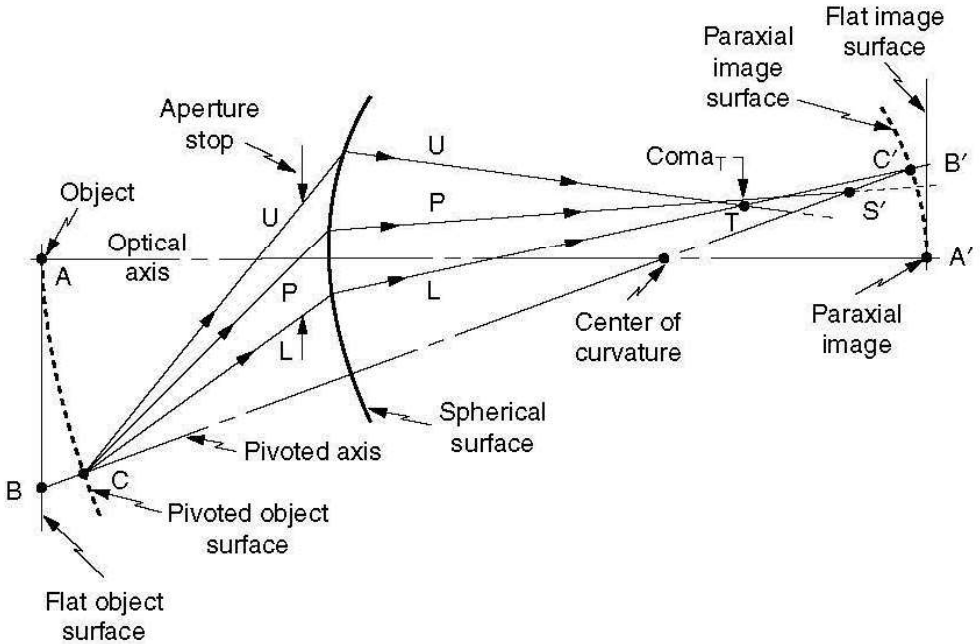



**Figure 1.1** A single spherical refracting surface with an aperture stop to its left

Three rays are shown traced from object point C through the aperture stop: an upper rim ray, U; a lower rim ray, L; and a principal ray, P. Spherical aberration causes them to intersect the pivoted axis (C-C′) at distances from the paraxial focus (C′), which increase as the square of the height at which they strike the surface. The intersection of rays U and L at point T is vertically displaced from the ray P; this is *tangential coma*. The focus of the upper and lower rays is at point T; this is the tangential (or meridional) focus of these marginal rays. The corresponding focus of the sagittal rays can be generated approximately by rotating the diagram of the principal ray about the pivoted axis (C-C′)—this focus is at point S. Points T and S are indicative of the *tangential* and *sagittal field curvature*, and the distance between them is the *astigmatism*. Note that T is three times as far from the curved image surface as is S. Object point B in the ideal flat object plane has an ideal image point at B′. The distance from B′ to the intersection of the principal ray with the ideal image plane is the *distortion*.

Until fairly recently the term "lens design" implied the design of telescope and camera lenses almost exclusively. Currently the applications of optics cast a much broader net. Nonetheless, we strongly urge a careful study of Kingslake's excellent *A History of the Photographic Lens*. Although you may not be designing a photographic lens, the techniques and principles that have been used to improve camera lens designs through the years are the same techniques and principles that may well allow you to improve whatever optical system on which you may be working, and Kingslake explains them very well (and interestingly).

Lens design is both a science and an art, possibly best learned at the feet of a master (lens designer). Although lens design today is carried out almost entirely on a personal computer, often with only a very minimal amount of preliminary manual calculation, a knowledge and understanding of the basic principles of geometrical optics is essential to success in this field. Learning how and why a successful lens design works and why a less successful design doesn't work is assuredly worth the effort. Many, foolishly and blindly, hope and believe that the optical software will miraculously find and deliver a great and practical solution to almost any lens problem. If only it were so! The best designs begin solidly grounded in a physically correct concept, and only then are they optimized and polished by the power of the computer programs.

There are many interesting and elegant analytical studies of optical system design problems, usually based on the third-order aberrations (see, for example, the equations in the formulary of Chap. 24). These *third-order aberration* (TOA) results can also be found using the optimization routines of a computer program without the labor of algebraic

derivations and the like. The equations of the formulary can be useful in getting to a good starting point for further computer work, but note that the equations also can yield a survey of the entire field (and not simply find a single local optimum as the optimization routines do). They can find a *good* solution, or multiple solutions, or show the impossibilities, and give a better understanding of the problem at hand. This is true for first-order imagery as well as for TOA, and the Chap. 24 formulary equations or the computer can equally well investigate the paraxial possibilities.

The basics should always be established before the actual design begins. As Bob Hopkins was fond of saying, "Keep your hands off the keyboard until you thoroughly understand the first- and third-order problem." Very wise advice indeed.

## 1.5 Lens Design Program Features

Most full-scale optical design programs have the following features or their equivalents. There are many specialized facilities in the major programs. We list here only those most widely utilized in ordinary practice.

**Variable parameters.** Any construction parameter, including:

Surface curvatures (radii)

Surface asphericities

Surface spacings (thickness and airspaces)

Materials (glass indices and dispersions)

Decentration or tilts

Vignetting

Variables may be subject to maximum and minimum bounds.

**Solves**

*Angle solve* for a curvature to get a desired ray slope after the surface. Applied to the last surface, it can establish the magnification or focal length.

*Height solve* to get a desired ray height at the next surface. If the axial ray height is zero it locates the paraxial focus.

*Pick up* of a previous data item

### Targets (operands)

Any construction value (e.g., $r$, $t$, $n$)

Any ray data (height, slope, exact, or paraxial)

First-, third-, and fifth-order aberration contributions or sums

Field curvature ($X_s$ and $X_t$)

Distortion

Offence against the sine condition (OSC)

S2T—aperture coefficient of coma

Optical path difference (OPD)

$(D - d)\delta n$

Derivatives of ray height or slope

Ray intersection differences

Averages, rms

Mathemetical combinations of targets (+, −, ×, ÷, **, >, <)

### Graphic representations (plots)

Ray intersections

Distortion versus field

Lateral chromatic versus field

Axial chromatic versus ray height

Spherical aberration versus ray height

OPD versus aperture or field

Spot size versus field

MTF versus frequency, focus, or field

## 1.6 About This Book

The book proper begins with Chap. 2 with a discussion of automatic lens design programs and how to use them effectively. Here the merit function, optimization, variables, and the various and several techniques that are useful in connection with a software program are covered. In Chap. 3 we discuss many specific improvement strategies that can be applied to an existing design to improve its performance. Chapter 4 deals with the evaluation of a design from the standpoint of both ray and wave aberrations, as well as such standard measures as spot size, MTF, and the Strehl ratio.

The presentation of the design data and performance in this book is explained in Chap. 5, as is the evaluation of the aberration data in estimating the potential of a particular design for a specific task. Chapters 6 through 22 discuss the major lens-design types and the manner of their design. For each there is the prescription, a drawing of the lens, and ray aberration analysis plots (with ray intercept plots, longitudinal plots of spherical, spherochromatic, and field curvature, plus transverse plots of distortion and lateral color). Chapter 23 on tolerance budgeting and Chap. 24, the Formulary, conclude the book.

Two particular design forms, the telescope objective in Chap. 6 and the Cooke triplet in Chap. 8, are given special and extensive attention. The design techniques used for these relatively simple designs are explained at length. These same techniques can be applied to more complex designs as well as to these more basic lens types.

*This page intentionally left blank*

Chapter

# 2

# Automatic Lens Design: Managing the Lens Design Program

## 2.1 Optimization

*"A lens is about as nonlinear as anything in physics."* RUDOLF KINGSLAKE

What is usually referred to as *automatic lens design* is, of course, nothing of the sort. The computer programs so described are actually optimization programs that drive an optical design to a local optimum, as defined by a *merit function* (which is not a true merit function, but actually a defect or error function). In spite of the preceding disclaimers, we will use these commonly accepted terms in the discussions that follow.

The lens design program typically operates this way: Each variable parameter is changed (one at a time) by a small increment whose size is chosen as a compromise between a large value (to get good numerical accuracy) and a small value (to get the local differential). The change produced in every item in the merit function is calculated. The result is a matrix of the partial derivatives of the defect items with respect to the parameters. Since there are usually many more defect items than variable parameters, the solution is a classical least-squares solution. It is based on the assumption that the relationships between the defect items and the variable parameters are linear. Since this is usually a false assumption, an ordinary least-squares solution will often produce an unrealizable lens or one that may in fact be worse than the starting design. The *damped least-squares* solution, in effect, adds the (suitably weighted) squares of the parameter changes to the merit function, heavily penalizing any large

changes and thus limiting the size of the changes in the solution. The mathematics of this process is described in Spencer, "A Flexible Automatic Lens Correction Program," *Applied Optics*, vol. 2, 1963, pp. 1257–1264, and in many other publications; it will not be repeated here.

The damped least-squares optimization process will pursue *any* improvement, no matter how small, and no matter the cost. If one arrives at a design with an overly thick element or a very big airspace, it's wise to test to see if it's really doing much good. Reoptimizing with the offending dimension fixed at a reasonable value and comparing the results will usually yield the answer to this question.

If the changes of the solution are small, the nonlinearities will not ruin the process, and the solution, although an approximate one, will be an improvement on the starting design. Continued repetition of the process will eventually drive the design to the nearest local optimum.

One can visualize the situation by assuming that there are only two variable parameters. Then the merit function space can be compared to a landscape where latitude and longitude correspond to the two variables and the elevation represents the value of the merit function. Thus the starting lens design is represented by a particular location in the landscape and the optimization routine will move the lens design downhill until a minimum elevation is found. Since there may be many depressions in the terrain of the landscape, this optimum may not be the best there is; it is a local optimum and there can be no assurance (except in very simple systems) that we have found a "global" optimum in the merit function. This simple topological analogy helps us to understand the dominant limitations of the optimization process: the program finds the nearest minimum in the merit function, and that minimum is uniquely determined by the design coordinates at which the process is begun. The landscape analogy is easy for the human mind to comprehend; when it is extended to a 10- or 20-dimension space, one can realize only that it is apt to be an extremely complex neighborhood.

One version of adaptive optimization adjusts the targets in the merit function so that they are an ever-decreasing fraction of the distance to the desired solution. This has the virtue of keeping the steps in the optimization process relatively small (and less nonlinear) and the progress relatively constant. This may be incorporated in a program, or it can be accomplished manually by examining the calculated values of the individual operands in the merit function and adjusting the target values accordingly.

The *multiconfiguration* feature in most programs allows the optimization of a system in two or more configurations, wherein all the configurations share common elements of construction but differ in other aspects such as spacings and added components. The classic example is the zoom lens where each configuration uses the same optical

elements, but several spaces differ (so as to change the focal length while maintaining focus and image quality). Other applications include: designing a system to operate at several different (or over a range of) object distances, as in a "macro" lens; increasing the depth of focus by designing configurations, each with slightly different image surface locations but the same optics; reducing sensitivity to tolerances by designing two configurations, each with a lens dimension (radius, thickness, alignment, and the like), which has a slightly different value in each configuration. Many programs perform a coarse search to determine the best value for the damping factor.

The result of a damped least-squares optimization run depends on the starting point given to the program. Obviously, changing the starting design form can produce a new result, for better or worse.

## 2.2 The Merit Function

The merit function is the most important aspect of an optical design program. As a result, the "care and feeding" of one's merit function may occupy more of the lens designer's time than anything else in the process.

Broadly speaking, the merit function can be described as a combination or function of calculated characteristics called "operands" or "targets," which is intended to completely describe, with a single number, the value and/or quality of a given lens design. This is obviously an exceedingly difficult thing to do. The typical merit function is the sum of the squares of many image defects; usually these image defects are evaluated for three locations in the field of view (unless the system covers a very large or a very small angular field). The squares of the defects are used so that a negative value of one defect does not offset a positive value of some other defect, and a large defect is doubly weighted by its size.

The defects may be of many different kinds; usually most are related to the quality of the image; however, any characteristic that can be calculated may be assigned a target value and its departure from that target regarded as a defect. Some rudimentary programs use only the third-order (Seidel) aberrations; these provide a rapid and efficient way of *adjusting* a design. These cannot be regarded as optimizing the image quality, but they do work well in *correcting* ordinary lenses. Another type of merit function traces a large number of rays from an object point. The radial distance of the image plane intersection of the ray from the centroid of all the ray intersections is then the image defect. Thus the merit function is effectively the sum of the root-mean-square (rms) spot sizes for several field angles. This type of merit function, while inefficient in that it requires many rays to be traced, has the advantage that it is both versatile and in some ways relatively foolproof. Some merit functions calculate the values of the classical aberrations, and convert (or weight)

them into their equivalent wavefront deformations. (See Chap. 24 for the conversion factors for several common aberrations.) This approach is very efficient as regards computing time, but requires careful and expert design of the merit function. Still another type of merit function uses the variance of the wavefront to define the defect items. The merit function used in the various David Grey programs is of this type, and is certainly one of the best commercially available merit functions in producing a good balance of the aberrations.

Characteristics that do not relate to image quality can also be controlled by the lens design program. Specific construction parameters, such as radii, thicknesses, and spaces, as well as focal length, working distance, magnification, numerical aperture, required clear apertures, and the like, can be controlled. Some programs include such items in the merit function along with the image defects. There are two drawbacks that somewhat offset the neat simplicity of this approach. One is that if the first-order characteristics that are targeted are not initially close to the target values, the program may correct the image aberrations without controlling these first-order characteristics; the result may be, for example, a well-corrected lens with the wrong focal length or numerical aperture. The program often finds this to be a local optimum and is unable to move away from it. The other drawback is that the inclusion of these items in the merit function has the effect of slowing the process of improving the image quality. An alternative approach is to use a system of constraints outside the merit function. Note also that many of these items can be controlled by features that are included in almost all programs, namely *angle-solves* and *height-solves*. These algebraically solve for a radius or space to produce a desired paraxial ray slope or height.

In any case, the merit function is a summation of suitably weighted defect items, which is hoped to describe in a single number the worth of the system. The smaller the value of the merit function, the better the lens. The numerical value of the merit function depends on the construction of the optical system; it is a function of the construction parameters, which are designated as variables. Without getting into the details of the mathematics involved, we can realize that the merit function is an $n$-dimensional space, where $n$ is the number of the variable constructional parameters in the optical system. The task of the design program is to find a location in this space (i.e., a lens prescription or a solution vector) that minimizes the size of the merit function. In general, for a lens of reasonable complexity there will a very large number of such locations in a typical merit function space. The automatic design program will simply drive the lens design to the nearest minimum in the merit function.

Optimally, the targets (or operands) in the merit function should be weighted so that all have a similar magnitude. This will avoid an implicit

heavier weighting of items that are dimensionally larger. For example, the *optical path difference* (OPD) in millimeters would not be as significantly weighted as the OPD in waves, and the transverse measure of an aberration would be less significant than the longitudinal measure. One way of handling this, where appropriate, is to convert everything to OPD in wavelengths.

Note that a merit function based on the third- and fifth-order aberration coefficients will be very robust. This is because no trigonometric rays are traced; the paraxial rays used to calculate the third- and fifth-order contributions are not subject to failure due to a ray missing a surface or encountering total internal reflection (no matter how unrealistic the calculated paraxial ray path may be).

A minimum ray-set at each field point for an aberration merit function would include the following:

Principal ray (or axis)

Upper margin ray (vignetted)

Lower margin ray (vignetted)

Sagittal ray

The merit function can have specific targets for spherical, coma, and so on—these can be the complete merit function, or they can be added to the default merit function to shift the design emphasis to a problem area. When adding to the merit function, be sure not to duplicate any existing operands; this may lead to an ill-conditioned matrix.

A merit function *samples* the aperture, field, and spectrum. The sampling is often minimal in order to save time and simplify matters. But remember that high speed, or wide-angle systems, may need a finer sampling than an ordinary system does. The more rays used in the merit function, the better the correlation with performance, but the slower the calculation.

A system with a small aperture and/or a narrow field of view will need fewer rays and field points. For example, it is common to use three field points (including the axis); two are sufficient for a narrow field, but four or five may be necessary for a wide field.

A typical aberration merit function may include the following operands:

- First order: *Effective focal length* (EFL), magnification, total track length, image height, angle solves, height solves, and lens vertex length. Avoid conflicts from multiple specifications.
- Aberrations: Marginal spherical, coma, astigmatism and Petzval field curvature, distortion, axial chromatic, lateral color, chromatic variations, and secondary spectrum.

- Zonal aberrations: Rays at 0.7 of the aperture can be used for zonal aberrations. The third-order aberration sums can also be used to reduce zonal residuals if the marginal correction is held, as can the Conrady *D-d, offence against the sine condition* (OSC), and aperture coefficient of coma (S2T) operands.

Another possible approach is to begin the optimization with an all third-order merit function (or one with both third and fifth), and then progress to one that uses the actual aberrations.

A feature common to many programs is to calculate the rms spot size (radius) or the rms OPD using a form of quadrature. This allows these values to be calculated quite accurately using only a few rays, which are spaced and weighted according to the quadrature rules (as are the field points). The Gaussian quadrature calculated value of the function will be exact if twice the number of rays minus one exceeds the order of the function that defines the ray pattern. Thus five rays along a radius can accurately evaluate a ninth-order function, e.g., spherical up to the ninth order.

Usually the rays are specified by the intersection of rings and radial spokes in the aperture. In general, the more rays, the more accurate the calculation and the use of twice as many spokes as rings is customary. For example, six spokes and three rings is a common default.

When the operands are rms spot size, the program tends to produce the smallest possible spot; when rms OPD is used, the tendency is to produce a tight, hard core with flare or wings. Since *modulation transfer function* (MTF) is a function of the variance of the OPD, an rms OPD merit function often produces a better MTF than rms spot merit function. Note that a minimum rms spot is not the same as correcting the marginal spherical to zero. If you want $TA_m$ equal to zero, rms spot won't do it; it tends toward undercorrected spherical accompanied by negative defocussing. To get $TA_m = 0.0$, use a target on the axial marginal ray.

Targets for first-order characteristics such as focal length, back focus, object to image distance, and magnification *may* be detrimental to the least-squares process. It may be beneficial to separate performance from such requirements. Consider using solves, Lagrange multiplier constraints, or even ad hoc intervention by the designer.

Systems that are not axially symmetric require more field points than the customary three. For example, a plane symmetric anamorphic system will need horizontal, vertical, and diagonal fields because the aberrations in each are different.

It is almost always necessary to tailor the design program's default merit function to suit the needs of the project at hand.

*Solves* available in almost all programs include:

1. The *angle solve*. This calculates (and adjusts) the curvature of a surface so as to produce a desired paraxial ray slope following the surface.

It can be applied to the axial ray or the chief ray. It is widely used to control focal length (or magnification) via the axial ray, and afocal magnification via the chief ray.

2. The *height solve*. This calculates (and adjusts) the distance to the next surface, which will produce the targeted ray height at that surface. It can be used on the chief ray to locate the pupil or on the axial ray to locate the paraxial focus position. A defocus from the paraxial focus is a much more stable variable than using the distance to the image surface as a variable.

3. The *pick up*. This will set the current curvature or thickness equal to a previous radius or thickness, with the same or a reversed sign. This is useful to maintain front-to-back symmetry or to use a surface twice, as in a mirror system.

Most programs also will allow a constant quantity to be added to the solved value.

Solves reduce the number of variables; this is good because it reduces the complexity of the solution. An angle solve to control focal length is exact only if the object distance is infinite; current programs use about $10^{20}$ units for the object distance. Most programs have a separate calculation for the exact value of the focal length, based on the data of two paraxial ray traces. If the tiny error in the focal length (approximately equal to $f/d^2$) because of the finite object distance is significant, this exact value can be used as a heavily weighted operand for the last step in the design process.

Some items that are commonly available for use in the merit function as operands or targets are as follows:

- Almost every constructional value, such as curvature, thickness, index, and aspheric coefficient
- Paraxial quantities, i.e., ray slopes and heights
- Aberration coefficients—chromatic, third- and fifth-order
- Ray trace values—intersections, angles (direction cosines and tangents), path length, field curvature ($X_s$ and $X_t$), derivatives of height versus pupil or versus object shift, displacement of a ray relative to a reference ray, OPD, Conrady *D-d*, OSC, S2T, and MTF
- Mathematical combinations of previously defined operands: +, –, ×, ÷, **, ave., and rms

The designer should occasionally stop and consider just what each item in the merit function is actually doing, and how it relates to the other operands. Be sure that something in the merit function, in the constraints, in the variable bounds, or in the solves is not inadvertently limiting the progress toward a solution.

A system that is nominally intended for a zero field size (i.e., to be used on axis only) should be corrected for coma if at all possible. This is a good precaution against the possibility that in use, a misalignment will actually cause the system to be used off axis.

A good approach is to take a default merit function and then customize it. After adjusting, enlarging, customizing, reweighting and/or rebalancing the merit/function, you have created a merit function for your system, which is presumably capable of directing the program to a good solution. At this point you can

a. Start over again
b. Start over from a different starting design
c. Start over with a different design form (e.g., Tessar versus triplet)
d. Go "global"

Targets/operands should be weighted to be of about equal magnitude. For example, a given amount of spherical aberration expressed as a longitudinal aberration is a bigger number than when expressed as a transverse aberration, which is a bigger number than when expressed as an OPD. The relationship expressed as an equation is: LA = TA/NA = $16 \cdot \text{OPD}/\text{NA}^2$. For an *f*/3 lens, NA = 0.1667 and LA = $6 \cdot$ TA = $576 \cdot$ OPD. Probably the simplest way to handle this in an aberration merit function is to convert the operands to an OPD expressed in units of the wavelength. See the equations in Chap. 24 for the conversions.

For a photographic application the corners should have a low weight; the objects in the corners are usually not of great interest and the fraction of the total area of the picture that is in the corner is a small one.

Don't try to control or fully correct an uncorrectable aberration, as for example, the astigmatism of a thin lens at the stop, which is totally independent of the lens shape or index.

Ideally, the merit function should correlate well with MTF (provided that a good MTF defines the desired state of correction). A merit function with this property is difficult to achieve. Some programs have the ability to use the MTF as an operand in the merit function, but this is slow at best, and unless the current design is close to the optimum, the process can be extremely volatile because the MTF is a very nonlinear function of the variables. An MTF merit function works well for a final balancing and adjustment of the design.

If non-image-quality items (e.g., EFL, back focal length (BFL), distances, thicknesses, and edge thicknesses) are put in the merit function as constraints (i.e., Lagrangian multipliers), this will eliminate any possible compromise between them and image quality.

An rms OPD merit function will usually produce a better MTF. The image is likely to have a good tight core with a flare or wings. A similar

correction can be obtained by heavily weighting rays in the center of the aperture. An rms spot size merit function will usually not produce as good an MTF, but it will have less flare and a more compact total spot.

## 2.3 Local Minima

Figure 2.1 shows a contour map of a hypothetical two-variable merit function, with three significant local minima at points *A*, *B*, and *C*; there are also three other minima at *D*, *E*, and *F*. It is immediately apparent that if we begin an optimization at point *Z*, the minimum at point *B* is the only one the routine can find. A start at *Y* on the ridge at the lower left will go to the minimum at *C;* however, a start at *X*, which is only a short distance away from *Y*, will find the best minimum of the three, at point *A*. If we had even a vague knowledge of the topography of the merit function, we could easily choose a starting point anywhere in the lower right quadrant of the map, which would guarantee finding point *A*. Note

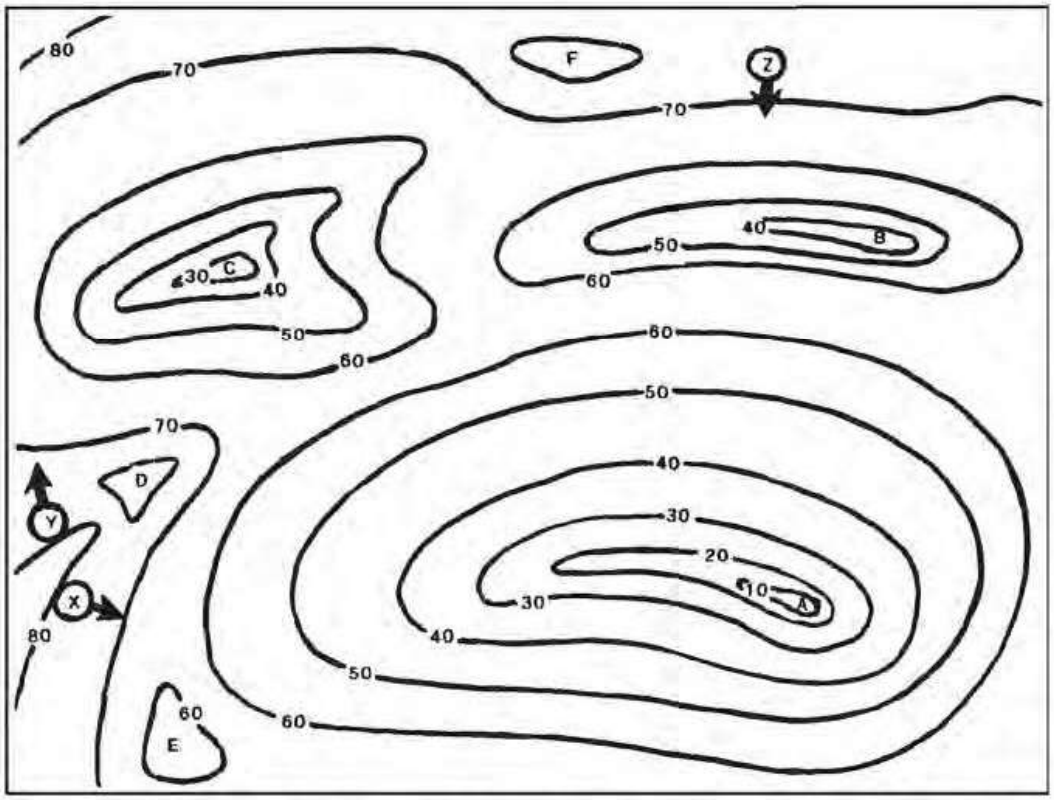



**Figure 2.1** Topography of a hypothetical two-variable merit function, with three significant minima (*A, B, C*) and three trivial minima (*D, E, F*). The minimum to which a design program will go depends on the point at which the optimization process is started. Starting points *X, Y,* and *Z* each lead to a different design minimum; other starting points can lead to one of the trivial minima.

also that a modest change in any of the three starting points could cause the program to stagnate in one of the trivial minima at *D*, *E*, or *F*. It is this sort of minimum from which one can escape by jolting the design, as described below.

The fact that the automatic design program is severely limited and can find only the nearest optimum emphasizes the need for a knowledge of lens designs, in order that one can select a starting design form that is close to a good optimum. This is the only way that an automatic program can systematically find a good design. If the program is started out near a poor local optimum, the result is a poor design.

The mathematics of the damped least-squares solution involves the inversion of a matrix. In spite of the damping action, the process can be slowed or aborted by either of the following conditions: (*1*) A variable that does not change (or that produces only a very small change in) the merit function items. (*2*) Two variables that have the same, nearly the same, or scaled effects on the items of the merit function. Fortunately, these conditions are rarely met exactly, and they can be easily avoided.

If the program settles into an unsatisfactory optimum (such as those at *D*, *E*, and *F* in Fig. 2.1) it can often be jolted out of it by manually introducing a significant change in one or more parameters. The trick is to make a change that is in the direction of a better design form. (Again, a knowledge of lens designs is helpful.) Sometimes simply freezing a variable to a desirable form or changing the weights in the merit function can be sufficient to force a move into a better neighborhood. The difficulty is that too big a change may cause rays to miss surfaces or to encounter total internal reflection, and the optimization process may break down. Conversely, too small a change may not be sufficient to allow the design to escape from a poor local optimum. Also, one should remember that if the program is one that adjusts (optimizes) the damping factor, the factor is usually made quite small near an optimum because the program is taking small steps and the situation looks quite linear; after the system is jolted, it is probably in a highly nonlinear region and a big damping factor may be needed to prevent a breakdown. A manual increase of the damping factor can often avoid this problem.

Another often-encountered problem is a design that persists in moving to an obviously undesirable form (when you *know* that there is a much better, very different one—the one that you really want). Freezing the form of one part of the lens for a few cycles of optimization will often allow the rest of the lens to settle into the neighborhood of the desired optimum. For example, if one were to try to convert a Cooke triplet into a split-front-crown form, the process might produce either a form like the original triplet with a narrow airspaced crack in the front

crown or a form with rather wild meniscus elements. A technique that will usually avoid these unfortunate local optima in this particular case is to freeze the front element to a plano-convex form by fixing the second surface to a plane for a few cycles of optimization. Again, one must know which lens forms are the good ones.

## 2.4 The Landscape Lens

As a real example to illustrate a simple merit function with two local optima, and to demonstrate the path of a typical damped least-squares optimization, we take a look at the *landscape lens*, which was probably the first real camera lens, i.e., a lens that covered a reasonably wide field. Its initial use was in the *camera obscura*, basically a dark room wherein the lens formed an image of the outside landscape on a wall or table for sketching, painting, or simply viewing.

The optical system of the landscape lens comprises a single positive element and a separate aperture stop. The stop is placed at the *natural* stop position where (assuming the lens has undercorrected third-order spherical aberration) the coma is zero and the tangential field is as backward curving (or the least inward curving) as possible. By bending the lens, one can determine the lens shape for which the field curvature is minimized when the coma is zero. There are two solutions: the lens is meniscus shaped in both, and the stop is located about 10 or 20 percent of the focal length away from the concave side of the lens. In the *front meniscus* the lens is convex toward the distant object and the stop is between the lens and the image; in the *rear meniscus* the lens is concave to the object and the stop is out in front of the lens.

To illustrate the effect of shifting the aperture stop in a lens with simple, undercorrected spherical aberration, Fig. 2.2 shows a ray intercept plot (i.e., an $H$-tan $U$ curve) with three positions for the stop indicated by the circled numbers. (Note that shifting the portion of the plot under consideration from left to right is equivalent to moving the stop from right to left, because by convention the intersection points of the upper rays of a beam are plotted at the right.) In the plot, when a straight line is drawn connecting the intercepts for the upper and lower rays of a beam, the sag of the plot equals the tangential coma, and the slope of the line ($\delta H/\delta$ tan U) equals the tangential field curvature, $X_t$. Thus for stop position no. 1, we see overcorrected (positive) coma and an inward-curving field. Moving the stop to position no. 2 corrects the coma to zero, and the field is now backward curving. By inspection it is quite apparent that this is the stop location at which the field is most backward curving; this is called the natural stop position. At stop location no. 3 we once again have an inward-curving field and coma, but with the stop in this position the coma is undercorrected (negative).

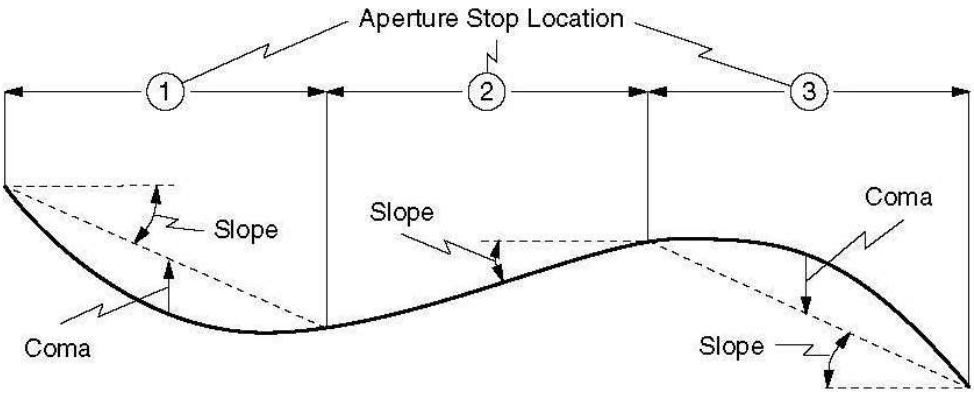



**Figure 2.2** A ray intercept plot of off-axis, undercorrected, third-order spherical, illustrating how moving the stop will sample different portions of the plot. The different sections show the change of coma and field curvature as the stop is shifted. It also indicates that the stop position with zero coma is also the stop position that produces the most backward-curving field. This is the natural stop position.

The landscape lens is an excellent example with which to demonstrate the existence of multiple optima in lens design. Figure 2.3 is a contour map of an optical software default rms-spot-size merit function, plotted against the lens shape (cv 1) and the stop location. The data are for a BK7 lens of 100-mm focal length at a speed of $f/16.7$ (NA = 0.03) and a half-field of 20°. The evaluation is monochromatic.

The two local minima in the merit function are immediately apparent. The paths of two *damped least-squares* (DLS) optimizations are shown by the sequential circled numbers. For both paths, the starting lens shape was plano-convex with the aperture stop placed at the plano surface. When the starting lens was oriented with the convex surface facing the distant object, the DLS routine found the front meniscus form. When the plano side faced the object, the rear meniscus form was the result. Although the optimization paths appear almost random, note that each step in the process did reduce the merit function. Interestingly, from any starting point in the upper-right or lower-left quadrants of the plot, after the first few steps the optimization always followed a path very similar to these.

This is one of the very few lenses that can be designed analytically, using the third-order aberration contribution equations (either the surface- or the thin lens-contribution equations—see Chap. 24).

As can be seen from the MTF and ray intercept plots in Figs. 2.4 to 2.6, the rear meniscus is opticaly the better of the two. Its monochromatic image quality is 25 or 30 percent better than the front meniscus, its surface curvatures are weaker, and its distortion is barrel (and thus somewhat less objectionable to the eye than the pincushion of the front meniscus). In spite of all this, the front

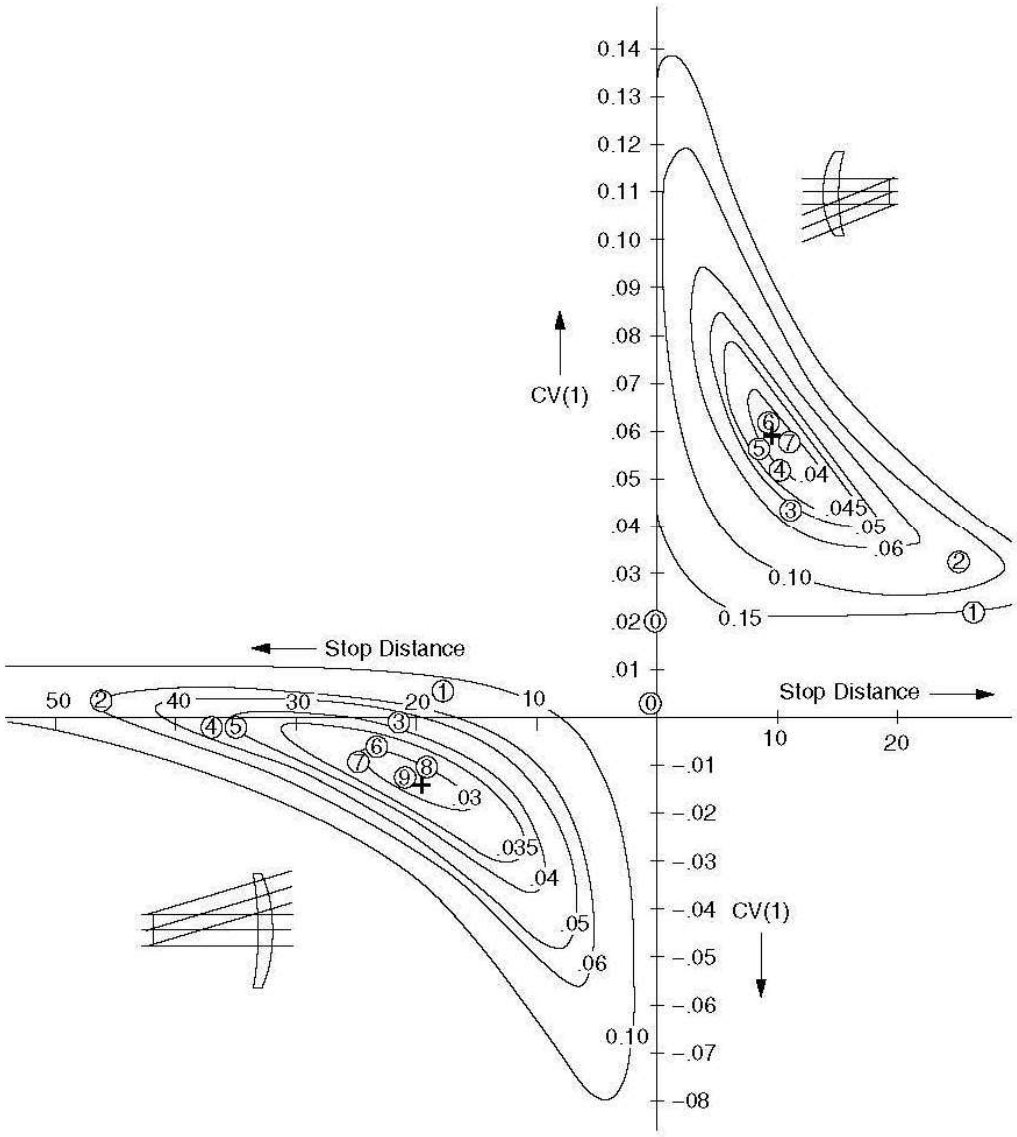
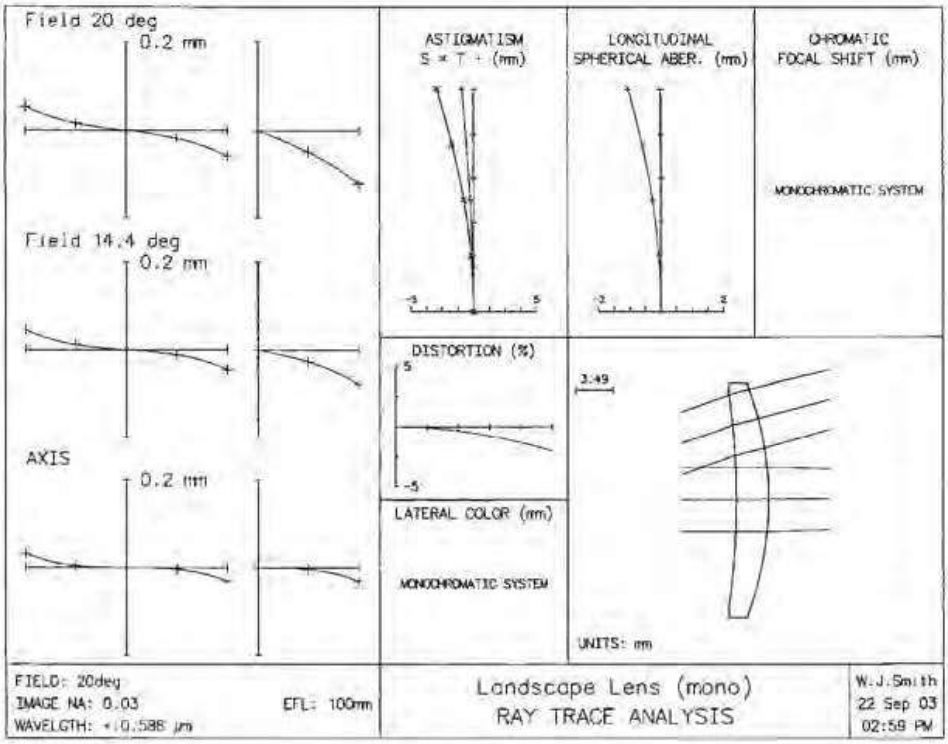



**Figure 2.3** A contour plot of the merit function (rms spot size) as a function of the lens shape [cv (1)] and the stop position. The encircled numbers show the progress of the steps in the optimization process, starting from a plano-convex lens with the stop in contact with the plane surface.

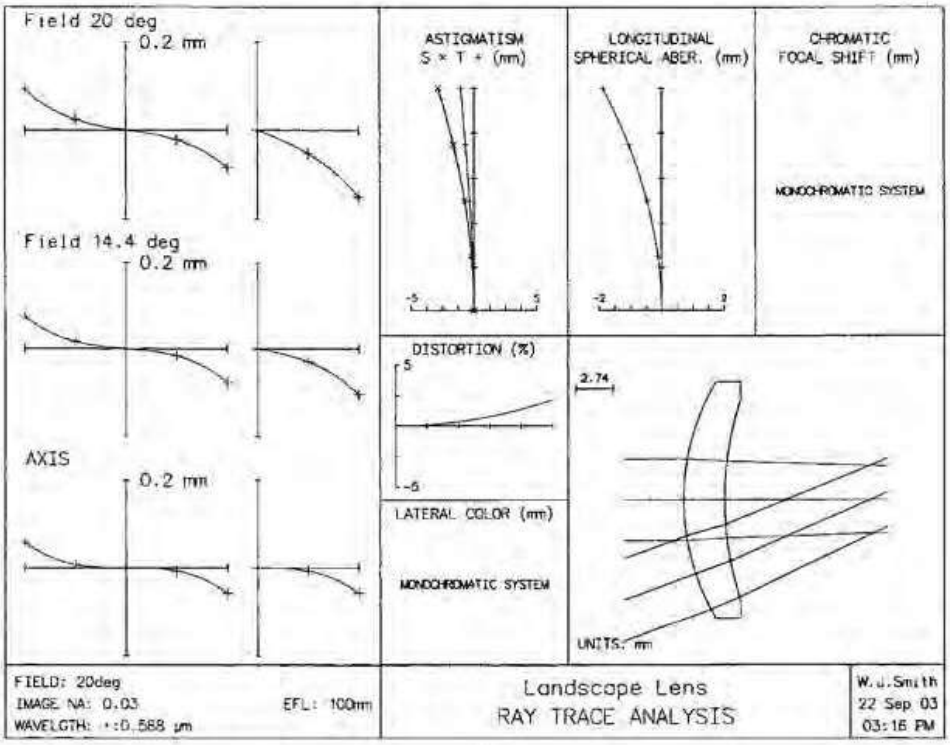
meniscus has been the lens of choice for inexpensive cameras for 50 or 60 years.

The reasons for prefering the front meniscus form are threefold:

1. The overall length of the camera with the front meniscus is about 20 percent shorter than with the rear meniscus. (In the rear meniscus the length of the camera is determined by the lens focal length plus the distance to the stop.)
2. The shutter mechanism is protected and kept dirt-free because it's behind the front meniscus lens.

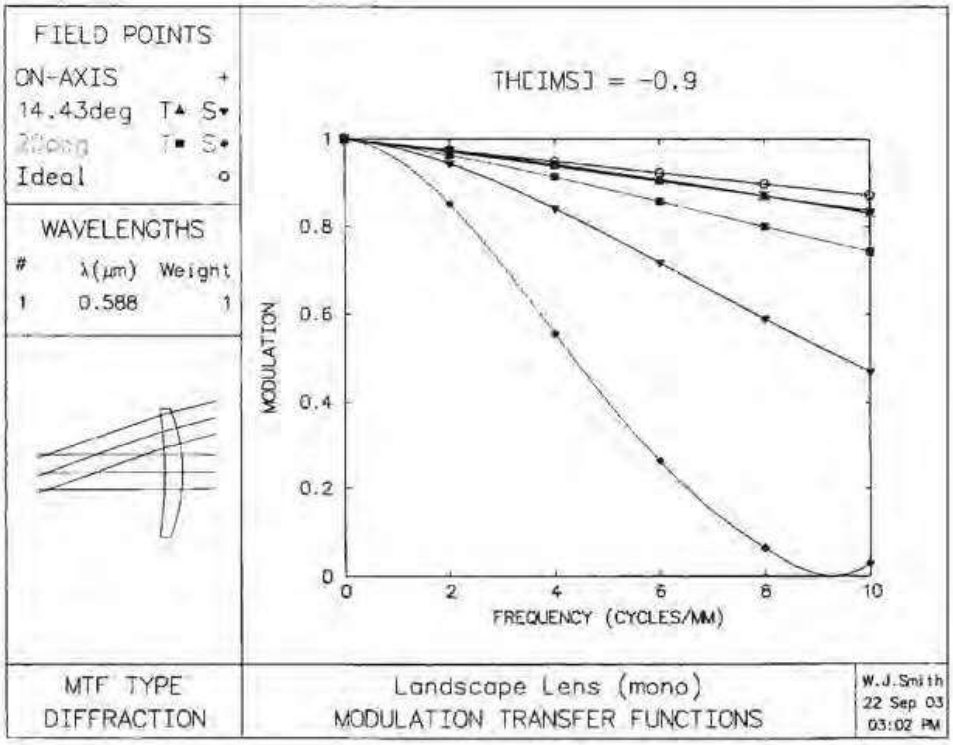



(*a*)

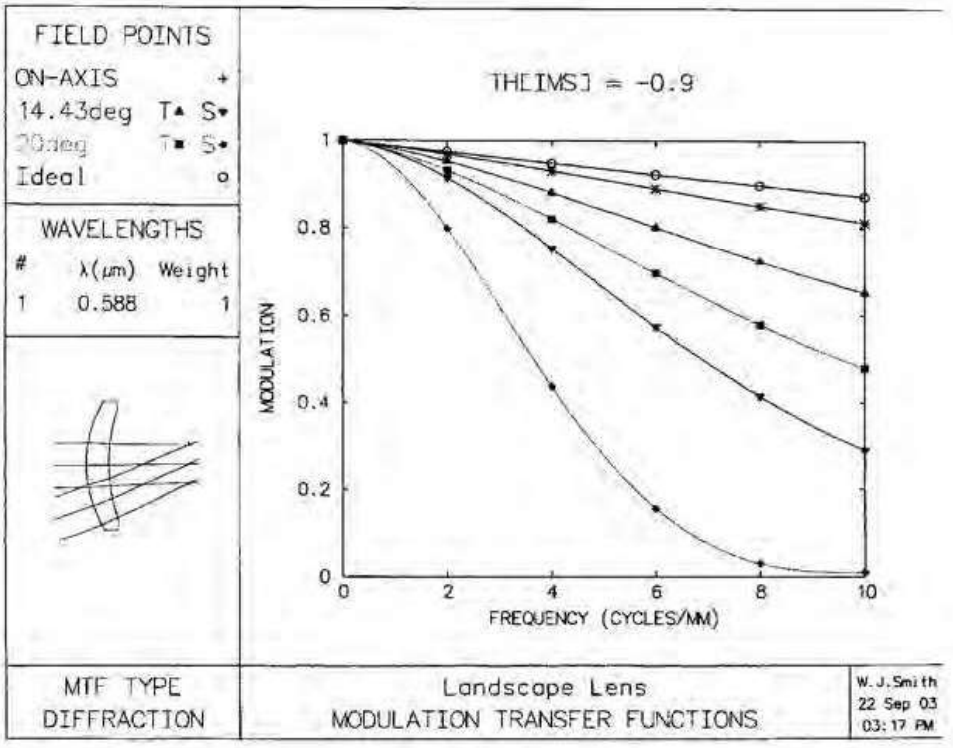



(*b*)

**Figure 2.4** (*a*) The aberration plots for the rear meniscus solution, and (*b*) for the front meniscus solution.

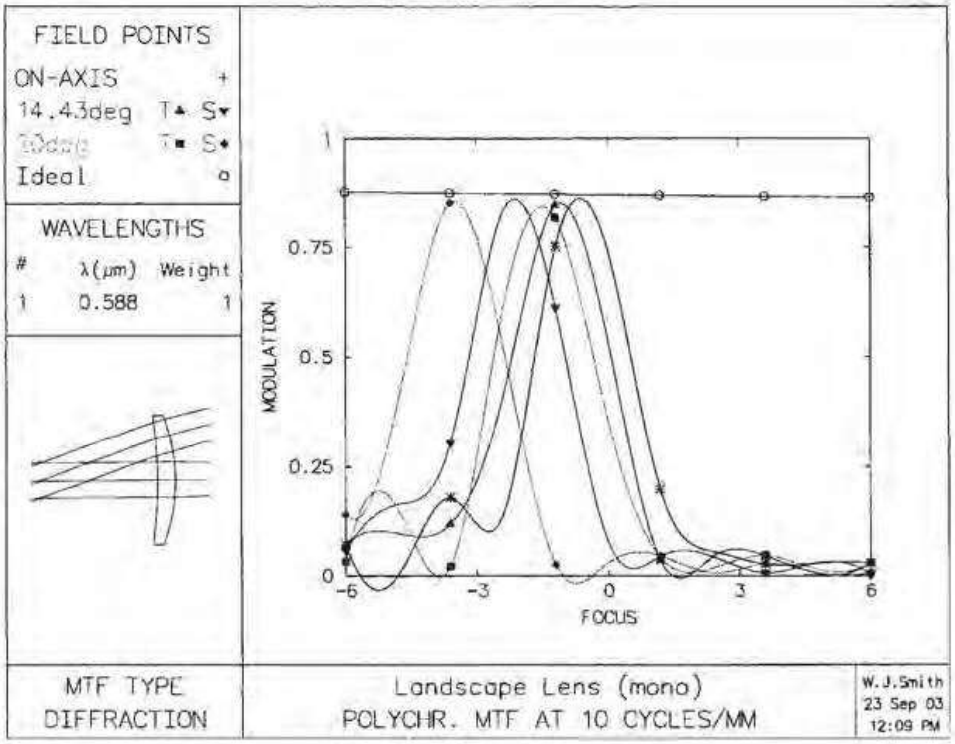



(*a*)

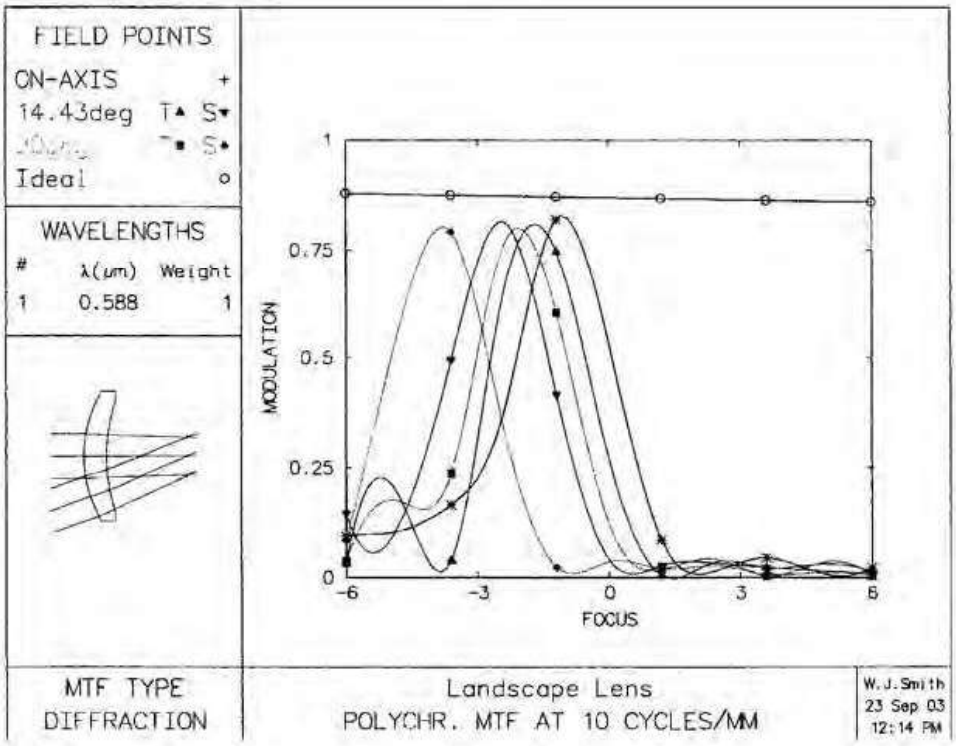



(*b*)

**Figure 2.5** (*a*) The MTF for the rear meniscus solution, and (*b*) the MTF for the front meniscus solution. Both are defocused 0.9 mm from the paraxial focus.




(*a*)




(*b*)

**Figure 2.6** (*a*) The through-focus MTF plot at 10 lpm for the rear meniscus, and (*b*) for the front meniscus.

3. The customer sees the lens, which is larger and far more appealing than the shutter mechanism and the small aperture stop.

And in addition, since inexpensive camera lenses are injection molded out of plastic, the stronger surface curvatures of the front meniscus are no more expensive to make than the flatter surfaces of the rear meniscus.

This design form is quite forgiving in that several different merit functions that we tried all yielded quite similar (but definitely different) lenses with similar levels of the primary aberrations; however, even in simple designs, there is often more than meets the eye. It regularly turns out that, even in simple systems, many forms exist that have a similar level of performance. One must look carefully for real-world considerations that may differentiate one design from another. In addition to the basic default (rms spot size) merit function evaluated at the paraxial focus and for a single wavelength, we also tried:

a. Allowing defocus

b. Simply optimizing for zero coma and a flat tangential field

c. Optimizing with third-order aberration contributions

d. Optimizing with third-order plus fifth-order aberrations

e. Including C- and F-light spot sizes

While each different merit function produced a different design, all the designs had very similar performance as regards the monochromatic Seidel aberrations. There was, however, a real difference between the optimum monochromatic designs and those produced when the merit function included chromatic aberration. In this lens lateral color has a significant and detrimental effect on the image quality, and the longer the distance from lens to stop, the larger the lateral chromatic. (This is one of the very few cases where a spot diagram—showing the lateral color—is actually of value to the design process.) Thus, when a polychromatic optimization was done, the solutions had stop distances that were much smaller than those of the monochromatic designs. This is of course a realistic consideration, since the polychromatic design will have a much better off-axis white light performance. Another real world consideration is the effect on costs resulting from the smaller diameter required when the stop distance is reduced.

Incidentally, the "descendants" of the landscape lens include the symmetrical doublet *periscopic*, the *hypergon*, the *topogon*, and the thick double meniscus anastigmats of a century ago (which are discussed in Chap. 11).

## 2.5 Types of Merit Functions

Many programs allow the user to define or modify the merit function. This can be a valuable feature because it is almost impossible to design a truly *universal* merit function. As an example, consider the design of a simple Fraunhofer telescope objective: a simple four-target merit function that controls the spherical and chromatic aberrations of the axial marginal ray and the coma of the oblique ray bundle (plus the focal length) is all that is necessary. If the design complexity is increased by allowing the airspace to vary and/or adding another element, the merit function may then profitably include entries that will control zonal spherical, spherochromatism, and/or fifth-order linear coma. But as long as the lens is thin and in contact with the aperture stop, it would be foolish to include in the merit function entries to control field curvature and astigmatism. There is simply no way that a thin stop-in-contact lens can have any control over the inherent large negative astigmatism; the presence of a target for this aberration in the merit function will simply slow down the solution process. It would be ridiculous to use a merit function of the type required for a photographic objective to design an ordinary telescope objective. (Indeed, an attempt to correct the field curvature may lead to a compromise design with a severely undercorrected axial spherical aberration, which, in combination with coma, may fool the computer program into thinking that it has found a useful optimum.)

There are many design tasks in this category, where the requirements are effectively limited in number and a simple, equally limited merit function is clearly the best choice. In such cases, it is usually obvious that some specific state of correction will yield the best results; there is no need to *balance* the correction of one aberration against another.

More often, however, the situation is not so simple; compromises and balances are required and a more complex, suitably weighted merit function is necessary. This can be a delicate and somewhat tricky matter. For example, in the design of a lens with a significant aperture and field, there is almost always a (poor) local optimum in which (1) the spherical aberration is left quite undercorrected, (2) a compromise focus is chosen well inside the paraxial focus, (3) the Petzval field is made inward-curving, and (4) overcorrected oblique spherical aberration is introduced to balance the design. A program that relies on the rms spot radius for its merit function is sometimes likely to fall into this trap. A better design usually results if the spherical (both axial and oblique) aberrations are corrected, the Petzval curvature is reduced, and a small amount of overcorrected astigmatism is introduced. When one recognizes this sort of situation, it is a simple matter to adjust the weighting of the appropriate targets in the merit function to force the design into a form with the type of aberration

balance that is desired. Another way to avoid this particular problem is to force the system to be evaluated/designed at the paraxial focus rather than at a compromise focus, i.e., to not allow (or to penalize) defocusing. As can be seen, the design of a general-purpose merit function that will optimally balance a wide variety of applications is not a simple matter.

Although it is not always necessary, there are occasions when it is helpful to begin the design process by controlling only the first-order properties (image size, image location, spatial limitations, and the like). Then one proceeds to control the chromatic and perhaps the Petzval aberrations. (Things may even go better if the first-order and the chromatic are fairly completely worked out by hand calculation before submitting the system to an automatic design process.) The next step in the sequence is to correct the primary aberrations (spherical, coma, astigmatism, and distortion), either directly or by using the Seidel coefficients, and finally proceed to balancing and correcting the higher-order residuals. This sort of ordered approach is sometimes useful (or even necessary) when one is exploring terra incognita, and, of course, it requires a user-defined merit function if it is to be implemented.

## 2.6 Stagnation

Sometimes the automatic design process will stagnate and the convergence toward a solution will become so slow as to be imperceptible. This can result from being in a very flat and broad optimum in the merit function. It can also result from an ill-designed merit function. Often first-order properties specified in the merit function are the cause of the problem. It is only too easy to require contradictory or redundant characteristics. This is especially true for zoom lenses or multiconfiguration systems, which can be confusingly complex. When stagnation occurs, or convergence is slower than you feel it should be, it is wise to stop and examine the merit function for problems. Look critically at every item in the merit function and consider what it is intended to be doing and what it *actually* does. Eliminate redundancies and try to make each entry in the merit function explicitly control its intended characteristic.

Another condition that may cause stagnation exists when the aberrations (or operands) of the initial design are extremely large. The small incremental changes in the variables used in the optimization process may produce changes in the merit function value, which are so tiny (relative to the large aberrations) that they are completely lost in the numerical noise of the computations. In this case a few manual changes to the initial design may reduce the merit function to a more reasonable level.

## 2.7 Generalized Simulated Annealing

The discussions above have centered on the standard DLS program or its equivalent. There have been several versions of *random search* programs proposed in the past. The most recent of these is quite sophisticated and is called *generalized simulated annealing*. In this, the computer *randomly* selects the lens dimensions (within a limited range and according to some probability distribution) and evaluates the resulting lens prescription. If the new version is better than the old, it is unconditionally accepted. If it is worse, it *may* be accepted, on the basis of random chance, weighted by a probability function, which reduces the chance of acceptance in proportion to the amount that the lens is worse than the original form. This sort of approach obviously allows the program an easy escape from the local minima described with Fig. 2.1, but it equally obviously requires a very large number of trials before a random chance can find a good combination of dimensions for the lens. Nonetheless, it does work, but not rapidly. As computers increase in speed, a program of this sort may displace or supplement the DLS as the routine of choice for automatic lens design.

In determining the acceptance of a new design that is worse than its predecessor, we define:

$$Z = \exp\left[\frac{-B \cdot \delta\phi}{(\phi_0 - \phi_{\text{best}})}\right]$$

where $\phi$ = value of the merit function
$B$ = constant (e.g., $\approx 3.5$)
$\delta\phi$ = increase in $\phi$
$\phi_0$ = $\phi$ before the changes
$\phi_{\text{best}}$ = best possible value of $\phi$

We randomly generate a number $\Gamma$ between 0 and 1. Then the new design is accepted if $Z > \Gamma$. Thus $Z$ approximates the chance of acceptance, and if:

$$\begin{aligned} \delta\phi/(\phi_0 - \phi_{\text{best}}) &= 10.0 \quad \text{then} \quad Z \approx 10^{-15} \\ &= 1.0 \qquad\quad Z \approx 0.03 \\ &= 0.1 \qquad\quad Z \approx 0.7 \\ &= 0.01 \qquad\quad Z \approx 0.97 \end{aligned}$$

These programs are often called *global optimization*. But consider that one can only say that such a solution is the optimum within a defined search area, after a certain number of hours of search.

Generalized simulated annealing in any of its forms is based on a specific merit function, a limited region of search, and a limited amount of search time. Given enough time, the probability is high that the global optimum within the search volume will be found.

## 2.8 Considerations about Variables for Optimization

The potential variables for use in optimization include: the surface curvatures, conic constants, and asphericities; the surface spacings; and the refractive characteristics of the materials involved. Occasionally tilts and decentrations are also included as variables.

### 2.8.1 Bounds on variables

There are several forms for bounds. The program may have a table (spreadsheet) of maximum and minimum values for the variables. There are different ways that these may be applied, i.e., soft versus hard versus elasitic versus rebound. *Constraints* are handled by Lagrangian multipliers and are solved for (as exactly as the ever-present nonlinearities allow) with a priority over the minimization operands in the merit function. A variable bound may be included in the merit function as a least-squares target, as an inequality (< or >), or it may be channeled (< and >). A merit function bound on a variable allows the variable to penetrate the boundary a little, hopefully temporarily. The designer's choice of bounds, constraints, operands, and weights may affect the direction and progress toward a design solution, and may define a different local optimum.

### 2.8.2 Materials

Although the material characteristics are not continuous variables, for optical glasses at least the index and dispersion (or $V$ value) can be varied within the boundaries of the glass map (Fig. 2.7) as if they were. The real glass nearest the optimized values can be substituted for the optimized glass to achieve nearly the same resultant design after another cycle or two of optimization with the real glass. Note that this is *not* true for partial dispersions, since there are relatively few glasses with partial dispersions unusual enough to be useful in the correction of secondary spectrum. Obviously, for applications outside the spectral regions where optical glass is usable, one cannot treat the refractive characteristics as variables, since the available materials tend to be few and far between.

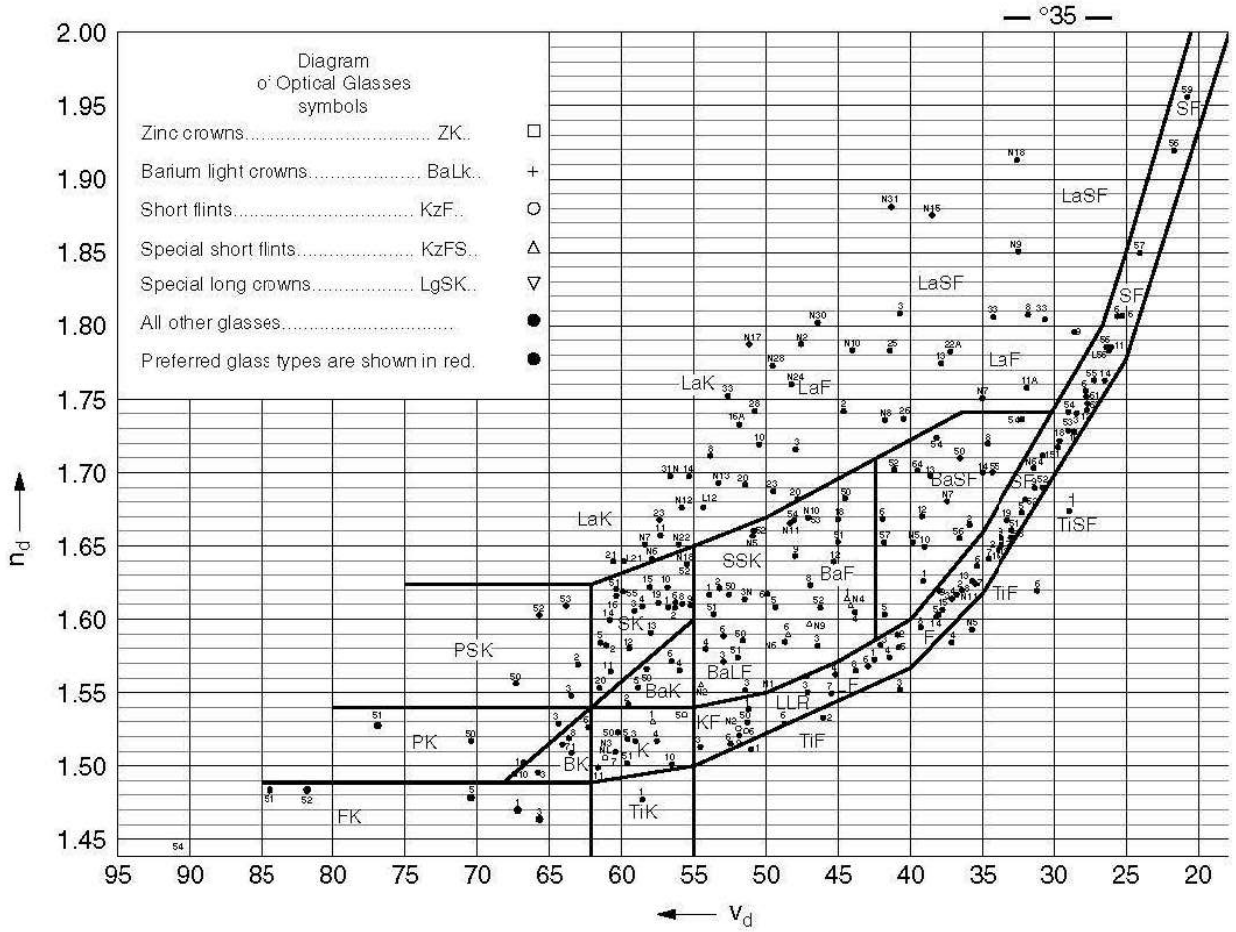



**Figure 2.7** The glass map or "glass veil." Index ($n_d$) plotted against reciprocal relative dispersion (Abbe $V$ value). The glass types are indicated by the letters in each area. The glass line is made up of the glasses of types K, KF, LLF, F, and SF, which are strung along the bottom of the veil. (Note that K stands for *kron*, German for crown, and S stands for *schwer*, or heavy or dense.) (*Courtesy of Schott Glass Technologies, Inc., Duryea, PA.*)

In many of the simpler types of designs it is absolutely essential to allow the glass characteristics to vary. In the Cooke triplet, for example, the relationship between the $V$-values of the crown and flint elements determines the overall length of the lens. As described in Sec. 8.3, the length of a triplet (and that of most airspaced anastigmats) determines the amount of higher-order spherical aberration and astigmatism; these in turn determine the aperture and field coverage capabilities of the lens. If the simpler types of lenses are to be fully optimized to suit the application at hand, the glass characteristics *must* be allowed to vary.

Some optimization programs have difficulty with the bounds of the glass map; if this is a problem, the optimization process is often facilitated by starting the variable glass well away from the boundary, so that it can find its best value before encountering the boundary problem.

It is often better to let the program vary the flint glasses rather than to vary the crowns. This is because the crowns usually tend to go toward the upper left corner (high index, high $V$-value) of the glass map. Flints head for the lower right corner, and are then, of course, constrained to lie on the glass line. The glasses along the glass line are numerous, inexpensive, and almost universally well behaved. On the other hand, the crown glasses in the upper left region of the map include in their number many that are expensive and/or easily attacked by the environment. Thus one might be willing to accept the computer's choice of a glass along the glass line, but would prefer to make a more discriminating selection from among the others.

Many programs allow the index to vary while holding the relative dispersion constant. For flint glasses, this amounts to varying the glass along the *glass line.* This is a very useful feature because this is often exactly the type of variation desired for flint glasses.

Be sure to bound the index and dispersion if they are allowed to vary; otherwise the index may well go to two or three, and the relative dispersion may go anywhere from more than one to less than zero.

### 2.8.3 Curvatures

In general, one would expect to want to make use of every available variable. This is almost always true regarding the curvatures, all of which, unless there is a reason to constrain the shape of an element, are usually allowed to vary. Unless the last surface is close to the image, the value of the last curvature can be "solved" to maintain $\mu'_k$; this controls the focal length or the magnification of the system.

### 2.8.4 Airspaces

Ordinarily, airspaces may be regarded in the same light as curvatures, since they are continuously variable and are very effective variables.

### 2.8.5 Defocusing

Although the distance by which the design image plane departs from the paraxial focus is usually an airspace and can be regarded as a variable, its effects can be insidious. If the image surface is allowed to depart from the paraxial focus from the beginning of the optimization process, an unfortunate lens may result. In some lenses, and with some optimization merit functions, the tendency is to produce a lens with:

1. The image plane well inside the paraxial focus
2. A large undercorrected spherical aberration
3. A strongly inward-curving field
4. A heavy overcorrecting oblique spherical

Although this combination occupies a local optimum in the merit function, this is often not the best state of correction. It fools the optimization program because the undercorrected spherical causes the best axial focus to lie to the left of the paraxial focus and the overcorrected oblique spherical causes the best off-axis focus to lie to the right of the inward-curving field; the net result is that, to the program, the field seems flat. One can usually avoid this pitfall by not allowing any defocus in the early stages of the optimization and/or putting a heavy penalty on the defocusing. Note that for *some* lenses, such as non-diffraction-limited systems, the correction described above may in fact be a good one. See the comments on aberration balance in Sec. 3.8.

Note that the defocus variable, as a departure from the paraxial focus (which can be set by a height solve), is much more stable than simply using the distance to the image surface as a variable. Including the defocus, suitably weighted, in the merit function will put a brake on a runaway defocus variable.

### 2.8.6 Thickness

Element thicknesses must be regarded quite differently than airspaces. They must of course be bounded by the necessity for a practical edge thickness for the positive elements and a reasonable center thickness for the negative elements. In many designs, element thickness is an insignificant and ineffective variable (and one whose effects are easily duplicated by an adjacent airspace). In this circumstance one can arbitrarily select a thickness on the basis of economy or ease of fabrication. The elements in such designs are typically quite thin; for example, consider a telescope objective or an ordinary Cooke triplet.

There are, however, many systems in which the element thickness is not only an effective variable, but one that is essential to the success of the design type. The older meniscus lenses (Protar, Dagor, and the like)

and the double-Gauss forms all depend on the separation of the concave and convex surfaces of their thick meniscus components to control the Petzval curvature and, in many instances, the higher-order aberrations as well. In lenses of this type it is absolutely essential that these glass thicknesses be allowed to vary.

One must be wary of, and skeptical toward, a thickness variable, which is very weakly effective. Occasionally an optimization program will produce a design with an overly thick element, where the large thickness produces only a very small improvement (which is not worth the added cost of producing the thick element). This occurs because the optimization routine will seek out any improvement that it can get, no matter how small, and without concern as to the cost. It is wise to test the value of a thick element if there is any doubt about its utility. This is readily accomplished with another optimization run, which fixes the thickness in question to a smaller value. Very often the performance of the thin version will not be noticeably different from that of the optimum thicker version. Although most significant with respect to lens thickness, this same rationale obviously applies to air-spaces as well. Another solution is to add the errant thickness to the merit function with a small or zero target.

### 2.8.7 Aspheric surfaces

Surface asphericity can be an extremely effective (if sometimes expensive) variable, but it is one that often requires a bit of finesse. On occasion, one may be ill-advised to begin an optimization with the conic constant and all the aspheric deformation coefficients used simultaneously as variables. The conic constant and the fourth-order deformation coefficient both affect the third-order aberrations in exactly the same way. Thus they are at least partially redundant. But more significantly, identical variables have an undesirable effect on the mathematics of the optimization process. It is often advisable to vary one or the other, but not both. A safe practice is to vary only the conic constant (or the fourth-order term) at first, and then add the higher-order terms (sixth, eighth, and tenth) one at a time, as necessary. The tenth-order term is, in many systems, totally unnecessary, adding little or nothing to the quality of the system; in fact, the eighth-order term is often something that can be done without.

A surface defined by a tenth-order polynomial can cause the spherical aberration to be corrected exactly to zero at four ray heights. If there are only four axial rays in the merit function, their ray intercept errors may all be brought to zero; the danger is that, between these rays, the residual aberration may be unacceptably large. A tenth-order surface can be a rather extreme shape. Thus the use of an aspheric surface sometimes calls for more rays in the merit function than one might otherwise expect to need. With a program that allows wavefront deformation or OPD targets in the merit function, the severity of this problem can be lessened somewhat.

Note that a conic constant is totally meaningless when applied to a plano surface. In order to aspherize a plane surface, a fourth-order deformation must be used. An aspheric located at the aperture stop can only (in the third order) affect the spherical aberration. If it is desired to affect coma, astigmatism, or distortion, the aspheric should be located well away from the stop. An aspheric does not affect Petzval or chromatic. An argument can be made that no more than three aspherics are necessary in a lens, one near the stop, plus one near each end of the lens. This is probably an oversimplification for complex lenses with significant higher-order aberrations. Note also that the third-order aberrations tend to be affected by the fourth-order surface deformations, the sixth-order deformations are effective against the fifth-order aberrations, and so on.

## 2.9 How to Increase the Speed or Field of a System and Avoid Ray Failure Problems

Very often, the lens designer is faced with the necessity of increasing the speed (i.e., relative aperture, numerical aperture, and the like) and/or the field of view of an existing optical system. There are two common reasons why this may be desirable. One may want to adapt an existing design (such as those in this book) to an application that requires a larger aperture or wider field than that for which the original lens has been configured. The other common situation is simply the creation of an entirely new system with a relatively large aperture and/or field. In either case the difficulty that can arise is that the rays that are needed to design the system may not be able to get through the initial lens prescription.

There are two reasons that a ray may not be able to get through. One reason is that the height of the ray at a surface may be greater than the radius of the surface; the ray simply misses the surface entirely and its path obviously cannot be calculated any further. The second reason is that the ray may encounter *total internal reflection* (TIR) in passing from a higher index to a lower; again, the ray path cannot be calculated. Each of these conditions represents a boundary that, if crossed, causes failure of the ray trace. Note also that as these boundaries are *approached*, the situation rapidly becomes very unstable. This is because close to the boundary, the angle of incidence (for the case of the ray height approaching the value of the radius) or the angle of refraction (for the case of the ray approaching TIR) is very near to 90°. Near this angle, Snell's law of refraction ($n \sin I = n' \sin I'$) becomes *very* nonlinear, producing a highly unstable situation, which often explodes as the lens construction parameters are incremented to calculate the partial derivative matrix in the course of the optimization.

A good way of dealing with this situation is simply to back off from the aperture and/or field requirement that is causing the problem. Most design programs have the capability to easily adjust or scale the aperture and field angle. If a change is made to smaller values of aperture or field, the rays will no longer be so near to the failure boundary. If the lens is now optimized, the program is very likely to adjust the lens parameters so as to reduce the angles of incidence, because this is usually a change that causes the aberrations to be reduced. The optimizing changes can thus be expected to pull the problem situation in the system further away from the ray failure boundary, *if this is possible.*

After the optimization has relaxed the problem, the field and/or the aperture can usually be adjusted to a moderately larger value without again encountering the failure boundary. Depending on just how sensitive the system is to the problem, an increase of about 10 to 50 percent may be appropriate. Now another cycle of optimization will strongly tend to again reduce the troublesome angles of incidence.

This process of adjusting the field or the aperture to larger values and then optimizing is continued until the desired aperture or field is attained without ray failures. This works well, *provided* that this desired result is possible for the lens configuration under study. It may be necessary to choose another configuration, usually one with more elements. When this is necessary, a drawing of the lens and rays (of the last design form that has successfully passed all the rays) will usually indicate which rays and which surfaces are causing the problem. One simply looks for angles of incidence or refraction that are large (and often near 90°). Then the offending element can be split into two (or more) elements shaped to reduce these angles. The scale-and-optimize process can now be repeated with a much improved chance of success. Note that the examination of the critical ray paths (typically those of the marginal rays) for large angles of incidence or refraction is a comple-tely general technique that will often indicate the source of a design problem.

## 2.10 Test Plate Fits, Melt Fits, Thickness Fits, and Reverse Aberration Fits

When the deleterious effects of fabrication tolerances become too large to bear, a technique commonly used to reduce these effects is to fit the lens design to the known values for the radius tooling and/or to the measured glass indices. The former is called a *test plate* (or *test glass* or *tooling*) fit; the latter is called a *melt fit.*

The *test plate fit* is begun by first obtaining a list of the available test plate radii from the shop that is scheduled to fabricate the lens. It is wise to ascertain that the radius values of the list are not just nominal values

but are based on accurate and recent measurements of the test plates, since there is often a significant difference between the two.

The fit is carried out as follows: A surface is selected at which to begin the process. This selection is based on one of the following criteria: (1) the surface most sensitive to change,* (2) the surface with the shortest radius, (3) the strongest surface (i.e., with the largest surface power $[n' - n]/r$), or (4) the surface that shows the largest curvature difference from the nearest available test plate radius. Very often most, if not all four, of these criteria will indicate the same surface; if not, the choice of which one to use is almost a matter of taste. The nearest radius on the test plate list is substituted for the selected surface and the lens is reoptimized, allowing all the variable parameters to change (except of course, the radius that has been set to the test plate value).

Another surface is then chosen and fixed to the nearest test plate radius; the lens is reoptimized again. This process is repeated until all the surfaces have been fitted to test plates. It is usually wise to avoid fitting both surfaces of a singlet (or all surfaces of a component) until all components have at least one fitted surface each. This allows the unfitted surface to vary and adjust for power, chromatic, and Petzval aberrations and the like for as long as possible.

With a reasonably complete test plate list, all radii can usually be fitted without significantly degrading the image quality (i.e., the merit function). In fact, the merit function is often slightly improved by this process because of the additional cycles of optimization that have been performed plus the small shocks or jolts the radius changes produce. If the test plate list is limited or has a gap in it, it may not be possible to fit all the design radii to test plates without degrading the performance. In such a case, one must either fabricate new tooling and test plates or seek a new vendor for the parts. One should *always* try for a 100 percent fit.

A *melt fit* reoptimizes the lens design by using measured data for the material indices instead of the nominal values from the glass catalog. This measured data comes in the form of what is called a *melt sheet* provided by the glass manufacturer or supplier. For noncritical applications, this data is usually sufficient. A worthwhile elaboration of the process is to determine the difference between the measured index values and the catalog values. These differences are then plotted against wavelength. This plot should be a smooth, relatively level line. A data point that does

*Note that the relative sensitivity of any dimension of the system can be determined quite easily by making an incremental change in the dimension and noting the effect it produces on the merit function. This, of course, assumes a merit function that represents the quality of the image reasonably well.

not plot smoothly is suspect; the measured data may well be in error. A slope of the line indicates a *V*-value difference from the catalogue values. A curvature of the line indicates a change in the partial dispersion. Next, a smoothly drawn curve through the points is used to determine improved values for the index differences. These differences are then applied to the catalog values to arrive at better values for the measured melt data, and the improved values are used in the melt fit reoptimization. The smoothed curve of differences can also be used to determine the index for wavelengths that are not included in the melt sheet.

An ordinary melt sheet will list the indices for the wavelengths of d, e, C, F, and g light. These are usually not individually measured. Instead, the index is measured for d light and the index difference between C and F light is measured. These two measurements are fed into a computer program, which uses the known characteristics of the glass type to calculate the indices for e, C, F, and g light. This, while not ideal, is adequate for many, even most, applications; however, for some critical applications, and for all applications in which an attempt has been made to reduce or correct the secondary spectrum, it may be quite unsatisfactory. For an additional charge, the glass manufacturer can provide what is usually called a *precision* melt sheet, for which the indices have been measured individually and to a greater precision. Index values for wavelengths specific to the application can also be measured. It is, of course, wise to subject even this data to the difference test and smoothing process described in the preceding paragraph.

We called the third of these fits a *thickness fit* in the heading for this section. This is a process that is carried out during the final assembly of the lens. In sum, one reoptimizes the lens by varying the airspaces of the system, using accurately measured data for the radii, thicknesses, and indices of the fabricated elements. If the melt data and the test plate data are available, this just amounts to adding the measured element thicknesses and reoptimizing.

In a *reverse aberration fit*, the aberrations (or wavefronts) of an assembled system are measured. Then the nominal design is optimized to match the aberration (or wavefront) values that were measured. If the system is then adjusted by the opposite of the changes found by the optimization, the expectation is that the system performance will be improved. This technique has been used with success in multimirror systems, which are not axially symmetric and in which the alignment and positioning of the mirrors in space is both difficult and critical.

As can be seen, all of the above procedures are designed to almost completely eliminate the effects of the fabrication tolerances on the performance of the system. What is left, instead of the tolerances, is the uncertainty or inaccuracy of the various measurements on which the fits are based. The effect is usually quite modest and, therefore, acceptable;

however, these uncertainties are, in fact, the exact equivalent of tolerances in determining the performance of the fabricated system.

## 2.11 Spectral Weighting

In any ray-tracing process, the index of refraction of the material is, of necessity, that corresponding to a single specific wavelength. Most lens *design* programs use three wavelengths to represent the spectral bandpass for which the system is to be designed. Some allow more. *Analysis* programs may allow the use of five or ten suitably weighted wavelengths in calculating such things as MTF, point spread functions, radial energy distributions, and the like.

In either case, the question becomes, "What wavelengths should I use, and how should they be weighted?" For visual systems and just three wavelengths, the classical answer is to use d (or e), C, and F light, with weightings typically set at 1.0, 0.5, and 0.5, respectively, or perhaps with a uniform weighting. For other applications, this approximately corresponds to using the central wavelength and the wavelengths 25 percent from the extreme edges of the passband. If the results of an image analysis need not be especially precise, three wavelengths may be sufficient. For calculations done in the midst of the design process, this is often good enough to enable a judgment as to the relative merit of, or the rate of improvement between, two stages in the design process.

However, what should one do *in general?* To immediately dispose of the obvious, it is apparent that the more wavelengths used, the more accurate the results can be. That said, how do we select the wavelengths to be used? There are three obvious choices. In order to assess the full effects of the chromatic aberrations, one would like to include the extreme long and short wavelengths. One would also probably like to consider dividing the spectral weighting function into increments of equal power or response, so that each wavelength would represent an equally weighted sector. But if this is done, the extremes of the spectral passband are not included. Another possible option is to choose wavelengths that are evenly distributed across the spectral band, and to weight them according to the spectral response function. One might choose an even distribution on a wavelength scale, or what could be a bit better, an even distribution on a wave number (reciprocal wavelength) scale.

From this it should be apparent that most wavelength and weighting choices are based on some sort of compromise. Often the outer two wavelengths are chosen to be fairly close to the ends of the passband and the intermediate wavelengths and weights are a compromise between an even power distribution and an even wavelength spacing. If there are peaks or bumps in the spectral response function, wavelengths are often selected to be at or near the peaks. Note that, to a limited extent, one's

choices will partially control the design process: heavier weighting at the ends of the spectrum will obviously emphasize both primary and secondary chromatic aberration, spherochromatism, and the like; heavy weights on the central wavelengths will emphasize the monochromatic aberrations at the expense of the chromatic aberrations.

## 2.12 How to Get Started

Experienced designers are often asked questions such as "How do you know where to start?" or "How did you decide that a (name of a design type) could meet the specs?" or "Why did you shift the components around?"

The answer to these questions is almost always "experience," which probably means that the full answer is different for every problem. It would be foolhardy to pretend to be able to give a complete, definitive answer or even a set of answers to such questions, but there are a few guidelines that should be reasonably dependable.

Figure 2.8 is, in effect, a compilation of some of that experience. If the system to be designed roughly corresponds to a photographic objective or to one of the other types indicated, the figure can be used as an easy guide to the selection of an appropriate form. In this plot the areas corresponding to various combinations of field and aperture are labeled to indicate the type of system commonly used there. Obviously the boundary of each area is fuzzy and ill-defined. In a presentation of this type there is also an implied level of performance within each area, which is presumably typical of each particular lens applications. In general, the performance (image quality, resolution, or whatever) is better when a given lens type is optimized for a smaller field and/or smaller aperture, that is, for a combination closer to the origin of the plot. Thus one can select the design type from Fig. 2.8 corresponding to the field-aperture combination required with reasonable assurance that it is an appropriate choice. But should the performance prove inadequate for the application at hand, one can move up to the design form above and to the right on the plot, which should have a greater performance capability. Of course, if the required level of performance is known in advance to be relatively high, one would select the more capable type to begin with. The same selection approach can be applied to the components of a more complex system. There are many design types and applications, which are not represented on this necessarily abbreviated chart; however, many of these are presented in the lens designs included in this book. The reader may wish to mark up Fig. 2.8 to add the types that are of particular interest to him or her.

For a more complex system, the approach must start on a more fundamental level. The first step is to collect and tabulate the requirements to be met; these may include such things as:

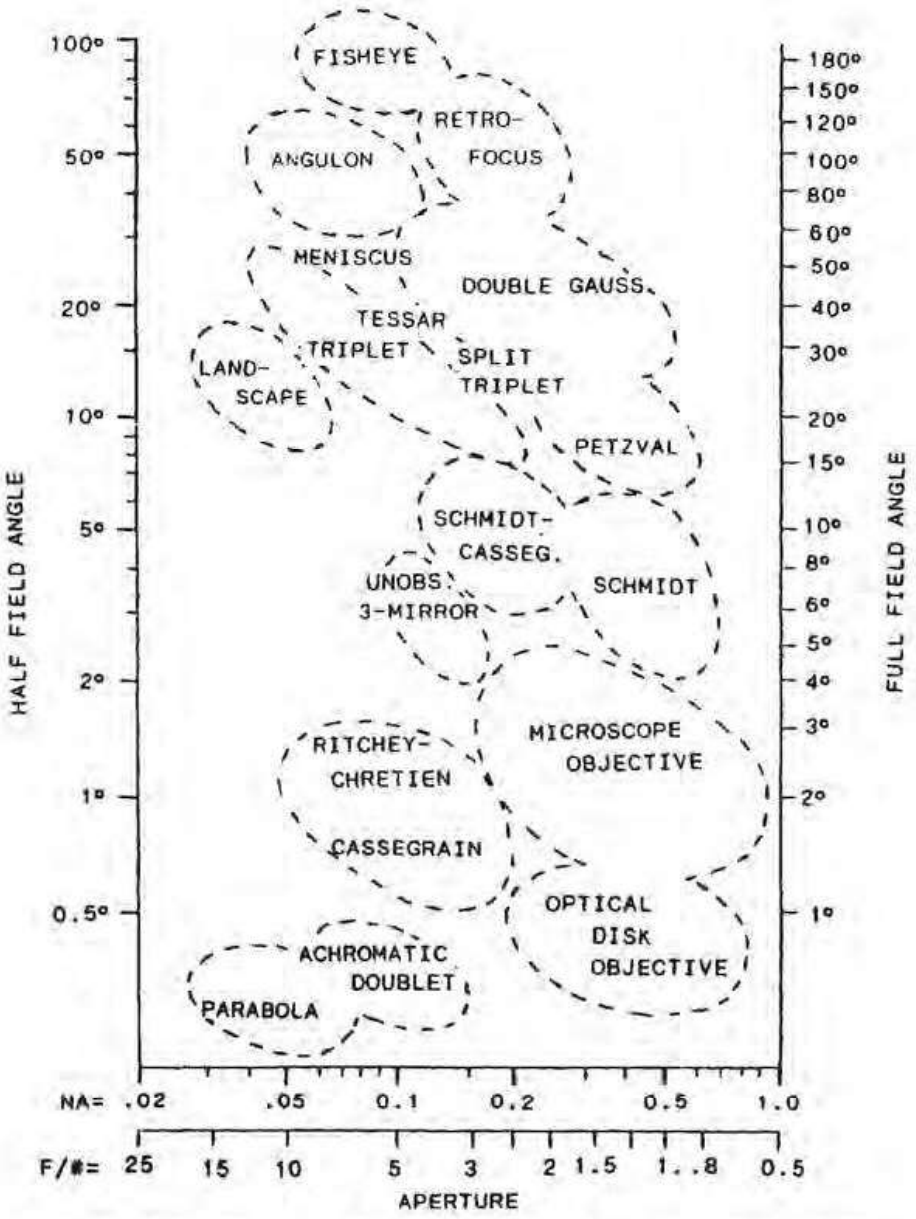



**Figure 2.8** Map showing the design types commonly used for various combinations of aperture and field of view.

Resolution or performance (versus diffraction limit)

Wavelength

Fields of view

Image size

Aperture, numerical aperture (NA), and *f* number

Vignetting or illumination uniformity

Focal lengths or magnifications

Space limitations, clearance, windows, and so on

(See Chap. 1 for a more extensive listing of possible specifications.)

The next step is to make a first-order layout of the system that will satisfy the requirements. The first-order layout is simply an arrangement of component powers (or focal lengths) and spacings that will produce an image in the required location, in the required orientation, and of the required size. At this stage, no consideration is given to the design type of each component; one is concerned only with its power, aperture, and field as first-order, i.e., paraxial, characteristics. For systems (or portions of systems) consisting of two components, the equations of Chap. 24 can be extremely useful. For more complex systems, the component by component ray-tracing equations may be used. A general approach is to trace paraxial rays that will define the required characteristics. The ray-trace results (ray heights, ray slopes, intersection lengths, and so on) can be expressed as equations with the component powers and spacings as unknowns to be solved for. An important facet of this stage is that one should try to find a layout that minimizes the component powers, or minimizes (or equalizes) the "work" (ray height times component power). Doing this will almost always produce a system with less aberration residuals, one that is less expensive to fabricate and less sensitive to fabrication and alignment errors. This process can be carried out with an optical design program as well.

Sometimes one can leap directly from the first-order layout to choosing the component design types, and thence to the optimization stage; however, if this is not the case, the next step is usually to analyze and/or correct the chromatic aberrations. Equations (24.85) and (24.97) can be used for the whole system, or the components can be individually achromatized. To this end, the element powers for a thin achromatic doublet are given by

$$\phi_A = \frac{V_A}{(V_A - V_B)F} \tag{2.1}$$

$$\phi_B = \frac{V_B}{(V_B - V_A)F} = \frac{1}{F} - \phi_A \tag{2.2}$$

and the element surface curvatures are determined from

$$C_1 - C_2 = \frac{\phi}{n-1} = \frac{1}{r_1} - \frac{1}{r_2} \tag{2.3}$$

At this point a sketch of the system is often helpful. Simply make a scale drawing, showing each element as either a plano-convex or an equi-convex form (plano-concave or equi-concave for negative elements). If the elements look too fat, they should be split into two or more elements. Be sure that the element diameters are sized properly for the rays that they must pass. At this stage the system should begin to look like a good lens. Again, the computer can be used to make the drawings.

The next step is to give some consideration to the Petzval curvature. Choosing suitable anastigmat types for the components is one way to handle this. Another is to use a field flattener in an appropriate location (i.e., near an image) in the system. And, of course, the usual device of configuring the system or component with separated positive and negative elements or surfaces to reduce the Petzval sum can always be utilized if a new design must be created from scratch.

Often many of these steps can be handled conveniently and expeditiously by the automatic design program. The first-order layout can be done with zero-thickness plano-convex or plano-concave elements, allowing the spaces and the curvatures of the curved surfaces (but not the plano surfaces) to vary. The merit function is a simple one, configured to define the required first-order characteristics. Note well the comments in Sec. 2.6 regarding stagnation and contradictory or redundant first-order entries in the merit function.

Since the chromatic aberrations and the Petzval curvature depend on element power and not on element shape, the lens design program can also be used to find a layout that is a preliminary solution with the chromatic and Petzval adjusted to desired, reasonable values, which typically should be small and negative.

Sometimes it is useful as a next step to allow the elements to bend and to correct the third-order aberrations. This can be done by putting an angle-solve on the second surface of each element so that the axial ray slope is maintained. More often than not, however, the next step will skip over the third-order and go directly to a full-dress thick lens optimization run.

A design start may be based on: (*1*) previous work from one's files; (*2*) an idea for the general configuration necessary; (*3*) published prescriptions or scalable drawings; (*4*) lens libraries; and (*5*) imagination and intuition.

Some starting points that have been suggested:

a. A series of plano–plano plates.

b. A series of nondispersive aspheric singlets, to be later converted to achromatic doublets after the first-order and the monochromatic aberrations have been satisfied.

The starting point (b) has, under the ministrations of a good designer, been successful. But the unguided start from plane parallel plates is

almost a parlor trick. Once the merit function is defined and the thickness, spacing, and glass of the plates are given, we must realize that, for a given program, the end result is then completely predetermined; there is no magic. A good design result is of course possible, and occasionally a pleasant surprise can surface, but this is unlikely unless there is a bit of chicanery or some steering of the process (the latter is, of course, why there are lens designers).

A general starting process:

1. Do a thin lens layout and drawing. Look for ridiculous powers, spacings, and the like.
2. Achromatize, do a drawing, and look again.
3. Revise the layout to reduce, add, or split elements.
4. Now start the program.

### 2.12.1 Hypothetical, generic design process

After selecting a design form or type:

1. Start by varying all the curvatures and the large air spaces. Manually adjust the center thicknesses of the positive elements for a suitable edge thickness. Check all bounds and first-order targets for compliance. Consider the general practicality of the result. Consider splitting elements that seem too strong.

2. Add all thicknesses and spaces to the variable list and optimize. Correct and temporarily freeze any problem areas (e.g., CT < 0.0, ETH < 0.0, or extreme thickness or spaces). If a ray fails (by missing a surface or by TIR), restart with only first-, third-, and fifth-order targets in the merit function. These cannot fail because they are not based on trigonometric (exact) rays.

3. Add glass to the variable list, either along the glass line, or vary both index and dispersion. Use tight bounds, because glass variables can easily get out of control. Sometimes varying only one or two glasses at a time is a useful limitation.

4. Adjust the target weights up or down or add targets as appropriate to control or attack trouble spots.

5. Fix the glasses, one at a time, to catalog types and vary all curvatures and thicknesses.

As the design progresses, occasionally check the MTF (both through focus and through frequency). Check the distortion and lateral color; it's easy to overlook them in the heat of battle. Draw the lens and the marginal rays at several field angles. Look for large incidence angles; enlarge the scale; and examine any trouble spots. Print out the third- and fifth-order

surface contributions and look for any surface with larger than average contributions; these often signal trouble.

In the best of all worlds, one simply sets up what seems to be a likely layout and proceeds directly to the automatic lens design program, which promptly turns out an excellent design. "Experience?" Possibly. But certainly, "Lots of luck!"

Chapter

# 3

# Improving a Design

## 3.1 Lens Design Tip Sheet: Standard Improvement Techniques

This is a listing of changes in the design form or the design approach that may significantly improve or change a lens design. Many changes can be incorporated without a major change in the system's power or magnification. They can be introduced blindly because they are expected to be beneficial.* Note that most of these changes require additional intervention, i.e., complementary changes to the balance of the system to restore (and hopefully improve) the design. In no particular order, the suggestions are:

1. Bend an element. Change its shape but maintain its power.
2. Add a field flattener or a concave surface close to the image.
3. Remove a weak element (add $cv_1$, $cv_2$, $cv_1 - cv_2$, and $th$ to the merit function with zero targets, and gradually increase the weighting).
4. Split a strong troublesome element into two weaker elements. Beware of *total internal reflection* (TIR) at the split; hold the incidence angles to less than the critical angle, $I_c = \arcsin(n'/n)$.
5. Replace a thick element with two thin elements and an airspace.
6. Add a plane parallel plate in front or in back of the lens, or in a large airspace, and vary its curvatures and thickness.
7. Add an element and later remove it (gradually, by targeting it in the merit function).

*This list was inspired in part by Ellis Betensky in "Postmodern Lens Design," *Optical Engineering*, Vol. 32, No. 8, pp. 1750–1756, Aug. 1993.

8. Replace a singlet made of expensive high-index glass with two elements of cheap glass.
9. Raise the index of the materials, maintaining element power.
10. Introduce unusual partial dispersion materials if secondary spectrum is a problem.
11. Convert a singlet into a cemented doublet with glasses of: (*a*) different indices, (*b*) different *V*-values, or (*c*) both.
12. Cement two adjacent surfaces that have nearly equal curvatures.
13. Reverse, i.e., turn around, a cemented doublet.
14. Change a thick, strong element into a cemented doublet.
15. Split (airspace) a cemented doublet (beware of TIR).
16. Insert a thick, low-power meniscus element before, after, or within the lens. A concave surface near the image can act as a field flattener. A *concentric* shell field flattener, which is also concentric with the focus, will not change the efl, focus position, spherical, coma, axial chromatic, lateral chromatic, or sagittal field curvature. It will affect the Petzval curvature, the distortion, and the tangential field curvature.
17. Reverse or flip a meniscus element, or a low-powered *pair* of elements. A concentric shell flipped about its center of curvature will not change *anything*, but it may lead the design to a quite different form.
18. Add thickness to a positive power meniscus to improve the Petzval field curvature by holding the power constant, which will require a stronger concave surface.
19. Add an aspheric to a surface and later remove it gradually by targeting it to zero in the merit function.
20. Insert or add a *disposable* zero power aspheric element, which will be removed later on.
21. Asphere a surface (near the stop for spherical, or away from the stop to also affect coma, astigmatism, and distortion).
22. Move the stop to a new location to use a better portion of the oblique beam (or avoid a worse portion).
23. Look for surfaces or elements with large third- or fifth-order contributions or a large angle of incidence. They are usually troublemakers.
24. Make a large change in a constructional parameter and freeze it for several optimization cycles, then release it. Choose a parameter that looks troublesome. Change it in the direction of reducing the angles of incidence.

25. Temporarily loosen (or remove) a bound or constraint, or reduce the weight on a target to allow the design to temporarily cross over the boundary to find a new solution, and then reimpose the limits.
26. Systematically vary the starting point for the optimization (in hopes of finding a new and better local optimum).
27. Reduce the weights on the off-axis targets in the merit function (or eliminate them entirely) for a few cycles, and then raise (or reapply) the weights.
28. Temporarily reduce the damping factor in *damped least squares* (DLS) to escape from a local optimum (by allowing changes that increase the value of the merit function—by a factor of 100 to 10,000 per Dave Shafer), then reoptimize.
29. After the DLS process has bottomed out, some hacking around can often squeeze a little more out of a design by making more or less random arbitrary changes and then reoptimizing; changing to a different merit function, e.g., OPD versus spot size versus an aberration merit function; and jolting the design and reoptimizing.
30. Try a reduced or limited set of variables—the important ones.
31. Remove the following: ineffective variables, variables that null each other, and targets that are redundant.
32. Modify the merit function to severely target the most bothersome aberrations by increased weighting or by added operands.
33. In general, add new rays and targets to the merit function to pressure the aberrations identified as problems in the ray intercept plots.
34. Add more rays, fields, and colors to the merit function toward the end of the design process to get more accurate evaluations and also to be sure that something has not slipped through the cracks.

A good part of the "art" in lens design consists of knowing when to add or delete an element. The simple, straightforward application of these techniques is no guarantee of improvement in a lens in that they do not automatically correct the defects that they are intended to address. In general, these changes tend to reduce the aberration contributions of the modified components; to take full advantage of this, the aberrations of the balance of the system must be reduced as well. The operative principle is this: if large amounts of aberrations are corrected or balanced by equally large amounts of aberrations of opposite sign, then the residual aberrations also tend to be large. Conversely, if the balancing aberrations are both small, then the residuals tend to be correspondingly small.

Although things may go smashingly well if you vary everything at once, it's not always the best approach to vary everything right from the

start. If there are many variables, it is sometimes useful to first vary only the curvatures and then add the thicknesses and spacings. Finally, allowing only one or two glasses at a time to vary is often wise.

When you have located the source of a troublesome aberration (perhaps by checking the third-order contributions) it should be corrected as soon (i.e., as close) as possible. Allowing it to propagate through the lens is a source of higher-order aberrations.

Always save any promising result you find in the course of a design as soon as you find it. Save it as a file or as a printout. After a long session of design changes and investigations, you may never be able to find it again, even if you try to retrace your steps from the start of the day.

In a finite conjugate system, do not constrain more than two of: (*a*) focal length, (*b*) object to image distance, or TTL, and (*c*) the magnification. Probably the best one to leave out is the focal length. If all three are constrained, a simultaneous solution may still be possible, but the design form will be severely limited and a better solution, which meets the magnification and TTL specifications, may be missed.

A good way of deciding where to attack a design is to draw the lens with marginal rays (and the upper and lower rays of the oblique beams). Look for large angles of incidence or refraction. This, combined with an analysis of the surface contributions of the third- (and fifth-) order aberrations will often pinpoint the source of troubles. A large ray height, either for the marginal or principal ray, may also indicate the surfaces at risk. If the angle of incidence or refraction of the ray is detrimentally large, or if the slope change (i.e., $u' - u$) is extreme, it can be added to the merit function to drive it down. The Aldis total aberration (see Ref. 3) is available in some programs; it is an all-inclusive representation of the surface contribution.

In most lenses there are many small variations of the design form with nearly equivalent performance or *modulation transfer function* (MTF). Note that the merit function and the MTF are very rarely well correlated (if at all). Poke around and look for something of interest that is off the beaten (DLS) path.

Other than at unit magnification (1:1), a fully symmetrical construction is usually not the best solution, although it may be a useful starting point. In fact, even at 1:1, the computer may find an unsymmetrical solution it likes better. Bear in mind that the improvement from the $f/4$ Plasmat lens to the $f/2$ double-Gauss lens was made possible by abandoning symmetry (and thickening the inner doublets). Although it's not a sufficient condition for a good design, a small Seidel sum is a necessary condition for a good design.

To take care of a troublesome aberration, target it in the merit function with a higher weighting, add the appropriate third- and fifth-order targets to the merit function, and add a ray (or field) to get it under control.

## 3.2 Glass Changes: Index and *V*-value

The refractive characteristics of the materials used in a lens are obviously significant and important to the design. In general, for a positive element, the higher the index the better. The higher index reduces the inward Petzval curvature, which plagues most lenses. It also tends to reduce most of the other aberrations as well. As an example, see Fig. 3.1, which clearly

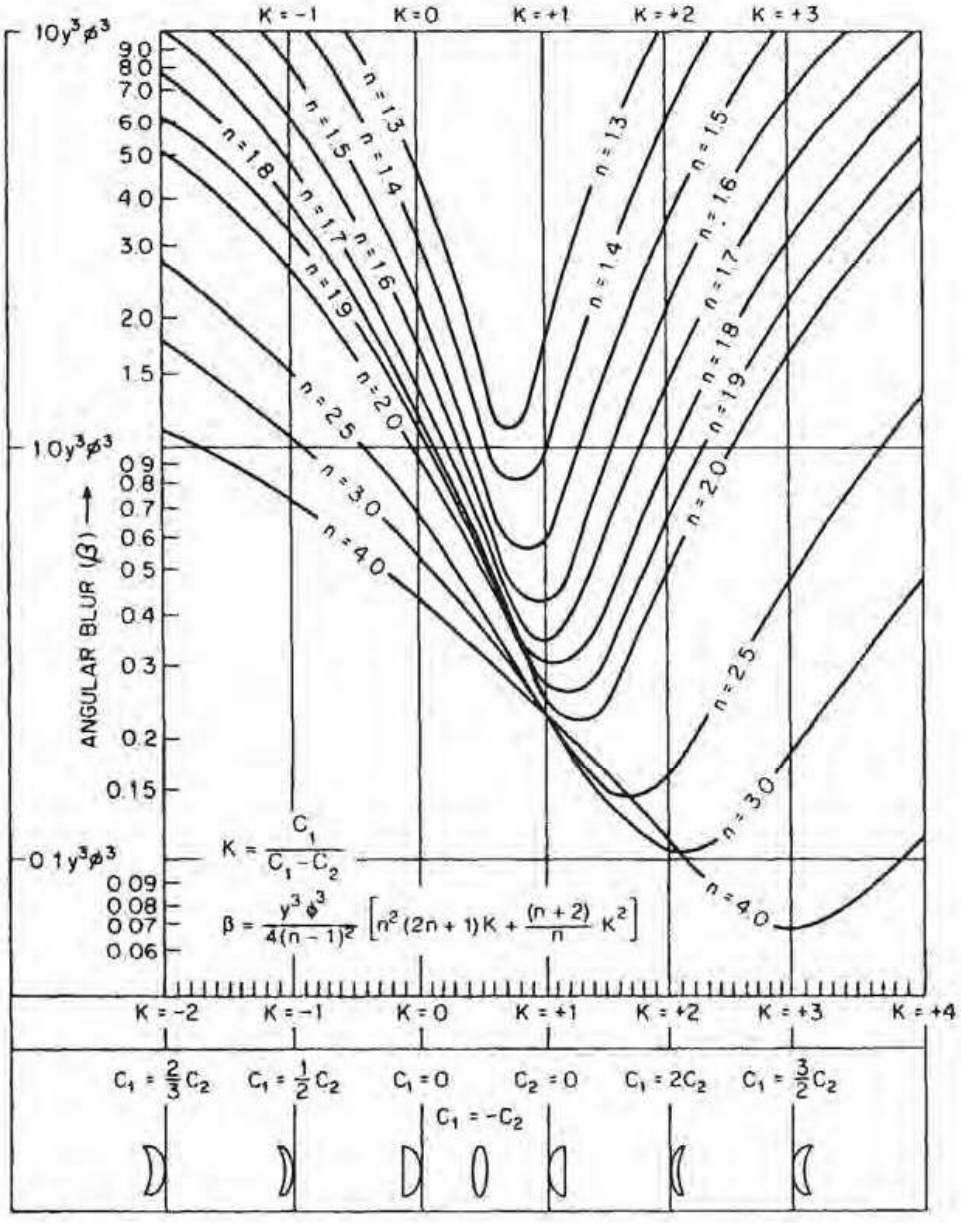



**Figure 3.1** The angular spherical aberration blur of a single lens element as a function of lens shape, for various values of the index of refraction; $\phi$ is the element power and $y$ is the semiaperture. The angular blur can be converted to longitudinal spherical aberration by $LA = 2\beta/y\phi^2$, or to transverse aberration by $TA = -2\beta/\phi$. (The object is at infinity.)

indicates the effect of higher index in reducing the spherical aberration of a single element. This sort of reduction is primarily a result of the fact that the surface curvature required to produce a given element power is inversely proportional to $(n - 1)$. The improvement also results from the reduction of the angles of incidence at the surfaces of the element.

In a negative element, the situation is less clear. From the standpoint of the Petzval correction, a low index would increase the overcorrecting contribution of a negative element. This can help to offset the (inward) undercorrection, which is a major problem in most lenses. On the other hand, a higher index would reduce the surface curvatures and have a generally desirable effect on the overall state of correction. The situation is usually resolved with the negative elements made from a glass along the "glass line" lower boundary of the glass map (Fig. 2.7).

A high $V$-value for the positive element and a low $V$-value for the negative element of an achromatic doublet reduce the element powers; this is ordinarily desirable. In lenses (such as the Cooke triplet) where the relative $V$-values of separated elements control the element spacing or the system length (see Chap. 8), this desideratum may be overridden by other concerns.

In general, for negative elements the flint glasses from along the glass line (see Fig. 2.7) are the best choice (except in a negative-power component). Again, in general, a higher index is better. Check the price and availability. Many glasses have recently been discontinued or reformulated to eliminate lead, arsenic, or cadmium. Some of the new "light" glasses are more attractive than their predecessors in many respects. Stain and weathering characteristics may be very important for some applications. Note that ordinarily bubbles are simply a cosmetic defect (unless the element is near an image). Striae are especially important in prisms or thick pieces. Incidentally, a system with many elements usually implies that a wide choice of glass types may be utilized.

Note that, as usual, when you are dealing with components of negative focal length, many of the considerations outlined above are reversed. In a negative achromatic doublet, the negative element is made of crown glass and the positive is made of flint. Here a high-index (flint) positive element will reduce the inward Petzval curvature, as will a low-index (crown) negative element.

## 3.3 Splitting Elements

Splitting an element into two (or more) approximately equal parts whose total power is equal to the power of the original element can reduce the aberration contribution by a significant factor. The reason why this reduces aberrations is that it allows the angles of incidence to be

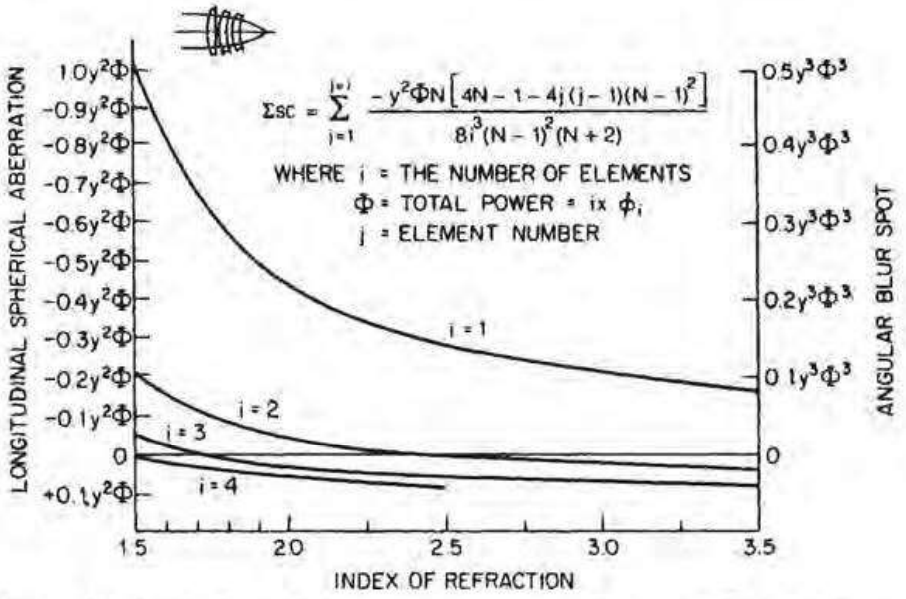



**Figure 3.2** The spherical aberration of one, two, three, and four thin positive elements, each bent for minimum spherical aberration, plotted as a function of the index of refraction, and showing the reduction in the amount of aberration produced by splitting a single element into two or more elements (of the same total power). Each plot is labeled with $i$, the number of elements in the set. (The object is at infinity.)

reduced; the nonlinearity of Snell's law means that smaller angles introduce less aberration than do large ones. This technique is often used in high-speed lenses to reduce the zonal spherical residual and in wide-angle lenses to control astigmatism, distortion, and coma.

Figure 3.2 shows the thin lens third-order spherical aberration for spherical-surfaced positive elements shaped (or bent) to minimize the undercorrected spherical. The upper plot shows the spherical as a function of the index of refraction for a single element with a distant object. The curve labeled $i = 2$ shows the spherical for two elements whose total power is equal to that of the single element. The best split is 50-50, i.e., the split elements have equal power; this minimizes the spherical. (The same is true for a split into more than two elements, i.e., three, four, and so on, as shown in the curves labeled $i = 3$ and $i = 4$.) The improvement produced by splitting an element in two can be seen to be a factor of about 5 for lenses of index equal to 1.5. The higher the index, the greater the reduction; for an index of 1.8, the factor is about 7. At an index of 2.5 or higher, the spherical can be brought to zero or even overcorrected with just two positive elements. Most other aberrations are similarly affected by splitting, although it should be obvious that neither Petzval nor chromatic is changed by splitting.

In high-speed lenses this technique is frequently used to reduce the residual zonal spherical; the positive elements are split. This illustrates

the basic idea. If the residual zonal spherical is negative (undercorrected), one splits a positive element; in the rare event that the zonal is positive (overcorrected), one would split a negative element. A similar philosophy can be applied for troublesome residuals of the other aberrations as well.

The choice of which element to split is often less apparent. The logical candidate would obviously seem to be the element that contributes most heavily to the problem aberration. (An examination of the third- and fifth-order surface contributions can often locate the source of the aberration. Alternately, look for large angles of incidence or refraction.) However, other considerations often become significant. For example, in the Cooke triplet, the rear element is the prime candidate for the split, and such a split is quite effective in reducing the zonal spherical, as Fig. 9.1 will attest. But the better choice for the split is the front element, not because it does a better job of reducing the zonal spherical, but because the resulting lens is better corrected for the other aberrations. This is because the meniscus-shaped second element reduces the Petzval sum and has very little spherical aberration. Figure 9.2 shows the simple split-front triplet. This is the ancestor of the Ernostar family of lenses; Figs. 9.6 to 9.10 illustrate designs that can be considered descendants from the split-front triplet. Although they have been largely superseded by the more powerful double-Gauss form, they are nonetheless excellent design types.

Many retrofocus and wide-angle lenses that use strong outer meniscus negative elements illustrate the use of this technique for the control of coma, astigmatism, and distortion by splitting these negative elements.

The implementation of this technique with an automatic design program is sometimes far from easy. For example, if one decides to split one of the crowns of a Cooke triplet and simply replaces one crown with two, after the computer optimization has run its course, the resultant lens may look like an ordinary triplet with a narrow cracklike airspace in the split element (a *cracked crown triplet*). The performance of the lens is the same as the original triplet; the split has not improved a thing. This is because the original lens was in a local optimum of the merit function. Aberrations other than the zonal spherical dominated the design; this caused the program to return the lens to its original design configuration. What is necessary in this situation is to force the split elements into a configuration that will accomplish the desired result.

Consider the split-front triplet. There are two ways to get to a design like Fig. 9.2. One approach is to make the lens so fast that the zonal spherical is by far the single, dominant aberration in the merit function. Then the program will probably choose a form that reduces the zonal spherical; the lens shapes in Fig. 9.2 are a likely result. A difficulty with

this approach may be that you aren't interested in a very fast lens, or if you are, the rays may miss the surfaces of the initial design completely, or encounter TIR. The alternative approach is to constrain the front elements to a configuration in which the spherical is minimized. Simply fixing the first element to a plano-convex form (by not allowing the plano surface curvature to vary) or holding the second element to an aplanatic meniscus shape is usually sufficient to obtain a stable design that is different enough from the cracked crown triplet. When this has been accomplished the constraint can be released and the automatic design routine allowed to find what is (one hopes) a new and better local optimum. The problem here is that this approach presupposes a knowledge of the configuration that will produce a good result. Obviously a knowledge of both aberration theory and of successful design forms is a useful tool to the designer.

## 3.4 Separating a Cemented Doublet

Airspacing a cemented doublet can provide two additional degrees of freedom: two bendings instead of one, plus an airspace. While this technique does not have the inherent aberration reduction capability that many other modifications possess, the extra variables may indirectly make a design improvement possible. A difficulty in implementing this is that the refraction at the cemented surface is apt to become much more abrupt when it is split into two glass-air interfaces than when it was a cemented surface; in fact, rays may encounter TIR if a simple split is attempted without a concomitant reduction of the angle of incidence. Manual intervention in the form of adjusting the radii to reduce the angle of incidence is often necessary.

## 3.5 Compounding an Element

Compounding a singlet to a doublet can be viewed in two different ways:

1. As a way of simulating a desirable but nonexistent glass type
2. As a way of introducing a cemented interface into the element in order to control the ray paths

The longitudinal axial chromatic of a singlet is given by $LA_{ch} = -f/V$. Thus a fully achromatized lens (with $LA_{ch} = 0.0$) has the chromatic characteristic of a lens made from a material with a $V$-value of infinity; a partially achromatized doublet acts like a singlet with a very high $V$-value.

The Petzval radius of a singlet is given by $\rho = -nf$, where $n$ is its index. An *old achromat* with a low-index crown and a higher-index flint has a shorter Petzval radius than a singlet of the crown glass. For

example, an achromat of BK7 (517:649) and SF1 (717:295) has a Petzval radius $\rho = -1.37f$; in other words, in regard to Petzval field curvature, it behaves like a singlet with an index of 1.37. A *new achromat* has a high-index crown and a lower-index flint. A new achromat of SSKN5 (658:509) and LF5 (581:409) has a Petzval radius $\rho = -2.19f$ and a Petzval curvature that is characteristic of a singlet with an index of 2.19.

Thus an achromatized (or partially achromatized) doublet with a high-index positive element and a low-index negative element has many of the characteristics of a lens made of a high-index, high-$V$-value crown glass. (Note that for a negative focal length doublet, the reverse is true.) Both conditions are usually to be desired to flatten the Petzval field and to achieve achromatism.

Note that in almost all examples of Tessar-type lenses (and other types that utilize compounded elements), the doublets have positive elements with high index and high $V$-values, while the negative element of the doublet has both a lower index and $V$-value. See Chap. 10 for examples.

Figure 3.3 shows a singlet, an old achromat, and a new achromat, each with the same focal length. The equivalent $V$-value of each achromat is equal to infinity. The Petzval radius for each is given in the figure caption.

The cemented interface of the doublet can be used for specific control of specific rays. In a lens such as the Tessar, where the doublet is located well away from the aperture stop, the upper and lower rim rays of the

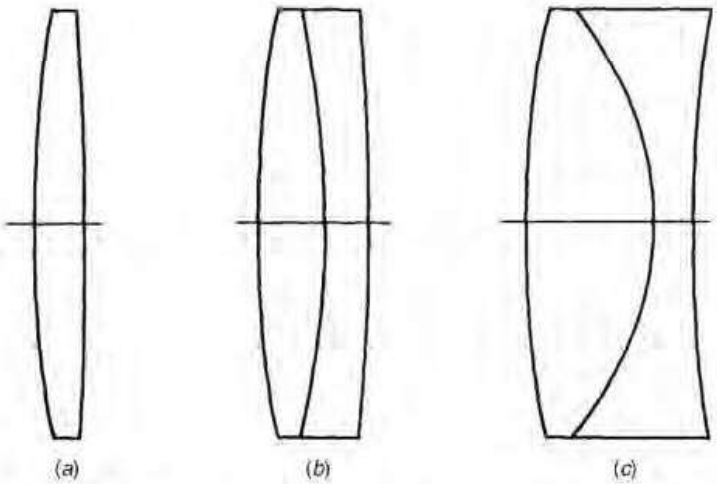



**Figure 3.3** Three lenses, each with the same focal length $f$. (*a*) A singlet of BK7 (517:642) glass; Petzval radius equals $-1.52f$. (*b*) An old achromat of BK7 (517:642) and SF1 (717:295) glasses; Petzval radius equals $-1.37f$. (*c*) A new achromat of SSKN5 (658:509) and LF5 (581:409); Petzval radius is $-2.19f$.

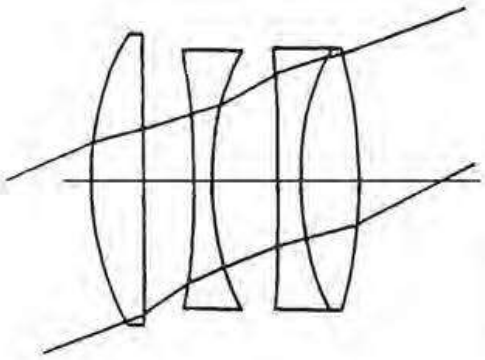


**Figure 3.4** The upper and lower rim rays have significantly different angles of incidence at the cemented interface in the rear doublet of this Tessar design. Properly handled, this difference can be used to modify the correction of the coma-type aberrations.

oblique fan have very different angles of incidence at the cemented surface. In Fig. 3.4 it can be seen that the angle of incidence at this surface is much larger for the upper ray than for the lower. In this type of lens the cemented surface is typically a convergent one, and the (trigonometric) nonlinear characteristic of Snell's law means that the upper ray is, in this case, refracted downward more than it would be were the refraction linear with angle. Thus the upper ray is deviated in such a way as to reduce any positive coma of this ray. This illustrates the manner in which a cemented surface can be used for an asymmetrical effect on an oblique beam.

The cemented surface in the Tessar doublet:

1. Reduces the zonal spherical aberration
2. Reduces the overcorrected oblique spherical
3. Reduces the astigmatism of the zonal field (i.e., the "belly" in the $X_s$ and $X_t$ curves)

### 3.5.1 The Merté surface

A strongly curved, collective cemented surface with a small index break (to the order of 0.05 to 0.15) has an effect that can be used to reduce the undercorrected zonal spherical aberration. The central doublet of the Hektor lens shown in Fig. 3.5 illustrates this principle. The cemented surface is a collective one (in that $[n' - n]/r$ is positive) and contributes undercorrected spherical aberration. For rays near the axis, the spherical aberration contribution of the surface is modest; however, when the ray intersection height increases and the angle of incidence becomes large, as shown in Fig. 3.5, the trigonometric nonlinearity of Snell's law causes the amount of ray deviation to be disproportionately increased. This causes the undercorrection from this surface to dominate the spherical aberration. The result is a spherical aberration characteristic like that shown in Fig. 3.5. The spherical in the central part of the aperture appears

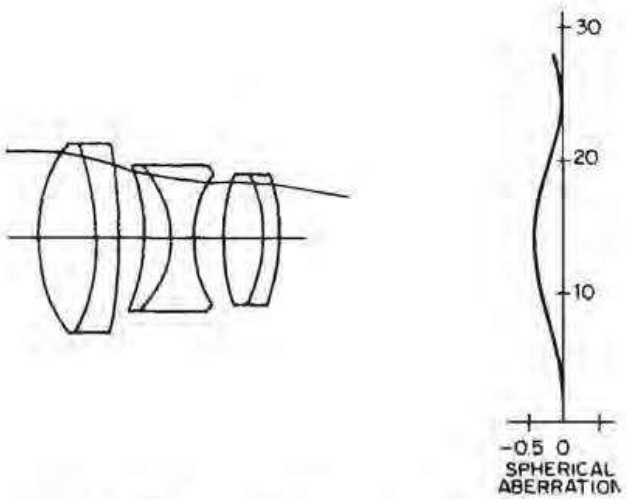



**Figure 3.5** The cemented surface in the center doublet of this Hektor lens is what is called a Merté surface. The index break $(n' - n)$ across the surface is small, but at the margin of the aperture the angle of incidence for the axial ray becomes quite large. This combination produces an undercorrecting seventh-order spherical aberration which, as the plot shows, dominates the spherical aberration at the margin of the aperture, causing the marginal spherical to be negative rather than the usual positive value.

quite typical—the undercorrected third-order dominates close to the axis and the overcorrecting fifth-order causes the plot to curve back as the ray height increases. However, toward the edge of the aperture the undercorrection of the Merté surface becomes dominant and the aberration plot reverses direction again. The net result is the equivalent of a reduced zonal spherical aberration.

It is rare to see as extreme an example of the Merté surface as that illustrated in the Hektor of Fig. 3.5. Such a surface is very sensitive to fabrication errors and is thus expensive to make. It is also often best used close to the aperture stop because, if it is located away from the stop, the asymmetrical effects described two paragraphs above can become quite undesirable; however, it is well worth noting that even an ordinary collective cemented surface has a tendency to behave as a mild Merté surface and to reduce the spherical zonal, at least somewhat.

## 3.6 Vignetting and Its Uses

Vignetting, which is simply the mechanical limitation or obstruction of an oblique beam, is usually regarded primarily as something that reduces the off-axis illumination in the image; however, vignetting often

plays an essential role in determining the off-axis image quality as well as the illumination. Of course there are many applications for which vignetting cannot be tolerated; the illumination must be as uniform as possible across the entire field of view. Thus for such applications, the complexity of the lens design must be sufficient to produce the required image quality at full aperture over the full field.

But for many applications, vignetting is quite tolerable. In commercial applications the clear apertures may well be established so as to be just sufficient to pass the full aperture rays for the axial image. It is not at all unusual for vignetting to exceed 50 percent at the edge of the field. For a camera lens, this vignetting will completely disappear when the iris of the lens is stopped down to an aperture below the vignetting level. Since camera lenses are most often used at less than full aperture, the vignetting is not as significant as it is in a lens that is always used at full aperture, such as a microscope or projection lens.

The *benefit* of vignetting is that it cuts off the upper and/or lower rim rays of the oblique tangential fan. Since these are ordinarily the most poorly behaved rays, the image quality may well be improved by their elimination. Most lenses that cover a significant field are afflicted with oblique spherical aberration, a fifth-order aberration that looks like third-order spherical aberration, but which varies as the square of the field angle. And since its magnitude is different for sagittal than for tangential rays, it can be seen to have characteristics of both astigmatism and spherical aberration. Oblique spherical aberration usually causes the rays at the edge of the oblique bundle to show strongly overcorrected spherical aberration; vignetting is a simple way to block these aberrant rays from the image.

Another factor favoring the use of vignetting is that it results from lens elements with small diameters. In general, one can count on a smaller-diameter lens being less costly to fabricate.

For a camera lens, one must be certain that the iris diaphragm is located centrally in the oblique beam so that, when the iris is closed down, the central rays of the beam are the ones which are passed. These are usually the best-corrected rays of the oblique beam. Also this location assures that the vignetting will be eliminated at the largest possible aperture.

A lens drawing with an oblique fan of 21 rays can yield a quick visual analysis of the vignetting; it can also show which parts of the system affect the aberrations and how they affect them. A fan of 21 rays is useful in checking the vignetting in increments of 10 percent of the semiaperture of the beam.

Note that vignetting is closely related to the overall length of the lens and if a vignetting problem seems intractable, a shorter lens may help. Unfortunately, lenses usually want to be long and large.

## 3.7 Eliminating a Weak Element—the Concentric Problem

Occasionally an automatic design program will produce a design with an element of very low power. Frequently this means that the element can be removed from the design without adversely affecting the quality of the design. Often a straightforward removal will not work; the design process may simply blow up. An approach that usually works (if anything will) is to add the thickness and the surface curvatures of the element to the merit function with target values of zero, allowing them to continue as variable parameters. Sometimes targeting the difference between the two curvatures is also useful. Usually, if the element isn't necessary to the design, a few cycles of optimization, possibly with gradually increasing weights on these targets, will change the element to a very thin, nearly plane parallel plate, which can then be removed without severe trauma to the design. If your design program will not accept curvatures and thicknesses as targets, an alternative technique is to remove the curvatures and the thickness as variables and to gradually weaken the curvatures and reduce the thickness (by hand) while continuing to optimize with the other variables.

An unfortunate form of the weak element is a fairly strongly bent meniscus, which the computer uses for a relatively important design function, such as the correction of spherical aberration or the reduction of the Petzval curvature. It is rarely possible to eliminate such an element because it is an integral part of the design. The unfortunate aspect of this situation is that, if the surfaces of the meniscus element are concentric or nearly so, the customary centering process used in optical manufacture is impossible or impractical, and the element is costly to fabricate. This situation can be ameliorated* by forcing the centers of curvature of the surfaces apart by a distance sufficient to allow the use of ordinary centering techniques. Again, including the required center-to-center spacing in the merit function and reoptimizing will usually modify the offending element to a more manufacturable form without significant damage to the system performance.

## 3.8 Balancing Aberrations

The optimum balance of the aberrations is not always the same in every case; the best balance varies with the application and depends on the size of the residual aberrations. In general, for well-corrected lenses, the aberrations should be balanced so as to minimize the optical path

*See: Smith, "The Problem of the Concentric Meniscus," *Optical Engineering*, Vol. 27, pp. 1039–1041, 1988.

difference (OPD), i.e., the wavefront variance, but there are significant exceptions.

### 3.8.1 Spherical aberration

If a lens is well-corrected and the high-order residual spherical aberration is small so that the OPD is to the order of a half- wave or less, then the best correction is almost always that with the marginal spherical corrected to zero, as illustrated in Fig. 3.6*b*; however, when the zonal spherical is large, there are two situations where one may want to depart from complete correction of the marginal spherical.

If the lens will always be used at full aperture (as a projection lens, for example), and if the spherical aberration residual is large (say to the order of a wave or so), the diffraction effects will be small when compared to the aberration blur; then the spherical aberration should be corrected to minimize the size of the blur spot rather than to minimize the OPD. This will produce the best contrast for an image with relatively coarse details, i.e., for a resolution well below the diffraction limit. As an example, at a speed of *f*/1.6, a 16-mm movie projection lens has a diffraction

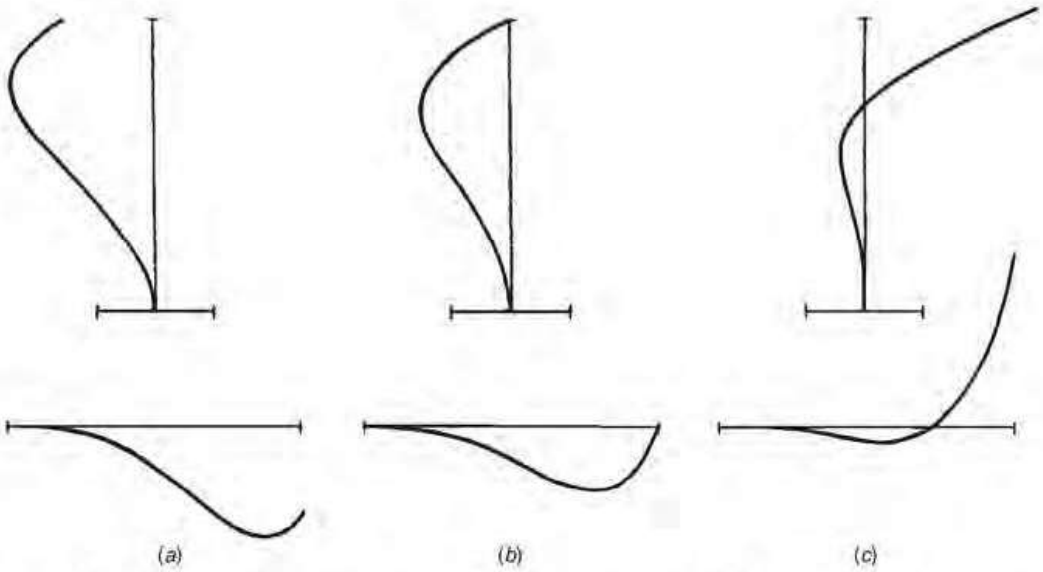



**Figure 3.6** Three states of correction of spherical aberration are shown. Each has the same amount of fifth-order spherical, but different amounts of third-order. (*a*) Spherical aberration balanced to give the smallest possible size blur spot when refocussed. This correction may be optimum when the aberration is large and the required level of resolution is low compared to the diffraction limit. (*b*) Spherical aberration balanced for minimum OPD. This is optimum when the system is diffraction-limited. (*c*) Spherical aberration balanced to minimize the focus shift as the lens aperture is stopped down. This correction is used in camera lenses when the residual spherical is large. The upper row is longitudinal spherical versus ray height; the lower row is transverse ray intercept plots.

cutoff frequency of about 1100 *line pairs per millimeter* (lpm). But its performance is considered quite good if it resolves 100 lpm, an order of magnitude less than the diffraction limit. Such a lens can advantageously be corrected for the minimum diameter geometrical blur spot. This state of correction occurs (for third- and fifth-order spherical) when $LA_z = 1.5LA_m$, or $TA_z = 1.05TA_m$, and the defocus is $\delta = 1.25LA_m = 0.83LA_z$; the result is a high-contrast, but low-resolution, image. This correction is illustrated in Fig. 3.6*a*. See also the comments on defocusing in Secs. 2.4 and 2.7.

For a lens used at varying apertures, as in a typical camera lens, it is important that the best focus position not shift as the size of the aperture stop is changed. If the spherical aberration is corrected at the margin of the aperture, or corrected as described in the paragraph above, the position of the best focus will shift as the aperture is changed. The best focus will move toward the paraxial focus as the aperture is reduced. The state of correction that is often used in such a case is overcorrection of the marginal spherical, as shown in Fig. 3.6*c* (assuming an undercorrected zonal residual). For example, when $LA_m = -LA_z$ (or $LA_m + LA_z = 0$) the result is a design in which the focus is reasonably stable as the lens is stopped down. The *resolution* is better than it would be otherwise, but, at full aperture, the *contrast* in the image is quite low. This works out reasonably well, especially in a high-speed camera lens because camera lenses are only infrequently used at full aperture. Typically, photographs are taken with the lens stopped down well below the full aperture, and, when the camera is stopped down, this state of correction yields a much better photograph than that produced by a balance like Fig. 3.6*a* or Fig. 3.6*b*.

Note that designing a camera lens in two configurations (wide open and stopped down about 50 percent) is one way of getting this "stable focus" kind of balance. Other ways are to design at the paraxial focus (i.e., allow no defocus) or to simply optimize the lens stopped down a bit from wide open.

The three correction states shown in Fig. 3.6 also indicate the manner in which the spherical aberration is changed when the third-order aberration is changed. This is a typical situation often encountered in lens design—the fifth- (and higher-) order aberrations are relatively stable and difficult to change, but the third order is easily modified (by bending an element, for example). In the figure, all three illustrations have exactly the same amount of fifth-order spherical; the difference between the three is solely in the amount of third-order. Note that in the (upper) longitudinal plots, the change of the aberration from one illustration to the next varies as $y^2$, whereas in the (lower) transverse plots the differences vary as $y^3$.

Although patent designs are customarily shown with the marginal spherical corrected to zero, it is apparent that this is not always the best state of correction.

Note that sagittal fan ray intercept plots are often very similar to the axial-ray plot. This is because the sagittal cross section remains relatively constant compared to the tangential cross section. Thus, depending on which section has the worse oblique problems, one should consider the changes of the appropriate beam cross sections.

Oblique spherical is often the aberration that limits the useful field of view. One may consider oblique spherical to be the result of the change in the beam shape from circular on axis to elliptical as the field obliquity increases—this upsets the axial balance of the spherical. The tangential and sagittal fans are affected differently, thus their defocusing effects may be balanced by third-order astigmatism and Petzval.

### 3.8.2 Chromatic aberration and spherochromatism

Here the question is how to balance the spherochromatism, which typically causes the spherical aberration at short (blue) wavelengths to be overcorrected and that at the long (red) wavelengths to be undercorrected. If the aberration is small (diffraction-limited), the best correction is probably with the chromatic aberration corrected at about the 0.7 zone of the aperture. This means that the central half of the aperture area is undercorrected for color and the outer half of the aperture is overcorrected, as shown in Fig. 3.7*a*. But if the amount of the aberration is large, the spherical overcorrection of the blue marginal ray causes a blue flare and a low contrast in the image. In these circumstances the correction zone can advantageously be moved to (or toward) the marginal zone, as shown in Fig. 3.7*b*. This will probably reduce the resolution somewhat because it increases the size of the core of the image blur, but it improves the contrast significantly and yields a more pleasing image, free of the blue flare and haze. This state of correction is accomplished by increasing the undercorrection of the chromatic aberration of the paraxial rays. This is good for a system where the resolution is much less than the diffraction cutoff frequency.

For camera lenses, Kingslake recommends correcting the chromatic at 0.4 to 0.7 of the aperture because in use they are usually stopped down enough to cut off the marginal blue flare produced by this state of correction. In most situations, this will result in better definition.

Secondary spectrum is discussed in Chap. 6.

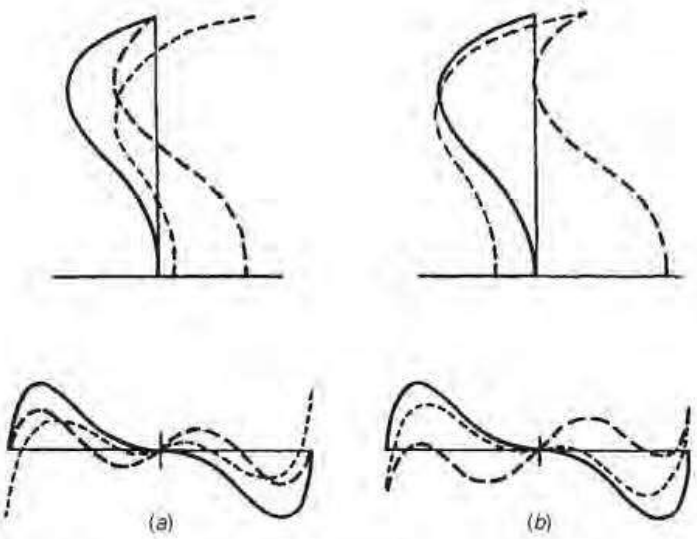



**Figure 3.7** Spherochromatism. (*a*) Chromatic aberration balanced so that the outer half of the aperture is overcorrected and the inner half is undercorrected. This may be best if the amount of aberration is small. (*b*) Chromatic aberration balanced so that it is corrected at the margin. If the aberration is large, this correction eliminates the blue flare that can result from the type of correction in (*a*). Note that the state of correction is more easily perceived in the upper, longitudinal aberration plots, whereas the effect on the blur spot size and flare is much more apparent in the lower transverse ray intercept plots.

### 3.8.3 Astigmatism and Petzval field curvature

In a typical anastigmat lens the fifth-order astigmatism tends to become significantly undercorrected (i.e., negative) as the field angle is increased. In order to minimize the astigmatism over the full field, the third-order astigmatism is overcorrected enough to balance the undercorrected fifth-order astigmatism. The result is the typical field curvature correction with the sagittal focal surface located inside the tangential focal surface in the central part of the field because of the overcorrected third-order astigmatism, with the reverse arrangement in the extreme outer portions. The field angle at which the *s* and *t* fields cross (i.e., where the astigmatism is zero) is called the *node*. Usually the two fields separate very rapidly outside the node, and the image quality quickly deteriorates, often suddenly. The Petzval curvature is usually made somewhat negative, so that both fields are slightly inward-curving and the *effective* field is as flat as possible. A typical state of correction is shown in Fig. 3.8.

Note that a field correction with the *s* and *t* focal surfaces spaced equally on either side of the focal plane (so that the compromise "smallest circle

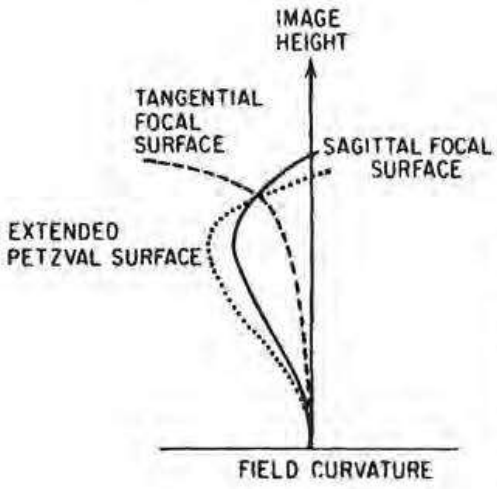



**Figure 3.8** This is the typical balance of astigmatism and Petzval curvature in the presence of undercorrecting fifth-order astigmatism and overcorrecting high-order Petzval curvature. This is achieved by leaving the third-order Petzval slightly inward-curving and overcorrecting the third-order astigmatism by a small amount. This is the usual aberration balance for most anastigmats.

of confusion" focal surface is flat) is definitely *not* the best state of correction. In considering the correction of the field curvature, one should also bear in mind that the oblique spherical aberration (a fifth-order aberration that varies as the cube of the aperture and the square of the field angle) typically goes overcorrected with increasing field angle. In addition, the oblique spherical is usually more significant for the tangential fan of rays than for the sagittal. Thus the *effective* field curvature for the full fan of rays is usually more backward-curving than the $X_s$ and $X_t$ field curves indicate and the *effective* astigmatism is greater. These field curves indicate the imagery of a very small bundle of rays close to the principal ray, and do not take the oblique spherical of the full aperture into account. Thus, for most designs, the astigmatism and field curvature are usually arranged somewhere between the state at which the $s$ and $t$ curves are superimposed (i.e., zero astigmatism) and that at which the $t$ field is approximately flat. Often a through-focus *modulation transfer function* (MTF) plot that includes both on-axis and off-axis plots will indicate quite clearly the effective field curvature; this is more informative than the $X_s$ and $X_t$ curves alone.

Note well that these discussions have assumed the type of higher-order residual aberrations one ordinarily finds—overcorrected fifth-order spherical aberration, undercorrected fifth-order astigmatism, and overcorrected spherochromatism for the shorter wavelengths. Although rare, the reverse is sometimes encountered. In such circumstances the obvious move is to apply the above advice in reverse.

As an additional consideration, note that the undercorrection of either the chromatic aberration or the Petzval curvature has the usually desirable side effect of reducing the power of the elements of the lens system. Thus a secondary benefit of this undercorrection is the reduction of

residual aberrations in general because a lower-power element produces less aberration, which means that there is less higher-order residual aberration left when the aberrations are balanced out.

The usual field curvature plots of $X_s$ and $X_t$ like those in Fig. 3.8 are based on paraprincipal rays. These plots effectively indicate the imagery through a pinhole aperture at the stop. The through-focus plot of the MTF for several field angles (e.g., Fig. 13.37) can be used to determine the *effective* field curvature for the lens at full (or any other) aperture.

Sometimes increasing the weight on the zonal field targets in the merit function will reduce the "belly" at the zonal field.

### 3.8.4 Coma

Both of the two fifth order coma aberrations (linear and elliptical) can be balanced by third-order coma, but in different ways. Figure 3.9*a* shows a typical balance of linear coma (which varies as $hy^4$), and Fig. 3.9*b* shows the balance of elliptical coma (which varies as $h^3y^2$). Note that linear coma is balanced across the aperture and elliptical coma is balanced across the field.

Although it is ordinarily not a dominant problem, field curvature ($X_s$ and $X_t$) can vary with wavelength. This can be detected by inspection of the off-axis ray intercept plots, or the field curvature can be plotted in different wavelengths. This is obviously a longitudinal chromatic which varies with field angle.

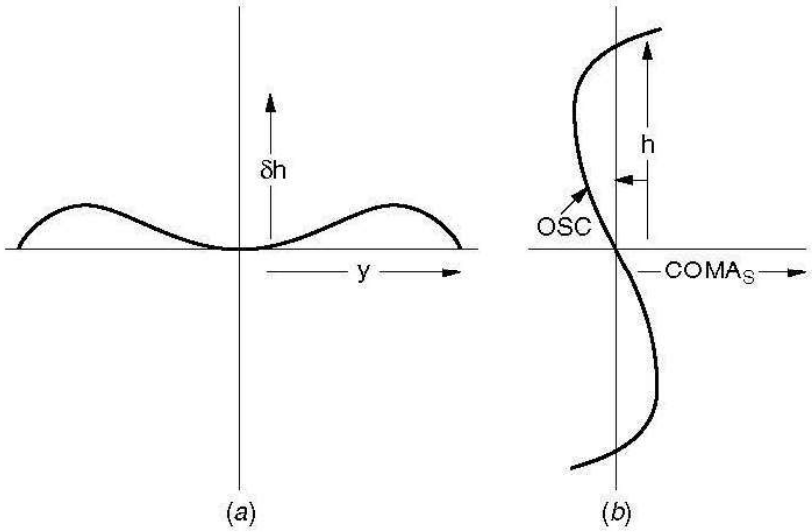



**Figure 3.9** (*a*) A ray intercept plot showing undercorrected fifth-order linear coma (which varies as $hy^4$) balanced by overcorrected third-order coma (which varies as $hy^2$). (*b*) A plot of coma versus field height showing overcorrected fifth-order elliptical coma (which varies as $h^3y^2$) balanced by undercorrected third-order coma (which varies as $hy^2$). Note that OSC = (sagittal coma)/$h$, as $h$ approaches zero.

## 3.9 The Symmetrical Principle

When an optical system has mirror (i.e., front-to-back) symmetry about the aperture stop (or a pupil), as shown in Fig. 3.10, the system is free of coma, distortion, and lateral color. This results from the fact that the components on one side of the stop have aberrations that exactly cancel the aberrations from the components on the other side of the stop. Obviously, to have mirror symmetry, the system must work at unit magnification, with equal object and image distances. For the symmetry to be *absolutely* complete, the object and image surfaces must be identical in shape; this then would imply separately curved sagittal and tangential surfaces at both object and image. (This is why the best 1:1 design is often not completely symmetrical.) However, the third-order coma, distortion, and lateral color are completely removed by symmetry, even with flat object and image surfaces.

Of course, most systems do not operate at unit magnification, and therefore a symmetrical construction of the lens will not completely eliminate these aberrations; however, even for a lens with an infinitely distant object, these aberrations are markedly reduced by symmetry, or even by an approximately symmetrical construction. This is why so many optical systems that cover a significant angular field display a rough symmetry of construction. Consider the Cooke triplet—it has outer crown elements that are similar but not identical in shape, and the center flint, while not equiconcave, is biconcave, and, except for slow-speed triplets, the airspaces are quite similar in size. The benefit of this is that the higher-order residuals of coma, distortion, and lateral color are markedly reduced by this symmetry. This is especially true for wide-angle lenses when good distortion correction is important.

A monocentric system, where all surfaces—refractive or reflective—have a common center and where the stop is located at the common center, has no coma, no astigmatism, and no lateral color. And if the object and image are both in a plane through the center, the spherical can also be zero (e.g., Offner and Dyson systems).

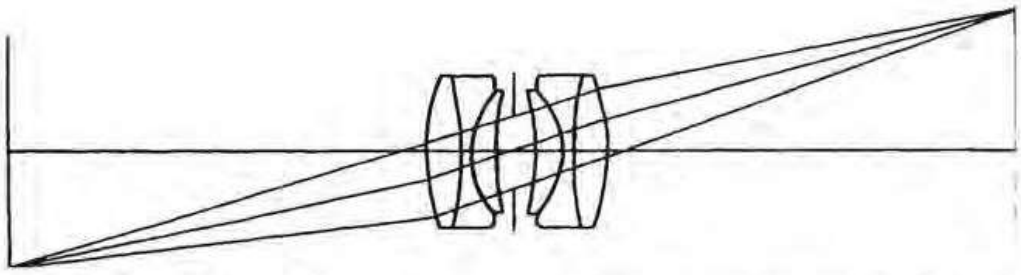


**Figure 3.10** A fully (left to right) symmetrical system is completely free of coma, distortion, and lateral color because the aberration in one half of the system exactly cancels out the aberration in the other half.

Before you decide that symmetry or monocentricity is always a blessing, remember that the big steps in progressing from the *f*/4 Plasmat lens to the *f*/2 double-Gauss lens were (*a*) dropping symmetry and (*b*) increasing the doublet thicknesses.

## 3.10 Aspheric Surfaces

Many designs can be improved by the use of one or more aspheric surfaces. Except for the case of a molded or diamond-turned element, an aspheric surface is many times more expensive to fabricate than a spherical surface. A conic aspheric is easier to test than a general aspheric and is therefore apt to be somewhat less costly. For many systems, e.g., mirror objectives, aspheric surfaces are essential to the design and cannot be avoided.

One technique for introducing an aspheric into an optical system is to first vary only the conic constant or the fourth-order deformation coefficient. (Note that the conic constant and the fourth-order deformation term have exactly the same effect on the third-order aberrations. Thus, allowing both to vary in an automatic design program may cause a slowing of the convergence or, in extreme cases, a failure of the process. Occasionally the difference between the effect of the conic and the fourth-order term on the fifth- and higher-order aberrations may be useful in a design, but more often than not one of the two is redundant.) If the effect of varying the conic constant alone is inadequate, one can then allow the sixth-order term to vary, then the eighth-order, and so on. Some designs have aspherics specified to the tenth-order term when just the sixth or eighth would suffice. It is a good idea to calculate the surface deformation caused by the highest-order term used; if it is a fraction of a wave at the edge of the surface aperture, its utility may well be totally imaginary.

Occasionally one encounters a design specification or print in which the aspheric is specified by a tabulation of sagittal heights instead of an equation. The optimization program can be used to fit the constants of the standard aspheric surface equation (Eq. (5.1)) to the tabulated data. The specification table is entered in the merit function as the sag of the intersections of (collimated) rays at the appropriate heights. The surface coefficients are allowed to vary, and the result is a least-squares fit to the sag table.

The equations of Chap. 24, Sec. 11 indicate the effects of a conic or a fourth-order aspheric term on the third-order aberrations. Several points are worthy of note. The aspheric has no effect on the Petzval curvature, on power, or on axial or lateral chromatic, although it can change the chromatic variation of aberrations. Further, if it is located at the aperture stop or at a pupil, then the principal ray height $y_p$ is zero

and the aspheric has no effect on third-order coma, astigmatism, or distortion; it can only affect spherical. In the Schmidt camera (Fig. 18.15) the aspheric surface is located at the stop because the coma and astigmatism are already zero (the stop being at the center of curvature of the spherical mirror). The purpose of the aspheric is to change *only* the spherical aberration. Conversely, if the purpose of an aspheric is to affect the third-order coma, astigmatism, or distortion, then it must be located at a significant distance from the aperture stop. Even if located at the stop, an aspheric can affect the high-order off-axis aberrations.

It is also worth noting that the primary effect of the conic, or fourth-order, deformation term is on the third-order aberrations. The primary effect of the sixth-order deformation term is on the fifth-order aberrations, and so on.

Note that a conic constant is meaningless on a plane (cv = 0.0) surface; a fourth-order deformation term must be used instead. Conics and the fourth-order aspheric terms mostly affect the third-order aberrations. Do not vary the conic constant and the fourth-order coefficient at the same time; they have exactly the same effect on the Seidel aberrations.

Common uses of aspheric surfaces:

1. To correct spherical aberration, as in the Schmidt telescope
2. To control astigmatism and distortion in wide angle eyepieces
3. To correct distortion in mapping or measuring systems
4. To correct pupil spherical in complex visual systems

In using an aspheric, more rays are needed than for a system with only spherical surfaces. Shannon (Ref. 22) suggests rays at 0.7, 0.85, 0.9, and 0.95 for a 10th order aspheric. In any case, one must watch out for big residuals between the rays.

embodiments now are quite good, and making minor design changes to conceal the best version is simply not worth the trouble (and there are legal reasons as well).

In most cases the published designs have been modified to some extent. For the majority of the designs, we have specified the optical glass as one of those from the Schott (Schott Glass Technologies, Inc.) catalog. We have chosen what we feel is the nearest Schott glass to that indicated in the source for the lens data. Occasionally this may constitute a significant change, but we have attempted to stay as close to the original data as possible. In a few designs, non-Schott glasses have been used. Note that many glass catalogs are currently in a state of flux; as the contents are changed, the number of glasses is reduced, and the glasses are reformulated to reduce density, improve durability, and eliminate lead, arsenic, and cadmium. For the newer designs we have used the standard six-digit code to identify glass types, rather than listing the index and *V*-value. For example, the six-digit code for LaK12 in Fig. 5.1 would be 678552, indicating an index of 1.678 and a *V*-value of 55.2.

In the first edition of this book the lens design figures were produced by a special, custom version of the lens design program GENII. This allowed us to produce all the significant aberration plots on one sheet. In this edition the new lens figures were done with the current version of OSLO, which includes all the features for which we needed a custom program just a decade ago. The new aberration plots are similar enough

F/4.5 25.2 deg Triplet US 1.987,878/1935 Schneider

| radius | thickness | mat'l | index | V-no | sa |
|---|---|---|---|---|---|
| 26.160 | 4.916 | LAK12 | 1.678 | 55.2 | 11.7 |
| 1201.700 | 3.988 | air | | | 11.7 |
| –83.460 | 1.038 | SF2 | 1.648 | 33.8 | 10.2 |
| 25.670 | 4.000 | air | | | 10.2 |
| | 6.925 | air | 1.651 | 55.9 | 9.2 |
| 302.610 | 2.567 | LAK22 | | | 10.3 |
| –54.790 | 81.433 | air | | | 10.3 |

| | | |
|---|---|---|
| EFL | = 98.56 | = Effective focal length |
| BFL | = 81.43 | = Back focal length |
| NA | = –0.1127 (F/4.4) | = Numerical aperture (F-number) |
| GIH | = 46.33 (HFOV = 25.17) | = Image height (half field in degrees) |
| PTZ/F | = –2.831 | = (Petzval radius) /F |
| VL | = 23.43 | = Vertex length |
| OD | infinite conjugate | = Object distance |

**Figure 5.1** Sample lens prescription.

to those in the style of the first edition that the interpretation should be obvious.

The aperture and field that are indicated for any given lens design are more a matter of taste than anything else. What constitutes an acceptable level of aberration depends mostly on the application to which the lens is to be put. Thus the values for field and aperture that accompany each design in this book have often been selected somewhat arbitrarily to yield a level of correction we thought reasonable.

The choice of the clear apertures for the lens elements is equally arbitrary. Obviously, the clear aperture of an element cannot be so large that the edge thickness at that diameter becomes negative or impractically thin. We have selected what seemed to be reasonable values for the clear apertures, based on both edge thickness considerations and the choice of a vignetting factor that allows a reasonably sized oblique beam through the lens and also trims the oblique beam to eliminate the worst-behaved rays.

## 5.2 Lens Prescriptions, Drawings, and Aberration Plots

The lens design data and the associated graphics for this book have been produced by computer. While data input errors and other glitches are always possible, by producing the lens data table, the lens drawing, and the aberration plots all from the same lens data file, we hope to prevent most of the errors that have afflicted some other efforts of this type. For the new designs in this edition (which represent a variety of applications), we have departed from the practice of always scaling the focal length of each lens to 100 mm. We have instead used focal lengths appropriate to the application that the design process story is illustrating. We have also incorporated the design data into the text because, for the most part, the new design data is integral to the description of the design process in the text.

### 5.2.1 Lens prescription

A sample lens data table is shown in Fig. 5.1. The lens construction data are tabulated in a quite straightforward way. The columns are headed *radius, thickness, mat'l, index, V-no,* and *sa* (for semiaperture); the meanings should be apparent. The radius value follows the usual sign convention that a positive radius has its center of curvature to the right of the surface. Plano surfaces (i.e., with infinite radius) are indicated by a blank entry in the radius column. The thickness and material following a surface are presented on the same line as the surface radius, and have the same number.

With few exceptions, the material names are those of Schott Glass Technologies, Inc. The index and $V$ number values correspond to the wavelengths given with the ray intercept plot (e.g., see Fig. 5.3); for most lenses we have used the $d$, $F$, and $C$ lines. The location of the aperture stop is indicated by a blank in the radius column with air on both sides of the surface. Aspheric surfaces are specified by the conic constant kappa and/or the aspheric deformation coefficients. The equation for the surface is

$$z = \frac{cy^2}{1+[1-(1+\kappa)c^2y^2]^{1/2}} + \mathrm{AD}y^4 + \mathrm{AE}y^6 + \mathrm{AF}y^8 + \mathrm{AG}y^{10} \qquad (5.1)$$

where $c$ = surface curvature ($c = 1/r$)
$y$ = radial distance from the axis
$\kappa$ = conic constant kappa
AD, AE, ... = fourth, sixth, ... order deformation coefficients

The conic constant is described as follows:

| | |
|---|---|
| $\kappa > 0$ | An oblate spheroid; an ellipse with the focal points not on the axis |
| $\kappa = 0$ | A sphere |
| $0 > \kappa > -1$ | A prolate spheroid; an ellipse with the focal points on the axis |
| $\kappa = -1$ | A paraboloid |
| $\kappa < -1$ | A hyperboloid |

Note that the first term of Eq. (5.1) is the equation of a conic section; if $\kappa = 0.0$, it is the equation of a sphere.

The data below the prescription tabulation has the following meanings:

| | |
|---|---|
| EFL | Effective focal length |
| BFL | Back focal length (the distance from the last surface to the paraxial focal point) |
| NA | Numerical aperture (the corresponding $f$ number is in parentheses) |
| GIH | Gaussian (paraxial) image height (half-field in degrees is in parentheses) |
| PTZ/$F$ | Petzval radius as a fraction of EFL |
| VL | Vertex length from first to last surface |
| OD | Object distance |

Equation (5.1) can simulate a cone surface (axicon) with an extreme hyperbola; use a very large curvature and set $\kappa = 1/\cos^2\theta$, where $\theta$ is the half angle of the cone.

### 5.2.2 Lens drawing

A sample lens drawing is shown in Fig. 5.2. The scale of the lens drawing is indicated by the dimensioned length of the line in the lens sketch. The rays in the sketch are the marginal and principal rays corresponding to the aperture and field angle, which are tabulated with the lens data. The aperture stop location is indicated by the point at which the principal ray crosses the optical axis. The lens elements are drawn to the clear apertures given in the prescription table as *sa*, the semiaperture.

Additional rays can easily be added to the lens drawing if desired by using a technique that is exact only for paraxial rays and is often accurate enough for use in estimating or drawing ray paths. A ray may be scaled by simply multiplying the heights at which it strikes the surfaces by a scaling constant. Also, rays may be added by adding their intersection heights together. In each case the result is a reasonable approximation to the path of another ray. Obviously, two rays can be scaled and then added. Thus any desired third ray can be drawn by determining its intersection heights from

$$Y_3 = AY_1 + BY_2 \tag{5.2}$$

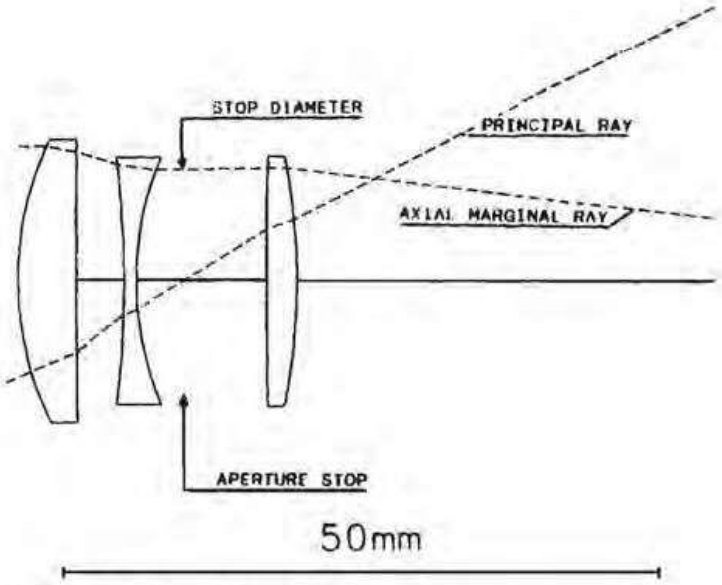



Figure 5.2 Sample lens drawing.

where $A$ and $B$ are scaling factors (see Chap. 24, Sec. 4) and $Y_1$ and $Y_2$ are the ray heights of the rays in the lens drawing. If one defines the desired third ray by its intersection with any two surfaces (which may include the object or image surface), then a simultaneous solution for $A$ and $B$ may be found from the two equations that result when the appropriate values of $Y_1$, $Y_2$, and $Y_3$ are substituted into the equation above. See Eqs. 24.40 through 24.43.

### 5.2.3 Aberration plots

A sample aberration plot is shown in Fig. 5.3. The aberration plots include both tangential and sagittal ray intercept plots (sometimes called $H$-tan $U$ curves) for the axis, 0.7 field, and full field. The ray displacements are plotted vertically, as a function of the position of the ray in the aperture. The vertical scale is given at the lower end of the vertical bar for the axial plot; the number given is the half length (i.e., from

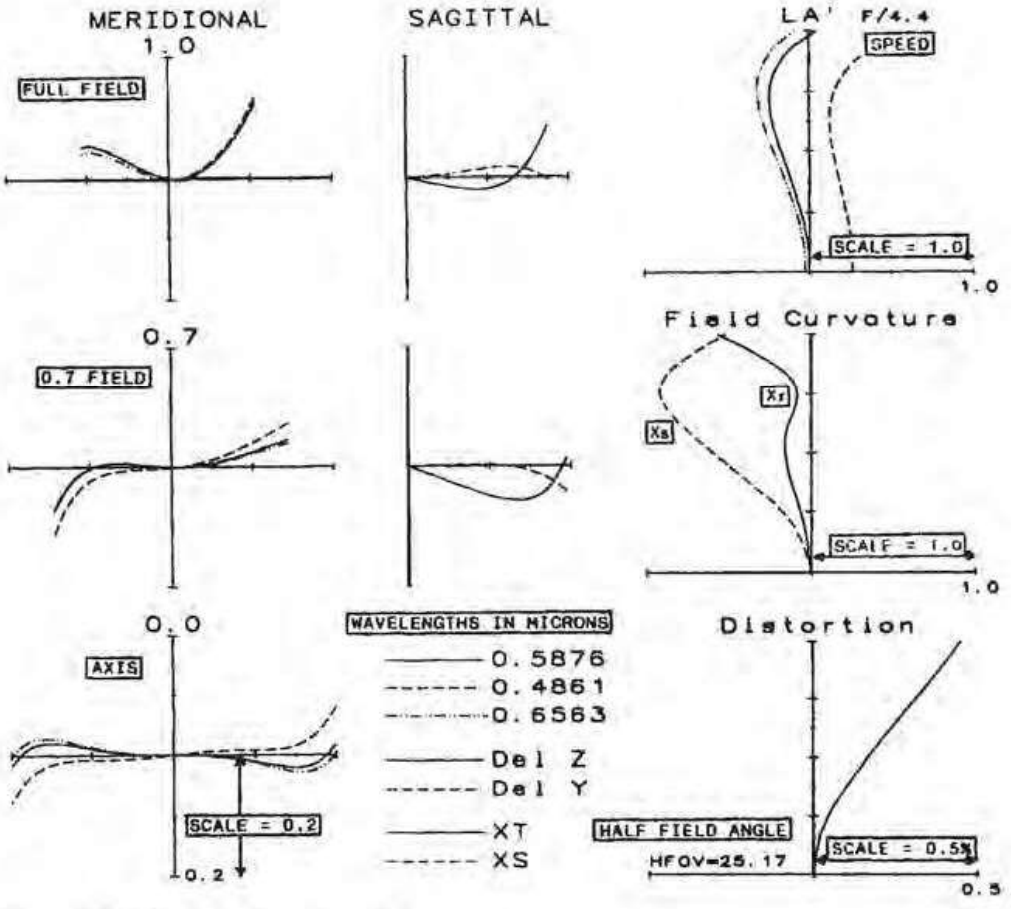



Figure 5.3 Sample aberration plot.

the origin to the end) of the vertical line in the plot. The horizontal scale is proportional to the tangent of the ray slope angle. Following the usual convention, the upper ray of the ray fan is plotted to the right. In the sagittal plots, the solid line is the transverse aberration in the $x$, or sagittal, direction and the dashed line is the ray displacement in the $y$ direction (which is the sagittal coma).

The ends of the ray intercept plot are determined either by the aperture stop or by the vignetting. An arrowhead on the horizontal scale line of the plot indicates that the plot was terminated because the intersection height went offscale.

In addition to the ray intercept plots (which are, in general, probably the most broadly useful presentation of the aberration characteristics of a design), two aberrations are also presented as longitudinal plots. The longitudinal representations of spherical aberration and field curvature have been the classical, conventional presentation for decades, despite the fact that they give a very incomplete picture of the state of correction of the lens; however, a longitudinal plot of the spherical aberration in three wavelengths does allow a much clearer understanding of the spherochromatism, as well as the secondary spectrum. The scale factor for this plot is the number given at the right end of the horizontal axis; the number is the half length of the horizontal line. The vertical dimension of the plot is the height of the ray at the pupil; the $f$ number is given at the top of the plot. This is the $f$ number of the imaging cone and is equal to 1/2NA. The longitudinal field curvature plots yield an excellent picture of the correction of the Petzval curvature and the astigmatism. The scale for $X_s$ and $X_t$ is given at the right end of the horizontal axis; again the number is the half length of the horizontal line. The solid line is $X_t$ and the dashed line is $X_s$. The vertical scale is the fraction of the *gaussian image height* (GIH); the half-field angle is given at the left side of the distortion plot. The scale for distortion is in percent, and the number indicates the half length of the horizontal line.

The field curvature values, $X_s$ and $X_t$, indicate the imagery of an infinitesimal fan of rays about the principal ray, i.e., paraprincipal rays. These are analogous to the paraxial rays about the optical axis. The slope of the ray intercept plots indicates the longitudinal defocus of a pinhole-sized beam of rays about the ray that has been plotted in the ray intercept plot. Note that in the ray intercept plot, a left- or right-shift along the meridional fan plot corresponds to a shift of the stop. See Fig. 2.2.

Spherical aberration is sometimes plotted longitudinally as $\mathrm{LA} = \delta L = (L' - l')$ versus the height of the ray in the pupil. With an object at infinity, the paraxial focal length is $f = -y_1/u_k'$. The "marginal focal length" is

given by $F = -Y_1/\sin U_k'$. The variation of focal length with ray height, $\delta F = (F - f)$, can be plotted in the same diagram as spherical, $\delta L = (L' - l')$. The space between the $\delta L$ and $\delta F$ plots is a measure of OSC, which is equal to $\delta$ (sagittal coma)/$\delta h$, and indicates the amount of coma at small field angles. This plot was widely used to display coma when most aberrations were presented longitudinally.

## 5.3 Estimating the Potential of a Redesign

For an existing design it is relatively easy to estimate the effect on the aberration plots caused by a modest redesign if we apply a knowledge of third-order aberration theory. This is possible because the third-order aberrations of a lens are easily adjusted by changing the spaces or the shapes of the elements, whereas the amount of higher-order aberration tends to be relatively stable and resistant to change.

Equations (5.3) and (5.4) are a power series expansion of the relationships between the ray intersection with the image plane $(x', y')$ as a function of the object height $h$ and the ray position in the pupil (defined in polar coordinates $s$ and $\theta$), as shown in Fig. 5.4.

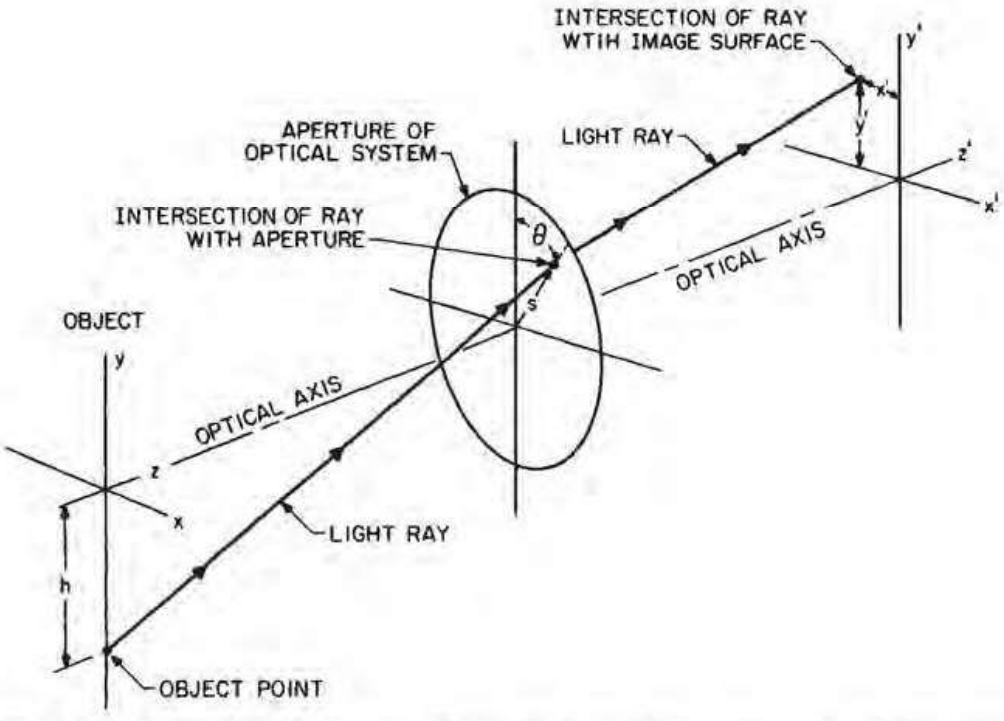



**Figure 5.4** A ray from the point $y = h$, $(x = 0)$ in the object passes through the optical system aperture at a point defined by its polar coordinates $(s, \theta)$ and intersects the image surface at $x', y'$.