Chapter 2

Optical Design Methods

2.1 Mathematical Preliminaries

Merit Function The defect vector $\mathbf{f}$ is a set of m functions f_i that depend on a set of n variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$:

$$\begin{aligned} f_1 &= f_1(x_1, x_2, \ldots, x_n) \\ f_2 &= f_2(x_1, x_2, \ldots, x_n) \\ \vdots &\quad \vdots \\ f_m &= f_m(x_1, x_2, \ldots, x_n) \end{aligned}$$

The merit function is of the type

$$\sigma^2 = \sum_{i=1}^{m} f_i^2 \tag{2.1}$$

or

$$\sigma^2 = \mathbf{f}^\dagger \mathbf{f} = \mathbf{f} \cdot \mathbf{f} \tag{2.2}$$

where $\mathbf{f}$ is a (m x 1) vector and $\mathbf{f}^\dagger$ is the (1 x m) transpose of $\mathbf{f}$. The first form of the expression uses the notation of matrix multiplication. The second form shows a vector dot (or inner) product.

Linear Defect Model Over a small region about the current design, the defects may be approximated by a Taylor series,

$$\mathbf{f} = \mathbf{f}_0 + \mathbf{A}\mathbf{s}. \tag{2.3}$$

where $\mathbf{A}$ is a (m x n) matrix of first derivatives:

$$A_{ij} = \frac{\partial f_i}{\partial x_j} \tag{2.4}$$

and $\mathbf{s}$ are changes in the variables from the current design.

Gradient The gradient $\mathbf{g}$ is a (n x 1) vector given by

$$\mathbf{g} = \frac{1}{2}\nabla\sigma^2 \tag{2.5}$$

Its components are

$$g_i = \frac{1}{2}\frac{\partial\sigma^2}{\partial x_i} = f_1\frac{\partial f_1}{\partial x_i} + f_2\frac{\partial f_2}{\partial x_i} + \cdots + f_M\frac{\partial f_M}{\partial x_i} \tag{2.6}$$

then

$$\mathbf{g} = \mathbf{A}^\dagger\mathbf{f} \tag{2.7}$$

Method of Least-Squares Using the linear model for the defects allows us to express the merit function as

$$\sigma^2 = (\mathbf{f}_0 + \mathbf{A}\mathbf{s})\cdot(\mathbf{f}_0 + \mathbf{A}\mathbf{s}) = \mathbf{f}_0\cdot\mathbf{f}_0 + 2\mathbf{g}_0\cdot\mathbf{s} + \mathbf{s}^\dagger\mathbf{C}\mathbf{s} \tag{2.8}$$

where

$$\begin{aligned}\mathbf{g}_0 &= \mathbf{A}^\dagger\mathbf{f}_0\\ \mathbf{C} &= \mathbf{A}^\dagger\mathbf{A}\end{aligned}$$

Let $\mathbf{a}_j$ represent column j of matrix $\mathbf{A}$. The matrix $\mathbf{C}$ is a symmetric (n x n) matrix, whose elements can be written as a sum over the defects,

$$\begin{aligned}g_j &= \sum_{i=1}^{m} A_{ij}(f_0)_i = \mathbf{a}_j\cdot\mathbf{f}_0\\ C_{jk} &= \sum_{i=1}^{m} A_{ij}A_{ik} = \mathbf{a}_j\cdot\mathbf{a}_k\end{aligned}$$

The matrix $\mathbf{C}$ is called the covariance array. The gradient is

$$\mathbf{g} = \mathbf{A}^\dagger\mathbf{f} = \mathbf{g}_0 + \mathbf{C}\mathbf{s} \tag{2.9}$$

The minimum of σ^2 is obtained by setting $\mathbf{g} = 0$ and solving for $\mathbf{s}$. The resulting matrix equation

$$\mathbf{g}_0 + \mathbf{C}\mathbf{s} = 0 \tag{2.10}$$

is a set of simultaneous linear equations known as the normal equations of least-squares. Providing that the matrix $\mathbf{C}$ is not singular, these equations can always be solved, and the formal solution $\mathbf{s}$ may be written

$$\mathbf{s} = -\mathbf{C}^{-1}\mathbf{g}_0 \tag{2.11}$$

2.2 Design Example 1

In our first numerical example we propose to design a thin lens with specified power and zero coma. We choose this example because we know that a solution exists and we could easily solve for it directly.

Specifications Make the lens focal length 20 mm with an f/2 aperture ($y_a = 5$ mm). Let the half field angle u_c be 0.1 (5.73°) and the wavelength be 0.55 μm. Let the glass index of refraction n_g be 1.5. Assume the object is at infinity ($M = 1$).

Defect Function The design variables are the two surface curvatures. The defect functions are power and coma. The wavefront errors introduced are given by

$$W_{020} = \frac{1}{2\lambda} y_a^2 \delta\phi \tag{2.12}$$

for a change in power $\delta\phi$ from the target value ϕ_o, and

$$W_{131} = \frac{1}{4\lambda} y_a^2 \phi^2 L \left(a_5 B - a_6 M\right) \tag{2.13}$$

for coma. We choose to scale the wavefront values by the common factor of $y_a^2/(2\lambda)$ to make the elements of the partial derivative matrix closer to unity. The defect functions are then given by

$$\begin{aligned} f_1 &= \phi - \phi_o \\ f_2 &= \frac{1}{2}\phi^2 L(a_5 B - a_6 M) \end{aligned} \tag{2.14}$$

The defect function for this design is contained in a Matlab function `sing1.m` which returns a (2x1) defect vector from a (2x1) input (variable) vector.

Derivatives Derivatives of the defect functions may be calculated either analytically or numerically. In optical design, numerical differences are commonly used. Our technique will be to use the following central difference formula.

$$A_{ij} = \frac{\partial f_i}{\partial x_j} \approx \frac{f_i(x_j + h_j) - f_i(x_j - h_j)}{2h_j} \tag{2.15}$$

where h_j is a small change in the variable x_j from its current value. In this example $h = 0.001$. The linear defect model can be obtained from the Matlab function `calculate_derivatives`. For example, the expression

```
v = [ 0.25 -0.15]';
[A fz] = calculate_derivatives(@sing1,v);
```

will calculate the linear defect model at the point $\mathbf{x} = \mathbf{v}$.

Iteration 1 We select a starting point of $c_1 = 0.25$ and c_2 = -0.15. The starting value of the merit function is $\sigma = 0.1511$. The linear defect model $\mathbf{f} = \mathbf{f}_0 + \mathbf{A}\mathbf{s}$ is

$$\mathbf{f} = \begin{pmatrix} 0.1500 \\ -0.0183 \end{pmatrix} + \begin{pmatrix} 0.5 & -0.5 \\ -0.029167 & 0.19583 \end{pmatrix} \mathbf{s} \tag{2.16}$$

We do not need to use the method of least-squares to solve this equation for $\mathbf{f} = 0$ because the number of variables equals the number of defects. The Matlab expression `s = A\fz` will give the solution in either case. The solution

$$\mathbf{s} = \begin{pmatrix} -0.2425 \\ 0.0575 \end{pmatrix} \tag{2.17}$$

is a vector from the starting point to the improved design. The end point, given by `x = x + s`, is then $c_1 = 0.0075$ and c_2 = -0.0925. The merit function at the end point is $\sigma = 0.003437$.

Iteration 2 The linear defect model for the second iteration is

$$\mathbf{f} = \begin{pmatrix} 0.0 \\ -0.0034 \end{pmatrix} + \begin{pmatrix} 0.5 & -0.5 \\ -0.030208 & 0.071875 \end{pmatrix} \mathbf{s}. \tag{2.18}$$

The solution is

$$\mathbf{s} = \begin{pmatrix} 0.0825 \\ 0.0825 \end{pmatrix}. \tag{2.19}$$

The end point is $c_1 = 0.09$ and c_2 = -0.01. The merit function at the end point is zero, so no further iterations are required.

Summary Fig. 2.1 shows a contour plot of the merit function in the vicinity of the solution. A logrithmic transformation of the merit function has been applied, of the form

$$\sigma' = \log_{10}(\sigma + \epsilon) - \log_{10}(\epsilon) \tag{2.20}$$

where ϵ prevents a negative infinity as $\sigma \to 0$. $\epsilon = 10^{-6}$ in Fig. 2.1. The contour lines are spaced by 5 dB, and the maximum contour range is 46 dB. The heavy line segments show the steps in the iterative solution.

2.3 Singular Value Decomposition (SVD)

Although we did not need to use to least-squares methods in the first example, *singular value decomposition* or SVD provides additional insight into the nature of the design process. SVD is a set of techniques for dealing with sets of equations that are nearly singular. It is the preferred method for solving most linear least squares problems.

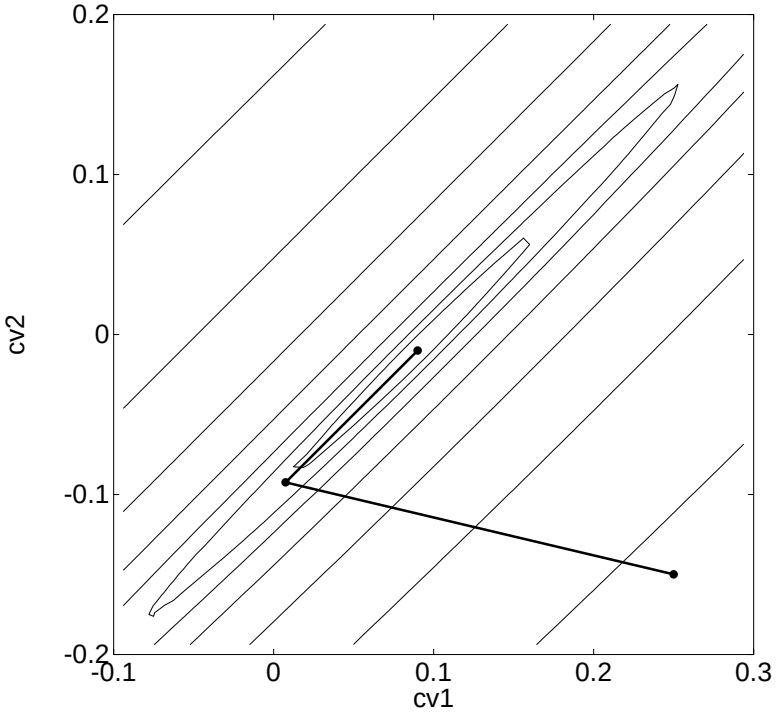

Figure 2.1: Merit function contours for design example 1. Heavy line segments show steps in iterative solution.

The SVD method is based on the following theorem of linear algebra. Any m x n matrix $\mathbf{A}$, where the number of rows m is greater than or equal to the number of columns n, can be written as the product of a m x n column-orthogonal matrix $\mathbf{U}$, a n x n diagonal matrix $\mathbf{W}$ with positive or zero elements, and the transpose of a n x n orthogonal matrix $\mathbf{V}$. This decomposition is given as

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^\dagger \tag{2.21}$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ are each orthognal in that the dot products of the columns are orthonormal. If $\mathbf{u}_i$ is a column of $\mathbf{U}$, for example, then

$$\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij} \tag{2.22}$$

or

$$\begin{aligned} \mathbf{U}^\dagger\mathbf{U} &= \mathbf{I} \\ (n \text{ x } m)(m \text{ x } n) &= (n \text{ x } n) \end{aligned} \tag{2.23}$$

where $\mathbf{I}$ is the identity matrix.

Example 1 The singular value decomposition of the $\mathbf{A}$ matrix given in Eq. 2.16 is found by the Matlab expression `[U W V] = svd(A,0)`, or

$$\mathbf{U}\mathbf{W}\mathbf{V}^\dagger = \begin{pmatrix} -0.9743 & 0.2252 \\ 0.2252 & 0.9743 \end{pmatrix} \begin{pmatrix} 0.7253 & 0 \\ 0 & 0.1149 \end{pmatrix} \begin{pmatrix} -0.6808 & 0.7325 \\ -0.7325 & -0.6808 \end{pmatrix}^\dagger \tag{2.24}$$

2.3.1 SVD Diagrams

The matrix $\mathbf{V}$ has the properties of a rotation matrix, converting the solution vector $\mathbf{s}$ into another vector $\mathbf{s}'$ by

$$\mathbf{s}' = \mathbf{V}^\dagger\mathbf{s}. \tag{2.25}$$

The matrix $\mathbf{U}$ can be interpreted as a projection matrix, converting the defect vector $\mathbf{f}$ of length m into another defect vector $\mathbf{f}'$ of reduced length n by

$$\mathbf{f}' = \mathbf{U}^\dagger\mathbf{f}. \tag{2.26}$$

A projective transformation of a vector reduces the length of the vector. Thus the merit function defined by $\mathbf{f}' \cdot \mathbf{f}'$ is smaller than the function defined by $\mathbf{f} \cdot \mathbf{f}$. The difference is the predicted residual value after optimization.

The transformed defect equations are now

$$\mathbf{f}' = \mathbf{a}'_0 + \mathbf{W}\mathbf{s}', \tag{2.27}$$

which is easy to solve because $\mathbf{W}$ is diagonal. If $\mathbf{w}$ is a n x 1 vector with the diagonal elements of $\mathbf{W}$, then the each element of the solution vector is given by

$$s = -\frac{a'_0}{w} \tag{2.28}$$

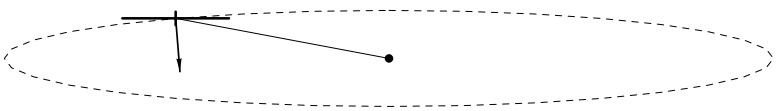

Figure 2.2: Construction of SVD diagram. Dashed line is merit function contour. Arrow is in direction of negative gradient. Medium line shows solution vector, and points to center of ellipse.

Furthermore each element of the gradient vector is given by

$$g = -2wf' \tag{2.29}$$

Example 2 Suppose that the transformed defect equations are given by

$$\mathbf{f}' = \begin{pmatrix} -2 \\ 3 \end{pmatrix} + \begin{pmatrix} 0.25 & 0 \\ 0 & 2 \end{pmatrix} \mathbf{x}'. \tag{2.30}$$

The merit function is then given by

$$\sigma^2 = (-2 + 0.25x)^2 + (3 + 2y)^2. \tag{2.31}$$

The contours of the merit function are ellipses passing through (x, y) and centered at (8, -1.5). The merit function is zero at the center of the ellipse. At the point (0, 0), $\sigma^2 = 13$.

Fig. 2.2 is a diagram showing SVD geometry. The dashed curve is a contour of the merit function. The principal axes, drawn as heavy lines, have a length proportional to $1/w$. The longer axis (major axis) is of fixed length, and the shorter axis is scaled proportionally. The ratio of axis lengths is 1:8 in this example, since that is the ratio of the smallest to largest diagonal elements. The arrow shows the direction of the negative gradient vector, and will always be perpendicular to the contour lines. The light line shows the solution vector, so it is drawn from the starting point to the center of the ellipse.

SVD diagrams like that in Fig. 2.2 can be added to a merit function contour plot. Fig. 2.3 is the same as Fig. 2.1 except that SVD diagrams have been added for each iteration.

2.4 Design Example 2

In this numerical example we add spherical aberration to the list of defects to be corrected. We choose this example because we know that no a solution exists that gives zero spherical aberration, so that defect is nonlinear, and the minimium in spherical aberration does not occur at the same design that gives zero coma, so a balance is implied. The spherical aberration defect is

$$f_3 = \frac{1}{16} y_a^2 \phi^3 (a_1 + a_2(B - a_3 M)^2 - a_4 M^2) \tag{2.32}$$

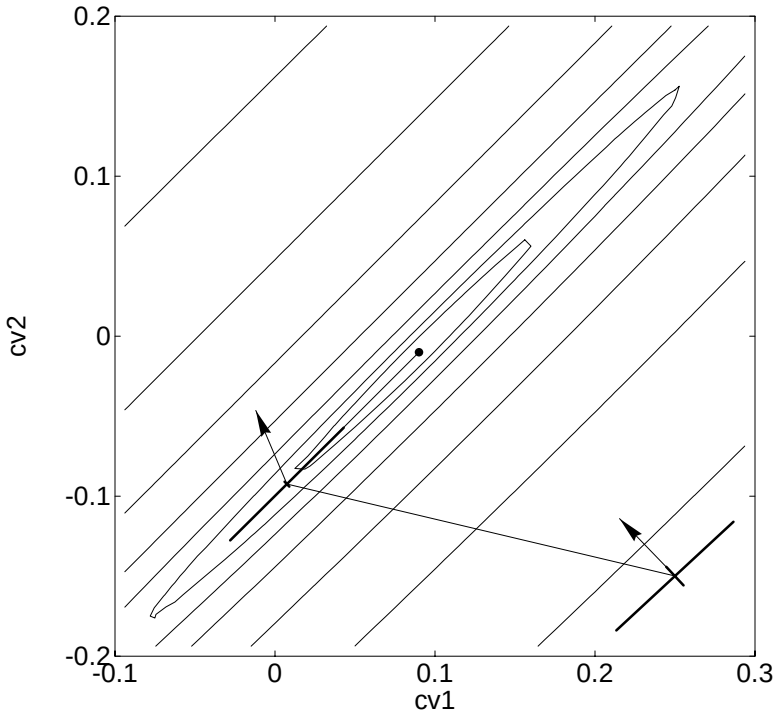

Figure 2.3: Merit function contours for design example 1. SVD diagrams show show steps in iterative solution.

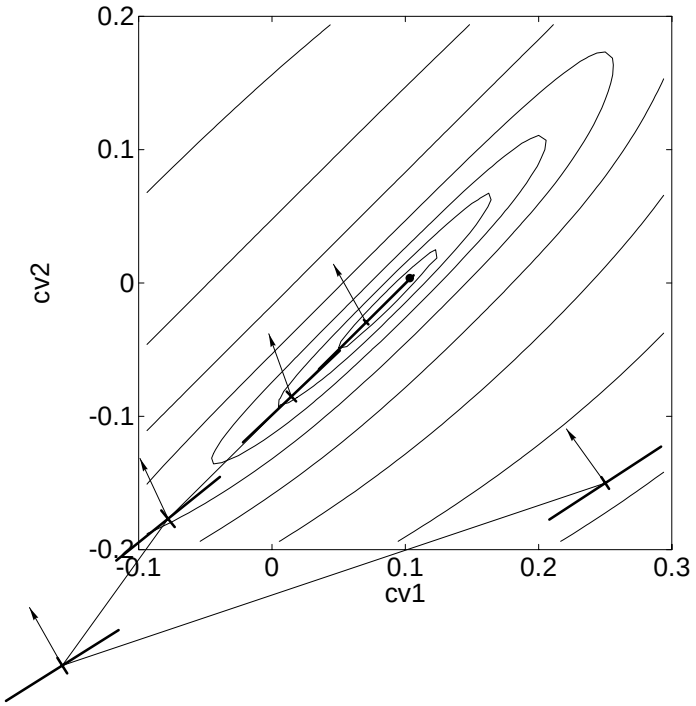

Figure 2.4: Merit function contours and SVD diagrams for design example 2.

A Matlab defect function for this example is contained in `sing2.m`.

Fig. 2.4 is a contour of the norm of the defect function showing an optimization trajectory with five iterations. The nonlinearity of the defect function is evident from the lack of symmetry of the contours. This can readily be seen by viewing the contours obliquely, sighting along the major axis. The following results summarize the SVD matrices for the first, third, and fifth iterations.

Iteration 1 At the starting point ($c_1 = 0.25$, c_2 = -0.15), the partial derivative matrix is

$$\mathbf{A} = \begin{pmatrix} 0.5 & -0.5 \\ -0.0292 & 0.1958 \\ 0.7891 & -1.3307 \end{pmatrix} \tag{2.33}$$

and the SVD equation is

$$\begin{pmatrix} -0.4098 & -0.8321 \\ 0.1059 & -0.4537 \\ -0.9078 & 0.3190 \end{pmatrix} \begin{pmatrix} 1.7030 & 0 \\ 0 & 0.1797 \end{pmatrix} \begin{pmatrix} -0.5416 & -0.8407 \\ -0.8407 & -0.5416 \end{pmatrix}^{\dagger} \mathbf{s} = \begin{pmatrix} -0.1500 \\ 0.0183 \\ -0.1323 \end{pmatrix} \tag{2.34}$$

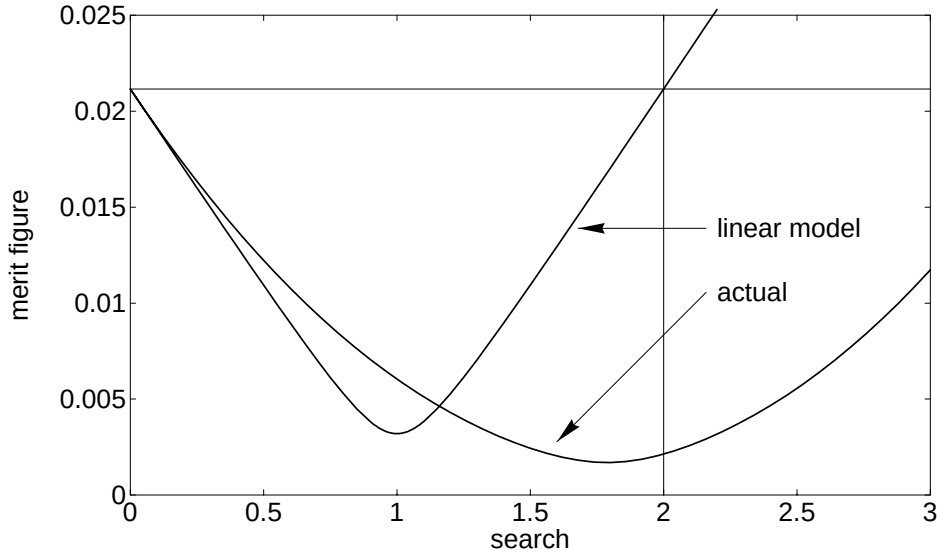

Figure 2.5: One-dimensional search along least-squares direction.

Iteration 3 The starting point is (c_1 = -0.074541, c_2 = -0.173431) and the SVD equation is

$$\begin{pmatrix} -0.8269 & -0.5597 \\ 0.1437 & 0.1159 \\ -0.5437 & 0.8205 \end{pmatrix} \begin{pmatrix} 0.8501 & 0 \\ 0 & 0.1375 \end{pmatrix} \begin{pmatrix} -0.6260 & -0.7799 \\ 0.7799 & -0.6260 \end{pmatrix}^{\dagger} \mathbf{s} = \begin{pmatrix} -0.0006 \\ 0.0067 \\ -0.0199 \end{pmatrix} \tag{2.35}$$

Iteration 5 The starting point is (c_1 = 0.071000, c_2 = -0.029008) and the SVD equation is

$$\begin{pmatrix} -0.9918 & 0.973 \\ 0.0490 & -0.8884 \\ 0.4486 & -0.1181 \end{pmatrix} \begin{pmatrix} 0.7129 & 0 \\ 0 & 0.0330 \end{pmatrix} \begin{pmatrix} -0.7039 & -0.7103 \\ 0.7103 & -0.7039 \end{pmatrix}^{\dagger} \mathbf{s} = \begin{pmatrix} 0.0 \\ -0.0008 \\ 0.0018 \end{pmatrix} \tag{2.36}$$

The linear defect model predicts a quadratic profile for the merit function. Along any one-dimensional projection this profile has the general form

$$\sigma^2 = c_0^2 + c_1^2(x_m - x)^2 \tag{2.37}$$

where c_0 and c_1 are constants. According to our model, the merit function must be symmetric about the minimum $x = x_m$. In practice, this is usually not the case. In Fig. 2.4, for example, the second and subsequent iterations are in the right direction but clearly underestimate the distance to the minimum. In other situations, the least-squares solution may overestimate the distance to a minimum.

The strategy usually employed next is to conduct a one-dimensional search for a minimum in the direction suggested by the least-squares solution and use the distance to the solution as a starting step size. Fig 2.5 shows a plot of such a one-dimensional search, starting at the third SVD diagram in Fig. 2.4. The vertical axis displays the rms value σ of the merit function. The horizonal axis is the search vector $\mathbf{s}$ given by

$$\mathbf{s} = p\mathbf{s}_m \tag{2.38}$$

where $\mathbf{s}_m$ is the least-squares solution vector and p is called the damping parameter. Setting p to 1 yields the least-squares solution vector. According to the linear model the merit function should attain its original value again at $p = 2$. The actual merit function value has its minimum value between p=1.5 and p=2 and has not returned to the original value by p=3.

The merit value σ_1 for the linear defect model as a function of the parameter p is given from the Matlab expression `sigma1 = norm(fz+p*A*s)`. The actual merit value σ_2 is given in Matlab by the expression `sigma2 = norm(sing2(x+p*s))`.

2.5 Damped Least Squares

The general mathematical technique of damped least squares is generally attributed to Levenberg [4] in 1944, but the method has been reinvented, modified, and adapted in various ways by numerous contributors since then.

The basic idea of damped least-squares is to start with the basic equation for the least-squares condition

$$\mathbf{g_0} + \mathbf{C}\mathbf{s} = 0 \tag{2.39}$$

from Eq. 2.10, where $\mathbf{g}_0$ is the gradient at the starting point, and augment the diagonal of the matrix $\mathbf{C}$ by the addition or factoring of a damping coefficient. Modifications of the form $c_{ii} + p$, for example, are called *additive* damping and those of the form $c_{ii}(1 + p)$ are called *multiplicative* damping. In the case of additive damping, the equation for the damped least-squares solution reduces to

$$\mathbf{g_0} + p\mathbf{s} + \mathbf{C}\mathbf{s} = 0 \tag{2.40}$$

As the damping factor p increases, the third term in the equation above becomes small and the solution vector becomes parallel to the gradient vector, that is

$$\mathbf{s} = -\frac{1}{p}\mathbf{g_0} \tag{2.41}$$

For small values of damping, the solution approaches undamped least-squares. Sufficiently large values of damping guarantee a non-singular solution to the often ill-conditioned least-squares system of equations. For large values of damping, the solution approaches a small step in the opposite direction to the gradient. The choice of the best value of p results from a one-dimensional search for a minimum.

Numerous schemes have been explored for improving the effectiveness of the damping coefficient. The most successful of these involve using different damping coefficients for each variable, and making the coefficients proportional to the second derivative.

A Taylor series expansion of a scalar function of several variables may be written as

$$\sigma^2 = \sigma_0^2 + \mathbf{g}_0 \cdot \mathbf{s} + \frac{1}{2}\mathbf{s}^\dagger \mathbf{H} \mathbf{s} \tag{2.42}$$

where $\mathbf{H}$ is the Hessian matrix of second derivatives, defined as

$$H_{jk} = \frac{\partial^2 \sigma^2}{\partial x_j \partial x_k} \tag{2.43}$$

For the least-squares merit function, the Hessian matrix may be expressed as

$$\begin{aligned} H_{jk} &= \sum_{i=1}^{m} \frac{\partial^2 f_i^2}{\partial x_j \partial x_k} \\ &= 2 \sum_{i=1}^{m} \left(\frac{\partial f_i}{\partial x_j} \frac{\partial f_i}{\partial x_j} + f_i \frac{\partial^2 f_i}{\partial x_j \partial x_k} \right) \end{aligned} \tag{2.44}$$

The linear defect model predicts that

$$\sigma^2 = \sigma_0^2 + \mathbf{g}_0 \cdot \mathbf{s} + \mathbf{s}^\dagger \mathbf{C} \mathbf{s} \tag{2.45}$$

and matches the first and second terms of the Taylor series, but it leaves out the second derivatives in the Hessian matrix, which is

$$H_{jk} = 2 \left(C_{jk} + \sum_{i=1}^{m} f_i \frac{\partial^2 f_i}{\partial x_j \partial x_k} \right) \tag{2.46}$$

As an alternative to performing lengthly calculations of the second derivative, a damping term proportional to the *diagonal* elements of the second derivative matrix may be incorporated into the damped least-squares equations. Dilworth [1], for example, has been very successful in using this technique in his optical design software.

2.6 Optimization Tactics

The strategy employed by the traditional damped least-squares method generates a solution vector which changes both in magnitude and in direction as a function of the damping coefficients. Our approach will be to fix the direction of the solution vector and vary its magnitude. We will select either the direction given by least-squares or the direction opposite the gradient.

A single optimization cycle will consist of the following steps.

1. Calculate the first-derivative matrix $\mathbf{A}$ and construct its singular-value decomposition.

2. Set any diagonal elements less than a given tolerance to zero. This eliminates the singular values from the matrix and prevents division by zero in the calculation of the least-squares solution.

3. Choose either the least-squares solution or the gradient solution vector as the test solution vector $\mathbf{s_m}$.

4. Calculate the defect vector for the solution vector $\mathbf{s} = p\mathbf{s}_m$ with $p = 1$. Then try to bracket the minimum by doing one of the following for a fixed maximum number of steps.

   (a) If the resulting merit figure is *less* than the starting merit figure, increase p by a factor of 1.6. Continue to increase p by factors of 1.6 until the merit figure begins to increase again.

   (b) If the resulting merit figure is *greater* than the starting merit figure, decrease p by a factor of 0.4. Continue to decrease p by factors of 0.4 until the merit figure is less than the starting value.

5. Using three values that bracket the minimum, construct a quadratic interpolation of the merit figure and calculate the damping factor corresponding to the vertex (minimum) of the resulting parabola. Evaluate the merit function at that point.

6. Choose the damping factor and corresponding solution vector for the smallest merit figure as the final solution. Return an error message if the merit function has not changed at all.

The choice of 1.6 and 0.4 as expansion and reduction factors, respectively, is not terribly important. These values do correspond approximately to those suggested by a "golden section" search. The exact values would be 1.61803 and 0.38197.

If the linear defect model is approximately correct, the merit function should be nicely parabolic. Then a single step past the predicted minimum will serve to bracket the minimum, and the parabolic interpolation ought to take us in one more step to the minimum itself (if the original solution step of p=1 were not already there).

An elaborate search for the exact minimum is not warranted. If the linear defect model is correct (at least locally), then a single parabolic interpolation should be sufficient. If the actual defect function is non-linear, then we should be seeking to change directions, which requires another SVD calculation.

2.7 Design Example 3

The third design example is the landscape lens, which we encountered earlier when we discussed the stop shift equations. This time we choose to correct the coma W_{131} and astigmatism W_{222} wavefront terms by varying the shape of the lens and the distance to the

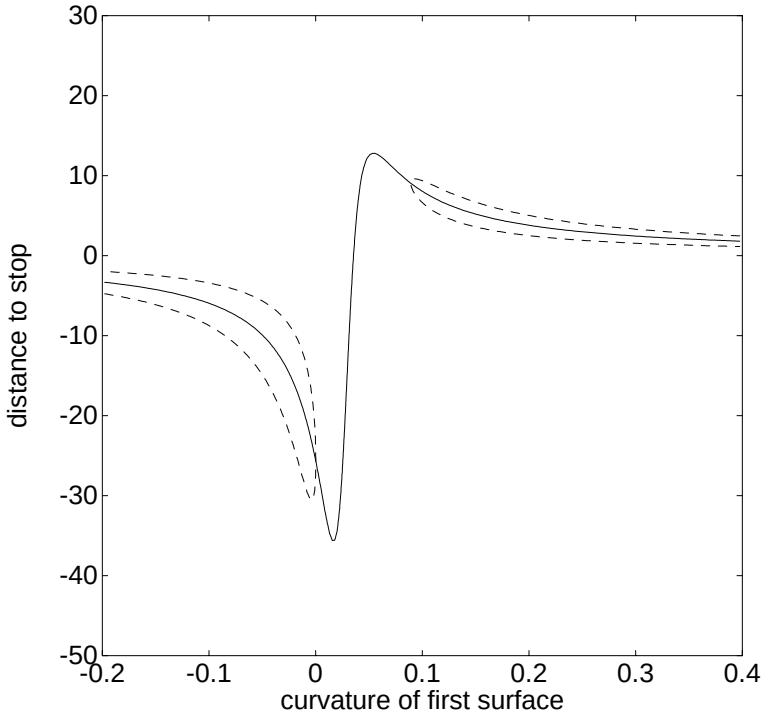

Figure 2.6: Zero contours of W_{131} (solid) and W_{222} (dashed) for landscape lens.

stop. The curvature of the second lens is adjusted by an angle solve appropriate for the desired power of the lens. The Matlab function `land.m` is a defect function for this design.

Fig 2.6 shows the zero contours of W_{131} (solid curve) and W_{222} (dashed curves). There are two solutions, one with the stop in front of the lens, and one with the stop behind the lens. The only control over spherical aberration is to reduce the aperture.

Fig 2.7 shows the merit function contours, which are spaced logarithmically. The variables are the curvature of the first surface (horizontal axis) and the distance from the lens to the stop (vertical axis). The curvature of the second surface is calculated using a paraxial solve to fix the focal length of the lens. The starting point for the optimization is (0.25, -20). The first iteration takes the design into a curved valley. The next few iterations track the valley to its lowest point. The last iteration is a very short correction step to the exact minimum. The progress of optimization is shown graphically in Fig 2.7. The dots are points where an optimization cycle (SVD calculation) is started. The heavy lines show the direction of the least-squares solution based on SVD and the distance to the minimum based on a one-dimensional search in that direction. The last dot is the final end-point.

The data and charts in Fig. 2.8 provide a more detailed record of the progress in optmization. The merit figure was 10^4 at the starting point. The first iteration underestimated the distance to a minimum, resulting in a damping factor of 2.6. There was a dramatic decrease in the merit figure of five orders of magnitude. The next few iterations were characterized

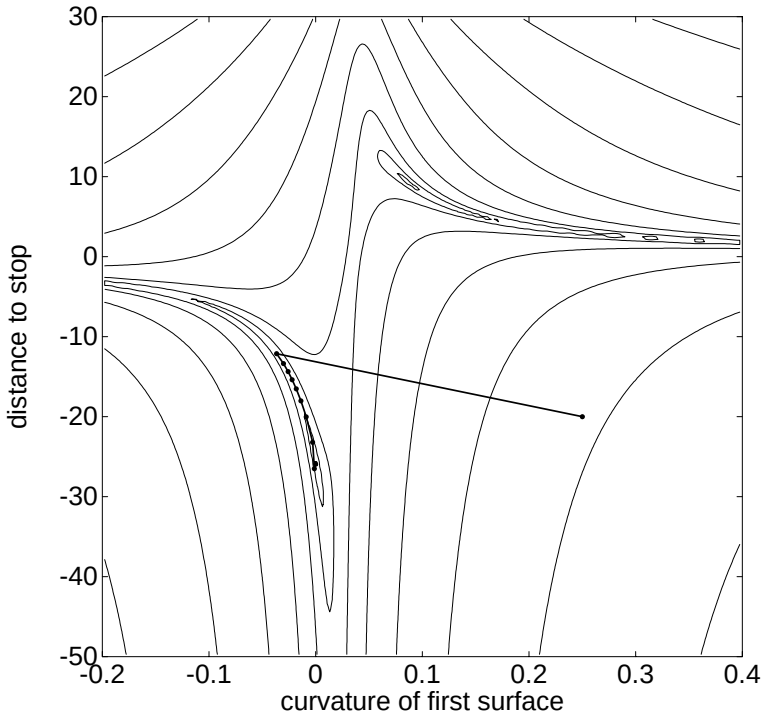

Figure 2.7: Merit function contours and optimization track for landscape lens.

#	damp	σ^2	$\log_{10}(\sigma^2)$ +15
0	0	2.11e04	19.32
1	2.62	6.61e-02	13.82
2	0.041	6.10e-02	13.79
3	0.048	5.71e-02	13.76
4	0.066	5.30e-02	13.72
5	0.085	4.80e-02	13.68
6	0.233	3.19e-02	13.50
7	0.509	1.69e-02	13.23
8	1.24	6.07e-04	11.78
9	1.04	2.95e-07	8.47
10	1.00	1.01e-15	0.00

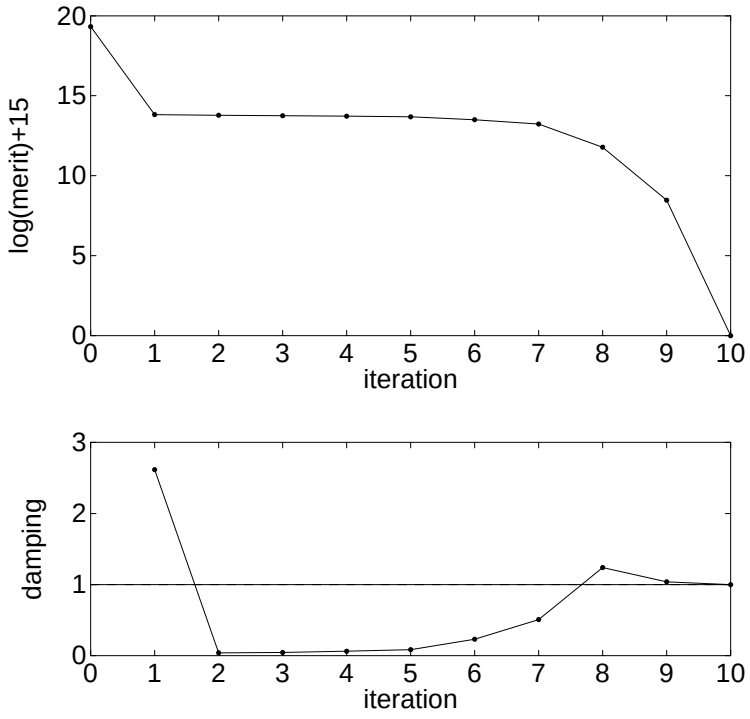

Figure 2.8: Merit function value and damping vs. iteration for landscape lens.

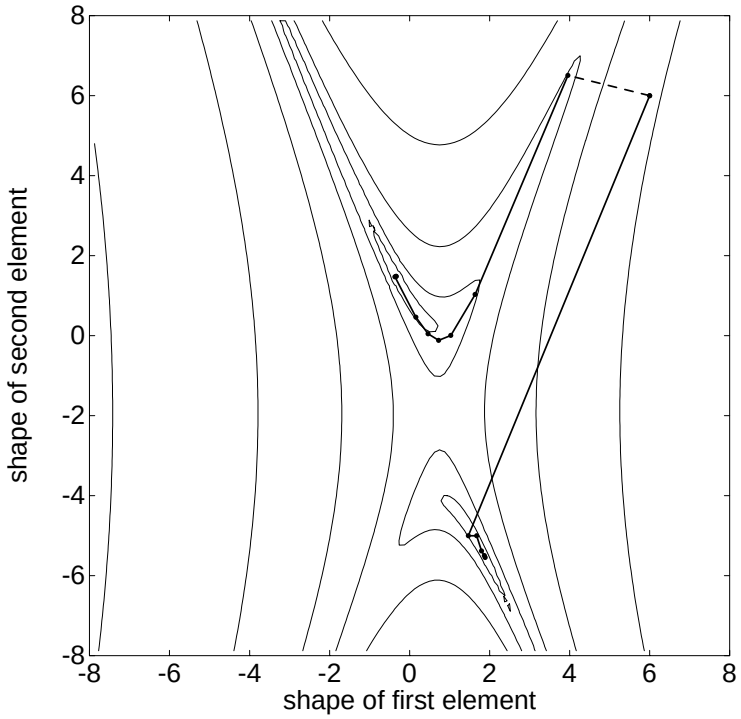

Figure 2.9: Merit function contours and optimization tracks for achromatic doublet lens.

by a high degree of damping (small values of p) and very slow changes in the merit figure. As the minimum was approached, the damping factor approached unity, an indication that the merit function is adequately represented by a linear defect model, and the merit function itself decreased more rapidly with each step. The merit figure at the end point was 10^{-15}.

2.8 Design Example 4

The fourth design example is the achromatic doublet lens, which we studied previously. The powers of the lenses are used to control the focal length and the longitudinal chromatic aberration. The resulting equations are linear and may be solved in one step. The Matlab function `achromat1.m` contains a defect function for the power variables. The shapes of the lenses are used to correct the coma W_{131} and the spherical aberration W_{040}. The Matlab function `achromat2.m` contains a defect function for the shape variables. The shape variables are orthogonal to the power variables. Changing the shape does not alter the solution for the powers.

Fig. 2.9 shows logarithmic merit function contours for the doublet lens as a function of the shapes of the elements. The starting point for optimization was (6, 6). The solid track from (6, 6) is in the direction of the least-squares solution and sends the optimization process

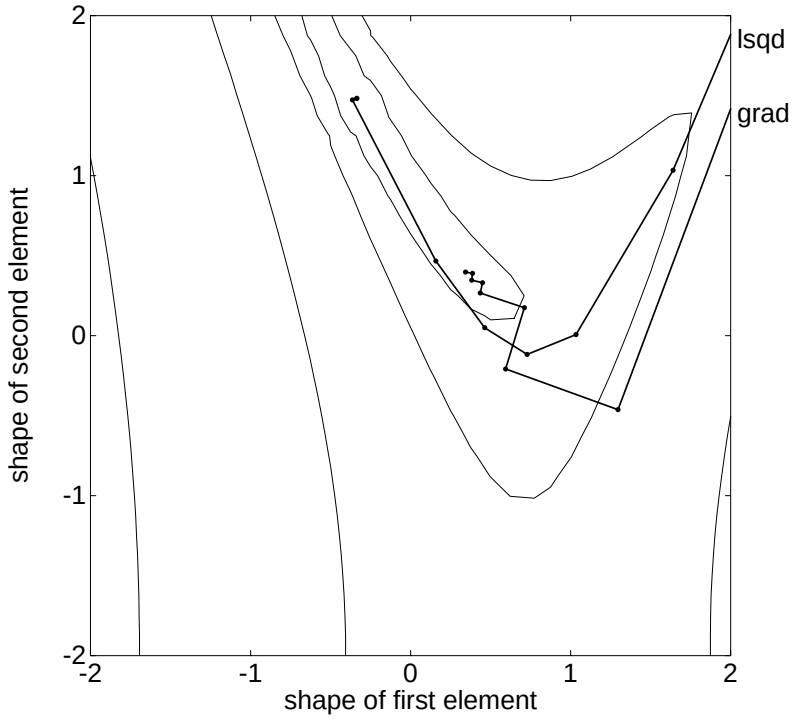

Figure 2.10: Optimization tracks for least-squares and gradient searches for achromatic doublet lens design.

toward the bottom solution. The dashed line from (6, 6) is in the direction of the gradient and sends the optimization process toward the top solution. The second solid track is an optimization sequence using least-squares vectors, but starting from the results of the first gradient step.

Fig. 2.10 magnifies the region of the contour diagram in the vicinity of the top minimum. There are two optimization tracks shown, one following the directions from least-squares solutions and the other directions from gradient solutions. Notice how the gradient search switches direction with each iteration, with each successive step becoming smaller and smaller. Eventually the process will reach the actual minimum, but we lose patience and stop looking long before the minimum is reached. It would be good fortune, indeed, if the gradient happened to point precisely in the direction of the actual minimum. As bad as continued use of the gradient is, however, we must not overlook its occasional usefulness. In this example it started us on the track to the top minimum.

2.9 Design of the Cooke Triplet Lens

The Cooke triplet lens is of interest because it represents the simplest lens form capable of correcting the five fourth-order wavefront aberrations: spherical aberration, coma, astigmatism, field curvature, and distortion; and both lateral and transverse chromatic aberration. The design variables are the power and shape of the three elements and the two element separations. The stop is set at the second lens. One power variable is used to set the focal length of the lens. The remaining seven variables can be used to control the aberrations.

An anomaly in the solution for power is shown in Fig. 2.11. If the power of the first lens ϕ_1 is given by

$$\phi_1 = \frac{1}{t_1} \tag{2.47}$$

where t_1 is the separation between the first and second lens, then the axial ray height $(y_a)_2$ will be zero at the second lens so that it can not be the stop location. Furthermore, wherever the power of the second lens ϕ_2 is given by

$$\phi_2 = \frac{1}{t_2}\frac{1-\phi_1(t_1+t_2)}{1-\phi_1 t_1} \tag{2.48}$$

where t_2 is the separation between the second and third lens, then the axial ray height $(y_a)_3$ will be zero at the third lens so that it can not contribute to the total power of the system.

In the limiting case of three lenses in contact (zero element separation), the distortion and transverse chromatic aberrations are zero. The three powers can be used to control the total power, longitudinal chromatic aberration, and Petzval curvature. This yields a set of three equations in three unknowns:

$$\begin{aligned} \phi_1 + \phi_2 + \phi_3 &= \phi \\ \frac{\phi_1}{n_1} + \frac{\phi_2}{n_2} + \frac{\phi_3}{n_3} &= 0 \\ \frac{\phi_1}{v_1} + \frac{\phi_2}{v_2} + \frac{\phi_3}{v_3} &= 0 \end{aligned} \tag{2.49}$$

where ϕ_i are the element powers, n_i the element indices of refraction, and v_i the corresponding Abbe numbers. The first equation determines the power, the second the Petzval curvature, and the third the longitudinal chromatic aberration. Solutions to this set of linear equations exist provided that three different glass types are chosen. Two of the shape variables can then be used to control spherical aberration and coma. We know this is possible from our study of the achromatic doublet. There is one shape variable left with which to control astigmatism, but the shape of a lens (with the stop at the lens) does not affect astigmatism.

The only way to control astigmatism is to invoke the stop shift equations, as we did in designing the landscape lens. If the stop is located at the second lens, then the outer two lenses form a pair of landscape lenses, one with the stop in front and the other with the stop behind, each of which can be designed to contribute zero astigmatism. All that is required, of course, is that the total astigmatism contribution be zero.

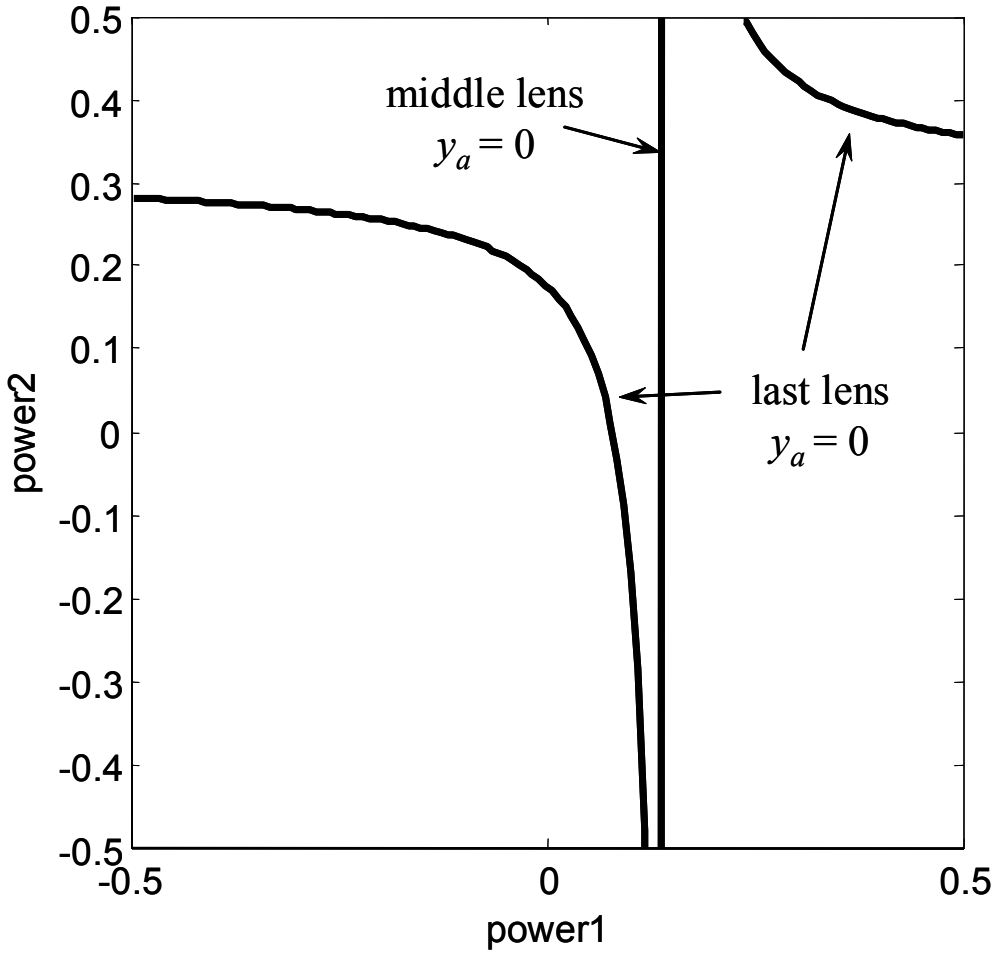

Figure 2.11: Anomalous conditions for power variables in triplet lens.

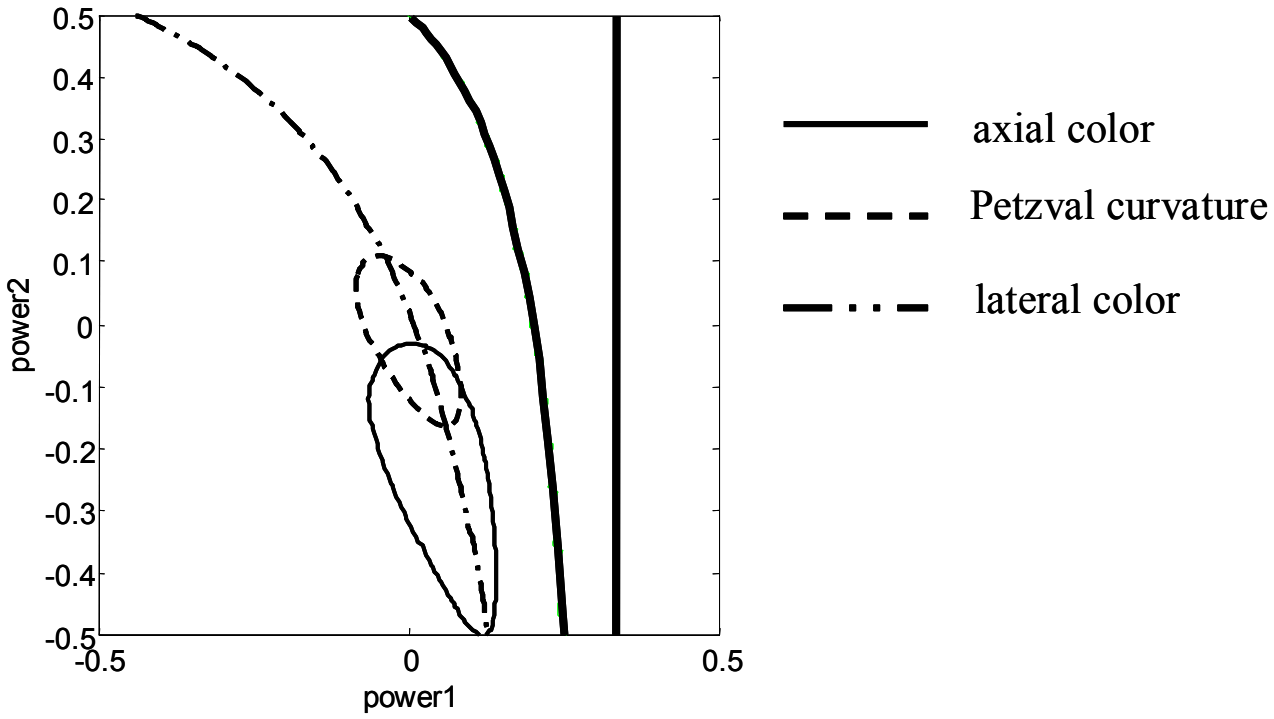

Figure 2.12: Zero-level contours for Petzval curvature and chromatic aberration for fixed lens separations.

In the case of separated lenses, the equations for Petzval curvature, power, and axial chromatic aberration can be written as

$$\begin{aligned} \left(\frac{\phi}{n}\right)_1 + \left(\frac{\phi}{n}\right)_2 + \left(\frac{\phi}{n}\right)_3 &= 0 \\ (y_a\phi)_1 + (y_a\phi)_2 + (y_a\phi)_3 &= y_a\phi \\ \left(\frac{y_a^2\phi}{V}\right)_1 + \left(\frac{y_a^2\phi}{V}\right)_2 + \left(\frac{y_a^2\phi}{V}\right)_3 &= 0 \end{aligned} \tag{2.50}$$

The non-linearity of these equations is shown in Fig. 2.12. The lens separations are fixed at arbitrary, different non-zero values. An angle solve is used to adjust the total power of the lens. Zero contours of Petzval curvature and chromatic aberration (axial and lateral) are shown on the figure. Also shown, as thick curves, are the anomalous conditions of zero y_a at either the second or third lens. Curved contours are an indication of nonlinearity. So is the existence of multiple solutions. In this case, there are six conditions in which two of the three aberrations are zero.

By introducing the stop shift we no longer have a compact triplet, so that distortion and transverse chromatic aberration is induced. Furthermore the focal length, Petzval cur-

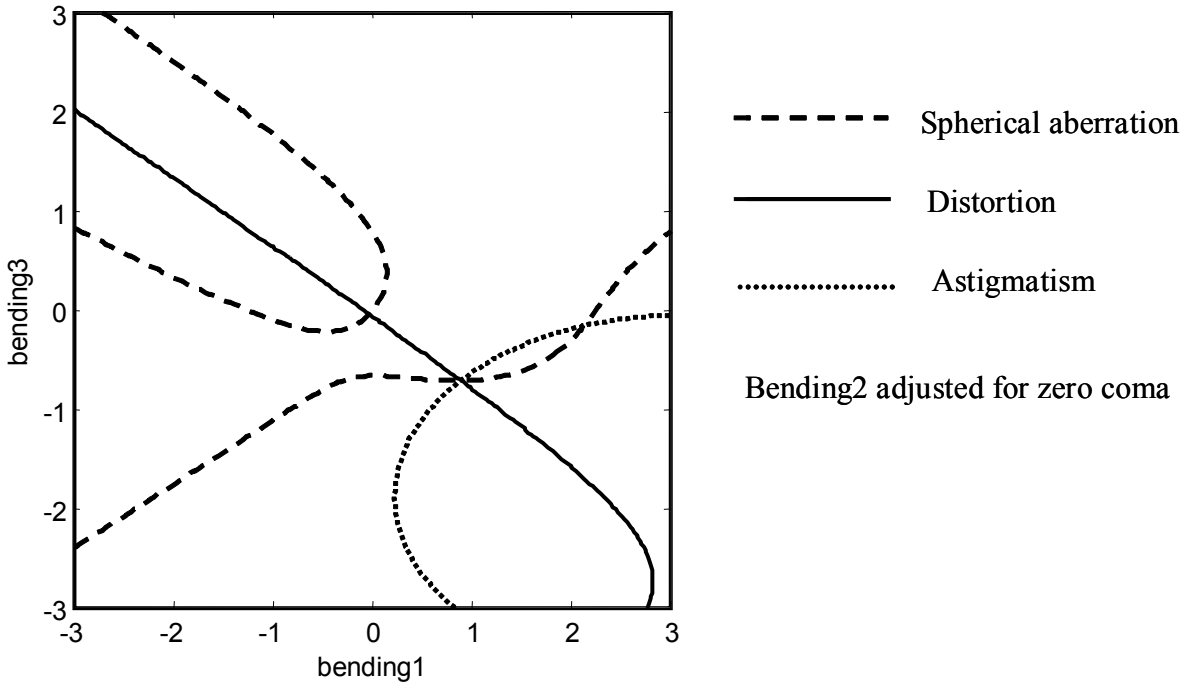

Figure 2.13: Zero-level contours for spherical aberration, distortion, and astigmatism.

vature and longitudinal chromatic aberration depend on the element separations. Finally, the element shapes must be adjusted to maintain zero spherical aberration and coma. As a result, a change in any variable will have some effect on almost all of the aberrations. We are compelled then to use an iterative process to arrive at a simultaneous solution for zero values of all aberrations.

In our previous examples we were able to reduce the designs to problems involving two variables, so we could draw contours of the merit function. In the case of the triplet, there is no true partitioning of variables and aberrations. We may selectively pair variables in various ways and try to plot their interactions, as shown in Fig. 2.12 and Fig. 2.13. The starting point in Fig. 2.13 is one of the two classical solutions to the thin-lens Cooke triplet. The bending of the second lens is adjusted to eliminate coma. Zero contours are shown for spherical aberration, distortion, and astigmatism. The variation in astigmatism with bending arises from the stop shifts for outer lenses. The coincidence of the three zero contours occurs because we started with solution values for the powers and lens separations.

A full parameter search generally finds one of the two classic fourth-order solutions to the triplet lens. These solutions do not make good lenses, however, because they neglect higher-order aberrations. The triplet design is of particular interest more for what it reveals about the design process. One usually sees a few iterations of rapid improvement in the figure of merit, followed by many steps in which the figure of merit changes slowly and the

search is highly damped. Near the final solution, the merit function decreases dramatically in two or three iterations, without damping, until it reaches zero.

A measure of the difficulty of the design can be obtained by looking at the singular value decomposition at the solution point. The singular values range from a maximum of 33000 to a minimum of 0.05 for one of the two classic triplet solutions found.

Finding triplet solutions is by no means a simple activity. From these diagrams alone, for example, can you deduce where to find both solutions to the Cooke triplet?

Bibliography

[1] Donald C. Dilworth, "Pseudo-second-derivative matrix and its application to automatic lens design," Applied Optics, **17**: 3372-3375 (1978).

[2] Donald P. Feder, "Automatic Optical Design," Applied Optics, **2** 1209-1226 (1963).

[3] Robert E. Hopkins, "Optical design 1937 to 1988 ... Where to from here?" Optical Engineering, **27**: 1019-1026 (1988).

[4] K. Levenberg, "A Method for the Solution of Certain Nonlinear Problems in Least Squares," Quart. Appl. Math. **2**: 164-168 (1944).

[5] T. H. Jamieson, *Optimization Techniques in Lens Design*, *Monographcs on Applied Optics*, No. 5, American Elsevier (1971).

[6] Abrahim Lavi and Thomas Vogl, Eds., *Recent Advances in Optimization Techniques*, John Wiley (1965).

[7] William G. Peck, "Automated Lens Design," in *Applied Optics and Engineering*, Vol VIII, Robert R. Shannon and James C. Wyant, (Eds.), Academic Press (1980).

[8] A. K. Rigler and R. J. Pegis, "Optimization Methods in Optics," in *The Computer in Optical Research*, B. R. Frieden (Ed.), *Topics in Applied Physics* Vol. 41, Springer-Verlag, (1980).

[9] David R. Shafer, "The triplet: an 'embarrassment of riches'," Optical Engineering, **27**: 1035-1038 (1988).