Chapter 5

Imaging Systems and Aberrations

At the surface, numerically evaluating imaging systems with monochromatic light is a simple extension of two-dimensional discrete convolution, as discussed in Sec. 3.1. This is because the response of light to an imaging system, whether the light is coherent or incoherent, can be modeled as a linear system. Determining the impulse response of an imaging system is more complicated, particularly when the system does not perfectly focus the image. This happens because of aberrations present in the imaging system. In this chapter, aberrations are treated first. Then, we show how aberrations affect the impulse response of imaging systems. Finally, the chapter finishes with a discussion of imaging system performance.

5.1 Aberrations

The light from an extended object can be treated as a continuum of point sources. Each point source emits rays in all directions as shown in Fig. 5.1. In geometric optics, the rays from a given object point that pass all the way through an ideal imaging system are focused to another point. Each point of the object emits (or reflects) an optical field which becomes a diverging spherical wave at the entrance pupil of the imaging system. To focus the light to a point in the image plane, the imaging system must apply a spherical phase delay to convert a diverging spherical

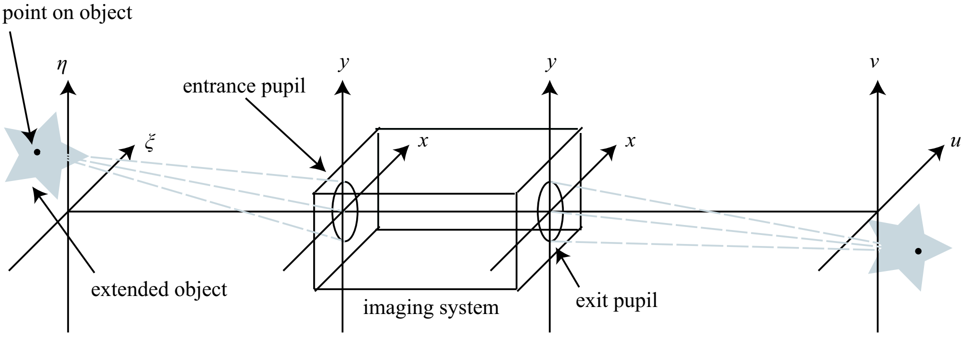


Figure 5.1 Basic model of an imaging system.

Table 5.1 Some Seidel aberration terms and their names.

Term	Name
A_0	piston
$A_1 r\cos\theta + A_2 r\sin\theta$	tilt
$A_3 r^2$	defocus
$A_4 r^2\cos(2\theta) + A_5 r^2\sin(2\theta)$	astigmatism
$A_6 r^3\cos\theta + A_7 r^3\sin\theta$	coma
$A_8 r^4$	spherical aberration

wavefront into a converging spherical wavefront. Aberrations are deviations from the spherical phase delay that cause the rays from a given object point to misfocus and form a finite-sized spot. When the image is viewed as a whole, the aberration manifests itself as blur. Light from different object points can experience different aberrations in the image plane depending on their distance from the optical axis. However, for the purposes of this book, we are not concerned with these field-angle-dependent aberrations but assume that they are constant.

With a detailed description of an imaging system, ray tracing can be used to determine the wavefront aberration for a given object point. Optical design software programs like CODE V,[16] OSLO,[17] and ZEMAX[18] are excellent for this task. In this book, we simply assume that ray tracing has been done already and use the resulting aberration as is. Aberrations can be expressed as a wavefront $W(x,y)$ measured in waves, or optical phase $\phi(x,y) = 2\pi W(x,y)$ measured in radians. Then, we can write a generalized pupil function $\mathcal{P}(x,y)$ by combining the effects of apodization and aberrations into one complex function:

$$\mathcal{P}(x,y) = P(x,y)\, e^{i2\pi W(x,y)}. \tag{5.1}$$

5.1.1 Seidel aberrations

It is common to write an arbitrary wavefront aberration as a polynomial expansion according to

$$\begin{aligned} W(x,y) = A_0 + A_1 r\cos\theta + A_2 r\sin\theta + A_3 r^2 + A_4 r^2\cos(2\theta) \\ + A_5 r^2\sin(2\theta) + A_6 r^3\cos\theta + A_7 r^3\sin\theta + A_8 r^4 + \ldots \end{aligned} \tag{5.2}$$

where r is a polar normalized pupil coordinate. The normalized coordinate is the physical radial coordinate divided by the pupil radius so that $r = 1$ at the edge of the aperture. These terms are classified as shown in Table 5.1. The A_i coefficients may be field-angle-dependent, but we assume that they are constant when imaging simulations are discussed in Sec. 5.2. If each object point experiences different aberrations, then each image of each object point must be simulated separately.

5.1.2 Zernike circle polynomials

The polynomial expansion from the previous section is convenient because of its simplicity, and it follows directly from use of ray tracing. However, its mathematical properties are lacking. When aberrations become complicated, it is better to use a representation that has completeness and orthogonality, so we describe such a representation here. Most of the time, we deal with circular apertures, and the above polynomial expansion is not orthogonal over a circular aperture. However, Zernike circle polynomials are complete and orthogonal over a circular aperture. Note that there are also Zernike annular polynomials that are orthogonal over an annular aperture, Zernike-Gauss circle polynomials that are orthogonal over a Gaussian aperture, and Zernike-Gauss annular polynomials that are orthogonal over Gaussian, annular apertures.[19] There are even Zernike vector polynomials whose dot product is orthonormal over a circular aperture.[20,21] These are all very interesting and useful, but we discuss only Zernike circle polynomials here.

There are several conventions and ordering schemes for defining Zernike circle polynomials.[4,19,22,23] This book uses the convention of Noll.[22] In this convention, the polynomials are defined as

$$Z_n^m\left(r,\theta\right)=\sqrt{2\left(n+1\right)}\,R_n^m\left(r\right)G^m\left(\theta\right), \tag{5.3}$$

where m and n are non-negative integers, and $m \leq n$. However, it is convenient to write $Z_n^m\left(r,\theta\right)$ with just one index

$$Z_i\left(r,\theta\right)=\begin{cases}\sqrt{2\left(n+1\right)}\,R_n^m\left(r\right)G^m\left(\theta\right) & m\neq 0\\ R_n^0\left(r\right) & m=0\end{cases}. \tag{5.4}$$

The mapping of $(n,m)\rightarrow i$ is complicated, but the ordering for the first 36 Zernike polynomials is given in Table 5.2. The radial and azimuthal factors $R_n^m\left(r\right)$ and $G^m\left(\theta\right)$ are given by[23]

$$R_n^m\left(r\right)=\sum_{s=0}^{(n-m)/2}\frac{\left(-1\right)^s\left(n-s\right)!}{s!\left(\frac{n+m}{2}-s\right)!\left(\frac{n-m}{2}-s\right)!}r^{n-2s} \tag{5.5a}$$

$$G^m\left(\theta\right)=\begin{cases}\sin\left(m\theta\right) & i\text{ odd}\\ \cos\left(m\theta\right) & i\text{ even.}\end{cases} \tag{5.5b}$$

Listing 5.1 gives the MATLAB function `zernike` that evaluates Eq. (5.4) given the mode number i and normalized polar coordinates on the unit circle. The reader should note that the factorials in Eq. (5.5) are coded in MATLAB as gamma functions [$s!=\Gamma\left(s+1\right)$] because the `gamma` function executes much faster than the `factorial` function.

Figure 5.2 shows an example of three different Zernike polynomials. The particular aberrations shown are three different orders of x primary astigmatism. In plot (a), $n=2$ and $m=2$; in plot (b), $n=4$ and $m=2$; and in plot (c), $n=6$ and $m=2$. Consequently, all three plots have the same azimuthal dependence, $\cos\left(2\theta\right)$, while the radial dependence is different for each. The largest power on

Table 5.2 The first 36 Zernike polynomials

n	m	i	$Z_n^m\,(r,\theta)$	Name
0	0	1	1	piston
1	1	2	$2\,r\cos\theta$	x tilt
1	1	3	$2\,r\sin\theta$	y tilt
2	0	4	$\sqrt{3}\,\left(2r^2-1\right)$	defocus
2	2	5	$\sqrt{6}\,r^2\sin(2\theta)$	y primary astigmatism
2	2	6	$\sqrt{6}\,r^2\cos(2\theta)$	x primary astigmatism
3	1	7	$\sqrt{8}\,\left(3r^3-2r\right)\sin\theta$	y primary coma
3	1	8	$\sqrt{8}\,\left(3r^3-2r\right)\cos\theta$	x primary coma
3	3	9	$\sqrt{8}\,r^3\sin(3\theta)$	y trefoil
3	3	10	$\sqrt{8}\,r^3\cos(3\theta)$	x trefoil
4	0	11	$\sqrt{5}\,\left(6r^4-6r^2+1\right)$	primary spherical
4	2	12	$\sqrt{10}\,\left(4r^4-3r^2\right)\cos(2\theta)$	x secondary astigmatism
4	2	13	$\sqrt{10}\,\left(4r^4-3r^2\right)\sin(2\theta)$	y secondary astigmatism
4	4	14	$\sqrt{10}\,r^4\cos(4\theta)$	x tetrafoil
4	4	15	$\sqrt{10}\,r^4\sin(4\theta)$	y tetrafoil
5	1	16	$\sqrt{12}\,\left(10r^5-12r^3+3r\right)\cos\theta$	x secondary coma
5	1	17	$\sqrt{12}\,\left(10r^5-12r^3+3r\right)\sin\theta$	y secondary coma
5	3	18	$\sqrt{12}\,\left(5r^5-4r^3\right)\cos(3\theta)$	x secondary trefoil
5	3	19	$\sqrt{12}\,\left(5r^5-4r^3\right)\sin(3\theta)$	y secondary trefoil
5	5	20	$\sqrt{12}\,r^5\cos(5\theta)$	x pentafoil
5	5	21	$\sqrt{12}\,r^5\sin(5\theta)$	y pentafoil
6	0	22	$\sqrt{7}\,\left(20r^6-30r^4+12r^2-1\right)$	secondary spherical
6	2	23	$\sqrt{14}\,\left(15r^6-20r^4+6r^2\right)\sin(2\theta)$	y tertiary astigmatism
6	2	24	$\sqrt{14}\,\left(15r^6-20r^4+6r^2\right)\cos(2\theta)$	x tertiary astigmatism
6	4	25	$\sqrt{14}\,\left(6r^6-5r^4\right)\sin(4\theta)$	y secondary tetrafoil
6	4	26	$\sqrt{14}\,\left(6r^6-5r^4\right)\cos(4\theta)$	x secondary tetrafoil
6	6	27	$\sqrt{14}\,r^6\sin(6\theta)$	
6	6	28	$\sqrt{14}\,r^6\cos(6\theta)$	
7	1	29	$4\left(35r^7-60r^5+30r^3-4r\right)\sin\theta$	y tertiary coma
7	1	30	$4\left(35r^7-60r^5+30r^3-4r\right)\cos\theta$	x tertiary coma
7	3	31	$4\left(21r^7-30r^5+10r^3\right)\sin(3\theta)$	
7	3	32	$4\left(21r^7-30r^5+10r^3\right)\cos(3\theta)$	
7	5	33	$4\left(7r^7-6r^5\right)\sin(5\theta)$	
7	5	34	$4\left(7r^7-6r^5\right)\cos(5\theta)$	
7	7	35	$4\,r^7\sin(7\theta)$	
7	7	36	$4\,r^7\cos(7\theta)$	
8	0	37	$3\,\left(70r^8-140r^6+90r^4-20r^2+1\right)$	tertiary spherical

Listing 5.1 Code for evaluating Zernike polynomials in MATLAB.

```
function Z = zernike(i, r, theta)
% function Z = zernike(i, r, theta)
% Creates the Zernike polynomial with mode index i,
% where i = 1 corresponds to piston
load('zernike_index'); % load the mapping of (n,m) to i
n = zernike_index(i,1);
m = zernike_index(i,2);
if m==0
    Z = sqrt(n+1)*zrf(n,0,r);
else
    if mod(i,2) == 0 % i is even
        Z = sqrt(2*(n+1))*zrf(n,m,r) .* cos(m*theta);
    else % i is odd
        Z = sqrt(2*(n+1))*zrf(n,m,r) .* sin(m*theta);
    end
end
return

% Zernike radial function
function R = zrf(n, m, r)
R = 0;
for s = 0 : ((n-m)/2)
    num = (-1)^s * gamma(n-s+1);
    denom = gamma(s+1) * gamma((n+m)/2-s+1) ...
        * gamma((n-m)/2-s+1);
    R = R + num / denom * r.^(n-2*s);
end
```

each is 2, 4, and 6 for primary, secondary, and tertiary astigmatism, respectively. As we follow the radial portion of each mode from the center to edge of the pupil, the higher-order modes have more peaks, troughs, and zero crossings.

With the modes completely defined, any wavefront $W(r,\theta)$ can be written as a Zernike series with coefficients a_i given by

$$W(r,\theta) = \sum_{i=1}^{\infty} a_i Z_i(r,\theta). \tag{5.6}$$

There are many benefits of this representation, and they are discussed below.

The key property of Zernike polynomials is that they are orthogonal over the unit circle. The orthogonality relationship for this convention of Zernike polynomials is

$$\int_0^1 R_n^m(r)\, R_{n'}^m(r)\, r\, dr = \frac{1}{2n+1}\delta_{nn'} \tag{5.7}$$

$$\int_0^{2\pi} G^m(\theta)\, G^{m'}(\theta)\; d\theta = \pi\delta_{mm'} \tag{5.8}$$

$$\Rightarrow \int_0^{2\pi}\int_0^1 Z_i(r,\theta)\, Z_{i'}(r,\theta)\, r\, dr\, d\theta = \pi\delta_{nn'}\delta_{mm'} = \pi\delta_{ii'}. \tag{5.9}$$

Using the orthogonality relationship, a given wavefront can be decomposed into its Zernike series by computing its Zernike coefficients with

$$a_i = \frac{\int_0^{2\pi}\int_0^1 W(r,\theta)\, Z_i(r,\theta)\, r\, dr\, d\theta}{\int_0^{2\pi}\int_0^1 Z_i^2(r,\theta)\, r\, dr\, d\theta}. \tag{5.10}$$

Often, we have a representation of a two-dimensional wavefront on a sampled two-dimensional Cartesian grid, either from a simulation or measurement. In that case, we can rewrite Eq. (5.10) as a discrete sum over Cartesian coordinates x_p and y_q given by

$$a_i = \frac{\sum_p \sum_q W(x_p, y_q)\, Z_i(x_p, y_q)}{\sum_p \sum_q Z_i^2(x_p, y_q)}. \tag{5.11}$$

In Eq. (5.11) the sums run over all p and q that are within the optical aperture. Notice that Eq. (5.11) does not actually depend on the values of x_p and y_q, only the

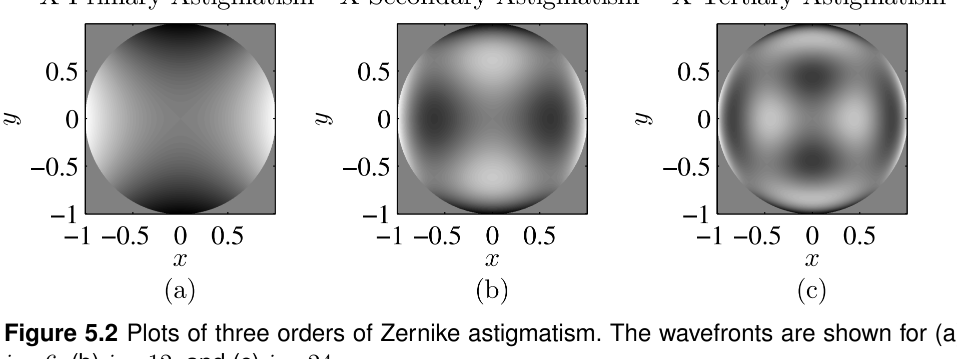


Figure 5.2 Plots of three orders of Zernike astigmatism. The wavefronts are shown for (a) $i = 6$, (b) $i = 12$, and (c) $i = 24$.

values of the wavefront and Zernike polynomials at the locations of x_p and y_q. To make this manifest, we define the notational changes

$$W_{pq} = W\left(x_p, y_q\right), \quad Z_{i,pq} = Z_i\left(x_p, y_q\right) \tag{5.12}$$

and use this new notation. This yields

$$a_i = \frac{\sum\limits_p \sum\limits_q W_{pq} Z_{i,pq}}{\sum\limits_p \sum\limits_q Z_{i,pq}^2}. \tag{5.13}$$

This notation can be simplified further by using only one index j to take the place of p and q. This means of referring to all wavefront and Zernike values within the aperture could be done in column-major, row-major, or any other order. The choice does not matter; however different programming (or scripting) languages handle certain orderings naturally. For example, C and C++ use row-major order, while MATLAB uses column-major order. Now, using just the index j for the different samples in the aperture gives

$$a_i = \frac{\sum\limits_j W_j\, Z_{i,j}}{\sum\limits_j Z_{i,j}^2}. \tag{5.14}$$

The same discretization and linear indexing could be applied to Eq. (5.6), leading to

$$W_j \cong \sum_{i=1}^{n_Z} Z_{i,j}\, a_i, \tag{5.15}$$

where n_Z is the number of modes being used. The reader should beware that the relationship is only approximate because of the discrete grid. The accuracy improves as more grid points are used.[24] This linear indexing now provides a new interpretation. We can treat Eq. (5.15) as a vector-matrix multiplication. Now, denote $\mathbf{W}$ as a column vector with elements W_i, Z as a matrix with elements Z_{ij}, and $\mathbf{A}$ as a column vector with elements A_i. To be explicit, the columns of Z are formed from individual Zernike polynomials evaluated at each aperture location such that

$$\mathsf{Z} = \left[Z_1 | Z_2 | \ldots | Z_{n_Z}\right], \tag{5.16}$$

where Z_1, Z_2, etc. are linear-indexed Zernike values. The number of rows in $\mathbf{W}$ is equal to the number of grid points within the aperture. The number of rows in $\mathbf{A}$ is equal to the number of modes being used. Correspondingly, the number of rows in Z is equal to the number of grid points, and the number of columns is equal to the number of modes. Finally, Eq. (5.15) compactly becomes

$$\mathbf{W} = \mathsf{Z}\mathbf{A}. \tag{5.17}$$

Listing 5.2 An example of computing Zernike coefficients from an arbitrary wavefront.

```
% example_zernike_projection.m

N = 32;     % number of grid points per side
L = 2;      % total size of the grid [m]
delta = L / N;  % grid spacing [m]
% cartesian & polar coordinates
[x y] = meshgrid((-N/2 : N/2-1) * delta);
[theta r] = cart2pol(x, y);
% unit circle aperture
ap = circ(x, y, 2);
% 3 Zernike modes
z2 = zernike(2, r, theta) .* ap;
z4 = zernike(4, r, theta) .* ap;
z21 = zernike(21, r, theta) .* ap;
% create the aberration
W = 0.5 *  z2 + 0.25 * z4 - 0.6 * z21;
% find only grid points within the aperture
idx = logical(ap);
% perform linear indexing in column-major order
W = W(idx);
Z = [z2(idx) z4(idx) z21(idx)];
% solve the system of equations to compute coefficients
A = Z \ W
```

Those familiar with linear algebra might recognize Eq. (5.14) as the Moore-Penrose pseudo-inverse (least-squares) solution to Eq. (5.17), written here in matrix notation as

$$\mathbf{A} = \left(\mathsf{Z}^T \mathsf{Z}\right)^{-1} \mathsf{Z}^T \mathbf{W}. \tag{5.18}$$

The vector-matrix forms here are compact in notation, and they can be implemented as a single line of code in many programming languages. For example, linear-algebra packages such as Linear Algebra PACKage (more commonly known as LAPACK)[25] and Basic Linear Algebra Subroutines (more commonly known as BLAS)[26,27], available for the C and FORTRAN languages, provide many fast-executing manipulations of matrices and vectors. Listing 5.2 gives a MATLAB example of projecting a complicated phase onto Zernike modes. The phase tested in the code is a weighted sum of modes 2, 4, and 21 with weights 0.5, 0.25, and -0.6, respectively. When the code is executed, the values in the array `A` are computed to be 0.5, 0.25, and -0.6, respectively.

5.1.2.1 Decomposition and mode removal

The previous subsection demonstrated how to compute the Zernike mode content of a phase map, given by its Zernike coefficients. Knowing this Zernike content can be quite useful. For example, we might have an optical system's measured aberration and wish to see what happens if we design an element to compensate for part of that aberration. As a practical instance, eye glasses and contact lenses often compensate for focus and astigmatism.

A real aberration $W\left(r,\theta\right)$ might contain a very large number of modes, but we may be interested in a mode-limited version $W'\left(r,\theta\right)$. Let us define

$$W'\left(r,\theta\right)=\sum_{i=1}^{n_Z}a_iZ_i\left(r,\theta\right) \tag{5.19}$$

as the mode-limited version of $W\left(r,\theta\right)$ such that

$$W\left(r,\theta\right)=W'\left(r,\theta\right)+\sum_{i=n_Z+1}^{\infty}a_iZ_i\left(r,\theta\right). \tag{5.20}$$

This is a good framework for partially corrected aberrations. With eye glasses and contact lenses, we ignore modes 1–3 because they do not affect visual image quality. Corrective lenses might compensate modes 4, 5, and 6. In that case, $n_Z=6$, and the eyes see images blurred by the residual aberration containing modes $i=7$ and up. Fortunately, the coefficients for these residual modes are usually much smaller than for the compensated modes.

An adaptive optics system is like a dynamically reconfigurable, high-resolution "contact lens" for imaging telescopes and cameras. A wavefront sensor is used to sense aberrations rapidly (sometimes over $10,000$ frames per second) and adjust the figure of a deformable mirror to compensate aberrations.[23] Many of today's astronomical telescopes use adaptive optics to compensate phase aberrations caused by imaging through Earth's turbulent atmosphere. Deformable mirrors can only reproduce a finite number of Zernike modes, so there is always some residual aberration uncorrected by the mirror. Listing 5.3 gives an example of generating a random draw of a turbulent aberration and producing a mode-limited version $W'\left(r,\theta\right)$ (generating the aberration is covered in Sec. 9.3). Figure 5.3 shows the original screen and versions limited to 3, 16, 36, and 100 modes. Notice how the mode-limited version increasingly resembles the original aberration as more modes are included in the Zernike series representation.

It is also interesting to examine the residual phase of mode-limited aberrations. Figure 5.4 shows the complement [remaining terms, i.e., the second term in Eq. (5.20)] to each of Fig. 5.3's mode-limited aberrations. Notice how the structures in the residual phase get finer as more modes are included in the Zernike series representation. Also, note that adaptive optics systems typically use a fast steering mirror to compensate turbulence-induced tilt, leaving modes 4 and higher to be compensated by the deformable mirror. Accordingly, the residual phase in the upper left corner of Figure 5.4 shows the aberration that the deformable mirror must compensate. For a deformable mirror that can represent up to the first 100 Zernike modes, the lower right corner of Figure 5.4 shows the residual aberration after the deformable mirror that still blurs the image. As one can see in the figure, if adaptive optics are designed properly, it usually reduces the aberration significantly

Listing 5.3 An example of synthesizing a mode-limited version of an arbitrary aberration. The aberration in this example is a random draw of an atmospheric phase screen, discussed in Sec. 9.3.

```
% example_zernike_synthesis.m

N = 40;     % number of grid points per side
L = 2;      % total size of the grid [m]
delta = L / N;  % grid spacing [m]
% cartesian & polar coordinates
[x y] = meshgrid((-N/2 : N/2-1) * delta);
[theta r] = cart2pol(x, y);
% unit circle aperture
ap = circ(x, y, 2);
% indices of grid points in aperture
idxAp = logical(ap);
% create atmospheric phase screen
r0 = L / 20;
screen = ft_phase_screen(r0, N, delta, inf, 0) ...
    / (2*pi) .* ap;
W = screen(idxAp);   % perform linear indexing

%%% analyze screen
nModes = 100;   % number of Zernike modes
% create matrix of Zernike polynomial values
Z = zeros(numel(W), nModes);
for idx = 1 : nModes
    temp = zernike(idx, r, theta);
    Z(:,idx) = temp(idxAp);
end
% compute mode coefficients
A = Z \ W;
% synthesize mode-limited screen
W_prime = Z*A;
% reshape mode-limited screen into 2-D for display
scr = zeros(N);
scr(idxAp) = W_prime;
```

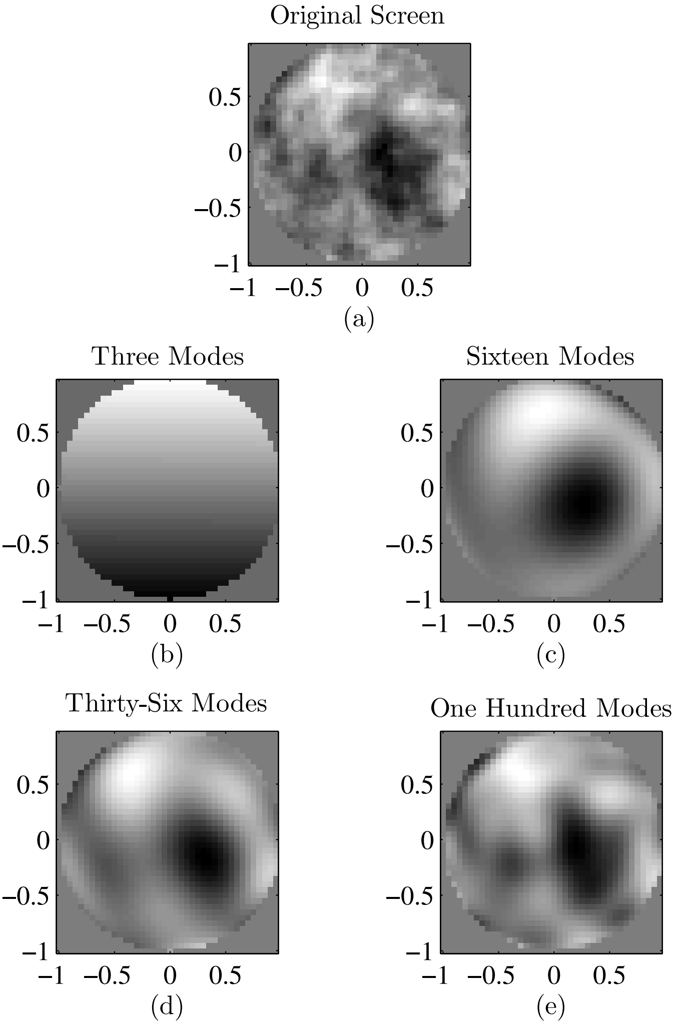


Figure 5.3 Plots of mode-limited phase screens. The original screen is at the top in plot (a). The four lower plots, (b)–(e) show the screen limited to 3, 16, 36, and 100 modes, respectively.

and provides greatly improved imagery.

5.1.2.2 RMS wavefront aberration

It is often handy to describe a wavefront aberration by its rms value σ averaged over the aperture. We compute the mean-square wavefront deviation straightforwardly via

$$\sigma^2 = \frac{1}{\pi} \int_0^{2\pi} \int_0^1 \left[W(r,\theta) - \overline{W} \right]^2 r \, dr \, d\theta, \tag{5.21}$$

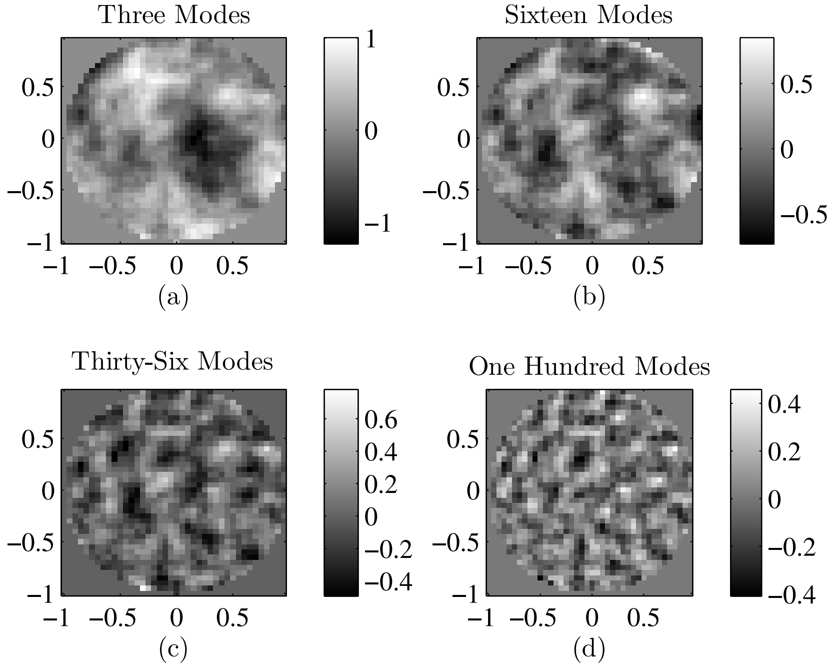


Figure 5.4 Plots of residual phase due to finite number of modes. These are the residuals for the mode limits in Fig. 5.3.

where $\overline{W}$ is the mean of W over the aperture. Note that in Eq. (5.21), the average is over the pupil area, which is π for a unit-radius circle. Writing the wavefront as a Zernike series yields

$$\sigma^2 = \frac{1}{\pi}\int_0^{2\pi}\int_0^1 \left[\sum_{i=2}^{\infty} a_i Z_i(r,\theta)\right]^2 r\,dr\,d\theta, \tag{5.22}$$

where the reader should note that the sum begins at $i = 2$ because $\overline{W}$ is the $i = 1$ term. We now factor the squared sum into an explicit product of two series so that

$$\sigma^2 = \frac{1}{\pi}\int_0^{2\pi}\int_0^1 \left[\sum_{i=2}^{\infty} a_i Z_i(r,\theta)\right]\left[\sum_{i'=2}^{\infty} a_{i'} Z_{i'}(r,\theta)\right] r\,dr\,d\theta \tag{5.23}$$

$$= \frac{1}{\pi}\sum_{i=2}^{\infty} a_i \sum_{i'=2}^{\infty} a_{i'} \int_0^{2\pi}\int_0^1 Z_i(r,\theta)\,Z_{i'}(r,\theta)\,r\,dr\,d\theta \tag{5.24}$$

$$= \frac{1}{\pi}\sum_{i=2}^{\infty} a_i \sum_{i'=2}^{\infty} a_{i'}\,\pi\delta_{ii'} \tag{5.25}$$

$$= \sum_{i=2}^{\infty} a_i^2. \tag{5.26}$$

This means that the wavefront variance can be found by simply summing the squares of the Zernike coefficients. This is a very convenient benefit of using an orthogonal basis set to describe aberrations.

5.2 Impulse Response and Transfer Function of Imaging Systems

Aberrations have a strong effect on the impulse response of an imaging system. Further, the imaging system model shown in Fig. 5.1 has different impulse responses depending on the coherence of the object's illumination. If the illumination is spatially coherent, the impulse response is called the amplitude spread function (or coherent spread function), and the system's frequency response is called the amplitude transfer function (or coherent transfer function).[5] This is discussed in Sec. 5.2.1. If the illumination is spatially incoherent, the impulse response is called the point spread function, and the system's frequency response is called the optical transfer function (OTF), and its magnitude is called the modulation transfer function (MTF). This is discussed in Sec. 5.2.2.

Note that wavefront aberrations are independent of the illumination. They only depend on the optical components of the imaging system. However, their effect on the image does depend on the coherence of the illumination.

5.2.1 Coherent imaging

When the light is coherent, imaging systems are linear in optical field. Accordingly, the image amplitude $U_i\left(u,v\right)$ is the convolution of the object amplitude $U_o\left(u,v\right)$ with the amplitude spread function $h\left(u,v\right)$ according to

$$U_i\left(u,v\right) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} h\left(u-\eta, v-\xi\right) U_o\left(\eta,\xi\right)\, d\xi\, d\eta \tag{5.27}$$

$$= h\left(u,v\right) \otimes U_o\left(u,v\right). \tag{5.28}$$

This assumes that the imaging system has unit magnification. Accounting for magnification just requires scaling of the object coordinates.[5] The amplitude spread function is given by

$$h\left(u,v\right) = \frac{1}{\lambda z_i}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \mathcal{P}\left(x,y\right) e^{-i\frac{2\pi}{\lambda z_i}\left(ux+vy\right)}\, dx\, dy \tag{5.29}$$

$$= \frac{1}{\lambda z_i}\mathcal{F}\left\{\mathcal{P}\left(x,y\right)\right\}_{f_x=\frac{u}{\lambda z_i}, f_y=\frac{v}{\lambda z_i}}, \tag{5.30}$$

where $\mathcal{P}\left(x,y\right)$ is the generalized pupil function defined in Eq. (5.1) and z_i is the image distance.

Listing 5.4 An example of coherent imaging in MATLAB.

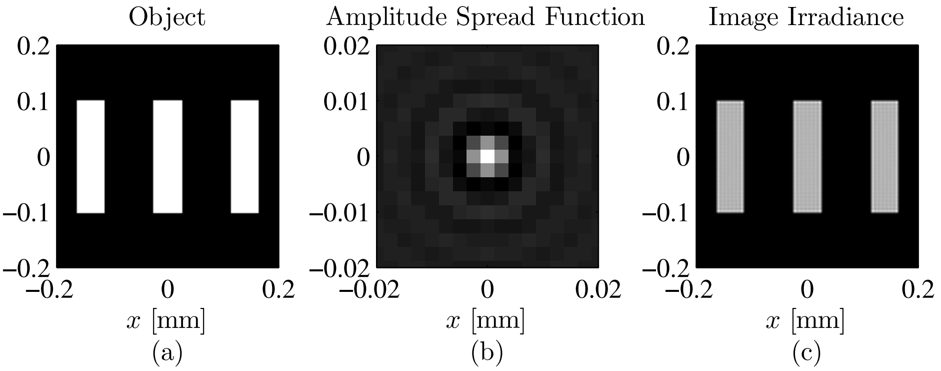
```
1  % example_coh_img.m
2
3  N = 256;    % number of grid points per side
4  L = 0.1;      % total size of the grid [m]
5  D = 0.07;   % diameter of pupil [m]
6  delta = L / N;  % grid spacing [m]
7  wvl = 1e-6; % optical wavelength [m]
8  z = 0.25;   % image distance [m]
9  % pupil-plane coordinates
10 [x y] = meshgrid((-N/2 : N/2-1) * delta);
11 [theta r] = cart2pol(x, y);
12 % wavefront aberration
13 W = 0.05 * zernike(4, 2*r/D, theta);
14 % complex pupil function
15 P = circ(x, y, D) .* exp(i * 2*pi * W);
16 % amplitude spread function
17 h = ft2(P, delta);
18 delta_u = wvl * z / (N*delta);
19 % image-plane coordinates
20 [u v] = meshgrid((-N/2 : N/2-1) * delta_u);
21 % object (same coordinates as h)
22 obj = (rect((u-1.4e-4)/5e-5) + rect(u/5e-5) ...
23     + rect((u+1.4e-4)/5e-5)) .* rect(v/2e-4);
24 % convolve the object with the ASF to simulate imaging
25 img = myconv2(obj, h, 1);
```

Listing 5.4 gives an example of how to compute a coherent image given the object and amplitude spread function of the imaging system. In the example, the object comprises three parallel rectangular slits as shown in Fig. 5.5(a). The aberration is 0.05 waves of Zernike defocus ($i = 4$), computed in line 13. The resulting generalized pupil function is computed in line 15. Line 17 computes the amplitude spread function using the `ft2` function, and it is shown in Fig. 5.5(b). Notice that is much narrower than the object. As noted in Sec. 3.1, this is typical of impulse responses in linear systems. Finally, the image field is formed by convolving the object field and amplitude spread function in line 25 using the `conv2` function. The resulting object intensity is shown in Fig. 5.5.

If the convolution theorem is applied to Eq. (5.27), the result is

$$\mathcal{F}\left\{U_i\left(u,v\right)\right\} = \mathcal{F}\left\{h\left(u,v\right)\right\}\mathcal{F}\left\{U_o\left(u,v\right)\right\}. \tag{5.31}$$

In this form, it is clear that the amplitude spread function's Fourier spectrum modulates the object's spectrum to yield the the diffraction image. This specifies how object's frequency spectrum is transferred through the imaging system to the diffrac-



(a) (b) (c)

Figure 5.5 Example of coherent imaging. Plot (a) shows the object, while plot (b) shows the amplitude spread function due to defocus, and plot (c) shows the coherent image blurred by 0.05 waves of defocus.

tion image, so we define this property of the system as the amplitude transfer function given by

$$\begin{aligned} H(f_x, f_y) &= \mathcal{F}\{h(u,v)\} && (5.32)\\ &= \mathcal{F}\left\{\frac{1}{\lambda z_i}\mathcal{F}\{\mathcal{P}(x,y)\}_{f_x=\frac{u}{\lambda z_i}, f_y=\frac{v}{\lambda z_i}}\right\} && (5.33)\\ &= \lambda z_i \mathcal{P}(-\lambda z_i f_x, -\lambda z_i f_y). && (5.34)\end{aligned}$$

In the last equation, Eq. (5.30) has been used to write the amplitude transfer function in terms of system's pupil function. The low-pass filtering property of imaging systems is now evident when we consider a common aperture like a circle. Eq. (5.34) indicates that a circular aperture with diameter D would pass all frequencies for which $\left(f_x^2 + f_y^2\right)^{1/2} < D/(2\lambda z_i)$ equally while filtering out all higher frequencies completely. In this way, image amplitude is a strictly bandlimited function.

5.2.2 Incoherent imaging

When the light is spatially incoherent, the image irradiance is the convolution of the object irradiance with the point spread function (PSF):

$$\begin{aligned} I_i(u,v) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(u-\eta, v-\xi)|^2 I(\eta,\xi)\, d\xi\, d\eta && (5.35)\\ &= |h(u,v)|^2 \otimes I(u,v). && (5.36)\end{aligned}$$

The point spread function is simply the squared magnitude of the amplitude spread function. Listing 5.5 gives an example of how to compute an incoherent image given the object and amplitude spread function of the imaging system. The

Listing 5.5 An example of incoherent imaging in MATLAB.

```
% example_incoh_img.m

N = 256;    % number of grid points per side
L = 0.1;      % total size of the grid [m]
D = 0.07;   % diameter of pupil [m]
delta = L / N;  % grid spacing [m]
wvl = 1e-6; % optical wavelength [m]
z = 0.25;   % image distance [m]
% pupil-plane coordinates
[x y] = meshgrid((-N/2 : N/2-1) * delta);
[theta r] = cart2pol(x, y);
% wavefront aberration
W = 0.05 * zernike(4, 2*r/D, theta);
% complex pupil function
P = circ(x, y, D) .* exp(i * 2*pi * W);
% amplitude spread function
h = ft2(P, delta);
U = wvl * z / (N*delta);
% image-plane coordinates
[u v] = meshgrid((-N/2 : N/2-1) * U);
% object (same coordinates as h)
obj = (rect((u-1.4e-4)/5e-5) + rect(u/5e-5) ...
    + rect((u+1.4e-4)/5e-5)) .* rect(v/2e-4);
% convolve the object with the PSF to simulate imaging
img = myconv2(abs(obj).^2, abs(h).^2, 1);
```

object and aberration are the same as those from the coherent example. The basic computations are the same, too, except that the object irradiance is convolved with the imaging system's point spread function. The results are shown in Fig. 5.6.

Like the coherent case, the convolution theorem can be applied to Eq. (5.35), and now the result is

$$\mathcal{F}\left\{I_i\left(u,v\right)\right\} = \mathcal{F}\left\{\left|h\left(u,v\right)\right|^2\right\}\mathcal{F}\left\{I_o\left(u,v\right)\right\}. \tag{5.37}$$

Again, we can see that the PSF's Fourier spectrum modulates the object irradiance's spectrum to yield the diffraction image. In the incoherent case, the filter function (called the optical transfer function) is defined as

$$\mathcal{H}\left(f_x,f_y\right) = \frac{\mathcal{F}\left\{\left|h\left(u,v\right)\right|^2\right\}}{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\left|h\left(u,v\right)\right|^2\,dudv}. \tag{5.38}$$

Similarly to the coherent case, we can relate this to the pupil function. Application

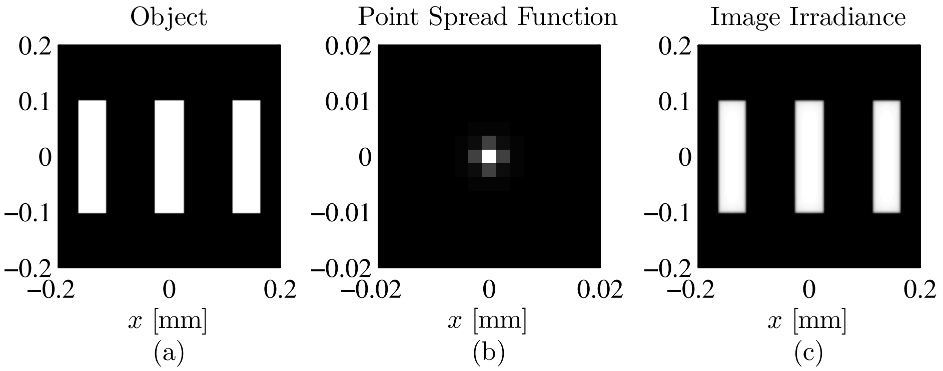


Figure 5.6 Example of incoherent imaging. Plot (a) shows the object, while plot (b) shows the point spread function due to defocus, and plot (c) shows the incoherent image blurred by 0.05 waves of defocus.

of the auto-correlation theorem and Parseval's theorem yields

$$\mathcal{H}(f_x, f_y) = \frac{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} H^*(p - f_x, q - f_y)\, H(p, q)\; dp\, dq}{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} |H(p, q)|^2\; dp\, dq} \tag{5.39}$$

$$= \frac{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \mathcal{P}^*(x - \lambda z_i f_x, y - \lambda z_i f_y)\, \mathcal{P}(x, y)\; dx\, dy}{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} |\mathcal{P}(x, y)|^2\; dx\, dy} \tag{5.40}$$

$$= \left.\frac{\mathcal{P}^*(x, y) \star \mathcal{P}(x, y)}{\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} |\mathcal{P}(x, y)|^2\; dx\, dy}\right|_{x=\lambda z_i f_x, y=\lambda z_i f_y} . \tag{5.41}$$

The example case of a circular aperture with diameter D is illustrative again. It can be shown that the OTF for a circular aperture is an azimuthally symmetric function of $f = \left(f_x^2 + f_y^2\right)^{1/2}$ given by

$$\mathcal{H}(f) = \begin{cases} \frac{2}{\pi}\left[\cos^{-1}\left(\frac{f}{2f_0}\right) - \frac{f}{2f_0}\sqrt{1 - \left(\frac{f}{2f_0}\right)^2}\right] & f \le 2f_0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.42}$$

where $f_0 = D/(2\lambda z_i)$. This quantity f_0 is the cutoff frequency for the coherent case, but as Eq. (5.42) indicates, frequencies up to $2f_0$ pass through (with some attenuation) when the light is incoherent. Still, incoherent images are strictly bandlimited. Another difference from the coherent case is that $\mathcal{H}(f) \ge 0$ for all frequencies.

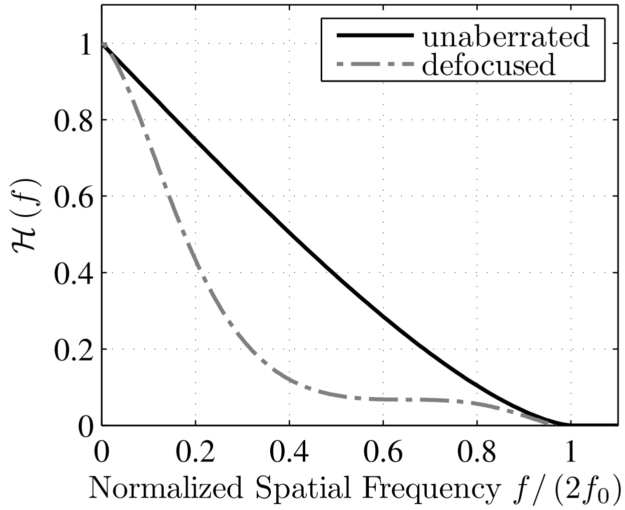


Figure 5.7 Optical transfer functions for unaberrated and defocused imaging systems.

Figure 5.7 shows a plot of two OTFs for imaging systems with circular apertures. The solid black line is the OTF for a system without aberrations as given in Eq. (5.42). The dash-dot gray line is the OTF for a system with defocus such that the wavefront error is 0.5 waves at the edge of the aperture (computed by numerical integration). Clearly, the defocused image would have many frequency components that are more attenuated than an aberration-free image. This is also characterized by a broader PSF, and results in a blurred image. The next subsection discusses a related metric for image quality.

5.2.3 Strehl ratio

Clearly, the performance of an imaging system is determined by its amplitude/-point spread function. It is handy to have a single-number metric to describe performance. The most common metric is Strehl ratio, which is the ratio of the on-axis actual point spread function value to the on-axis ideal point spread function value. Typically, this is a comparison of an aberrated system to an almost identical but unaberrated system. The on-axis value of a PSF is computed by using Eq. (5.29) at the origin:

$$|h(0,0)|^2 = \frac{1}{\lambda^2 z_i^2}\left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{P}(x,y)\, e^0\, dx\, dy\right|^2 \tag{5.43}$$

$$= \frac{1}{\lambda^2 z_i^2}\left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{P}(x,y)\ dx\, dy\right|^2 . \tag{5.44}$$

Because the only contribution to non-zero phase in the generalized pupil function $\mathcal{P}(x,y)$ is caused by aberrations, $P(x,y)$ is the unaberrated pupil function. As a result, the Strehl ratio $\mathcal{S}$ is computed as

$$\mathcal{S} = \frac{\left| \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \mathcal{P}(x,y)\ dx\,dy \right|^2}{\left| \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} P(x,y)\ dx\,dy \right|^2}. \tag{5.45}$$

To make the aberration phase $\phi(x,y)$ more manifest, we can rewrite Eq. (5.45) as

$$\mathcal{S} = \frac{\left| \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} P(x,y)\, e^{i\phi(x,y)}\ dx\,dy \right|^2}{\left| \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} P(x,y)\ dx\,dy \right|^2} \tag{5.46}$$

$$= \frac{\int\limits_{-\infty}^{\infty} \mathcal{H}(f_x,f_y)\ df_x\,df_y}{\int\limits_{-\infty}^{\infty} \mathcal{H}_{dl}(f_x,f_y)\ df_x\,df_y}, \tag{5.47}$$

where Eqs. (5.30) and (5.38) have been applied to obtain the latter equation and $\mathcal{H}_{dl}(f_x,f_y)$ is the OTF of an unaberrated (or diffraction-limited) system.

For a perfectly unaberrated system, $\mathcal{S} = 1$, and this is the maximum possible value of the Strehl ratio. Aberrations and amplitude variations in the pupil (for example, an annular aperture) always reduce the Strehl ratio.[19] Consequently, low Strehl ratio indicates poor image quality, i.e, coarse resolution and low contrast.

For small aberrations, the Strehl ratio of an image is determined by the variance of the pupil phase. To show this, we can rewrite Eq. (5.46) in the abbreviated form

$$\mathcal{S} = \left| \left\langle e^{i\phi} \right\rangle \right|^2, \tag{5.48}$$

where the angle brackets $\langle \ldots \rangle$ indicate a spatial average over the amplitude-weighted pupil. For example, the amplitude-weighted average phase is given by[19]

$$\langle \phi \rangle = \frac{\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} P(x,y)\,\phi(x,y)\ dx\,dy}{\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} P(x,y)\ dx\,dy}. \tag{5.49}$$

Multiplying Eq. (5.48) by $\left| e^{-i\langle\phi\rangle} \right|^2 = 1$ yields

$$\mathcal{S} = \left| \left\langle e^{i(\phi - \langle\phi\rangle)} \right\rangle \right|^2 \tag{5.50}$$

$$= \langle \cos(\phi - \langle\phi\rangle)\rangle^2 + \langle \sin(\phi - \langle\phi\rangle)\rangle^2 . \tag{5.51}$$

Taking the first terms up to second order of the Taylor-series expansions gives

$$\mathcal{S} \simeq \left\langle 1 - \frac{(\phi - \langle\phi\rangle)^2}{2} \right\rangle^2 + \langle \phi - \langle\phi\rangle\rangle^2 \tag{5.52}$$

$$\simeq \left(1 - \frac{\sigma_\phi^2}{2}\right)^2 . \tag{5.53}$$

Carrying out the multiplication and keeping only the first two terms leads to

$$\mathcal{S} \simeq 1 - \sigma_\phi^2, \tag{5.54}$$

where $\sigma_\phi^2 = 4\pi^2\sigma^2$ is the variance of the phase, measured in rad^2. This result is the same as writing

$$\mathcal{S} \simeq e^{-\sigma_\phi^2} \tag{5.55}$$

and keeping only the first two terms in its Taylor series expansion. Eqs. (5.53)–(5.55) all represent commonly used approximations for computing Strehl ratio. Eq. (5.53) is the Maréchal formula. Eq. (5.55), while it is presented here as an approximation to Eq. (5.54), actually is an empirical formula that gives the best fit to numerical results for various aberrations.[19]

5.3 Problems

1. The Sellmeier equation is an empirical relationship between optical wavelength and refractive index for glass. It is given by

$$n^2(\lambda) = 1 + \sum_i \frac{B_i \lambda^2}{\lambda^2 - C_i} \tag{5.56}$$

Each type of glass has its own measured set of Sellmeier coefficients B_i and C_i.

(a) Find the Sellmeier coefficients for borosilicate crown glass (more commonly called BK7) and compute the standard refractive indices

$$n_F = n\,(486.12\,\text{nm}) \quad \text{blue Hydrogen line} \tag{5.57}$$

$$n_d = n\,(587.56\,\text{nm}) \quad \text{yellow Helium line} \tag{5.58}$$

$$n_C = n\,(656.27\,\text{nm}) \quad \text{red Hydrogen line} \tag{5.59}$$

to six significant digits.

(b) You are given a thin plano-convex lens made of BK7 glass. The convex side is spherical with a 51.68-mm radius of curvature, and the lens diameter is 12.7 mm. Compute the focal lengths and diffraction-limited spot diameters corresponding to each of the standard wavelengths from part (a).

(c) Follow the coherent-imaging example of Sec. 5.2.1 to compute each diffraction-limited PSF. Add several different levels of defocus aberration and compute the resulting PSFs. For all wavelengths, plot the $v = 0$ slice of each PSF to demonstrate how the focal spot evolves near the geometric focal plane. Use these PSF-slice plots to show that you have computed the correct spot diameters. Use 1024 grid points per side and a grid spacing of 0.199 mm.

2. For a lens that is aberrated with one wave of Zernike primary astigmatism, add several different levels of defocus aberration and compute the resulting PSFs. Show images of these PSFs to demonstrate how the focal spot evolves near the geometric focal plane. Use a grid size = 4 m, aperture diameter = 2 m, with 512 points per side, optical wavelength $= 1\mu$m, and focal length $= 16$ m.

3. For a lens that is aberrated with one wave of Zernike primary spherical aberration, add several different levels of defocus aberration and compute the resulting PSFs. Show images of these PSFs to demonstrate how the focal spot evolves near the geometric focal plane. Use a grid size $= 4$ m, aperture diameter $= 2$ m, with $= 512$ points per side, optical wavelength $= 1\mu$m, and focal length $= 16$ m.

4. Given

$$W\left(x,y\right) = 0.07\,Z_4 + 0.05\,Z_5 - 0.05\,Z_6 + 0.03\,Z_7 - 0.03\,Z_8, \qquad (5.60)$$

compute the Strehl ratio

(a) using Eqs. (5.26) and (5.55),

(b) and using a simulation to compute the aberrated and diffraction-limited PSFs (similar to the example of Sec. 5.2.1). Use a grid size $= 8$ m, aperture diameter $= 2$ m, with $= 512$ points per side, optical wavelength $= 1\mu$m, and focal length $= 64$ m.

5. Numerically compute the PSF of an annular aperture whose inner and outer diameters are 1 m and 2 m, respectively. Also compute the PSF of a filled 2 m circular aperture. Use a grid size $= 8$ m, with $= 512$ points per side, optical wavelength $= 1\mu$m, and focal length $= 64$ m. Provide displays of both PSFs and compute the Strehl ratio of the annular aperture as the ratio of the peaks of the PSFs. Confirm your numerical results with analytic calculations.

6. Numerically compute the PSF of a sparse (or aggregate) aperture composed of three 1-m-diameter circular apertures each centered at coordinates $(0.6, 0.6)$ m, $(-0.6, 0.6)$ m, and $(0, 0.6)$ m. Use a grid size $=$ 8 m, grid size $=$ 512 points per side, optical wavelength $= 1\mu$m, and focal length $= 64$ m. Provide displays of the aperture and PSF. Confirm your numerical results with analytic calculations.