Chapter

3

Aberrations

3.1 Introduction

In Chap. 2 we discussed the image-forming characteristics of optical systems, but we limited our consideration to an infinitesimal thread-like region about the optical axis called the paraxial region. In this chapter we will consider, in general terms, the behavior of lenses with *finite* apertures and fields of view. It has been pointed out that well-corrected optical systems behave nearly according to the rules of paraxial imagery given in Chap. 2. This is another way of stating that a lens without aberrations forms an image of the size and in the location given by the equations for the paraxial or first-order region. We shall measure the aberrations by the amount by which rays miss the paraxial image point.

It can be seen that aberrations may be determined by calculating the location of the paraxial image of an object point and then tracing a large number of rays (by the exact trigonometrical ray-tracing equations of Chap. 10) to determine the amounts by which the rays depart from the paraxial image point. Stated this baldly, the mathematical determination of the aberrations of a lens which covered any reasonable field at a real aperture would seem a formidable task, involving an almost infinite amount of labor. However, by classifying the various types of image faults and by understanding the behavior of each type, the work of determining the aberrations of a lens system can be simplified greatly, since only a few rays need be traced to evaluate each aberration; thus the problem assumes more manageable proportions.

Seidel investigated and codified the primary aberrations and derived analytical expressions for their determination. For this reason,

the primary image defects are usually referred to as the *Seidel aberrations.*

3.2 The Aberration Polynomial and the Seidel Aberrations

With reference to Fig. 3.1, we assume an optical system with symmetry about the optical axis so that every surface is a figure of rotation about the optical axis. Because of this symmetry, we can, without any loss of generality, define the object point as lying on the y axis; its distance from the optical axis is $y = h$. We define a ray starting from the object point and passing through the system aperture at a point described by its polar coordinates (s, θ). The ray intersects the image plane at the point x', y'.

We wish to know the form of the equation which will describe the image plane intersection coordinates y' and x' as a function of h, s, and θ; the equation will be a power series expansion. While it is impractical to derive an exact expression for other than very simple systems or for more than a few terms of the power series, it *is* possible to determine the general form of the equation. This is simply because we have assumed an axially symmetrical system. For example, a ray which

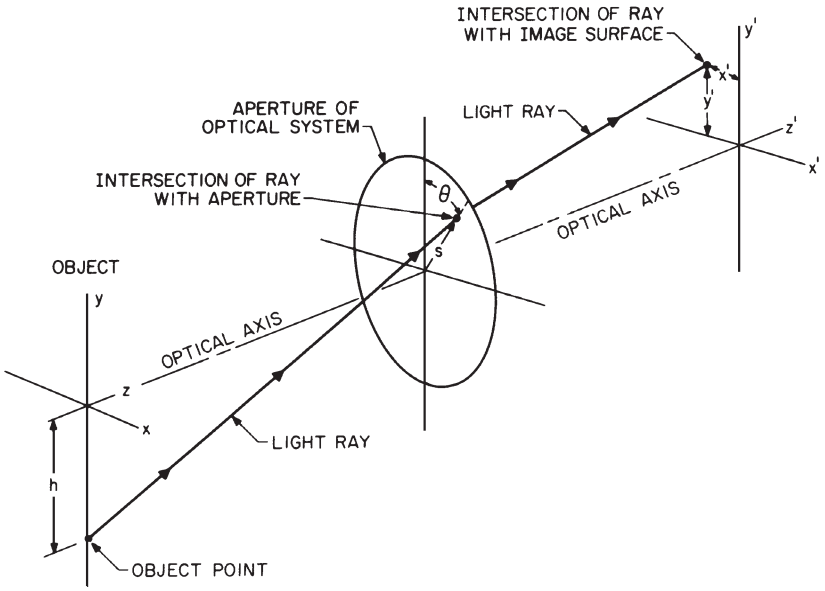


Figure 3.1 A ray from the point $y = h$, $(x = 0)$ in the object passes through the optical system aperture at a point defined by its polar coordinates, (s, θ), and intersects the image surface at x', y'.

intersects the axis in object space must also intersect it in image space. Every ray passing through the same axial point in object space and also passing through the same annular zone in the aperture (i.e., with the same value of s) must pass through the same axial point in image space. A ray in front of the meridional (y, z) plane has a mirror-image ray behind the meridional plane which is identical except for the (reversed) signs of x' and θ. Similarly, rays originating from $\pm h$ in the object and passing through corresponding upper and lower aperture points must have identical x' intersections and oppositely signed y' values. With this sort of logic one can derive equations such as the following:

$$\begin{aligned} y' = {} & A_1 s \cos\theta + A_2 h \\ & + B_1 s^3 \cos\theta + B_2 s^2 h(2 + \cos 2\theta) + (3B_3 + B_4) s h^2 \cos\theta + B_5 h^3 \\ & + C_1 s^5 \cos\theta + (C_2 + C_3 \cos 2\theta) s^4 h + (C_4 + C_6 \cos^2\theta) s^3 h^2 \cos\theta \\ & + (C_7 + C_8 \cos 2\theta) s^2 h^3 + C_{10} s h^4 \cos\theta + C_{12} h^5 + D_1 s^7 \cos\theta + \cdots \end{aligned} \tag{3.1}$$

$$\begin{aligned} x' = {} & A_1 s \sin\theta \\ & + B_1 s^3 \sin\theta + B_2 s^2 h \sin 2\theta + (B_3 + B_4) s h^2 \sin\theta \\ & + C_1 s^5 \sin\theta + C_3 s^4 h \sin 2\theta + (C_5 + C_6 \cos^2\theta) s^3 h^2 \sin\theta \\ & + C_9 s^2 h^3 \sin 2\theta + C_{11} s h^4 \sin\theta + D_1 s^7 \sin\theta + \cdots \end{aligned} \tag{3.2}$$

where A_n, B_n, etc., are constants, and h, s, and θ have been defined above and in Fig. 3.1.

Notice that in the A terms the exponents of s and h are unity. In the B terms the exponents total 3, as in s^3, s^2h, sh^2, and h^3. In the C terms the exponents total 5, and in the D terms, 7. These are referred to as the first-order, third-order, and fifth-order terms, etc. There are 2 first-order terms, 5 third-order, 9 fifth-order, and

$$\frac{(n + 3)(n + 5)}{8} - 1$$

nth-order terms. In an axially symmetrical system there are no even-order terms; only odd-order terms may exist (unless we depart from symmetry as, for example, by tilting a surface or introducing a toroidal or other nonsymmetrical surface).

It is apparent that the A terms relate to the paraxial (or first-order) imagery discussed in Chap. 2. A_2 is simply the magnification (h'/h), and A_1 is a transverse measure of the distance from the paraxial focus to our "image plane." All the other terms in Eqs. 3.1 and 3.2 are called

transverse aberrations. They represent the distance by which the ray misses the ideal image point as described by the paraxial imaging equations of Chap. 2.

The B terms are called the third-order, or Seidel, or primary aberrations. B_1 is spherical aberration, B_2 is coma, B_3 is astigmatism, B_4 is Petzval, and B_5 is distortion. Similarly, the C terms are called the fifth-order or secondary aberrations. C_1 is fifth-order spherical aberration; C_2 and C_3 are linear coma; C_4, C_5, and C_6 are oblique spherical aberration; C_7, C_8, and C_9 are elliptical coma; C_{10} and C_{11} are Petzval and astigmatism; and C_{12} is distortion.

The 14 terms in D are the seventh-order or tertiary aberrations; D_1 is the seventh-order spherical aberration. A similar expression for OPD, the wave front deformation, is given in Chap. 11.

As noted above, the Seidel aberrations of a system in monochromatic light are called spherical aberration, coma, astigmatism, Petzval curvature, and distortion. In this section we will define each aberration and discuss its characteristics, its representation, and its effect on the appearance of the image. Each aberration will be discussed as if it alone were present; obviously in practice one is far more likely to encounter aberrations in combination than singly. The third-order aberrations can be calculated using the methods given in Chap. 10.

Spherical aberration

Spherical aberration can be defined as the variation of focus with aperture. Figure 3.2 is a somewhat exaggerated sketch of a simple lens forming an "image" of an axial object point a great distance away. Notice that the rays close to the optical axis come to a focus (intersect the axis) very near the paraxial focus position. As the ray height at the lens increases, the position of the ray intersection with the optical axis moves farther and farther from the paraxial focus. The distance from the paraxial focus to the axial intersection of the ray is called longitudinal spherical aberration. Transverse, or lateral, spherical aberration is the name given to the aberration when it is measured in the "vertical" direction. Thus, in Fig. 3.2 AB is the longitudinal, and AC the transverse spherical aberration of ray R.

Since the magnitude of the aberration obviously depends on the height of the ray, it is convenient to specify the particular ray with which a certain amount of aberration is associated. For example, marginal spherical aberration refers to the aberration of the ray through the edge or margin of the lens aperture. It is often written as LA_m or TA_m.

Spherical aberration is determined by tracing a paraxial ray and a trigonometric ray from the same axial object point and determining their final intercept distances l' and L'. In Fig. 3.2, l' is distance OA

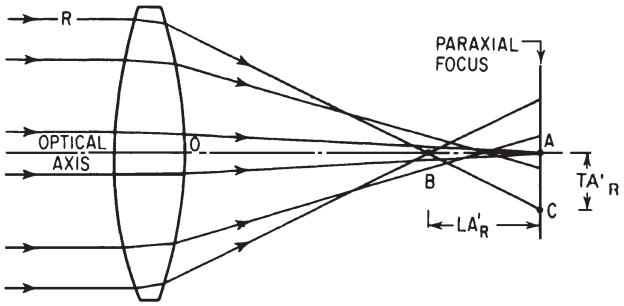


Figure 3.2 A simple converging lens with undercorrected spherical aberration. The rays farther from the axis are brought to a focus nearer the lens.

and L' (for ray R) is distance OB. The longitudinal spherical aberration of the image point is abbreviated LA' and

$$LA' = L' - l' \tag{3.3}$$

Transverse spherical aberration is related to LA' by the expression

$$TA'_R = -LA' \tan U'_R = -(L' - l') \tan U'_R \tag{3.4}$$

where U'_R is the angle the ray R makes with the axis. Using this sign convention, spherical aberration with a negative sign is called *undercorrected spherical,* since it is usually associated with simple uncorrected positive elements. Similarly, positive spherical is called *overcorrected* and is generally associated with diverging elements.

The spherical aberration of a system is usually represented graphically. Longitudinal spherical is plotted against the ray height at the lens, as shown in Fig. 3.3a, and transverse spherical is plotted against the final slope of the ray, as shown in Fig. 3.3b. Figure 3.3b is called a *ray intercept curve.* It is conventional to plot the ray through the top of the lens on the right in a ray intercept plot regardless of the sign convention used for ray slope angles.

For a given aperture and focal length, the amount of spherical aberration in a simple lens is a function of object position and the shape, or bending, of the lens. For example, a thin glass lens with its object at infinity has a minimum amount of spherical at a nearly plano-convex shape, with the convex surface toward the object. A meniscus shape, either convex-concave or concave-convex has much more spherical aberration. If the object and image are of equal size (each being two focal lengths from the lens), then the shape which gives the minimum spherical is equiconvex. Usually, a uniform distribution of the amount that a ray is "bent" or deviated will minimize the spherical.

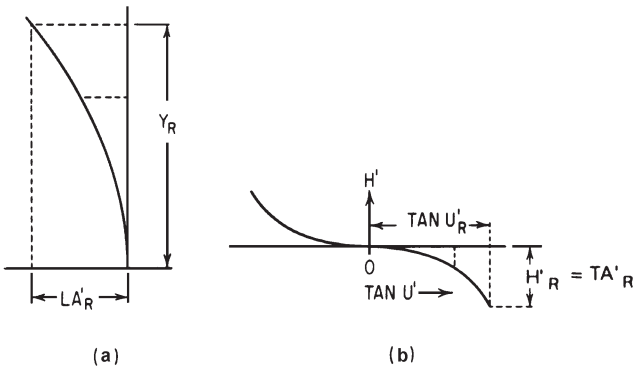


Figure 3.3 Graphical representation of spherical aberration. (a) As a longitudinal aberration, in which the longitudinal spherical aberration (LA') is plotted against ray height (Y). (b) As a transverse aberration, in which the ray intercept height (H') at the paraxial reference plane is plotted against the final ray slope (tan U').

The image of a point formed by a lens with spherical aberration is usually a bright dot surrounded by a halo of light; the effect of spherical on an extended image is to soften the contrast of the image and to blur its details.

In general, a positive, converging lens or surface will contribute undercorrected spherical aberration to a system, and a negative lens or divergent surface, the reverse, although there are certain exceptions to this.

Figure 3.3 illustrated two ways to present spherical aberration, as either a longitudinal or a transverse aberration. Equation 3.4 showed the relation between the two. The same relationship is also appropriate for astigmatism and field curvature (Sec. 3.2.3) and axial chromatic (Sec. 3.3). Note that coma, distortion, and lateral chromatic do not have a longitudinal measure. All of the aberrations can also be expressed as angular aberrations. The angular aberration is simply the angle subtended from the second nodal (or in air, principal) point by the transverse aberration. Thus

$$AA = \frac{TA}{s'} \tag{3.5}$$

Yet a fourth way to measure an aberration is by OPD, the departure of the actual wave front from a perfect reference sphere centered on the ideal image point, as discussed in Sec. 3.6 and Chap. 11.

The transverse measure of an aberration is directly related to the size of the image blur. Graphing it as a ray intercept plot (e.g., Fig. 3.3b and Fig. 3.24) allows the viewer to identify the various types of aberration afflicting the optical system. This is of great value to the

lens designer, and the ray intercept plot of the transverse aberrations is an almost universally used presentation of the aberrations. As discussed later (in Chap. 11), the OPD, or wave-front deformation, is the most useful measure of image quality for well-corrected systems, and a statement of the amount of the OPD is usually accepted as definitive in this regard. The longitudinal presentation of the aberrations is most useful in understanding field curvature and axial chromatic, especially secondary spectrum.

Coma

Coma can be defined as the variation of magnification with aperture. Thus, when a bundle of oblique rays is incident on a lens with coma, the rays passing through the edge portions of the lens may be imaged at a different height than those passing through the center portion. In Fig. 3.4, the upper and lower rim rays A and B, respectively, intersect the image plane above the ray P which passes through the center of the lens. The distance from P to the intersection of A and B is called the tangential coma of the lens, and is given by

$$\text{Coma}_T = H'_{AB} - H'_P \tag{3.6}$$

where H'_{AB} is the height from the optical axis to the intersection of the upper and lower rim rays, and H'_P is the height from the axis to the intersection of the ray P with the plane perpendicular to the axis and passing through the intersection of A and B. The appearance of a point image formed by a comatic lens is indicated in Fig. 3.5. Obviously the aberration is named after the comet shape of the figure.

Figure 3.6 indicates the relationship between the position at which the ray passes through the lens aperture and the location which it occupies in the coma patch. Figure 3.6a represents a head-on view of the lens aperture, with ray positions indicated by the letters A through

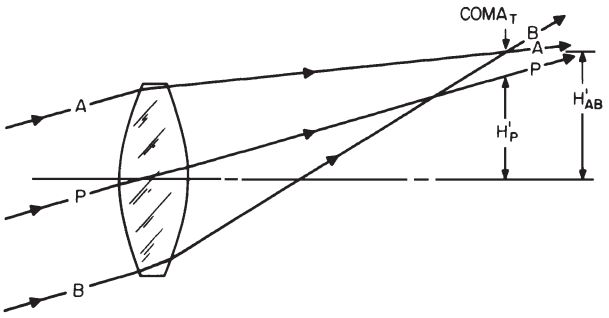


Figure 3.4 In the presence of coma, the rays through the outer portions of the lens focus at a different height than the rays through the center of the lens.

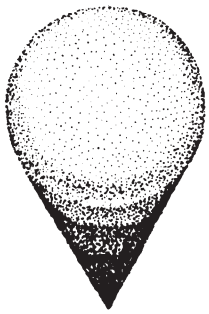


Figure 3.5 The coma patch. The image of a point source is spread out into a comet-shaped flare.

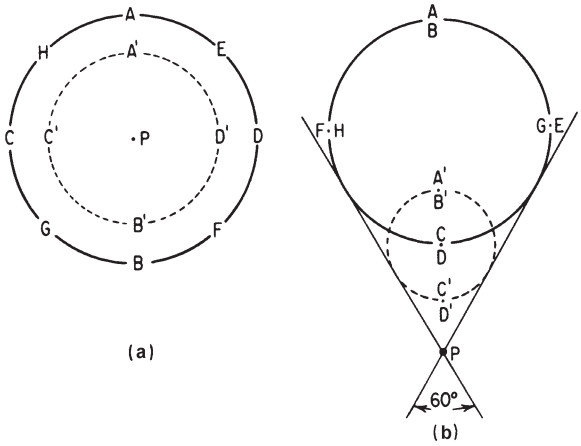


Figure 3.6 The relationship between the position of a ray in the lens aperture and its position in the coma patch. (a) View of the lens aperture with rays indicated by letters. (b) The letters indicate the positions of the corresponding rays in the image figure. Note that the diameters of the circles in the image are proportional to the *square* of the diameters in the aperture.

H and A' through D', with the primed rays in the inner circle. The resultant coma patch is shown in Fig. 3.6b with the ray locations marked with corresponding letters. Notice that the rays which formed a circle on the aperture also form a circle in the coma patch, but as the rays go around the aperture circle once, they go around the image circle twice in accord with the B_2 terms in Eqs. 3.1 and 3.2. The primed rays of the smaller circle in the aperture also form a correspondingly smaller circle in the image, and the central ray P is at the point of the figure. Thus the comatic image can be viewed as being made up of a series of different-sized circles arranged tangent to a 60° angle. The size of the image circle is proportional to the square of the diameter of the aperture circle.

In Fig. 3.6b the distance from P to AB is the tangential coma of Eq. 3.6. The distance from P to CD is called the sagittal coma and is one-third as large as the tangential coma. About half of all the energy in the coma patch is concentrated in the small triangular area between P and CD; thus the sagittal coma is a somewhat better measure of the *effective* size of the image blur than is the tangential coma.

Coma is a particularly disturbing aberration since its flare is nonsymmetrical. Its presence is very detrimental to accurate determination of the image position since it is much more difficult to locate the "center of gravity" of a coma patch than for a circular blur such as that produced by spherical aberration.

Coma varies with the shape of the lens element and also with the position of any apertures or diaphragms which limit the bundle of rays forming the image. In an axially symmetrical system there is no coma on the optical axis. The size of the coma patch varies linearly with its distance from the axis. The offense against the Abbe sine condition (OSC) is discussed in Chap. 10.

Astigmatism and field curvature

In the preceding section on coma, we introduced the terms "tangential" and "sagittal"; a fuller discussion of these terms is appropriate at this point. If a lens system is represented by a drawing of its axial section, rays which lie in the plane of the drawing are called *meridional* or *tangential* rays. Thus rays *A, P,* and *B* of Fig. 3.6 are tangential rays. Similarly, the plane through the axis is referred to as the *meridional* or *tangential* plane, as may *any* plane through the axis.

Rays which do not lie in a meridional plane are called *skew* rays. The oblique meridional ray through the center of the aperture of a lens system is called the *principal,* or *chief, ray.* If we imagine a plane passing through the chief ray and perpendicular to the meridional plane, then the (skew) rays from the object which lie in this sagittal plane are sagittal rays. Thus in Fig. 3.6 all the rays except A, A', P, B', and B are skew rays, and the sagittal rays are C, C', D', and D.

As shown in Fig. 3.7, the image of a point source formed by an oblique fan of rays in the tangential plane will be a line image; this line, called the *tangential image,* is perpendicular to the tangential plane; i.e., it lies in the sagittal plane. Conversely, the image formed by the rays of the sagittal fan is a line which lies in the tangential plane.

Astigmatism occurs when the tangential and sagittal (sometimes called radial) images do not coincide. In the presence of astigmatism, the image of a point source is not a point, but takes the form of two separate lines as shown in Fig. 3.7. Between the astigmatic foci the image is an elliptical or circular blur. (Note that if diffraction effects are significant, this blur may take on a square or diamond characteristic.)

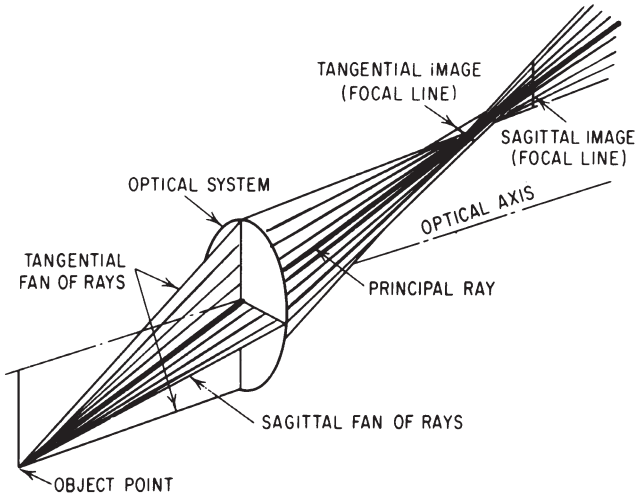


Figure 3.7 Astigmatism.

Unless a lens is poorly made, there is no astigmatism when an *axial* point is imaged. As the imaged point moves further from the axis, the amount of astigmatism gradually increases. Off-axis images seldom lie exactly in a true plane; when there is primary astigmatism in a lens system, the images lie on curved surfaces which are paraboloid in shape. The shape of these image surfaces is indicated for a simple lens in Fig. 3.8.

The amount of astigmatism in a lens is a function of the power and shape of the lens and its distance from the aperture or diaphragm which limits the size of the bundle of rays passing through the lens. In the case of a simple lens or mirror whose own diameter limits the size of the ray bundle, the astigmatism is equal to the square of the distance from the axis to the image (i.e., the image height) divided by the focal length of the element, i.e., $-h^2/f$.

Every optical system has associated with it a sort of basic field curvature, called the Petzval curvature, which is a function of the index of refraction of the lens elements and their surface curvatures. When there is no astigmatism, the sagittal and tangential image surfaces coincide with each other and lie on the Petzval surface. When there is primary astigmatism present, the tangential image surface lies three times as far from the Petzval surface as the sagittal image; note that both image surfaces are on the same side of the Petzval surface, as indicated in Fig. 3.8.

When the tangential image is to the left of the sagittal image (and both are to the left of the Petzval surface) the astigmatism is called negative, undercorrected, or inward-(toward the lens) curving. When the

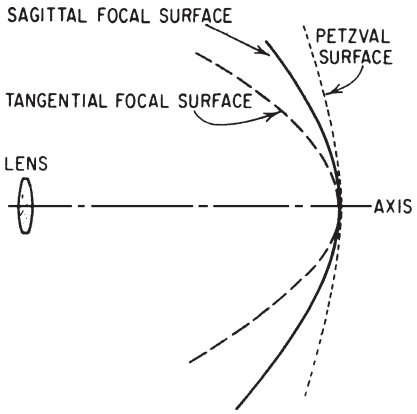


Figure 3.8 The primary astigmatism of a simple lens. The tangential image is three times as far from the Petzval surface as the sagittal image. Note that the figure is to scale.

order is reversed, the astigmatism is overcorrected, or backward-curving. In Fig. 3.8, the astigmatism is undercorrected and all three surfaces are inward-curving. It is possible to have overcorrected (backward curving) Petzval and undercorrected (inward) astigmatism, or vice versa.

Positive lenses introduce inward curvature of the Petzval surface to a system, and negative lenses introduce backward curvature. The Petzval curvature (i.e., the *longitudinal* departure of the Petzval surface from the ideal flat image surface) of a thin simple element is equal to one-half the square of the image height divided by the focal length and index of the element, $-h^2/2nf$. Note that "field curvature" means the *longitudinal* departure of the focal surfaces from the ideal image surface (which is usually flat) and not the reciprocal of the radius of the image surface.

Distortion

When the image of an off-axis point is formed farther from the axis or closer to the axis than the image height given by the paraxial expressions of Chap. 2, the image of an extended object is said to be distorted. The amount of distortion is the displacement of the image from the paraxial position, and can be expressed either directly or as a percentage of the ideal image height, which, for an infinitely distant object, is equal to $h' = f \tan \theta$.

The amount of distortion ordinarily increases as the image size increases; the distortion itself usually increases as the cube of the image height (percentage distortion increases as the square). Thus, if a centered rectilinear object is imaged by a system afflicted with distortion, it can be seen that the images of the corners will be displaced more (in proportion) than the images of the points making up the sides. Figure 3.9 shows the appearance of a square figure imaged by a

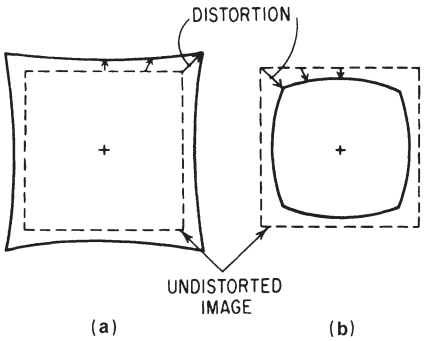


Figure 3.9 Distortion. (a) Positive, or pincushion, distortion. (b) Negative, or barrel, distortion. The sides of the image are curved because the amount of distortion varies as the cube of the distance from the axis. Thus, in the case of a square, the corners are distorted $2\sqrt{2}$ as much as the center of the sides.

lens system with distortion. In Fig. 3.9a the distortion is such that the images are displaced outward from the correct position, resulting in a flaring or pointing of the corners. This is overcorrected, or pincushion, distortion. In Fig. 3.9b the distortion is of the opposite type and the corners of the square are pulled inward more than the sides; this is negative, or barrel, distortion.

A little study of the matter will show that a system which produces distortion of one sign will produce distortion of the opposite sign when object and image are interchanged. Thus a camera lens with barrel distortion will have pincushion distortion if used as a projection lens (i.e., when the film is replaced by a slide). Obviously if the same lens is used both to photograph and to project the slide, the projected image will be rectilinear (free of distortion) since the distortion in the slide will be canceled out upon projection.

3.3 Chromatic Aberrations

Because of the fact that the index of refraction varies as a function of the wavelength of light, the properties of optical elements also vary with wavelength. Axial chromatic aberration is the longitudinal variation of focus (or image position) with wavelength. In general, the index of refraction of optical materials is higher for short wavelengths than for long wavelengths; this causes the short wavelengths to be more strongly refracted at each surface of a lens so that in a simple positive lens, for example, the blue light rays are brought to a focus closer to the lens than the red rays. The distance along the axis between the two focus points is the longitudinal axial chromatic aberration. Figure 3.10 shows the chromatic aberration of a simple positive element. When the short-wavelength rays are brought to a focus to the left of the long-wavelength rays, the chromatic is termed undercorrected, or negative.

The image of an axial point in the presence of chromatic aberration is a central bright dot surrounded by a halo. The rays of light which are in focus, and those which are nearly in focus, form the bright dot.

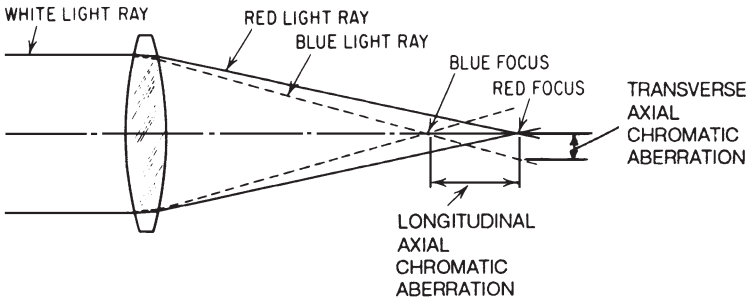


Figure 3.10 The undercorrected longitudinal chromatic aberration of a simple lens is due to the blue rays undergoing a greater refraction than the red rays.

The out-of-focus rays form the halo. Thus, in an undercorrected visual instrument, the image would have a yellowish dot (formed by the orange, yellow, and green rays) and a purplish halo (due to the red and blue rays). If the screen on which the image is formed is moved toward the lens, the central dot will become blue; if it is moved away, the central dot will become red.

When a lens system forms images of different sizes for different wavelengths, or spreads the image of an off-axis point into a rainbow, the difference between the image heights for different colors is called *lateral color,* or *chromatic difference of magnification.* In Fig. 3.11 a simple lens with a displaced diaphragm is shown forming an image of an off-axis point. Since the diaphragm limits the rays which reach the lens, the ray bundle from the off-axis point strikes the lens above the axis and is bent downward as well as being brought to a focus. The blue rays are bent downward more than the red and thus form their image nearer the axis.

The chromatic variation of index also produces a variation of the monochromatic aberrations discussed in Sec. 3.2. Since each aberration results from the manner in which the rays are bent at the surfaces of the optical system, it is to be expected that, since rays of different color are bent differently, the aberrations will be somewhat different for each color. In general this proves to be the case, and these effects are of practical importance when the basic aberrations are well corrected.

3.4 The Effect of Lens Shape and Stop Position on the Aberrations

A consideration of either the thick-lens focal length equation

$$\frac{1}{f} = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2} + \frac{n - 1}{n}\,\frac{t}{R_1 R_2}\right)$$

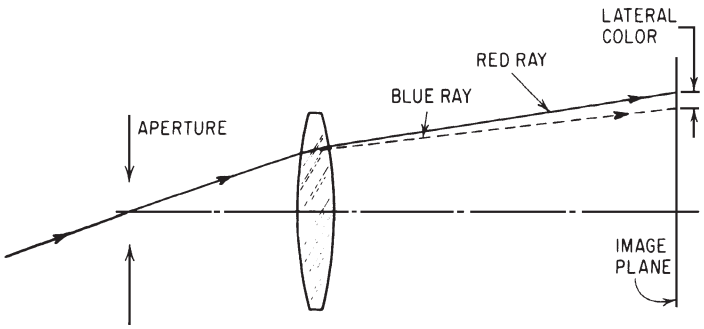


Figure 3.11 Lateral color, or chromatic difference of magnification, results in different-sized images for different wavelengths.

or the thin-lens focal length equation

$$\frac{1}{f} = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = (n - 1)\,(C_1 - C_2)$$

reveals that for a given index and thickness, there is an infinite number of combinations of R_1 and R_2 which will produce a given focal length. Thus a lens of some desired power may take on any number of different shapes or "bendings." The aberrations of the lens are changed markedly as the shape is changed; this effect is the basic tool of optical design.

As an illustrative example, we will consider the aberrations of a thin positive lens made of borosilicate crown glass with a focal length of 100 mm and a clear aperture of 100 mm (a speed of *f*/10) which is to image an infinitely distant object over a field of view of ±17°. A typical borosilicate crown is 517:642, which has an index of 1.517 for the helium *d* line ($\lambda = 5876$ Å), an index of 1.51432 for *C* light ($\lambda = 6563$ Å), and an index of 1.52238 for *F* light ($\lambda = 4861$ Å).

(The aberration data presented in the following paragraphs were calculated by means of the thin-lens third-order aberration equations of Chap. 10.)

If we first assume that the stop or limiting aperture is in coincidence with the lens, we find that several aberrations do *not* vary as the lens shape is varied. Axial chromatic aberration is constant at a value of −1.55 mm (undercorrected); thus the blue focus (*F* light) is 1.55 mm nearer the lens than the red focus (*C* light). The astigmatism and field curvature are also constant. At the edge of the field (30 mm from the axis) the sagittal focus is 7.5 mm closer to the lens than the paraxial focus, and the tangential focus is 16.5 mm inside the paraxial focus. Two aberrations, distortion and lateral color, are zero when the stop is at the lens.

Spherical aberration and coma, however, vary greatly as the lens shape is changed. Figure 3.12 shows the amount of these two aberrations plotted against the curvature of the first surface of the lens. Notice that coma varies linearly with lens shape, taking a large positive value when the lens is a meniscus with both surfaces concave toward the object. As the lens is bent through plano-convex, convex-plano, and convex meniscus shapes, the amount of coma becomes more negative, assuming a zero value near the convex-plano form.

The spherical aberration of this lens is always undercorrected; its plot has the shape of a parabola with a vertical axis. Notice that the spherical aberration reaches a minimum (or more accurately, a maximum) value at approximately the same shape for which the coma is zero. This, then, is the shape that one would select if the lens were to be used as a telescope objective to cover a rather small field of view. Note that if both object and image are "real" (i.e., not virtual), the spherical aberration of a positive lens is always negative (undercorrected).

Let us now select a particular shape for the lens, say, $C_1 = -0.02$ and investigate the effect of placing the stop away from the lens, as shown in Fig. 3.13. The spherical and axial chromatic aberrations are

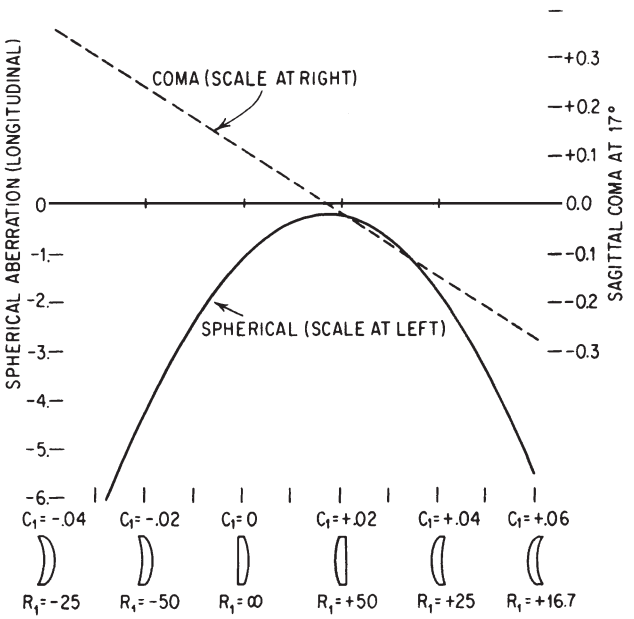


Figure 3.12 Spherical aberration and coma as a function of lens shape. Data plotted are for a 100-mm focal length lens (with the stop at the lens) at *f*/10 covering ±17° field.

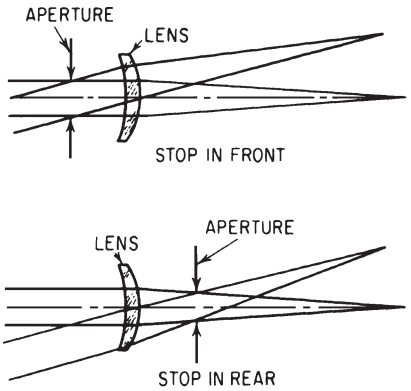


Figure 3.13 The aperture stop away from the lens. Notice that the oblique ray bundle passes through an entirely different part of the lens when the stop is in front of the lens than when it is behind the lens.

completely unchanged by shifting the stop, since the axial rays strike the lens in exactly the same manner regardless of where the stop is located. The lateral color and distortion, however, take on positive values when the stop is behind the lens and negative when it is before the lens. Figure 3.14 shows a plot of lateral color, distortion, coma, and tangential field curvature as a function of the stop position. The most pronounced effects of moving the stop are found in the variations of coma and astigmatism. As the stop is moved toward the object, the coma decreases linearly with stop position, and has a zero value when the stop is about 18.5 mm in front of the lens. The astigmatism becomes less negative so that the position of the tangential image approaches the paraxial focal plane. Since astigmatism is a quadratic function of stop position, the tangential field curvature (x_t) plots as a parabola. Notice that the parabola has a maximum at the same stop position for which the coma is zero. This is called the *natural* position of the stop, and for all lenses with undercorrected primary spherical aberration, the natural, or coma-free, stop position produces a more backward curving (or less inward curving) field than any other stop position.

Figure 3.12 showed the effect of lens shape with the stop fixed in contact with the lens, and Fig. 3.14 showed the effect of stop position with the lens shape held constant. There is a "natural" stop position for each shape of the simple lens we are considering. In Fig. 3.15, the aberrations of the lens have again been plotted against the lens shape; however, in this figure, the aberration values are those which occur when the stop is in the natural position. Thus, for each bending the coma has been removed by choosing this stop position, and the field is as far backward curving as possible.

Notice that the shape which produces minimum spherical aberration also produces the maximum field curvature, so that this shape,

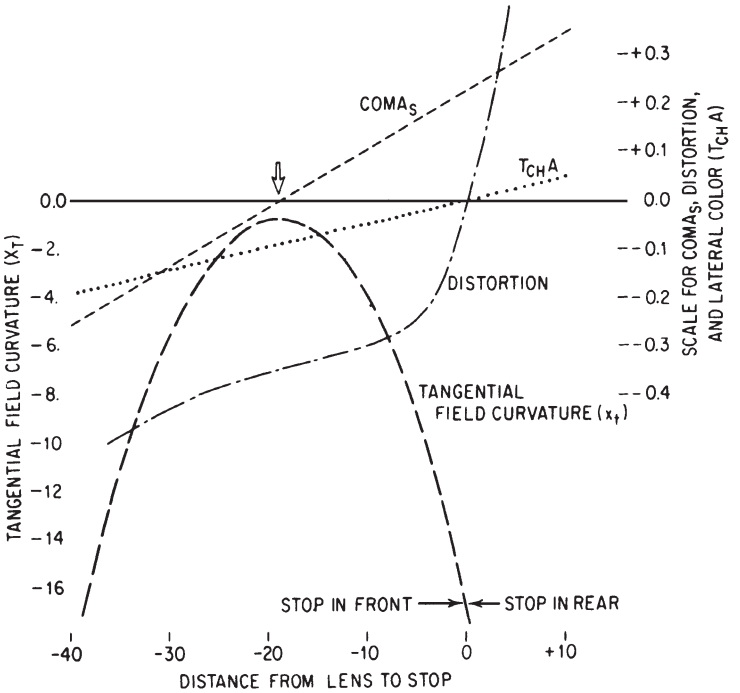


Figure 3.14 Effect of shifting the stop position on the aberrations of a simple lens. The arrow indicates the "natural" stop position where coma is zero. (efl = 100, $C_1 = -0.02$, speed = *f*/10, field = ±17°.)

which gives the best image near the axis, is not suitable for wide field coverage. The meniscus shapes at either side of the figure represent a much better choice for a wide field, for although the spherical aberration is much larger at these bendings, the field is much more nearly flat. This is the type of lens used in inexpensive cameras at speeds of *f*/11 or *f*/16.

3.5 Aberration Variation with Aperture and Field

In the preceding section, we considered the effect of lens shape and aperture position on the aberrations of a simple lens, and in that discussion we assumed that the lens operated at a fixed aperture of *f*/10 (stop diameter of 10 mm) and covered a fixed field of ±17° (field diameter of 60 mm). It is often useful to know how the aberrations of such a lens vary when the size of the aperture or field is changed.

Figure 3.16 lists the relationships between the primary aberrations and the semi-aperture y (in column one) and the image height

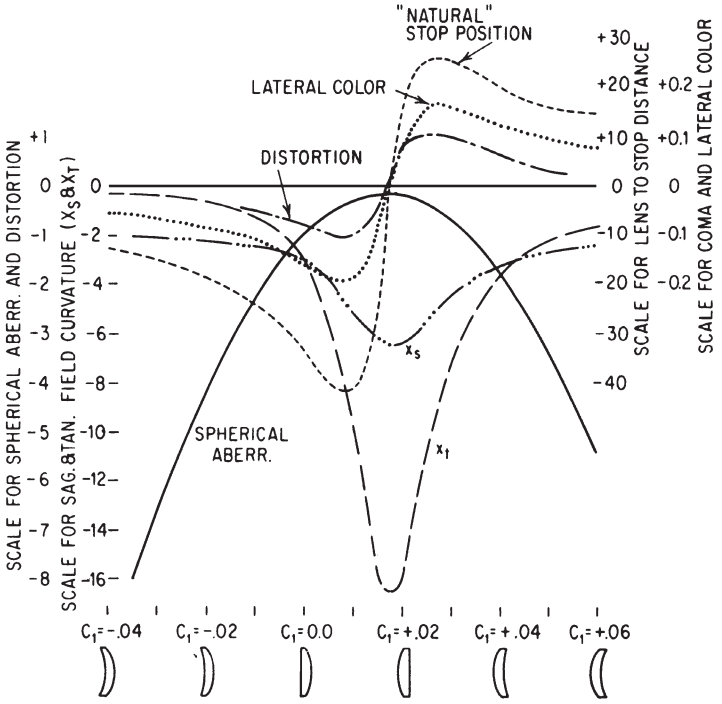


Figure 3.15 The variation of the aberrations with lens shape when the stop is located in the "natural" (coma free) position for each shape. Data is for 100-mm *f*/10 lens covering ±17° field, made from BSC-2 glass (517:645).

h (in column two). To illustrate the use of this table, let us assume that we have a lens whose aberrations are known; we wish to determine the size of the aberrations if the aperture diameter is increased by 50 percent and the field coverage reduced by 50 percent. The new y will be 1.5 times the original, and the new h will be 0.5 times the original.

Since longitudinal spherical aberration is shown to vary with y^2, the 1.5 times increase in aperture will cause the spherical to be $(1.5)^2$, or 2.25, times as large. Similarly transverse spherical, which varies as y^3, will be $(1.5)^3$, or 3.375, times larger (as will the image blur due to spherical).

Coma varies as y^2 and h; thus, the coma will be $(1.5)^2 \times 0.5$, or 1.125, times as large. The Petzval curvature and astigmatism, which vary with h^2, will be reduced to $(0.5)^2$, or 0.25, of their previous value, while the blurs due to astigmatism or field curvature will be $1.5(0.5)^2$, or 0.375, of their original size.

The aberrations of a lens also depend on the position of the object and image. A lens which is well corrected for an infinitely distant

Aberration	vs. Aperture	vs. Field Size or Angle
Spherical (longitudinal)	y^2	—
Spherical (transverse)	y^3	—
Coma	y^2	h
Petzval curvature (longitudinal)	—	h^2
Petzval curvature (transverse)	y	h^2
Astigmatism and field curvature (longitudinal)	—	h^2
Astigmatism and field curvature (transverse)	y	h^2
Distortion (linear)	—	h^3
Distortion (in percent)	—	h^2
Axial chromatic (longitudinal)	—	—
Axial chromatic (transverse)	y	—
Lateral chromatic	—	h
Lateral chromatic (CDM)	—	—

Figure 3.16 The variation of the primary aberrations with aperture and field.

object, for example, may be very poorly corrected if used to image a nearby object. This is because the ray paths and incidence angles change as the object position changes.

It should be obvious that if *all* the dimensions of an optical system are scaled up or down, the *linear* aberrations are also scaled in exactly the same proportion. Thus if the simple lens used as the example in Sec. 3.4 were increased in focal length to 200 mm, its aperture increased to 20 mm, and the field coverage increased to 120 mm, then the aberrations would all be doubled. Note, however, that the speed, or *f*/number, would remain at *f*/10 and the angular coverage would remain at ±17 . The *percentage* distortion would not be changed.

Aberrations are occasionally expressed as angular aberrations. For example, the transverse spherical aberration of a system subtends an angle from the second principal point of the system; this angle is the angular spherical aberration. Note that the angular aberrations are not changed by scaling the size of the optical system.

3.6 Optical Path Difference (Wave Front Aberration)

Aberrations can also be described in terms of the wave nature of light. In Chap. 1, it was pointed out that the light waves converging to form a "perfect" image would be spherical in shape. Thus when aberrations are present in a lens system, the waves converging on an image point are deformed from the ideal shape (which is a sphere centered on the image point). For example, in the presence of undercorrected spherical aberration the wave front is curled inward at the edges, as shown in Fig. 3.17. This can be understood if we remember that a ray is the path

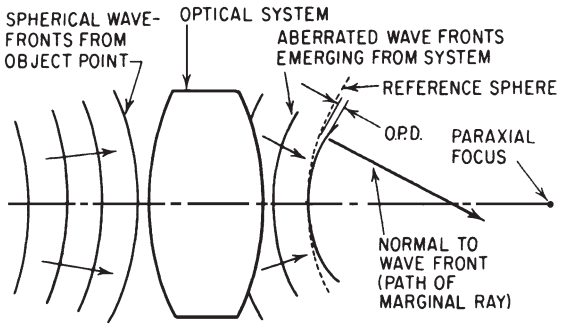


Figure 3.17 The optical path difference (OPD) is the distance between the emerging wave front and a reference sphere (centered in the image plane) which coincides with the wave front at the axis. The OPD is thus the difference between the marginal and axial paths through the system for an axial point.

of a point on the wave front and that the ray is also normal to the wave front. Thus, if the ray is to intersect the axis to the left of the paraxial focus, the section of the wave front associated with the ray must be curled inward. The wave front shown is "ahead" of the reference sphere; the distance by which it is ahead is called the *optical path difference,* or OPD, and is customarily expressed in units of wavelengths. The wave fronts associated with axial aberrations are symmetrical figures of rotation, in contrast to the off-axis aberrations such as coma and astigmatism. For example, the wave front for astigmatism would be a section of a torus (the outer surface of a doughnut) with different radii in the prime meridians. For off-axis imagery, the reference sphere is chosen to pass through the center of the exit pupil (in some calculations, the reference sphere has an infinite radius, for convenience in computing).

3.7 Aberration Correction and Residuals

Section 3.4 indicated two methods which are used to control aberrations in simple optical systems, namely lens shape and stop position. For many applications a higher level of correction is needed, and it is then necessary to combine optical elements with aberrations of opposite signs so that the aberrations contributed to the system by one element are cancelled out, or corrected, by the others. A typical example is the achromatic doublet used for telescope objectives, shown in Fig. 3.18. A single positive element would be afflicted with both undercorrected spherical aberration and undercorrected chromatic aberration. In a negative element, in the other hand, both aberrations are overcorrected. In the doublet a positive element is combined with a less

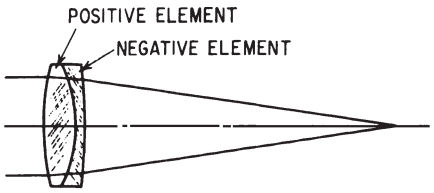


Figure 3.18 Achromatic doublet telescope objective. The powers and shapes of the two elements are so arranged that each cancels the aberrations of the other.

powerful negative element in such a way that the aberrations of each balance out. The positive lens is made of a (crown) glass with a low chromatic dispersion, and the negative element of a (flint) glass with a high dispersion. Thus, the negative element has a greater amount of chromatic aberration *per unit of power,* by virtue of its greater dispersion, than the crown element. The relative powers of the elements are chosen so that the chromatic exactly cancels while the focusing power of the crown element dominates.

The situation with regard to spherical aberration is quite analogous except that element power, shape, and index of refraction are involved instead of power and dispersion as in chromatic. If the index of the negative element is higher than the positive, the inner surface is divergent, and will contribute overcorrected spherical to balance the undercorrection of the outer surfaces.

Aberration correction usually is exact only for one zone of the aperture of a lens or for one angle of obliquity, because the aberrations of the individual elements do not balance out exactly for all zones and angles. Thus, while the spherical aberration of a lens may be corrected to zero for the rays through the edge of the aperture, the rays through the other zones of the aperture usually do not come to a focus at the paraxial image point. A typical longitudinal spherical aberration plot for a "corrected" lens is shown in Fig. 3.19. Notice that the rays through only one zone of the lens intersect the paraxial focus. Rays through the smaller zones focus nearer the lens system and have undercorrected spherical; rays above the corrected zone show overcorrected spherical. The undercorrected aberration is called residual, or zonal, aberration; Fig. 3.19 would be said to show an undercorrected zonal aberration. This is the usual state of affairs for most optical systems. Occasionally a system is designed with an overcorrected spherical zone, but this is unusual.

Chromatic aberration has residuals which take two different forms. The correction of chromatic aberration is accomplished by making the foci of two different wavelengths coincide. However, due to the nature of the great majority of optical materials, the nonlinear dispersion characteristics of the positive and negative elements used in an achromat do not "match up," so that the focal points of other wavelengths do

not coincide with the common focal point of the two selected colors. This difference in focal distance is called secondary spectrum. Figure 3.20 shows a plot of back focal distance versus wavelength for a typical achromatic lens, in which the rays for *C* light (red) and *F* light (blue) are brought to a common focus. The yellow rays come to a focus about 1/2400th of the focal length ahead of the *C*-*F* focal point.

The second major chromatic residual may be regarded as a variation of chromatic aberration with ray height, or as a variation of spherical aberration with wavelength, and is called spherochromatism. In ordinary spherochromatism, the spherical aberration in blue light is overcorrected and the spherical in red light is undercorrected (when the spherical aberration for the yellow light is corrected). Figure 3.21 is a spherical aberration plot in three wavelengths for a typical achromatic doublet of large aperture. The correction has been adjusted so that the red and blue rays striking the lens at a height of 0.707 of the marginal ray height are brought to a common focus. The distance between the yellow focus and the red-blue focus at this height is, of course, the

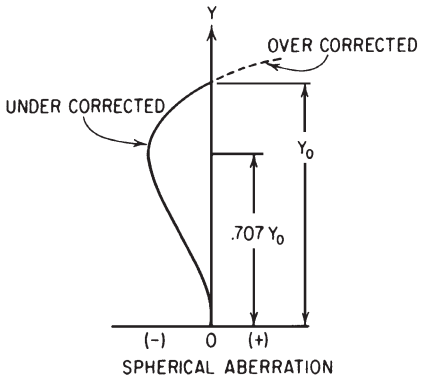


Figure 3.19 Plot of longitudinal spherical aberration versus ray height for a "corrected" lens. For most lenses, the maximum undercorrection occurs for the ray whose height is 0.707 that of the ray with zero spherical.

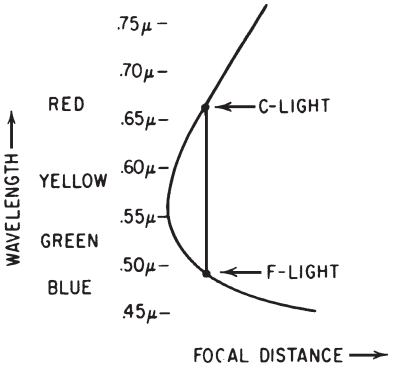


Figure 3.20 The secondary spectrum of a typical doublet achromat, corrected so that *C* and *F* light are joined at a common focus. The distance from the common focus of *C* and *F* to the minimum of the curve (in the yellow green at about 0.55 μ) is called the *secondary spectrum.*

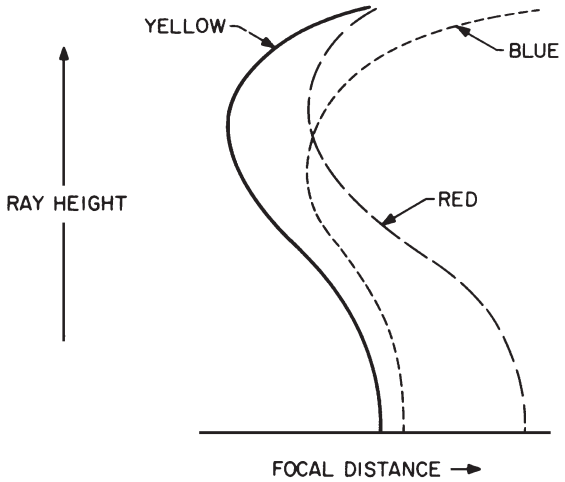


Figure 3.21 Spherochromatism. The longitudinal aberration of a "corrected" lens is shown for three wavelengths. The marginal spherical for yellow light is corrected but is overcorrected for blue light and undercorrected for red. The chromatic aberration is corrected at the zone but is overcorrected above it and undercorrected below. A transverse plot of these aberrations is shown in Fig. 3.24k.

secondary spectrum discussed above. Notice that above this 0.707 zone the chromatic is overcorrected and below it is undercorrected so that one half of the area of the lens aperture is overconnected and one half undercorrected.

The other aberrations have similar residuals. Coma may be completely corrected for a certain field angle, but will often be overcorrected above this obliquity and undercorrected below it. Coma may also undergo a change of sign with aperture, with the central part of the aperture overcorrected and the outer zone undercorrected.

Astigmatism usually varies markedly with field angle. Figure 3.22 shows a plot of the sagittal and tangential field curvatures for a typical photographic anastigmat, in which the astigmatism is zero for one zone of the field. This point is called the node, and typically the two focal surfaces separate quite rapidly beyond the node.

3.8 Ray Intercept Curves and the "Orders" of Aberrations

When the image plane intersection heights of a fan of meridional rays are plotted against the slope of the rays as they emerge from the lens, the resultant curve is called a *ray intercept curve* or an H'–tan U' curve. The shape of the intercept curve not only indicates the amount of spreading or blurring of the image directly, but also can serve to

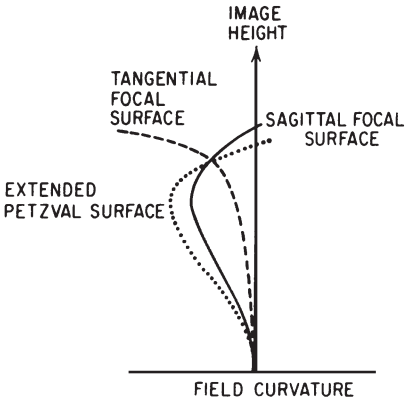


Figure 3.22 Field curvature of a photographic anastigmat. The astigmatism has been corrected for one zone of the field but is overcorrected inside this zone and undercorrected beyond it.

indicate which aberrations are present. Figure 3.3b, for example, shows simple undercorrected spherical aberration.

In Fig. 3.23, an oblique fan of rays from a distant object point is brought to a perfect focus at point *P*. If the reference plane passes through *P*, it is apparent that the H'–tan U' curve will be a straight horizontal line. However, if the reference plane is behind *P* (as shown) then the ray intercept curve becomes a tilted straight line since the height, H', decreases as tan U' decreases. Thus it is apparent that shifting the reference plane (or focusing the system) is equivalent to a rotation of the H'–tan U' coordinates. A valuable feature of this type of aberration representation is that one can immediately assess the effects of refocusing the optical system by a simple rotation of the abscissa of the figure. Notice that the slope of the line ($\Delta H'/\Delta$ tan U') is exactly equal to the distance (δ) from the reference plane to the point of focus, so that for an oblique ray fan the tangential field curvature is equal to the slope of the ray intercept curve.

The accepted convention for plotting the ray intercepts is that (1) they are plotted for positive image heights (i.e., above the axis) and (2) that the ray through the top of the lens is plotted at the right end of the plot. For compound systems, where the image is relayed by a second component, the ray plotted to the right is the one with the most negative slope, i.e., the one through the bottom of the first component. The result of this is that the sign of the aberrations shown in the ray intercept plot can be instantly recognized. For example, the plot for undercorrected spherical always curves down at the right end and up at the left, and a line connecting the ends of a plot showing positive coma always passes above the point representing the principal ray. Note that in an H'–tan U' plot, this plotting convention violates the convention for the sign of the ray slope. This seeming contradiction is the result of the change from the historical optical ray slope sign convention which occurred several decades ago.

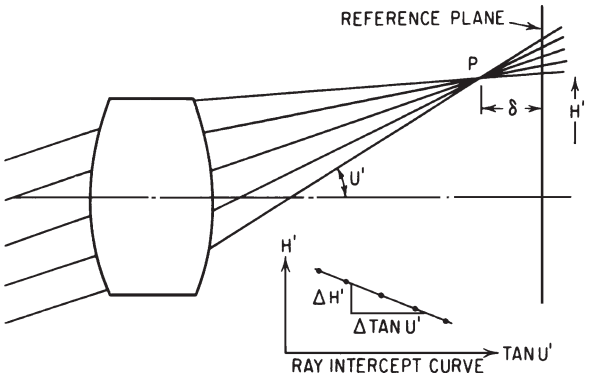


Figure 3.23 The ray intercept curve (H' – tan U') of a point which does not lie in the reference plane is a tilted straight line. The slope of the line ($\Delta H'/\Delta \tan U'$) is mathematically identical to δ, the distance from the reference plane to the point of focus P. Note that δ is equal to X_T, the tangential field curvature, when the paraxial focal plane is chosen as the reference plane.

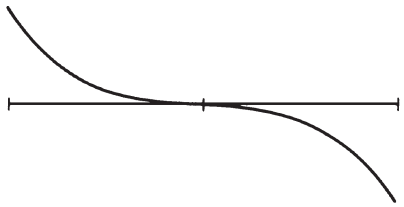
Figure 3.24 shows a number of intercept curves, each labeled with the aberration represented. The generation of these curves can be readily understood by sketching the ray paths for each aberration and then plotting the intersection height and slope angle for each ray as a point of the curve. Distortion is not shown in Fig. 3.24; it would be represented as a vertical displacement of the curve from the paraxial image height h'. Lateral color would be represented by curves for two colors which were vertically displaced from each other. The ray intercept curves of Fig. 3.24 are generated by tracing a fan of meridional or tangential rays from an object point and plotting their intersection heights versus their slopes. The imagery in the other meridian can be examined by tracing a fan of rays in the sagittal plane (normal to the meridional plane) and plotting their x-coordinate intersection points against their slopes in the sagittal plane (i.e., the slope relative to the principal ray lying in the meridional plane). Note that Fig. 3.24k is for the same lens as the longitudinal plot in Fig. 3.21.

It is apparent that the ray intercept curves which are "odd" functions, that is, the curves which have a rotational or point symmetry about the origin, can be represented mathematically by an equation of the form

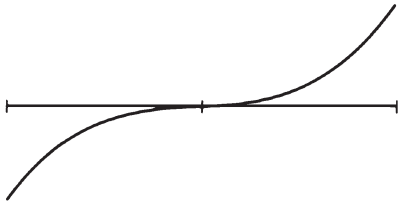
$$y = a + bx + cx^3 + dx^5 + \cdots$$

or

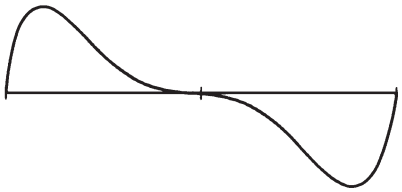
$$H' = a + b \tan U' + c \tan^3 U' + d \tan^5 U' + \cdots \qquad (3.7)$$



(a) UNDERCORRECTED SPHERICAL ABERRATION



(b) OVERCORRECTED SPHERICAL ABERRATION



(c) UNDERCORRECTED ZONAL SPHERICAL ABERRATION

Figure 3.24 The ray intercept plots for various aberrations. The ordinate for each curve is *H,* the height at which the ray intersects the (paraxial) image plane. The abscissa is tan *U,* the final slope of the ray with respect to the optical axis. Note that it is conventional to plot the ray through the top of the lens at the right of the figure, and that curves for image points above the axis are customarily shown. [Figure continues with parts (d) to (k).]

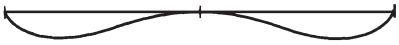
All the ray intercept curves for *axial* image points are of this type. Since the curve for an axial image must have $H' = 0$ when $\tan U' = 0$, it is apparent that the constant a must be a zero. It is also apparent that the constant b for this case represents the amount the reference plane is displaced from the paraxial image plane. Thus the curve for lateral spherical aberration plotted with respect to the paraxial focus can be expressed by the equation

$$TA' = c \tan^3 U' + d \tan^5 U' + e \tan^7 U' + \cdots \qquad (3.8)$$

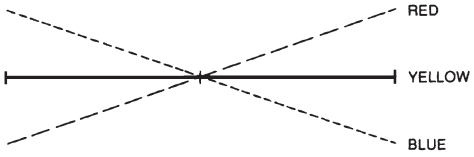
It is, of course, possible to represent the curve by a power series expansion in terms of the final angle U', or $\sin U'$, or the ray height at the lens (Y), or even the initial slope of the ray at the object (U_0) instead of $\tan U'$. The constants will, of course, be different for each.

For simple uncorrected lenses the first term of Eq. 3.8 is usually adequate to describe the aberration. For the great majority of "corrected" lenses the first two terms are dominant; in a few cases

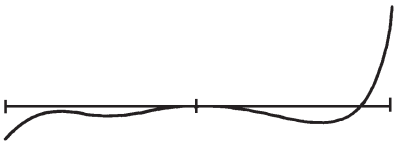
(d) UNDERCORRECTED COMA



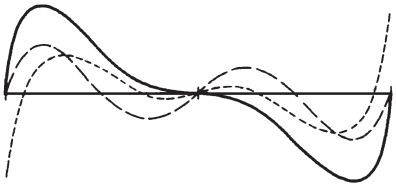
(e) COMA --- THIRD AND FIFTH ORDER



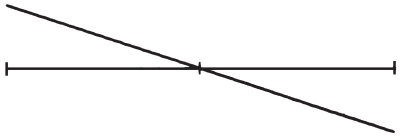
(f) UNDERCORRECTED AXIAL CHROMATIC ABERRATION



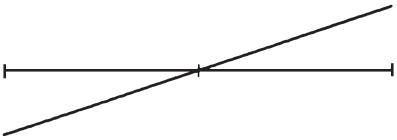
(j) A "TYPICAL" OFF-AXIS CURVE



(k) ON-AXIS PLOT FOR AN ACHROMATIC DOUBLET, SHOWING ZONAL SPHERICAL, SECONDARY SPECTRUM, AND SPHERO-CHROMATISM



(g) INWARD CURVING TANGENTIAL FIELD



(h) BACKWARD CURVING TANGENTIAL FIELD

(i) OVERCORRECTED SPHERICAL REFOCUSSED FOR MINIMUM BLUR

Figure 3.24 (*Continued*)

three terms (and rarely four) are necessary to satisfactorily represent the aberration. As examples, Figs. 3.3, 3.24a, and 3.24b can be represented by $TA' = c \tan^3 U'$, and this type of aberration is called third-order spherical. Figure 3.24c, however, would require two terms of the expansion to represent it adequately; thus $TA' = c \tan^3 U' + d \tan^5 U'$. The amount of aberration represented by the second term is called the fifth-order aberration. Similarly, the aberration represented by the third term of Eq. 3.8 is called the seventh-order aberration. The fifth-, seventh-, ninth-, etc., order aberrations are collectively referred to as higher-order aberrations.

As will be shown in Chap. 10, it is possible to calculate the amount of the primary, or third-order, aberrations without trigonometric raytracing, that is, by means of data from a paraxial raytrace. This type of aberration analysis is called *third-order theory.* The name "first-order optics" given to that part of geometrical optics devoted to locating the paraxial image is also derived from this power series expansion, since the first-order term of the expansion results purely from a longitudinal displacement of the reference plane from the paraxial focus.

Notes on the interpretation of ray intercept plots

The ray intercept plot is subject to a number of interesting interpretations. It is immediately apparent that the top-to-bottom extent of the plot gives the size of the image blur. Also, a rotation of the horizontal (abscissa) lines of the graph is equivalent to a refocusing of the image and can be used to determine the effect of refocusing on the size of the blur.

Figure 3.23 shows that the ray intercept plot for a defocused image is a sloping line. If we consider the slope of the curve at any point on an H–tan U ray intercept plot, the slope is equal to the defocus of a small-diameter bundle of rays centered about a ray represented by that point. In other words, this would represent the focus of the rays passing through a pinhole aperture which was so positioned as to pass the rays at that part of the H–tan U plot. Similarly, since shifting an aperture stop along the axis is, for an oblique bundle of rays, the equivalent of selecting one part or another of the ray intercept plot, one can understand why shifting the stop can change the field curvature, as discussed in Section 3.4.

The OPD (optical path difference) or wave-front aberration can be derived from an H–tan U ray intercept plot. The area under the curve between two points is equal to the OPD between the two rays which correspond to the two points. Ordinarily, the reference ray for OPD is either the optical axis or the principal ray (for an oblique bundle).

Thus the OPD for a given ray is usually the area under the ray intercept plot between the center point and the ray.

Mathematically speaking, then, the OPD is the integral of the H–tan U plot and the defocus is the first derivative. The coma is related to the curvature or second derivative of the plot, as a glance at Fig. 3.24d will show.

It should be apparent that a more general ray intercept plot for a given object point can be considered as a power series expansion of the form

$$H' = h + a + bx + cx^2 + dx^3 + ex^4 + fx^5 + \cdots \tag{3.9}$$

where h is the paraxial image height, a is the distortion, and x is the aperture variable (e.g., tan U'). Then the art of interpreting a ray intercept plot becomes analogous to decomposing the plot into its various terms. For example, cx^2 and ex^4 represent third- and fifth-order coma, while dx^3 and fx^5 are the third- and fifth-order spherical. The bx term is due to a defocusing from the paraxial focus and could be due to curvature of field. Note that the constants *a, b, c,* etc., will be different for points of differing distances from the axis. For the primary aberrations, the constants will vary according to the table of Fig. 3.16, and in general per Eqs. 3.1 and 3.2.

Bibliography

Note: Titles preceded by an asterisk are out of print.

Smith, W., "Image Formation: Geometrical and Physical Optics" in W. Driscoll (ed.), *Handbook of Optics*, New York, McGraw-Hill, 1978.

Smith, W., "Optical Design" (Chap. 8) and "Optical Elements—Lenses and Mirrors" (Chap. 9) in W. Wolfe and G. Zissis (eds.), *The Infrared Handbook*, Arlington, Va., Office of Naval Research, 1985.

*Welford, W., *Aberrations of the Symmetrical Optical System*, New York, Academic, 1974.

Exercises

1 The longitudinal spherical aberration of two rays which have been traced through a system is -1.0 and -0.5; the ray slopes (tan U') are -0.5 and -0.35, respectively. What are the transverse aberrations (a) in the paraxial plane, and (b) in a plane 0.2 before the paraxial plane?

ANSWER: (a) -0.5, -0.175; (b) -0.4, -0.105

2 A lens has $\text{coma}_T = 1$. Plot the focal plane intercepts of 12 rays equally spaced around (a) the marginal zone, (b) the 0.707 zone, and (c) the 0.5 zone (see Fig. 3.6).

3 A certain type of lens has the following primary aberrations at a focal length of 100, an aperture of 10, and a field of ±5°: longitudinal spherical = 1.0; $\text{coma}_T = 1.0$; $X_T = 1.0$. What are the aberrations of this type of lens when: (a) $f = 200$, aperture = 10, field = ±2.50? (b) $f = 50$, aperture = 10, field = ±10°?

ANSWER: (a) LA = 0.5, $\text{Coma}_T = 0.25$, $X_T = 0.5$, (b) LA = 2.0, $\text{Coma}_T = 4.0$, $X_T = 2.0$

4 Plot the ray intercept curve for a lens with transverse spherical, coma_T, and X_T, each equal to 1.0. Assume third-order aberrations and also that $\tan U'_m = 1.0$.