NONIMAGING
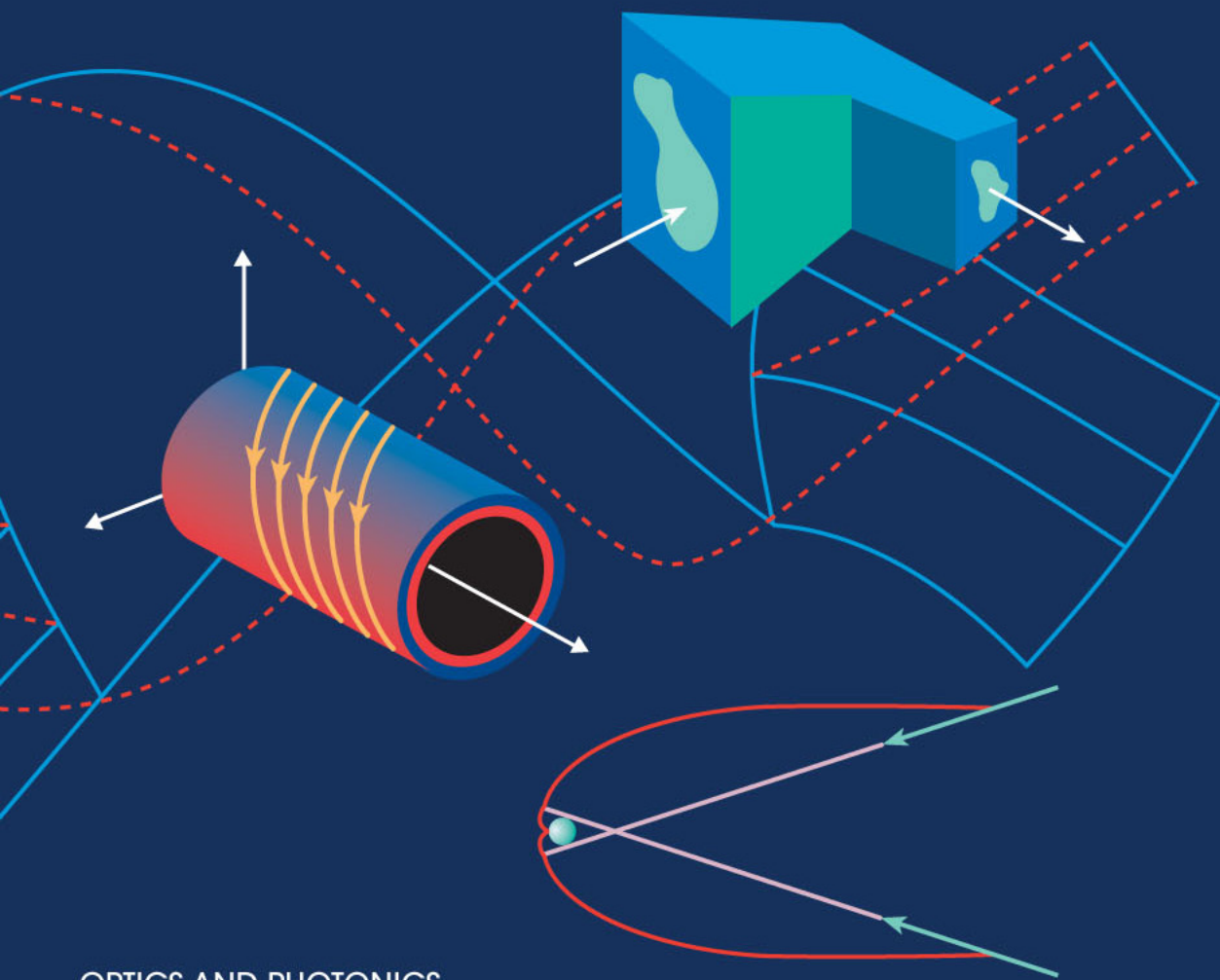# OPTICS

ROLAND WINSTON
JUAN C.MIÑANO
PABLO BENÍTEZ

With contributions by
Narkis Shatz *and* John C. Bortz

OPTICS AND PHOTONICS

NONIMAGING OPTICS

NONIMAGING OPTICS

Roland Winston

University of California, Merced, CA

Juan C. Miñano and *Pablo Benítez*

Technical University of Madrid UPM, CEDINT, Madrid,
Spain and Light Prescriptions Innovators LLC, Irvine, CA

With contributions by

Narkis Shatz and *John C. Bortz*

Science Applications International Corporation, San Diego, CA

This book is printed on acid-free paper. ∞

CONTENTS

APPENDICES

PREFACE

This book is the successor to *High Collection Nonimaging Optics*, published by Academic Press in 1989, and *Optics of Nonimaging Concentrators*, published 10 years earlier, by W. T. Welford and R. Winston. Walter Welford was one of the most distinguished optical scientists of his time. His work on aberration theory remains the definitive contribution to the subject. From 1976 until his untimely death in 1990, he took on the elucidation of nonimaging optics with the same characteristic vigor and enthusiasm he had applied to imaging optics. As a result, nonimaging optics developed from a set of heuristics to a complete subject. We dedicate this book to his memory.

It incorporates much of the pre-1990 material as well as significant advances in the subject. These include elaborations of the flow-line method, designs for prescribed irradiance, simultaneous multiple surface method, optimization, and symmetry breaking. A discussion of radiance connects theory with measurement in a physical way.

We will measure our success by the extent to which our readers advance the subject over the next 10 years.

RW
JCM
PB
NS
JB

Photograph of W. T. Welford
(*courtesy of Jacqueline Welford*)

1

NONIMAGING OPTICAL SYSTEMS AND THEIR USES

1.1 NONIMAGING COLLECTORS

Nonimaging concentrators and illuminators have several actual and some potential applications, but it is best to explain the general concept of a nonimaging concentrator by highlighting one of its applications; its use of solar energy. The radiation power density received from the sun at the earth's surface, often denoted by S, peaks at approximately $1\,\mathrm{kWm^{-2}}$, depending on many factors. If we attempt to collect this power by absorbing it on a perfect blackbody, the equilibrium temperature T of the blackbody will be given by[1]

$$\sigma T^4 = S \tag{1.1}$$

where σ is the Stefan Boltzmann constant, $5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,{}^\circ K^{-4}}$. In this example, the equilibrium temperature would be $364\,^\circ\mathrm{K}$, or just below the boiling point of water.

For many practical applications of solar energy this is sufficient, and it is well known that systems for domestic hot water heating based on this principle are available commercially for installation in private dwellings. However, for larger-scale purposes or for generating electric power, a source of heat at $364\,^\circ\mathrm{K}$ has a low thermodynamic efficiency, since it is not practicable to get a very large temperature difference in whatever working fluid is being used in the heat engine. If we wanted, say, $\geq 300\,^\circ\mathrm{C}$—a useful temperature for the generation of motive power—we should need to increase the power density S on the absorbing blackbody by a factor C of about 6 to 10 from Eq. (1.1).

This, briefly, is one use of a concentrator—to increase the power density of solar radiation. When it is stated plainly like that, the problem sounds trivial. The principles of the solution have been known since the days of Archimedes and his burning glass:[2] we simply have to focus the image of the sun with an image-forming

---

[1]   Ignoring various factors such as convection and conduction losses and radiation at lower effective emissivities.

[2]   For an amusing argument concerning the authenticity of the story of Archimedes, see Stavroudis (1973).

system—a lens—and the result will be an increased power density. The problems to be solved are technical and practical, but they also lead to some interesting pure geometrical optics. The first question is that of the maximum concentration: How large a value of C is theoretically possible? The answer to this question is simple in all cases of interest. The next question—can the theoretical maximum concentration be achieved in practice?—is not as easy to answer. We shall see that there are limitations involving materials and manufacturing, as we should expect. But there are also limitations involving the kinds of optical systems that can actually be designed, as opposed to those that are theoretically possible. This is analogous to the situation in classical lens design. The designers sometimes find that a certain specification cannot be fulfilled because it would require an impractically large number of refracting or reflecting surfaces. But sometimes they do not know whether it is in principle possible to achieve aberration corrections of a certain kind.

   The natural approach of the classical optical physicist is to regard the problem as one of designing an image-forming optical system of very large numerical aperture—that is, small aperture ratio or f-number. One of the most interesting results to have emerged in this field is a class of very efficient concentrators that would have very large aberrations if they were used as image-forming systems. Nevertheless, as concentrators, they are substantially more efficient than image-forming systems and can be designed to meet or approach the theoretical limit. We shall call them *nonimaging concentrating collectors*, or *nonimaging concentrators* for short. Nonimaging is sometimes substituted by the word *anidolic* (from the Greek, meaning "without image") in languages such as Spanish and French because it's more specific. These systems are unlike any previously used optical systems. They have some of the properties of light pipes and some of the properties of image-forming optical systems but with very large aberrations. The development of the designs of these concentrators and the study of their properties have led to a range of new ideas and theorems in geometrical optics. In order to facilitate the development of these ideas, it is necessary to recapitulate some basic principles of geometrical optics, which is done in Chapter 2. In Chapter 3, we look at what can be done with conventional image-forming systems as concentrators, and we show how they necessarily fall short of ideal performance. In Chapter 4, we describe one of the basic nonimaging concentrators, the compound parabolic concentrator, and we obtain its optical properties. Chapter 5 is devoted to several developments of the basic compound parabolic concentrator: with plane absorber, mainly aimed at decreasing the overall length; with nonplane absorber; and with generalized edge ray wavefronts, which is the origin of the tailored designs. In Chapter 6, we examine in detail the Flow Line approach to nonimaging concentrators both for 2D and 3D geometries, and we include the description of the Poisson brackets design method. At the end of this chapter we introduce elliptic bundles in the Lorentz geometry formulation. Chapter 7 deals with a basic illumination problem: designing an optical system that produces a prescribed irradiance with a given source. This problem is considered from the simplest case (2D geometry and point source) with increasing complexity (3D geometry, extended sources, free-form surfaces). Chapter 8 is devoted specifically to one method of design called Simultaneous Multiple Surfaces (SMS) method, which is the newest and is more powerful for high concentration/collimation applications. Nonimaging is not the opposite of imaging. Chapter 9 shows imaging applications of nonimaging designs. Sometimes

the performance of some devices is theoretically limited by the use of rotational or linear symmetric devices, Chapters 10 and 11 discuss the problem of improving this performance by using free-form surfaces departing from symmetric designs that are deformed in a controlled way. The limits to concentration or collimation can be derived from Chapter 12, which is devoted to the physical optics aspects of concentration and in particular to the concept of radiance in the physical optics. Chapters 13 and 14 are devoted to the main applications of nonimaging optics: illumination and concentration (in this case of solar energy). Finally, in Chapter 15 we examine briefly several manufacturing techniques. There are several appendixes in which the derivations of the more complicated formulas are given.

1.2 DEFINITION OF THE CONCENTRATION RATIO; THE THEORETICAL MAXIMUM

From the simple argument in Section 1.1 we see that the most important property of a concentrator is the ratio of area of input beam divided by the area of output beam; this is because the equilibrium temperature of the absorbing body is proportional to the fourth root of this ratio. We denote this ratio by C and call it the concentration ratio. Initially we model a concentrator as a box with a plane entrance aperture of area A and a plane exit aperture of area A' that is just large enough to allow all transmitted rays to emerge (see Figure 1.1). Then the concentration ratio is

$$C = A/A' \qquad (1.2)$$

In the preceding definition, it was tacitly assumed that compression of the input beam occurred in both the dimensions transverse to the beam direction, as in ordinary lens systems. In solar energy technology there is a large class of systems in which the beam is compressed in only one dimension. In such systems
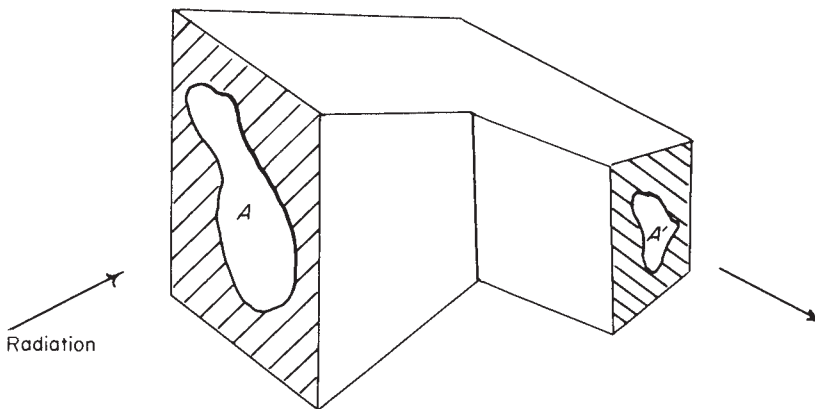
Figure 1.1   Schematic diagram of a concentrator. The input and output surfaces can face in any direction; they are drawn in the figure so both can be seen. It is assumed that the aperture A' is just large enough to permit all rays passed by the internal optics that have entered within the specified collecting angle to emerge.

Radiation
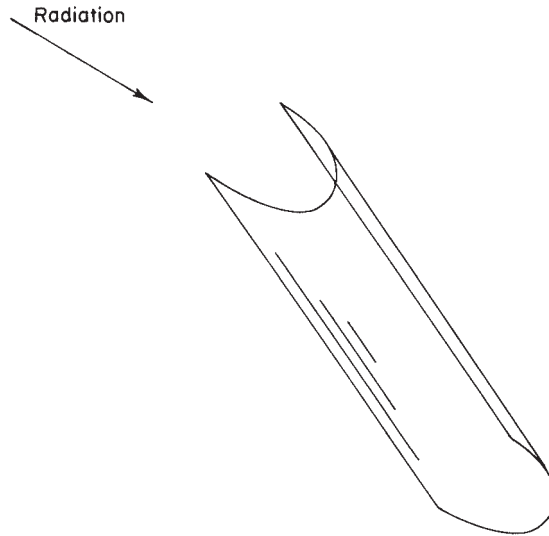
Figure 1.2   A trough concentrator; the absorbing element is not shown.

all the operative surfaces, reflecting and refracting, are cylindrical with parallel generators (but not in general circular cylindrical). Thus, a typical shape would be as in Figure 1.2, with the absorbing body (not shown) lying along the trough. Such long trough collectors have the obvious advantage that they do not need to be guided to follow the daily movement of the sun across the sky. The two types of concentrator are sometimes called three- and two-dimensional, or 3D and 2D, concentrators. The names 3D and 2D are also used in this book (from Chapter 6 to the end) to denote that the optical device has been designed in 3D geometry or in 2D geometry (in the latter case, the real concentrator, which of course exists in a 3D space, is obtained by rotational or translational symmetry from the 2D design). In these cases we will use the name 2D design or 3D design to differentiate from a 2D or a 3D concentrator. The 2D concentrators are also called linear concentrators. The concentration ratio of a linear concentrator is usually given as the ratio of the transverse input and output dimensions, measured perpendicular to the straight-line generators of the trough.

The question immediately arises whether there is any upper limit to the value of C, and we shall see that there is. The result, proved later, is very simple for the 2D case and for the 3D case with an axis of revolution symmetry (rotational concentrator). Suppose the input and output media both have a refractive index of unity, and let the incoming radiation be from a circular source at infinity subtending a semiangle θ_i. Then the theoretical maximum concentration in a rotational concentrator is

$$C_{\mathrm{max}} = 1/\sin^2 \theta_i \qquad (1.3)$$

Under this condition the rays emerge at all angles up to $\pi/2$ from the normal to the exit face, as shown in Figure 1.3. For a linear concentrator the corresponding value will be $1/\sin \theta_i$.

The next question that arises is, can actual concentrators be designed with the theoretically best performance? In asking this question we make certain ide-
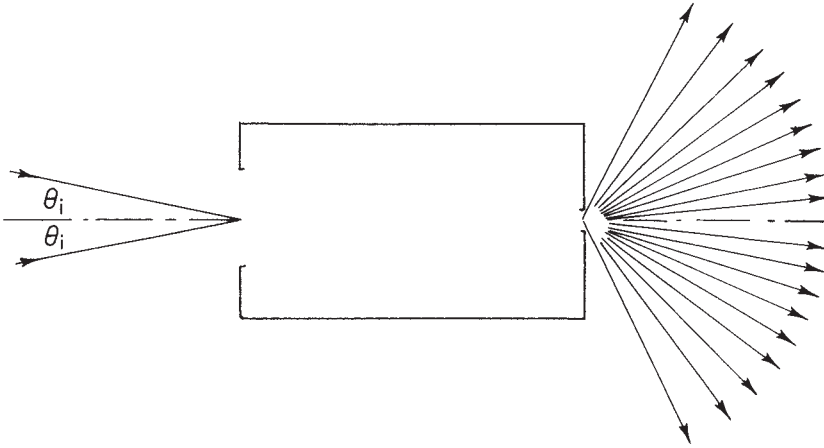
Figure 1.3   Incident and emergent ray paths for an ideal 3D concentrator with symmetry about an axis of revolution. The exit aperture diameter is $\sin \theta_i$ times the exit aperture diameter; the rays emerge from all points in the exit aperture over a solid angle 2π.

alizing assumptions—for example, that all reflecting surfaces have 100% reflectivity, that all refracting surfaces can be perfectly antireflection coated, that all shapes can be made exactly right, and so forth. We shall then see that the following answers are obtained: (1) 2D concentrators can be designed with the theoretical maximum concentration; (2) 3D concentrators can also have the theoretical maximum concentration if they use variable refractive index material or a pile of infinitely thin surface waveguides properly shaped; and (3) some rotational symmetric concentrators can have the theoretical maximum concentration. In case (3) it appears for other types of design that it is possible to approach indefinitely close to the theoretical maximum concentration either by sufficiently increasing the complexity of the design or by incorporating materials that are in principle possible but in practice not available. For example, we might specify a material of very high refractive index—say, 5—although this is not actually available without large absorption in the visible part of the spectrum.

1.3   USES OF CONCENTRATORS

The application to solar energy utilization just mentioned has, of course, stimulated the greatest developments in the design and fabrication of concentrators. But this is by no means the only application. The particular kind of nonimaging concentrator that has given rise to the greatest developments was originally conceived as a device for collecting as much light as possible from a luminous volume (the gas or fluid of a Čerenkov counter) over a certain range of solid angle and sending it onto the cathode of a photomultiplier. Since photomultipliers are limited in size and the volume in question was of order $1\,m^3$, this is clearly a concentrator problem (Hinterberger and Winston, 1966a,b).

Subsequently the concept was applied to infrared detection (Harper et al., 1976), where it is well known that the noise in the system for a given type of detector increases with the surface area of the detector (other things being equal).

Another type of application was to the optics of visual receptors. It has been noted (Winston and Enoch, 1971) that the cone receptors in the human retina have a shape corresponding approximately to that of a nonimaging concentrator designed for approximately the collecting angle that the pupil of the eye would subtend at the retina under dark-adapted conditions.

1.4   USES OF ILLUMINATORS

Nonimaging collectors are also used in illumination. The source (a filament, an LED, etc.) is in general emitting in a wide angular spread at low intensity, and the problem consists of designing an optical device that efficiently collimates this radiation so it is emitted in a certain angular emitting region, which is smaller than the angular emitting region of the source. The problem is conceptually similar to the concentrating problem, substituting aperture areas for angular regions sizes. We will see soon that both statements are equivalent.

There are several other possible applications of nonimaging concentrators, and these will be discussed in Chapters 9, 13, and 14.

REFERENCES

Harper, D. A., Hildebrand, R. H., Pernic, R., and Platt, S. R. (1976). Heat trap: An optimised far infrared field optics system. *Appl. Opt.* **15**, 53–60.

Hinterberger, H., and Winston, R. (1966a). Efficient light coupleer for threshold Černkov counters. *Rev. Sci. Instrum.* **37**, 1094–1095.

Winston, R., and Enoch, J. M. (1971). Retinal cone receptor as an ideal light collector. *J. Opt. Soc. Am.* **61**, 1120–1121.

2

SOME BASIC IDEAS IN GEOMETRICAL OPTICS

2.1   THE CONCEPTS OF GEOMETRICAL OPTICS

Geometrical optics is used as the basic tool in designing almost any optical system, image forming or not. We use the intuitive ideas of a ray of light, roughly defined as the path along which light energy travels, together with surfaces that reflect or transmit the light. When light is reflected from a smooth surface, it obeys the well-known law of reflection, which states that the incident and reflected rays make equal angles with the normal to the surface and that both rays and the normal lie in one plane. When light is transmitted, the ray direction is changed according to the law of refraction: Snell's law. This law states that the sine of the angle between the normal and the incident ray bears a constant ratio to the sine of the angle between the normal and the refracted ray; again, all three directions are coplanar.

A major part of the design and analysis of concentrators involves ray tracing—that is, following the paths of rays through a system of reflecting and refracting surfaces. This is a well-known process in conventional lens design, but the requirements are somewhat different for concentrators, so it will be convenient to state and develop the methods ab initio. This is because in conventional lens design the reflecting or refracting surfaces involved are almost always portions of spheres, and the centers of the spheres lie on one straight line (axisymmetric optical system) so that special methods that take advantage of the simplicity of the forms of the surfaces and the symmetry can be used. Nonimaging concentrators do not, in general, have spherical surfaces. In fact, sometimes there is no explicitly analytical form for the surfaces, although usually there is an axis or a plane of symmetry. We shall find it most convenient, therefore, to develop ray-tracing schemes based on vector formulations but with the details covered in computer programs on an ad hoc basis for each different shape.

In geometrical optics we represent the power density across a surface by the density of ray intersections with the surface and the total power by the number of rays. This notion, reminiscent of the useful but outmoded "lines of force" in electrostatics, works as follows. We take N rays spaced uniformly over the entrance aperture of a concentrator at an angle of incidence θ, as shown in Figure 2.1.

7
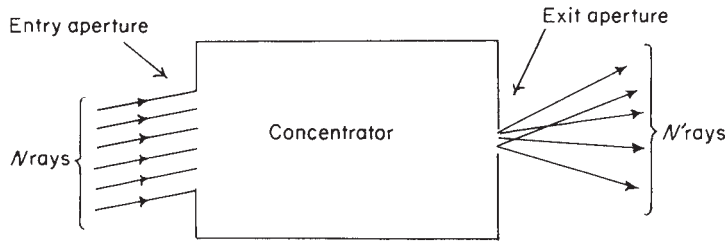
Figure 2.1   Determining the transmission of a concentrator by ray tracing.
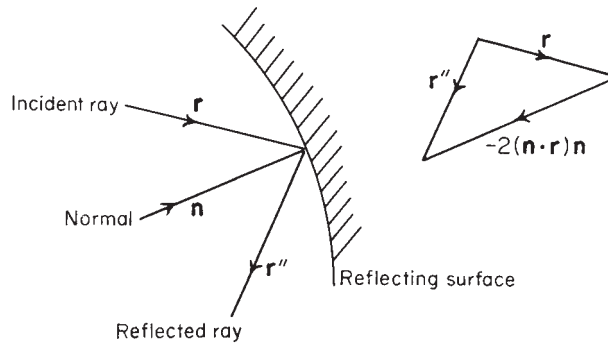
Figure 2.2   Vector formulation of reflection. $\mathbf{r}$, $\mathbf{r}''$, and $\mathbf{n}$ are all unit vectors.

Suppose that after tracing the rays through the system only N' emerge through the exit aperture, the dimensions of the latter being determined by the desired concentration ratio. The remaining $N - N'$ rays are lost by processes that will become clear when we consider some examples. Then the power transmission for the angle θ is taken as N'/N. This can be extended to cover a range of angle θ as required. Clearly, N must be taken large enough to ensure that a thorough exploration of possible ray paths in the concentrator is made.

2.2   FORMULATION OF THE RAY-TRACING PROCEDURE

To formulate a ray-tracing procedure suitable for all cases, it is convenient to put the laws of reflection and refraction into vector form. Figure 2.2 shows the geometry with unit vectors $\mathbf{r}$ and $\mathbf{r}''$ along the incident and reflected rays and a unit vector $\mathbf{n}$ along the normal pointing into the reflecting surface. Then it is easily verified that the law of reflection is expressed by the vector equation

$$\mathbf{r}'' = \mathbf{r} - 2(\mathbf{n} \cdot \mathbf{r})\mathbf{n} \tag{2.1}$$

as in the diagram.

Thus, to ray-trace "through" a reflecting surface, first we have to find the point of incidence, a problem of geometry involving the direction of the incoming ray and
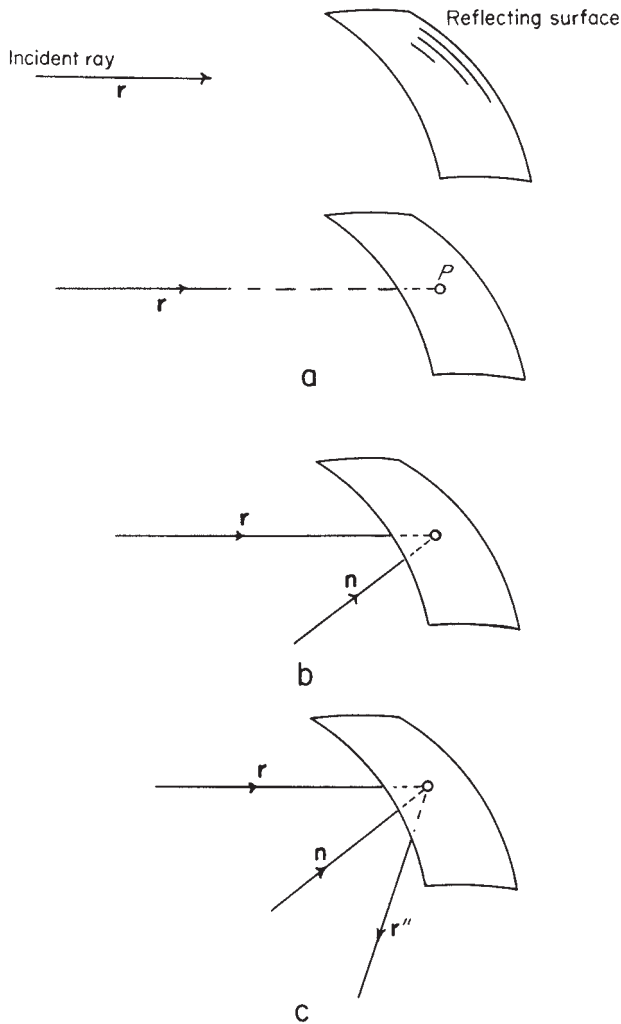
Figure 2.3   The stages in ray tracing a reflection. (a) Find the point of incidence P. (b) Find the normal at P. (c) Apply Eq. (2.1) to find the reflected ray **r″**.

the known shape of the surface. Then we have to find the normal at the point of incidence—again a problem of geometry. Finally, we have to apply Eq. (2.1) to find the direction of the reflected ray. The process is then repeated if another reflection is to be taken into account. These stages are illustrated in Figure 2.3. Naturally, in the numerical computation the unit vectors are represented by their components—that is, the direction cosines of the ray or normal with respect to some Cartesian coordinate system used to define the shape of the reflecting surface.

Ray tracing through a refracting surface is similar, but first we have to formulate the law of refraction vectorially. Figure 2.4 shows the relevant unit vectors. It is similar to Figure 2.2 except that **r′** is a unit vector along the refracted ray. We denote by n, $n′$ the refractive indexes of the media on either side of the refracting boundary; the refractive index is a parameter of a transparent medium related
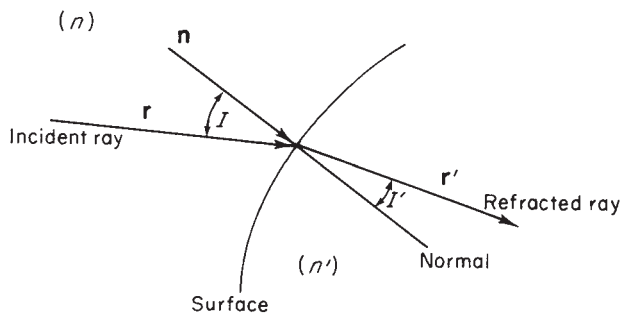
Figure 2.4   Vector formulation of refraction.

to the speed of light in the medium. Specifically, if c is the speed of light in a vacuum, the speed in a transparent material medium is c/n, where n is the refractive index. For visible light, values of n range from unity to about 3 for usable materials in the visible spectrum. The law of refraction is usually stated in the form

$$n' \sin I' = n \sin I \qquad (2.2)$$

where I and I' are the angles of incidence and refraction, as in the figure, and where the coplanarity of the rays and the normal is understood. The vector formulation

$$n' \mathbf{r}' \times \mathbf{n} = n \mathbf{r} \times \mathbf{n} \qquad (2.3)$$

contains everything, since the modulus of a vector product of two unit vectors is the sine of the angle between them. This can be put in the form most useful for ray tracing by multiplying through vectorially by n to give

$$n' \mathbf{r}' = n \mathbf{r} + (n' \mathbf{r}' \cdot \mathbf{n} - n \mathbf{r} \cdot \mathbf{n})\mathbf{n} \qquad (2.4)$$

which is the preferred form for ray tracing.[1] The complete procedure then parallels that for reflection explained by means of Figure 2.3. We find the point of incidence, then the direction of the normal, and finally the direction of the refracted ray. Details of the application to lens systems are given, for example, by Welford (1974, 1986).

   If a ray travels from a medium of refractive index n toward a boundary with another of index $n' < n$, then it can be seen from Eq. (2.2) that it would be possible to have $\sin I'$ greater than unity. Under this condition it is found that the ray is completely reflected at the boundary. This is called total internal reflection, and we shall find it a useful effect in concentrator design.

---

[1]   The method of using Eq. (2.4) numerically is not so obvious as for Eq. (2.2), since the coefficient of $\mathbf{n}$ in Eq (2.4) is actually $n' \cos I' - n \cos I$. Thus, it might appear that we have to find $\mathbf{r}'$ before we can use the equation. The procedure is to find $\cos I'$ via Eq. (2.2) first, and then Eq. (2.4) is needed to give the complete three-dimensional picture of the refracted ray.
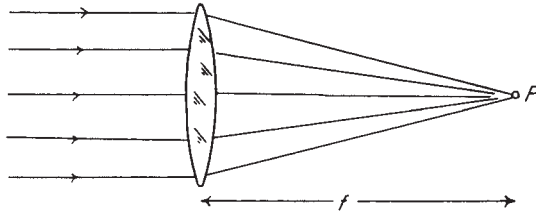
Figure 2.5   A thin converging lens bringing parallel rays to a focus. Since the lens is technically "thin," we do not have to specify the exact plane in the lens from which the focal length f is measured.

2.3   ELEMENTARY PROPERTIES OF IMAGE-FORMING OPTICAL SYSTEMS

In principle, the use of ray tracing tells us all there is to know about the geometrical optics of a given optical system, image forming or not. However, ray tracing alone is of little use for inventing new systems having properties suitable for a given purpose. We need to have ways of describing the properties of optical systems in terms of general performance, such as, for example, the concentration ratio C introduced in Chapter 1. In this section we shall introduce some of these concepts.

Consider first a thin converging lens such as one that would be used as a magnifier or in eyeglasses for a farsighted person (see Figure 2.5). By "thin" we mean that its thickness can be neglected for the purposes of our discussion. Elementary experiments show us that if we have rays coming from a point at a great distance to the left, so that they are substantially parallel as in the figure, the rays meet approximately at a point F, the focus. The distance from the lens to F is called the focal length, denoted by f. Elementary experiments also show that if the rays come from an object of finite size at a great distance, the rays from each point on the object converge to a separate focal point, and we get an image. This is, of course, what happens when a burning glass forms an image of the sun or when the lens in a camera forms an image on film. This is indicated in Figure 2.6, where the object subtends the (small) angle 2θ. It is then found that the size of the image is $2f\theta$. This is easily seen by considering the rays through the center of the lens, since these pass through undeviated.

Figure 2.6 contains one of the fundamental concepts we use in concentrator theory, the concept of a beam of light of a certain diameter and angular extent. The diameter is that of the lens—say, $2a$—and the angular extent is given by 2θ. These two can be combined as a product, usually without the factor 4, giving θa, a quantity known by various names including extent, étendue, acceptance, and Lagrange invariant. It is, in fact, an invariant through the optical system, provided that there are no obstructions in the light beam and provided we ignore certain losses due to properties of the materials, such as absorption and scattering. For example, at the plane of the image the étendue becomes the image height θf multiplied by the convergence angle a/f of the image-forming rays, giving again θa. In discussing 3D systems—for example, an ordinary lens such as we have supposed Figure 2.6 to represent—it is convenient to deal with the square of this quantity, $a^2\theta^2$. This is also sometimes called the étendue, but generally it is clear from
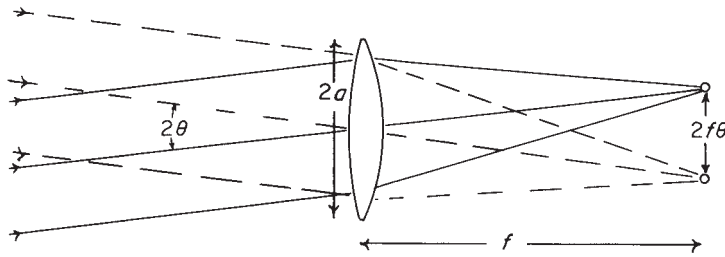
Figure 2.6   An object at infinity has an angular subtense 2θ. A lens of focal length f forms an image of size $2f\theta$.
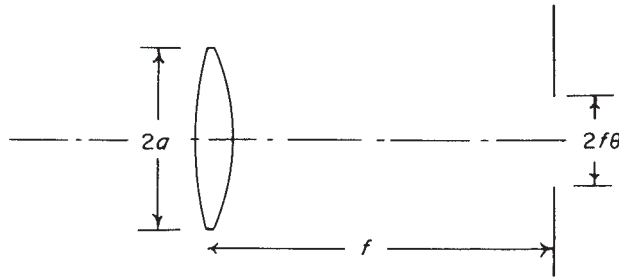
Figure 2.7   An optical system of acceptance, throughput, or étendue $a^2\theta^2$.

the context and from dimensional considerations which form is intended. The 3D form has an interpretation that is fundamental to the theme of this book. Suppose we put an aperture of diameter $2f\theta$ at the focus of the lens, as in Figure 2.7. Then this system will only accept rays within the angular range $\pm\theta$ and inside the diameter $2a$. Now suppose a flux of radiation B (in $\mathrm{W\,m^{-2}sr^{-1}}$) is incident on the lens from the left.[2] The system will actually accept a total flux $B\pi^2\theta^2 a^2\mathrm{W}$; thus, the étendue or acceptance $\theta^2 a^2$ is a measure of the power flow that can pass through the system.

The same discussion shows how the concentration ratio C appears in the context of classical optics. The accepted power $B\pi^2\theta^2 a^2\mathrm{W}$ must flow out of the aperture to the right of the system, if our preceding assumptions about how the lens forms an image are correct[3] and if the aperture has the diameter $2f\theta$. Thus, our system is acting as a concentrator with concentration ratio $C = (2a/2f\theta)^2 = (a/f\theta)^2$ for the input semiangle θ.

Let us relate these ideas to practical cases. For solar energy collection we have a source at infinity that subtends a semiangle of approximately 0.005 rad (1/4°) so that this is the given value of θ, the collection angle. Clearly, for a given diameter of lens we gain by reducing the focal length as much as possible.

---

[2]   In full, B watts per square meter per steradian solid angle.
[3]   As we shall see, these assumptions are only valid for limitingly small apertures and objects.

2.4   ABERRATIONS IN IMAGE-FORMING OPTICAL SYSTEMS

According to the simplified picture presented in Section 2.3, there is no reason why we could not make a lens system with an indefinitely large concentration ratio by simply decreasing the focal length sufficiently. This is, of course, not so, partly because of aberrations in the optical system and partly because of the fundamental limit on concentration stated in Section 1.2.

We can explain the concept of aberrations by looking again at our example of the thin lens in Figure 2.5. We suggested that the parallel rays all converged after passing through the lens to a single point F. In fact, this is only true in the limiting case when the diameter of the lens is taken as indefinitely small. The theory of optical systems under this condition is called paraxial optics or Gaussian optics, and it is a very useful approximation for getting at the main large-scale properties of image-forming systems. If we take a simple lens with a diameter that is a sizable fraction of the focal length—say, $f/4$—we find that the rays from a single point object do not all converge to a single image point. We can show this by ray tracing. We first set up a proposed lens design, as shown in Figure 2.8. The lens has curvatures (reciprocals of radii) c_1 and c_2, center thickness d, and refractive index n. If we neglect the central thickness for the moment, then it is shown in specialized treatment (e.g., Welford, 1986) that the focal length f is given in paraxial approximation by

$$1/f = (n - 1)(c_1 - c_2) \qquad (2.5)$$

and we can use this to get the system to have roughly the required paraxial properties.

Now we can trace rays through the system as specified, using the method outlined in Section 2.2 (details of ray-tracing methods for ordinary lens systems are given in, for example, Welford, 1974). These will be exact or finite rays, as opposed to paraxial rays, which are implicit in the Gaussian optics approximation. The results for the lens in Figure 2.8 would look like Figure 2.9. This shows rays traced from an object point on the axis at infinity—that is, rays parallel to the axis.

In general, for a convex lens the rays from the outer part of the lens aperture meet the axis closer to the lens than the paraxial rays. This effect is known as
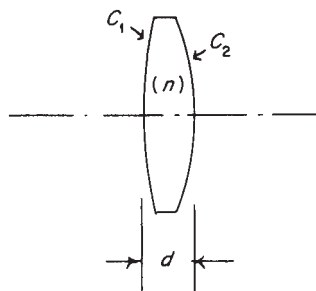
Figure 2.8   Specification of a single lens. The curvature c_1 is positive as shown, and c_2 is negative.
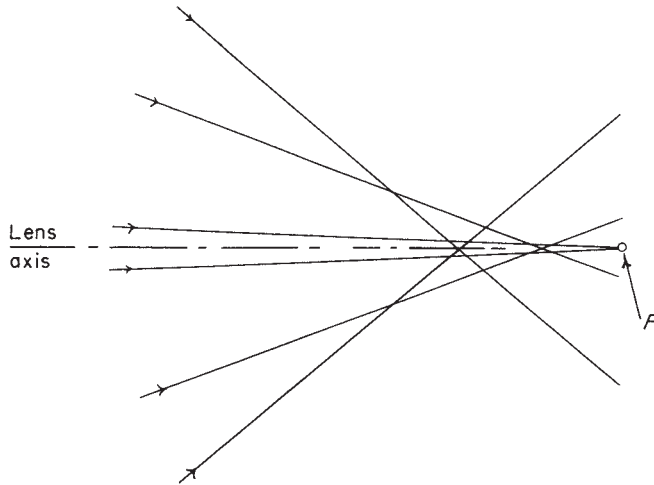
Figure 2.9   Rays near the focus of a lens showing spherical aberration.

spherical aberration. (The term is misleading, since the aberration can occur in systems with nonspherical refracting surfaces, but there seems little point in trying to change it at the present advanced state of the subject.)

Spherical aberration is perhaps the simplest of the different aberration types to describe, but it is just one of many. Even if we were to choose the shapes of the lens surfaces to eliminate the spherical aberration or were to eliminate it in some other way, we would still find that the rays from object points away from the axis did not form point images—in other words, there would be oblique or off-axis aberrations. Also, the refractive index of any material medium changes with the wavelength of the light, and this produces chromatic aberrations of various kinds. We do not at this stage need to go into the classification of aberrations very deeply, but this preliminary sketch is necessary to show the relevance of aberrations to the attainable concentration ratio.

2.5   THE EFFECT OF ABERRATIONS IN AN IMAGE-FORMING SYSTEM ON THE CONCENTRATION RATIO

Questions regarding the extent to which it is theoretically possible to eliminate aberrations from an image-forming system have not yet been fully answered. In this book we shall attempt to give answers adequate for our purposes, although they may not be what the classical lens designers want. For the moment, let us accept that it is possible to eliminate spherical aberration completely, but not the off-axis aberrations, and let us suppose that this has been done for the simple collector of Figure 2.7. The effect will be that some rays of the beam at the extreme angle θ will fall outside the defining aperture of diameter $2f\theta$. We can see this more clearly by representing an aberration by means of a spot diagram. This is a diagram in the image plane with points plotted to represent the intersections of
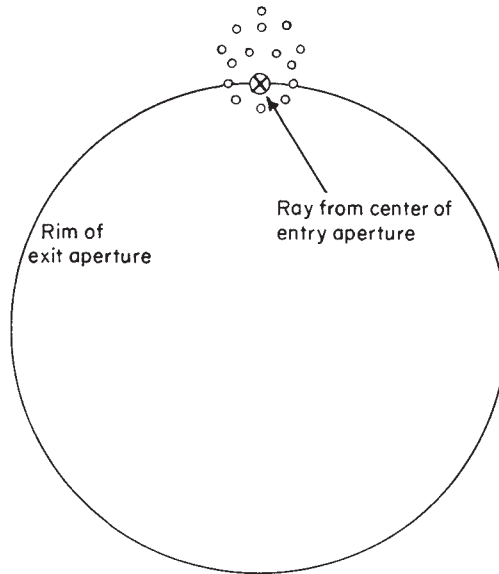
Figure 2.10   A spot diagram for rays from the beam at the maximum entry angle for an image-forming concentrator. Some rays miss the edge of the exit aperture due to aberrations, and the concentration is thus less than the theoretical maximum.

the various rays in the incoming beam. Such a spot diagram for the extreme angle θ might appear as in Figure 2.10. The ray through the center of the lens (the principal ray in lens theory) meets the rim of the collecting aperture by definition, and thus a considerable amount of the flux does not get through. Conversely, it can be seen (in this case at least) that some flux from beams at a larger angle than θ will be collected.

We display this information on a graph such as in Figure 2.11. This shows the proportion of light collected at different angles up to the theoretical maximum, θ_{max}. An ideal collector would behave according to the full line—that is, it would collect all light flux within θ_{max} and none outside. At this point it may be objected that all we need to do to achieve the first requirement is to enlarge the collecting aperture slightly, and the second requirement does not matter. However, we recall that our aim is to achieve maximum concentration because of the requirement for high operating temperature so that the collector aperture must not be enlarged beyond $2f\theta$ diameter.

Frequently in discussions of aberrations in books on geometrical optics, the impression is given that aberrations are in some sense "small." This is true in optical systems designed and made to form reasonably good images, such as camera lenses. But these systems do not operate with large enough convergence angles (a/f in the notation for Figure 2.6) to approach the maximum theoretical concentration ratio. If we were to try to use a conventional image-forming system under such conditions, we should find that the aberrations would be very large and that they would severely depress the concentration ratio. Roughly, we can say that this is one limitation that has led to the development of the new, nonimaging concentrators. Nevertheless, we cannot say that imaging-forming is incom-
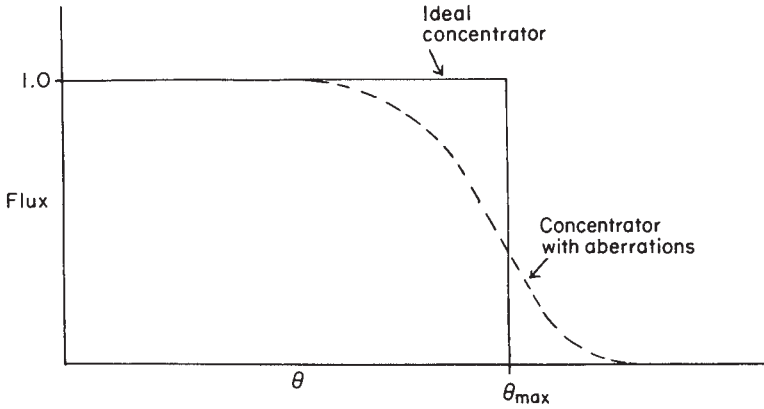
Figure 2.11   A plot of collection efficiency against angle. The ordinate is the proportion of flux entering the collector aperture at angle θ that emerges from the exit aperture.
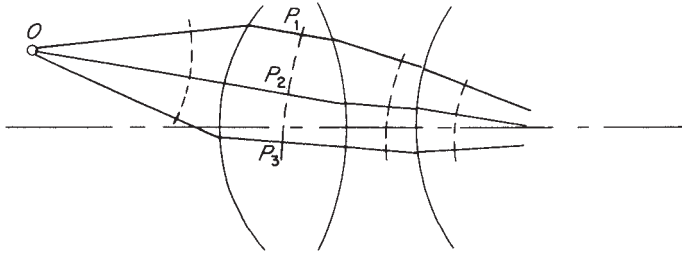
Figure 2.12   Rays and (in broken line) geometrical wave fronts.

patible with attaining maximum concentration. We will show later examples in which both properties are combined.

2.6   THE OPTICAL PATH LENGTH AND FERMAT'S PRINCIPLE

There is another way of looking at geometrical optics and the performance of optical systems, which we also need to outline for the purposes of this book. We noted in Section 2.2 that the speed of light in a medium of refractive index n is c/n, where c is the speed in a vacuum. Thus, light travels a distance s in the medium in time $s/v = ns/c$; that is, the time taken to travel a distance s in a medium of refractive index n is proportional to ns. The quantity ns is called the optical path length corresponding to the length s. Suppose we have a point source O emitting light into an optical system, as in Figure 2.12. We can trace any number of rays through the system, as outlined in Section 2.2, and then we can mark off along these rays points that are all at the same optical path length from O—say, P_1, P_2. ... We do this by making the sum of the optical path lengths from O in each medium the same—that is,
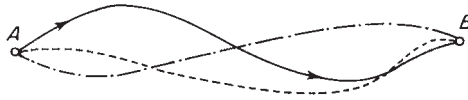
Figure 2.13   Fermat's principle. It is assumed in the diagram that the medium has a continuously varying refractive index. The solid line path has a stationary optical path length from A to B and is therefore a physically possible ray path.

$$\sum ns = \text{const.} \qquad (2.6)$$

in an obvious notation. These points can be joined to form a surface (we are supposing rays out of the plane of the diagram to be included), which would be a surface of constant phase of the light waves if we were thinking in terms of the wave theory of light.[4] We call it a geometrical wave front, or simply a wave front, and we can construct wave fronts at all distances along the bundle of rays from O.

We now introduce a principle that is not so intuitive as the laws of reflection and refraction but that leads to results that are indispensable to the development of the theme of this book. It is based on the concept of optical path length, and it is a way of predicting the path of a ray through an optical medium. Suppose we have any optical medium that can have lenses and mirrors and can even have regions of continuously varying refractive index. We want to predict the path of a light ray between two points in this medium—say, A and B in Figure 2.13. We can propose an infinite number of possible paths, of which three are indicated. But unless A and B happen to be object and image—and we assume they are not—only one or perhaps a small finite number of paths will be physically possible—in other words, paths that rays of light could take according to the laws of geometrical optics. Fermat's principle in the form most commonly used states that a physically possible ray path is one for which the optical path length along it from A to B is an extremum as compared to neighboring paths. For "extremum" we can often write "minimum," as in Fermat's original statement. It is possible to derive all of geometrical optics—that is, the laws of refraction and reflection from Fermat's principle. It also leads to the result that the geometrical wave fronts are orthogonal to the rays (the theorem of Malus and Dupin); that is, the rays are normal to the wave fronts. This in turn tells us that if there is no aberration—if all rays meet at one point—then the wave fronts must be portions of spheres. So if there is no aberration, the optical path length from object point to image point is the same along all rays. Thus, we arrive at an alternative way of expressing aberrations: in terms of the departure of wave fronts from the ideal spherical shape. This concept will be useful when we come to discuss the different senses in which an image-forming system can form "perfect" images.

---

[4]   This construction does not give a surface of constant phase near a focus or near an edge of an opaque obstacle, but this does not affect the present applications.
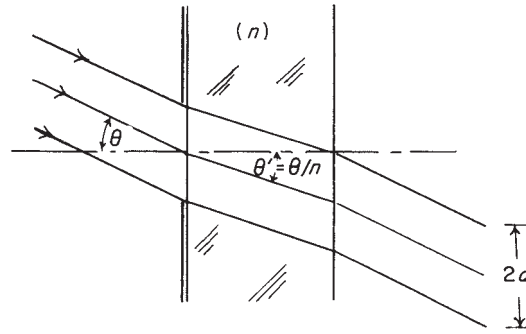
Figure 2.14   Inside a medium of refractive index n the étendue becomes $n^2 a^2 \theta'^2$.

2.7   THE GENERALIZED ÉTENDUE OR LAGRANGE INVARIANT AND THE PHASE SPACE CONCEPT

We next have to introduce a concept that is essential to the development of the principles of nonimaging concentrators. We recall that in Section 2.3 we noted that there is a quantity $a^2\theta^2$ that is a measure of the power accepted by the system, where a is the radius of the entrance aperture and θ is the semiangle of the beams accepted. We found that in paraxial approximation for an axisymmetric system this is invariant through the optical system. Actually, we considered only the regions near the entrance and exit apertures, but it is shown in specialized texts on optics that the same quantity can be written down for any region inside a complex optical system. There is one slight complication: If we are considering a region of refractive index different from unity—say, the inside of a lens or prism—the invariant is written $n^2a^2\theta^2$. The reason for this can be seen from Figure 2.14, which shows a beam at the extreme angle θ entering a plane-parallel plate of glass of refractive index n. Inside the glass the angle is $\theta' = \theta/n$, by the law of refraction[5] so that the invariant in this region is

$$\text{étendue} = n^2a^2\theta'^2 \tag{2.7}$$

We might try to use the étendue to obtain an upper limit for the concentration ratio of a system as follows. We suppose we have an axisymmetric optical system of any number of components—that is, not necessarily the simple system sketched in Figure 2.7. The system will have an entrance aperture of radius a, which may be the rim of the front lens or, as in Figure 2.15, possibly some limiting aperture inside the system. An incoming parallel beam may emerge parallel, as indicated in the figure, or not, and this will not affect the result. But to simplify the argument it is easier to imagine a parallel beam emerging from an aperture of radius a'. The concentration ratio is by definition $(a/a')^2$, and if we use the étendue invariant and assume that the initial and final media are both air or vacuum—refractive index unity—the concentration ratio becomes $(\theta'/\theta)^2$. Since

---

[5]   The paraxial approximation is implied so that $\sin\theta \sim \theta$.
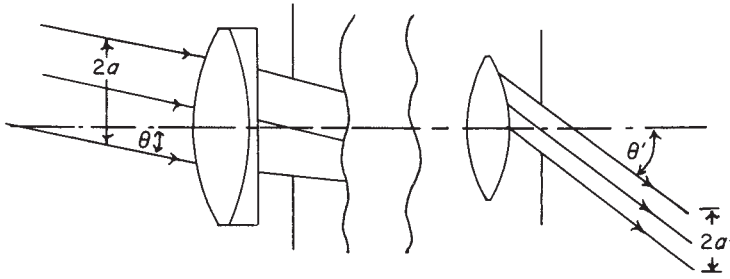
Figure 2.15   The étendue for a multielement optical system with an internal aperture stop.
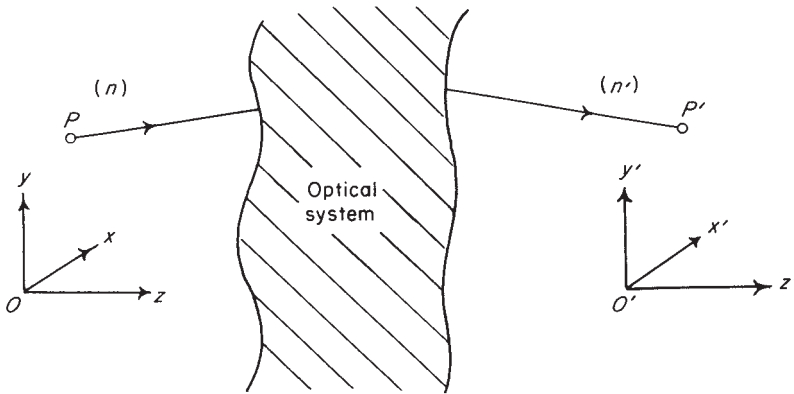
Figure 2.16   The generalized étendue.

from obvious geometrical considerations θ' cannot exceed $\pi/2$, this suggests $(\pi/2\theta)^2$ as a theoretical upper limit to the concentration.

Unfortunately, this argument is invalid because the étendue as we have defined it is essentially a paraxial quantity. Thus, it is not necessarily an invariant for angles as large as $\pi/2$. In fact, the effect of aberrations in the optical system is to ensure that the paraxial étendue is not an invariant outside the paraxial region so that we have not found the correct upper limit to the concentration.

There is, as it turns out, a suitable generalization of the étendue to rays at finite angles to the axis, and we will now explain this. The concept has been known for some time, but it has not been used to any extent in classical optical design, so it is not described in many texts. It applies to optical systems of any or no symmetry and of any structure—refracting, reflecting, or with continuously varying refractive index.

Let the system be bounded by homogeneous media of refractive indices n and n' as in Figure 2.16, and suppose we have a ray traced exactly between the points P and P' in the respective input and output media. We wish to consider the effect of small displacements of P and of small changes in direction of the ray segment through P on the emergent ray so that these changes define a beam of rays of a certain cross section and angular extent. In order to do this we set up a Cartesian coordinate system $Oxyz$ in the input medium and another, $O'x'y'z'$, in the output
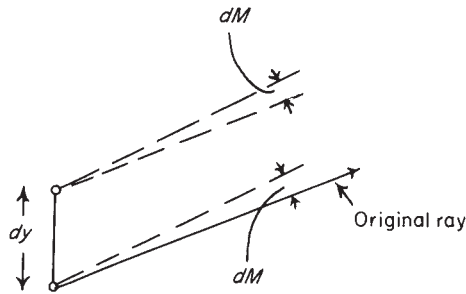
Figure 2.17   The generalized étendue in the y section.

medium. The positions of the origins of these coordinate systems and the directions of their axes are quite arbitrary with respect to each other, to the directions of the ray segments, and, of course, to the optical system. We specify the input ray segment by the coordinates of $P(x, y, z)$, and by the direction cosines of the ray (L, M, N). The output segment is similarly specified. We can now represent small displacements of P by increments dx and dy to its x and y coordinates, and we can represent small changes in the direction of the ray by increments dL and dM to the direction cosines for the x and y axes. Thus, we have generated a beam of area $dxdy$ and angular extent defined by $dLdM$. This is indicated in Figure 2.17 for the y section.[6] Corresponding increments dx', dy', dL', and dM' will occur in the output ray position and direction.

Then the invariant quantity turns out to be $n^2\, dx\, dy\, dL\, dM$—that is, we have

$$n'^2 dx' dy' dL' dM' = n^2 dx\, dy\, dL\, dM \qquad (2.8)$$

The proof of this theorem depends on other concepts in geometrical optics that we do not need in this book. We have therefore given proof in Appendix A, where references to other proofs of it can also be found.

The physical meaning of Eq. (2.8) is that it gives the changes in the rays of a beam of a certain size and angular extent as it passes through the system. If there are apertures in the input medium that produce this limited étendue, and if there are no apertures elsewhere to cut off the beam, then the accepted light power emerges in the output medium so that the étendue as defined is a correct measure of the power transmitted along the beam. It may seem at first remarkable that the choice of origin and direction of the coordinate systems is quite arbitrary. However, it is not very difficult to show that the generalized étendue or Lagrange invariant as calculated in one medium is independent of coordinate translations and rotations. This, of course, must be so if it is to be a meaningful physical quantity.

The generalized étendue is sometimes written in terms of the optical direction cosines $p = nL$, $q = nM$, when it takes the form

$$dx\, dy\, dp\, dq \qquad (2.9)$$

---

[6]   It is necessary to note that the increments dL and dM are in direction cosines, not angles. Thus, in Figure 2.17 the notation on the figure should be taken to mean not that dM is the angle indicated, but merely that it is a measure of this angle.
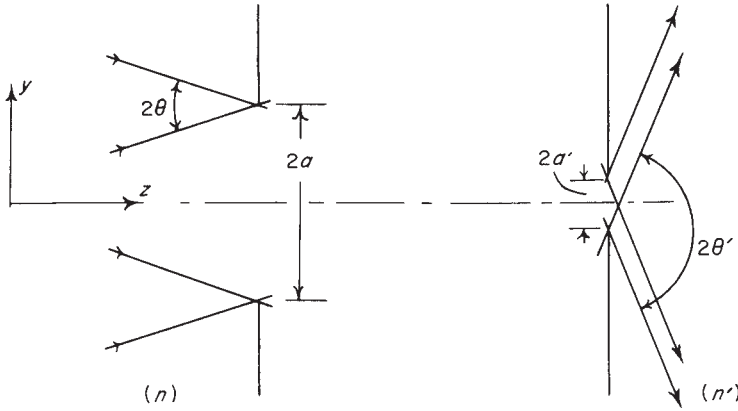
Figure 2.18   The theoretical maximum concentration ratio for a 2D optical system.

An étendue value is associated to any 4-parameter bundle of rays. Each combination of the four parameters defines one single ray. In the example of Figure 2.16, the four parameters are x, y, L, M (or x', y', L', M'), but there are many other possible sets of 4 parameters describing the same bundle. For the cases in which the rays are not described at a z = constant (or z' = constant planes), then the following generalized expression can be used to calculate the differential of étendue of the bundle dE:

$$dE = dx\, dy\, dp\, dq + dy\, dz\, dq\, dr + dz\, dx\, dr\, dp \qquad (2.10)$$

The total étendue is obtained by integration of all the rays of the bundles. In what follows we will assume that the bundle can be described at a z = constant plane.

In 2D geometry, when we only consider the rays contained in a plane, we can also define an étendue for any 2-parameter bundle of rays. If the plane in which all the rays are contained is a x = constant plane, then the differential of étendue can be written as $dE = n\, dy\, dM$. As in the 3D case, the étendue is an invariant of the bundle, and the same result is obtained no matter where it is calculated. For instance, it can be calculated at z' = constant, and the result should be the same: $n'\, dy'\, dM' = n\, dy\, dM$, or, in terms of the optical direction cosines, $dy'\, dq' = dy\, dq$.

We can now use the étendue invariant to calculate the theoretical maximum concentration ratios of concentrators. Consider first a 2D design, as in Figure 2.18. We have for any ray bundle that transverses the system

$$n\, dy\, dM = n'\, dy'\, dM' \qquad (2.11)$$

and integrating over y and M we obtain

$$4na \sin\theta = 4n'a' \sin\theta' \qquad (2.12)$$

so that the concentration ratio is

$$\frac{a}{a'} = \frac{n' \sin\theta'}{n \sin\theta} \qquad (2.13)$$

In this result a' is a dimension of the exit aperture large enough to permit any ray that reaches it to pass, and θ' is the largest angle of all the emergent

rays. Clearly θ' cannot exceed $\pi/2$, so the theoretical maximum concentration ratio is

$$C_{\max} = \frac{n'}{n \sin \theta} \tag{2.14}$$

Similarly, for the 3D case we can show that for an axisymmetric concentrator the theoretical maximum is

$$C_{\max} = \left(\frac{a}{a'}\right)^2 = \left(\frac{n'}{n \sin \theta}\right)^2 \tag{2.15}$$

where again θ is the input semiangle.

The results in Eqs. (2.14) and (2.15) are maximum values, which may or may not be attained. We find in practice that if the exit aperture has the diameter given by Eq. (2.15), some of the rays within the incident collecting angle and aperture do not pass it. We sometimes also find in a number of the systems to be described that some of the incident rays are actually turned back by internal reflections and never reach the exit aperture. In addition, there are losses due to absorption, imperfect reflectivity, and so forth, but these do not represent fundamental limitations. Thus, Eqs. (2.14) and (2.15) give theoretical upper bounds on performance of concentrators.

Our results so far apply to linear concentrators [Eq. (2.14)] with rectangular entrance and exit apertures and to rotational concentrators with circular entrance and exit apertures [Eq. (2.15)]. We ought, for completeness, to discuss briefly what happens if the entrance aperture is not circular but the concentrator itself still has an axis of symmetry. The difficulty with this case is that it depends on the details of the internal optics of the concentrator. It may happen that the internal optical system forms an image of the entrance aperture on the exit aperture—in which case it would be correct to make them similar in shape. For an entry aperture of arbitrary shape but uniform entry angle $\pm\theta_i$ all that can be said in general is that for an ideal concentrator the area of the exit aperture must equal that of the entry aperture multiplied by $\sin^2\theta_i$. We will see in Chapter 6 that such concentrators can be designed.

2.8   THE SKEW INVARIANT

There is an invariant associated with the path of a skew ray through an axisymmetric optical system. Let S be the shortest distance between the ray and the axis—that is, the length of the common perpendicular—and let γ be the angle between the ray and the axis. Then the quantity

$$h = nS \sin \gamma \tag{2.16}$$

is an invariant through the whole system. If the medium has a continuously varying refractive index, the invariant for a ray at any coordinate z_1 along the axis is obtained by treating the tangent of the ray at the z value as the ray and using the refractive index value at the point where the ray cuts the transverse plane z_1. The skew-invariant formula will be proved in Appendix C.

If we use the dynamical analogy described in Appendix A, then h corresponds to the angular momentum of a particle following the ray path, and the skew-

invariant theorem corresponds to conservation of angular momentum. In terms of the Hamilton's equations, the skew invariant is just a first integral that derives from the symmetry condition.

2.9 DIFFERENT VERSIONS OF THE CONCENTRATION RATIO

We now have some different definitions of concentration ratio. It is desirable to clarify them by using different names. First, in Section 2.7 we established upper limits for the concentration ratio in 2D and 3D systems, given respectively by Eqs. (2.14) and (2.15). These upper limits depend only on the input angle and the input and output refractive indices. Clearly we can call either expression the theoretical maximum concentration ratio.

Second, an actual system will have entry and exit apertures of dimensions $2a$ and $2a'$. These can be width or diameter for linear or rotational systems, respectively. The exit aperture may or may not transmit all rays that reach it, but in any case the ratios (a/a') or $(a/a')^2$ define a geometrical concentration ratio.

Third, given an actual system, we can trace rays through it and determine the proportion of incident rays within the collecting angle that emerge from the exit aperture. This process will yield an optical concentration ratio.

Finally, we could make allowances for attenuation in the concentrator by reflection losses, scattering, manufacturing errors, and absorption in calculating the optical concentration ratio. We could call the result the optical concentration ratio with allowance for losses. The optical concentration ratio will always be less than or equal to the theoretical maximum concentration ratio. The geometrical concentration ratio can, of course, have any value.

REFERENCE

Welford, W. T. (1986). "Aberrations of Optical Systems." Hilger, Bristol, England.

3

SOME DESIGNS OF IMAGE-FORMING CONCENTRATORS

3.1 INTRODUCTION

In this chapter we examine image-forming concentrators of conventional form—paraboloidal mirrors, lenses of short focal length, and so forth—and estimate their performance. Then we shall show how the departure from ideal performance suggests a principle for the design of nonimaging concentrators—the "edge-ray principle," as we shall call it.

3.2 SOME GENERAL PROPERTIES OF IDEAL IMAGE-FORMING CONCENTRATORS

In order to fix our ideas we use the solar energy application to describe the mode of action of our systems. The simplest hypothetical image-forming concentrator would then function as in Figure 3.1. The rays are coded to indicate that rays from one direction from the sun are brought to a focus at one point in the exit aperture—that is, the concentrator images the sun (or other source) at the exit aperture. If the exit medium is air, then the exit angle θ' must be $\pi/2$ for maximum concentration. Such a concentrator may in practice be constructed with glass or some other medium of refractive index greater than unity forming the exit surface, as in Figure 3.2. Also, the angle θ' in the glass would have to be such that $\sin\theta' = 1/n$ so that the emergent rays just fill the required $\pi/2$ angle. For typical materials the angle θ' would be about 40°.

Figure 3.2 brings out an important point about the objects of such a concentrator. We have labeled the central or principal ray of the two extreme angle beams a and b, respectively, and at the exit end these rays have been drawn normal to the exit face. This would be essential if the concentrator were to be used with air as the final medium, since, if rays a and b were not normal to the exit face, some of the extreme angle rays would be totally internally reflected (see Section 2.2), and thus the concentration ratio would be reduced. In fact, the condition that the exit principal rays should be normal to what, in ordinary lens design, is termed
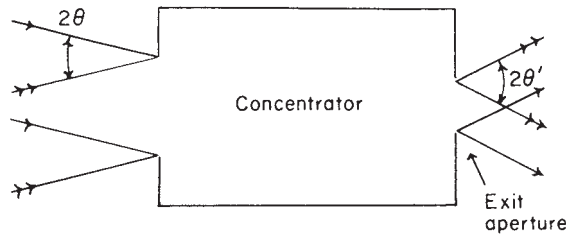
Figure 3.1   An image-forming concentrator. An image of the source at infinity is formed at the exit aperture of the concentrator.
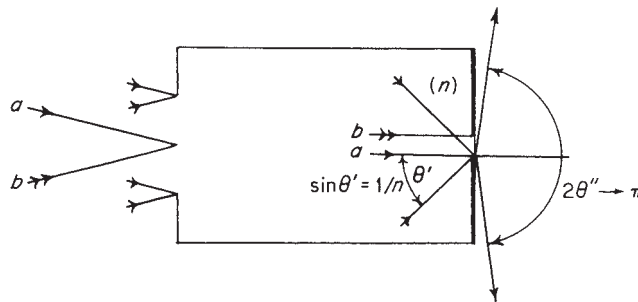
Figure 3.2   In an image-forming concentrator of maximum theoretical concentration ratio the final medium in the concentrator would have to have a refractive index n greater than unity. The angle θ' in this medium would be arcsin $(1/n)$, giving an angle $\pi/2$ in the air outside.

the image plane is not usually fulfilled. Such an optical system, called telecentric, needs to be specially designed, and the requirement imposes constraints that would certainly worsen the attainable performance of a concentrator. We shall therefore assume that when a concentrator ends in glass of index n, the absorber or other means of utilizing the light energy is placed in optical contact with the glass in such a way as to avoid potential losses through total internal reflection.

An alternative configuration for an image-forming concentrator would be as in Figure 3.3. The concentrator collects rays over θ_{max} as before, but the internal optics form an image of the entrance aperture at the exit aperture, as indicated by the arrow coding of the rays. This would be in optics terminology a telescopic or afocal system. Naturally, the same considerations about using glass or a similar material as the final medium holds as for the system of Figures 3.1 and 3.2, and there is no difference between the systems as far as external behavior is concerned.

If the concentrator terminates in a medium of refractive index n, we can gain in maximum concentration ratio by a factor n or n^2, depending on whether it is a 2D or 3D system, as can be seen from Eqs. (2.13) and (2.14). This corresponds to having an extreme angle $\theta' = \pi/2$ in this medium. We then have to reinstate the requirement that the principal rays be normal to the exit aperture, and we also have to ensure that the absorber can utilize rays of such extreme angles.

In practice there are problems in using extreme collection angles approaching $\theta' = \pi/2$ whether in air or a higher-index medium. There has to be very good match-
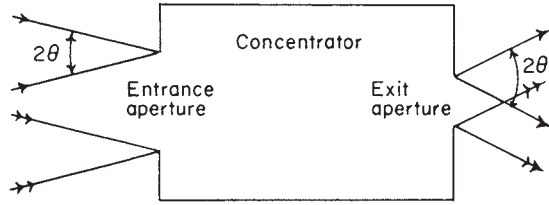
Figure 3.3 An alternative configuration of an image-forming concentrator. The rays collected from an angle $\pm\theta$ form an image of the entrance aperture at the exit aperture.
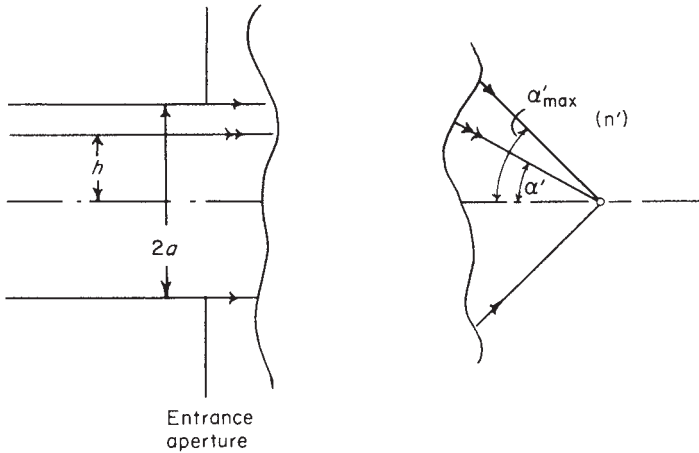
Figure 3.4 The definition of the numerical aperture of an image forming system. The NA is $n' \sin \alpha'$.

ing at the interface between glass and absorber to avoid large reflection losses of grazing-incidence rays, and irregularities of the interface can cause losses through shadowing. Therefore, we may well be content with values of θ' of, say, 60°. This represents only a small decrease from the theoretical maximum concentration, as can be seen from Eqs. (2.14) and (2.15).

Thus, in speaking of ideal concentrators we can also regard as ideal a system that brings all incident rays within θ_{max} out within θ'_{max} and inside an exit aperture a' given by Eq. (2.12)—that is, $a' = na \sin\theta_{max}/n' \sin\theta'_{max}$. Such a concentrator will be ideal, but it will not have the theoretical maximum concentration.

The concentrators sketched in Figures 3.1 and 3.2 clearly must contain something like a photographic objective with very large aperture (small f-number), or perhaps a high-power microscope objective used in reverse. The speed of a photographic objective is indicated by its f-number or aperture ratio. Thus, an $f/4$ objective has a focal length four times the diameter of its entrance aperture. This description is not suitable for imaging systems in which the rays form large angles approaching $\pi/2$ with the optical axis for a variety of reasons. It is found that in discussing the resolving power of such systems the most useful measure of performance is the numerical aperture or NA, a concept introduced by Ernst Abbe in connection with the resolving power of microscopes. Figure 3.4 shows an optical system with entrance aperture of diameter $2a$. It forms an image of the axial object

point at infinity and the semiangle of the cone of extreme rays is α'_{max}. Then the numerical aperture is defined by

$$NA = n' \sin \alpha'_{max} \qquad (3.1)$$

where n' is the refractive index of the medium in the image space. We assume that all the rays from the axial object point focus sharply at the image point—that is, there is (to use the terminology of Section 2.4) no spherical aberration. Then Abbe showed that off-axis object points will also be sharply imaged if the condition

$$h = n' \sin \alpha' \times \text{const.} \qquad (3.2)$$

is fulfilled for all the axial rays. In this equation h is the distance from the axis of the incoming ray, and α' is the angle at which that ray meets the axis in the final medium. Equation (3.2) is a form of the celebrated Abbe sine condition for good image formation. It does not ensure perfect image formation for all off-axis object points, but it ensures that aberrations that grow linearly with the off-axis angle are zero. These aberrations are various kinds of coma. The condition of freedom from spherical aberration and coma is called aplanatism.

Clearly, a necessary condition for our image-forming concentrator to have the theoretical maximum concentration—or even for it to be ideal as an image-forming system (but without theoretical maximum concentration)—is that the image formation should be aplanatic. This is not, unfortunately, a sufficient condition.

The constant in Eq. (3.2) has the significance of a focal length. The definition of focal length for optical systems with media of different refractive indices in the object and image spaces is more complicated than for the thin lenses discussed in Chapter 2. In fact, it is necessary to define two focal lengths, one for the input space and one for the output space, where their magnitudes are in the ratio of the refractive indices of the two media. In Eq. (3.2) it turns out that the constant is the input side focal length, which we shall denote by f.

From Eq. (3.2) we have for the input semiaperture

$$a = f \cdot NA \qquad (3.3)$$

and also, from Eq. (2.13),

$$a' = \frac{a \sin \theta_{max}}{NA} \qquad (3.4)$$

By substituting from Eq. (3.3) into Eq. 3.4 we have

$$a' = f \sin \theta_{max} \qquad (3.5)$$

where θ_{max} is the input semiangle. To see the significance of this result we recall that we showed that in an aplanatic system the focal length is a constant, independent of the distance h of the ray from the axis used to define it. Here we are using the generalized sense of "focal length" meaning the constant in Eq. (3.2), and aplanatism thus means that rays through all parts of the aperture of the system form images with the same magnification. Thus, Eq. (3.5) tells us that in an imaging concentrator with maximum theoretical concentration the diameter of the exit aperture is proportional to the sine of the input angle. This is true even if the concentrator has a numerical exit aperture less than the theoretical maximum, n', provided it is ideal in the sense just defined.
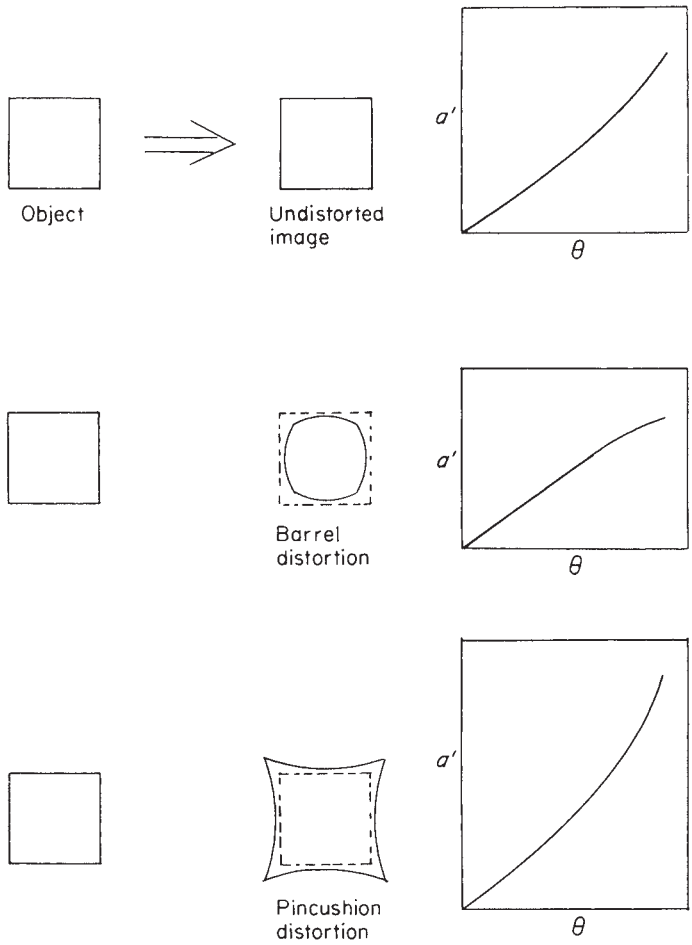
Figure 3.5   Distortion in image-forming systems. The optical systems are assumed to have symmetry about an axis of rotation.

From the point of view of conventional lens optics, the result of Eq. (3.3) is well known. It is simply another way of saying that the aplanatic lens with largest aperture and with air as the exit medium is $f/0.5$, since Eq. (3.5) tells us that $a = f$. The importance of Eq. (3.5) is that it tells us something about one of the shape-imaging aberrations required of the system—namely distortion. A distortion-free lens imaging onto a flat field must obviously have an image height proportional to $\tan \theta$, so our concentrator lens system is required to have what is usually called barrel distortion. This is illustrated in Figure 3.5.

Our picture of an imaging concentrator is gradually taking shape, and we can begin to see that certain requirements of conventional imaging can be relaxed. Thus, if we can get a sharp image at the edge of the exit aperture and if the diameter of the exit aperture fulfils the requirement of Eqs. 3.3–3.5, we do not need perfect image formation for object points at angles smaller than θ_{max}. For example, the image field perhaps could be curved, provided we take the exit aperture in the plane of the circle of image points for the direction θ_{max}, as in Figure 3.6. Also, the
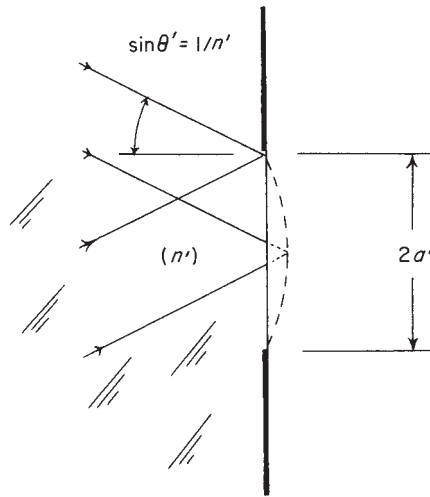
Figure 3.6   A curved image field with a plane exit aperture.
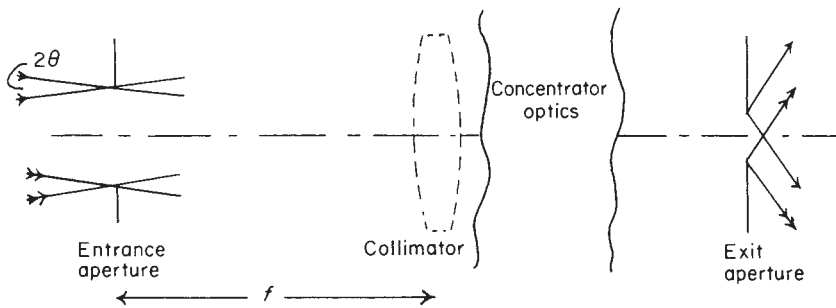
Figure 3.7   An afocal concentrator shown as two image-forming systems.

inner parts of the field could have point-imaging aberrations, provided these were not so large as to spill rays outside the circle of radius a'. Thus, we see that an image-forming concentrator need not, in principle, be so difficult to design as an imaging lens, since the aberrations need to be corrected only at the edge of the field. In practice this relaxation may not be very helpful because the outer part of the field is the most difficult to correct. However, this leads us to a valuable principle for nonimaging concentrators. Not only is it unnecessary to have good aberration correction except at the exit rim, but we do not even need point imaging at the rim itself. It is only necessary that rays entering at the extreme angle θ_{max} should leave from some point at the rim and that the aberrations inside should not be such as to push rays outside the rim of the exit aperture. We shall return to this edge-ray principle later in connection with nonimaging concentrators.

   The above arguments need only a little modification to apply to the alternative configuration of imaging concentrator in Figure 3.3, in which the entrance aperture is imaged at the exit aperture. Referring to Figure 3.7, we can imagine that the optical components of the concentrator are forming an image at the exit aperture of an object at a considerable distance, rather than at infinity, and that
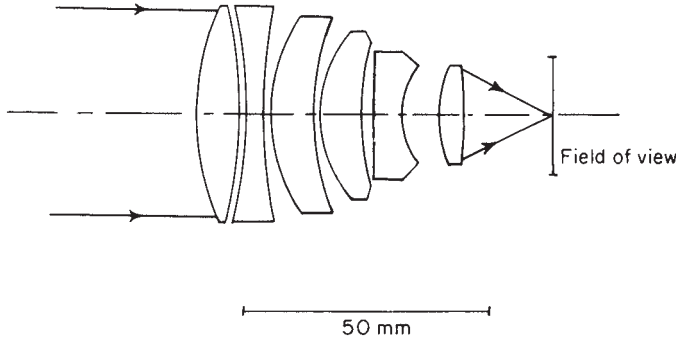
Figure 3.8   A high-aperture camera objective. The drawing is to scale for a 50-mm focal length. The emerging cone of rays has a semiangle of 26° at the center of the field of view.

this object is the entrance aperture. Alternatively, we can imagine that part of the concentrator is a collimating lens of focal length f, shown in broken line in the figure, and that this projects the entrance aperture to infinity with an angle subtending $2a/f$. The same considerations as before then apply to the aberration corrections.

3.3   CAN AN IDEAL IMAGE-FORMING CONCENTRATOR BE DESIGNED?

In Section 3.2 we outlined some requirements for ideal image-forming concentrators, and we now have to ask whether they can be designed to fulfill these requirements with useful collection angles.

High-aperture camera lenses are made at about $f/1.0$, but these are complex structures with many components. Figure 3.8 shows a typical example with a focal length of 50 mm. Such a system is by no means aberration-free, and the cost of scaling it up to a size useful for solar work would be prohibitive. Anyway, its numerical aperture is still only about 0.5. The only systems with numerical apertures approaching the theoretical limit are microscope objectives. Figure 3.9 shows one of the simplest designs of microscope objective of numerical aperture about 1.35, drawn in reverse and with one conjugate at infinity. The image or exit space has a refractive index of 1.52, since it is an oil immersion objective. Such systems have good aberration correction only to about 3° from the axis. Beyond this the aberrations increase rapidly, and also there is less light transmission because of vignetting.[1] The collecting aperture would be about 4 mm in diameter. Again, it would be impracticable to scale up such a system to useful dimensions.

Thus, a quick glance at the state of the art in conventional lens design suggests that imaging concentrators in the form of lens systems will not be very efficient on a practical scale. Nevertheless, it is interesting to see what might be done with the classical imaging design techniques if practical limitations are ignored.

---

[1]   Vignetting is caused by rims of components at either end of a long system shearing against each other as the system is turned off-axis.
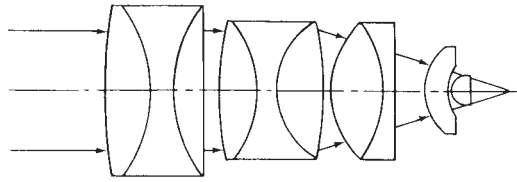
Figure 3.9   An oil-immersion microscope objective of high numerical aperture. Such systems can have a convergence angle of up to 60° with an aberration-free field of about ±3°. However, they can only be designed aberration free for focal lengths up to 2 mm—that is, an actual field diameter of about 200 μm.
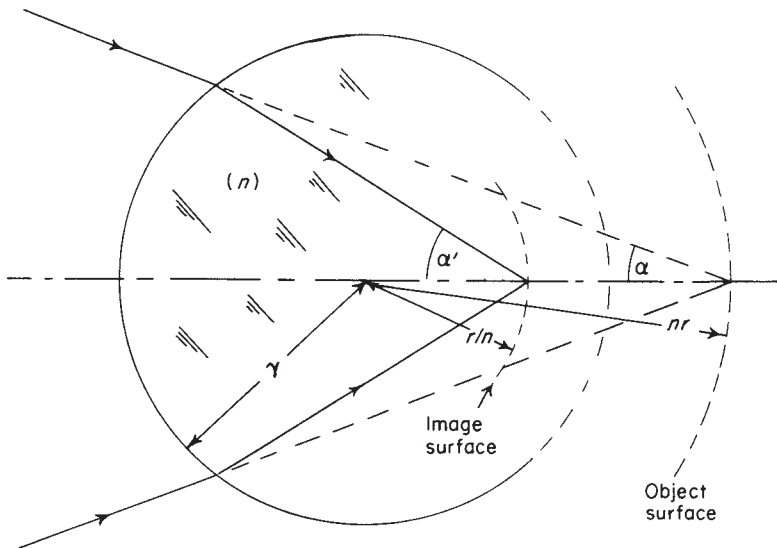
Figure 3.10   The aplanatic surfaces of a spherical refracting surface.

Roughly, the position seems to be that we cannot design an ideal concentrator—one with the theoretical maximum collection efficiency, using a finite number of lens elements. But, by increasing the number of elements sufficiently or by postulating sufficiently extreme optical properties, we can approach indefinitely close to the ideal. Exceptions to the preceding proposed rule occur in optical systems with spherical symmetry. It has been known since the time of Huygens that a spherical lens element images a concentric surface, as shown in Figure 3.10.

The two conjugate surfaces have radii r/n and nr, respectively. The configuration is used in microscope objectives having a high numerical aperture, as in Figure 3.9. Unfortunately, one of the conjugates must always be virtual (the object conjugate as the figure is drawn), so the system alone would not be very practical as a concentrator. It seems to be true, although this has not been proven, that no combination of a finite number of concentric components can form an aberration-free real image of a real object. However, as we shall see, this can be done with media of continuously varying refractive index. The system of Figure 3.10 would clearly be useful as the last stage of an imaging concentrator. It can easily be shown that the convergence angles are related by the equation
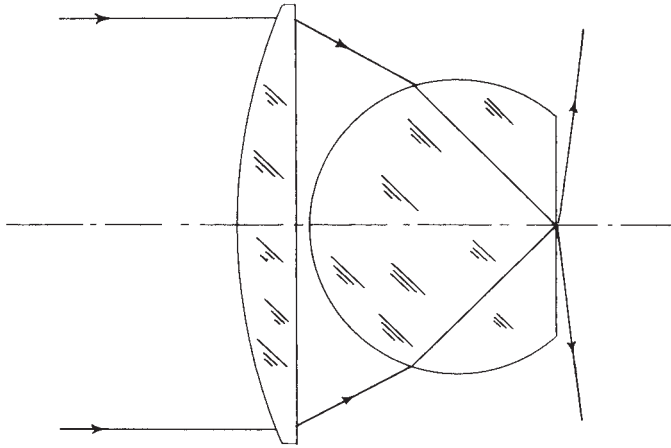
Figure 3.11   An image-forming concentrator with an aplantic component.
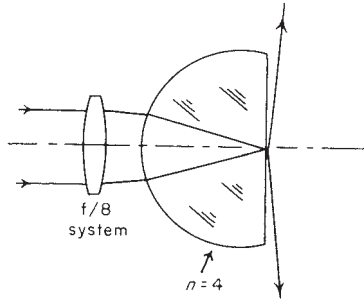
Figure 3.12   Use of an aplanatic component of high refractive index to produce a well-corrected optical system.

$$\sin \alpha' = n \sin \alpha \qquad (3.6)$$

Also, if there is a plane surface terminating in air, the final emergent angle α'' is given by

$$\sin \alpha'' = n^2 \sin \alpha \qquad (3.7)$$

Thus, the system could be used in conjunction with another system of relatively low numerical aperture, as in Figure 3.11, to form a fairly well-corrected concentrator. This is, of course, merely a reinvention of the microscope objective of Figure 3.9, and the postulated additional system still needs to operate at about $f/1$ if ordinary materials are used. If we assume some extreme material qualities—say, a refractive index of 4 with adequate antireflection coating for the aplanatic component—then the auxiliary system only needs to be $f/8$ to give $\pi/2$ emergent angles in air, as in Figure 3.12.

This is not a difficult requirement. In fact, it is probably true that if we ignore chromatic aberration—that is, if we assume our postulated material of index 4 has no dispersion—this system could be designed in moderate sizes for which the aberrations are indefinitely small over a reasonable acceptance angle.

Ultimately the ray aberrations of optical systems become negligible because the performance, as an imaging system, is limited by diffraction effects. Thus, if an imaging optical system of a certain numerical aperture is used to form an image of a point source and if there are precisely no ray aberrations, the image of the point will not be indefinitely small. It is shown in books on physical optics[2] that the point image will be a blurred diffraction pattern in which most of the light flux falls inside a circle of radius

$$0.61\lambda/NA \tag{3.8}$$

where λ is the wavelength of the light. This provides us with a tolerance level for ray aberrations below which we can say the aberrations are negligible. This is sometimes expressed in the form that all of the points in the spot diagram for the aberrations (see Figure 2.10) must fall within a circle of radius given by Eq. (3.8). Another way of setting a tolerance is to say that the wave-front shape as determined by the methods outlined in Section 2.6 would not depart from the ideal spherical shape by more than a specified amount, usually $\lambda/4$. Other tolerance systems are also described by Born and Wolf (1975).

Thus, in some way we could arrive at tolerances for the geometrical aberrations such that an imaging concentrator with aberrations inside these tolerances would have ideal performance. Any lens system made with available materials would be impractically complicated and costly if scaled up to a size suitable for solar concentration. But hypothetical materials could be used to bring the aberrations within the diffraction limit with a simple design such as in Figure 3.12.

At this point it might be interesting to consider a different example of a lens system as a concentrator. It is well known that an ellipsoidal solid lens will focus parallel light without spherical aberration, as in Figure 3.13, provided the eccentricity is equal to $1/n$. In order to use this as a concentrator we put the entry aperture in front of the ellipsoid, as in Figure 3.14, so that the principal ray emerges parallel to the axis. We find that the ellipsoid has very strong coma of such a sign that all other rays meet the image plane nearer the axis than the principal ray. Thus, all rays strike within this circle. However, this is not an ideal concentrator, since, as can be seen from the diagram, the radius of this circle is proportional to $\tan(\theta_{max})$ whereas, according to Eq. (3.5), all rays should strike within a circle of radius proportional to $\sin(\theta_{max})$.[3]

3.4   MEDIA WITH CONTINUOUSLY VARYING REFRACTIVE INDICES

We stated in Section 3.3 that it is thought to be impossible to design an ideal imaging concentrator with a finite number of reflecting or refracting surfaces, even with spherical symmetry, although the example of Figure 3.10 shows that perfect imagery is possible if one conjugate surface is virtual. It has long been known that if we admit continuously varying refractive index, then perfect imagery between surfaces in a spherically symmetric geometry is possible. James Clerk Maxwell (1854) showed that if a medium had the refractive index distribution

---

[2]   See, for example, Born and Wolf, 1975.
[3]   In fact, on account of the curvature of the ellipse, the radius is even slightly greater than a value proportional to $\tan(\theta_{max})$.
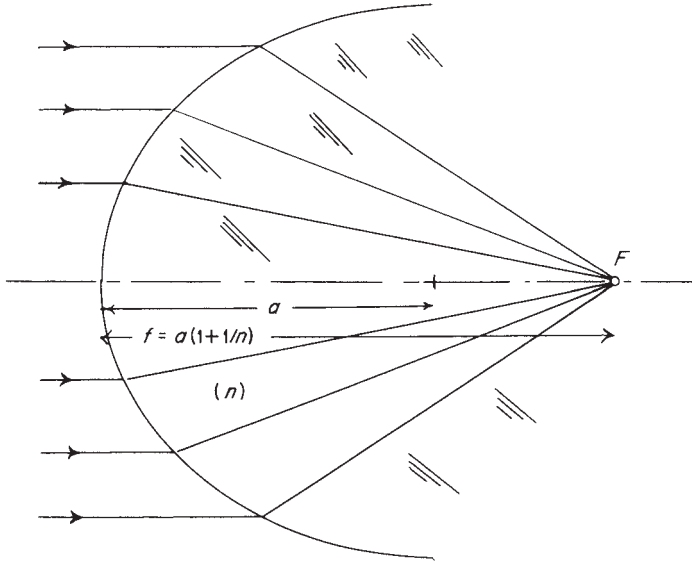
Figure 3.13 A portion of an ellipsoid of revolution as a single refracting surface free from spherical aberration. The generating ellipse has eccentricity $1/n$ and semimajor axis a. The figure is drawn to scale with $n = 1.81$.
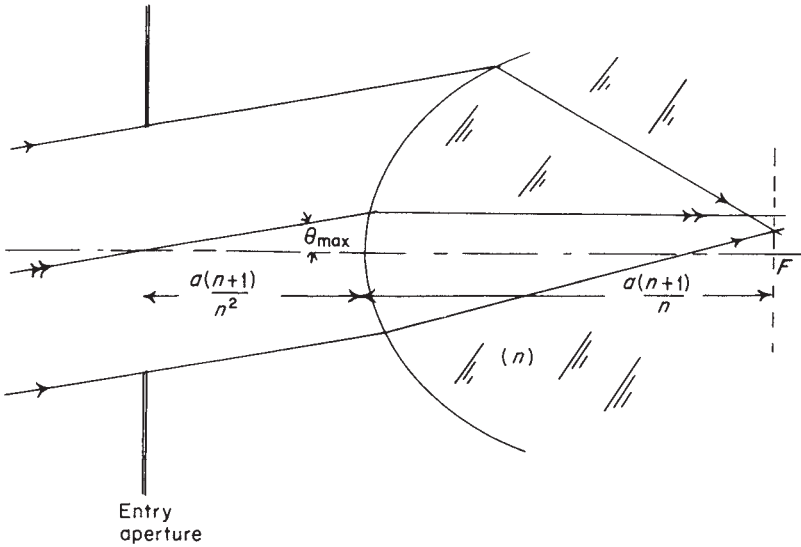
Figure 3.14 An ellipsoid of revolution as a concentrator. The entry aperture is set at the first focus of the system.

$$n = \frac{a^2}{b^2 + r^2} \tag{3.9}$$

where a and b are constants and r is a radial coordinate, then any point would be perfectly imaged at another point on the opposite side of the origin. If $a = b = 1$, the distances of conjugate points from the origin are related by
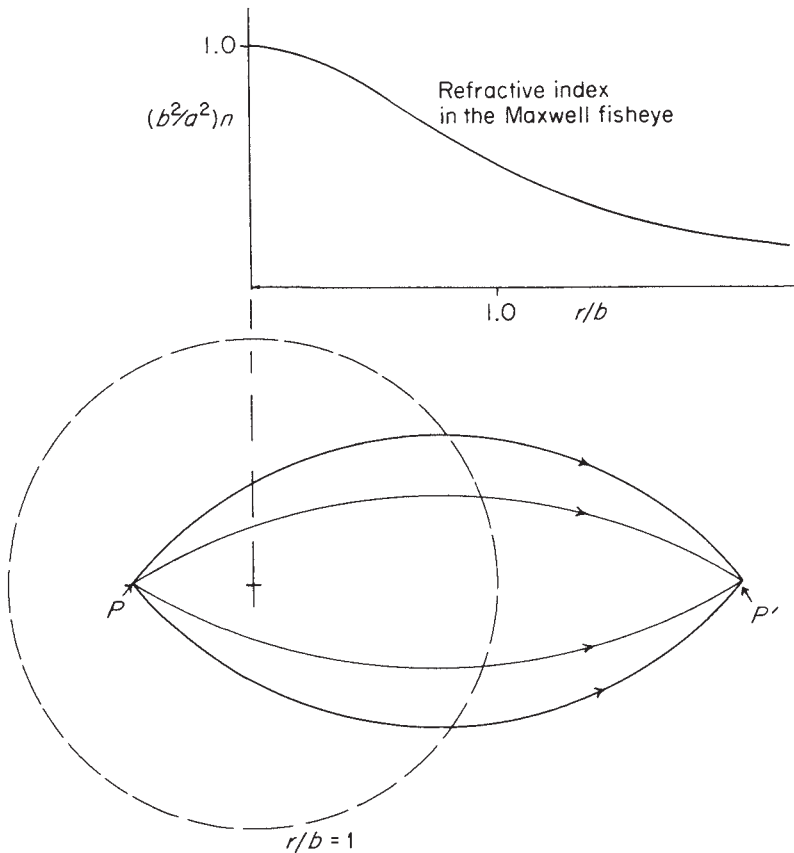
Figure 3.15   Rays in the Maxwell fisheye. The rays are arcs of circles.

$$rr' = 1 \tag{3.10}$$

and a typical set of imaging rays would be as in Figure 3.15. This system, known as Maxwell's fisheye lens, is not very useful for our purposes, since both object and image have to be immersed in the medium. Luneburg (1964) gave several more examples of media of spherical symmetry with ideal imaging properties. In particular, he found an example, now known in the literature as the Luneburg lens, where the index distribution extends over a finite radius only and where the object conjugate is at infinity. The index distribution is

$$n(r) = \begin{cases} \sqrt{2 - \dfrac{r^2}{a^2}}, & r < 1 \\ 1, & r \geq 1 \end{cases} \tag{3.11}$$

This distribution forms a perfect point image with numerical aperture unity, as in Figure 3.16, and on account of the spherical symmetry, it can be shown to form an ideal concentrator of maximum theoretical concentration.

    Appendix F presents a more detailed treatment that shows how the ray paths are calculated. You can see that the Luneburg lens satisfied Abbe's sine condition [Eq.(3.2)]. Also, it follows from the spherical symmetry that perfect point images
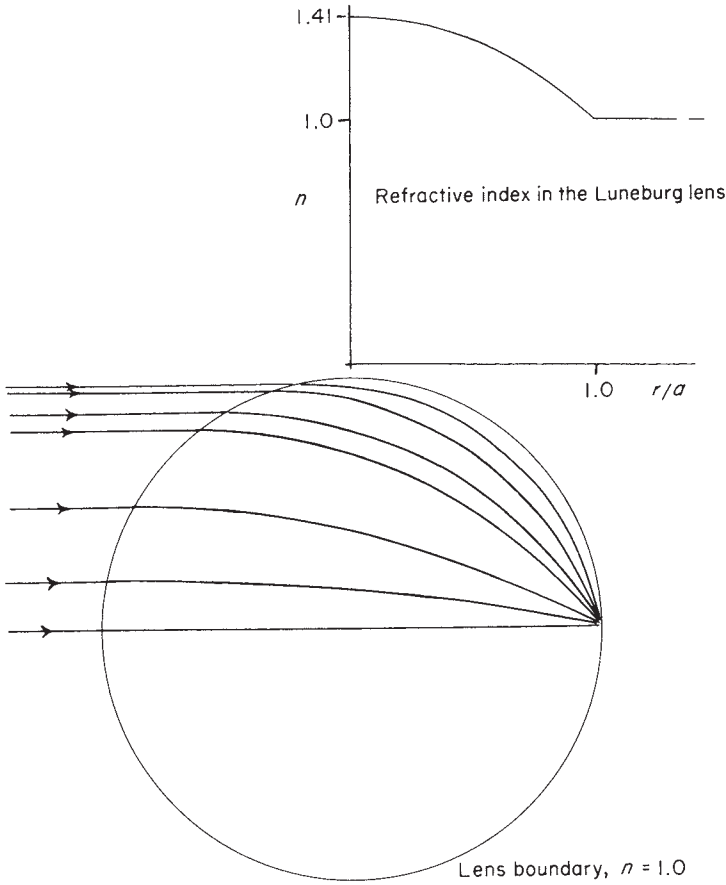
Figure 3.16   The Luneburg lens.

are formed from parallel rays coming in all directions. It is then possible to consider the Luneburg lens as having theoretical maximum concentration ratio for any desired collection angle $\theta_{\max}$ up to $\pi/4$ but collecting from a concave spherical source at infinity onto a concave spherical absorber attached to the lens. Apart from the practical problem of making the lens this is rather an artificial configuration, since up until now we have been considering plane entry and exit apertures. Yet, the Luneburg lens would have an exit aperture in the form of a spherical cap and an entrance aperture that changes in shape with the angle of the rays. Nevertheless, we show in Appendix F that with reasonable and consistent interpretations of "entrance aperture" and "exit aperture" the Luneburg lens has an optical concentration ratio equal to the theoretical maximum.

3.5   ANOTHER SYSTEM OF SPHERICAL SYMMETRY

The discussion in the last section, in which it was suggested that the ideas of concentration could be extended to nonplane absorbers, suggests a way in which the
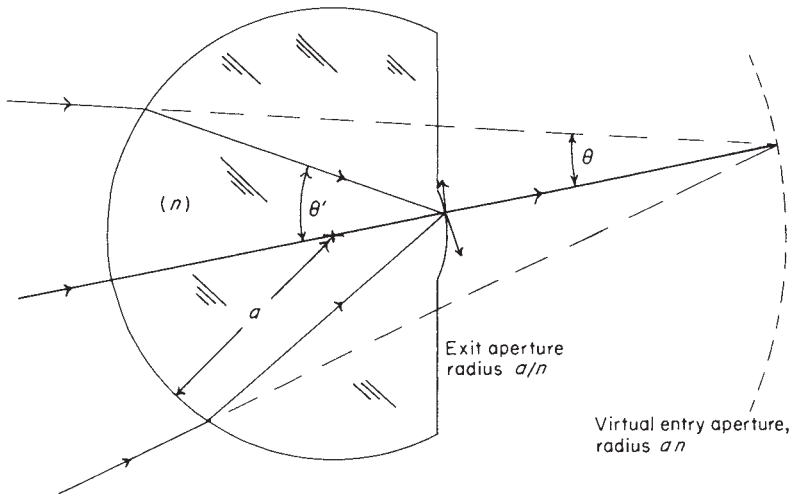
Figure 3.17   The aplanatic spherical lens as an ideal concentrator; the diagram is to scale for refractive index $n = 2$; the rays shown emerge in air as the extreme rays in a solid angle 2π.

aplanatic imaging system of Figure 3.10 could be used by itself as a concentrator, as in Figure 3.17. The surface of radius a/n forms the spherical exit surface, and the internal angle 2α of the cone meeting this face is such that $\sin\alpha = 1/n$. Thus, the emerging rays cover a solid angle 2π, as with the Luneburg lens. The entry aperture is now a virtual aperture on the surface of radius an. The collecting angle $\theta_{\max}$ is thus given by $\sin\theta_{\max} = 1/n^2$. The concentration ratio from air to air is n^4 for a 3D system—that is, it is determined only by the refractive index. Similarly, the collecting angle is fixed. Thus, this is not a very flexible system, apart from the fact that it has a virtual collecting aperture. But it does have the theoretical maximum concentration ratio, and, if we admit such systems, it is another example of an ideal system.

3.6   IMAGE-FORMING MIRROR SYSTEMS

In this section we examine the performance of mirror systems as concentrators. Concave mirrors have, of course, been used for many years as collectors for solar furnaces and the like. Historical material about such systems is given by Krenz (1976). However, little seems to have been published in the way of angle-transmission curves for such systems. Consider first a simple paraboloidal mirror, as in Figure 3.18. As is well known, this mirror focuses rays parallel to the axis exactly to a point focus, or in our terminology, it has no spherical aberration. However, the off-axis beams are badly aberrated. Thus, in the meridian section (the section of the diagram) it is easily shown by ray tracing that the edge rays at angle θ meet the focal plane further from the axis than the central ray, so this cannot be an ideal concentrator even for emergent rays at angles much less than $\pi/2$. An elementary geometrical argument (see, e.g., Harper et al., 1976) shows how big the exit aperture must be to collect all the rays in the meridian section.
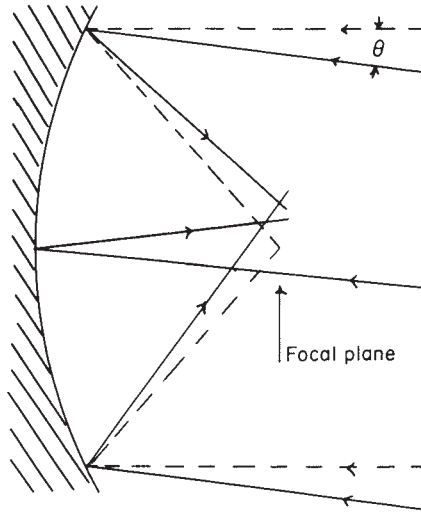
Figure 3.18 Coma of a paraboloidal mirror. The rays of an axial beam are shown in broken line. The outer rays from the oblique beams at angle θ meet the focal plane further from the axis than the central ray of this beam.
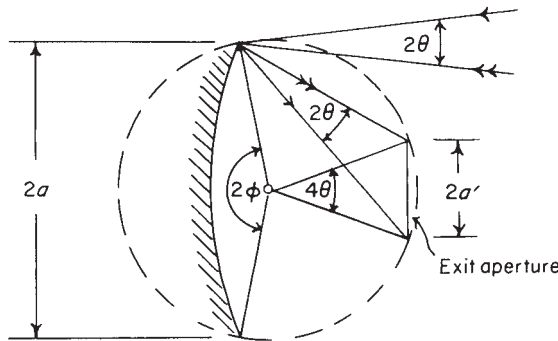
Figure 3.19 Collecting all the rays from a concave mirror.

Referring to Figure 3.19, we draw a circle passing through the ends of the mirror and the absorber (i.e., exit aperture). Then, by a well-known property of the circle, if the absorber subtends an angle $4\theta_{\max}$ at the center of the circle, it subtends $2\theta_{\max}$ at the ends of the mirror, so the collecting angle is $2\theta_{\max}$. The mirror is not specified to be of any particular shape except that it must reflect all inner rays to the inside of the exit aperture. Then if the mirror subtends 2ϕ at the center of the circle, we find

$$\frac{a'}{a} = \frac{\sin 2\theta_{\max}}{\sin \phi} \tag{3.12}$$

and the minimum value of a' is clearly attained when $\phi = \pi/2$. At this point the optical concentration ratio is, allowing for the obstruction caused by the absorber,

$$\left(\frac{a'}{a}\right)^2 - 1 = \frac{1}{4\sin^2 \theta_{\max}} \cdot \frac{\cos^2 2\theta_{\max}}{\cos^2 \theta_{\max}} \tag{3.13}$$
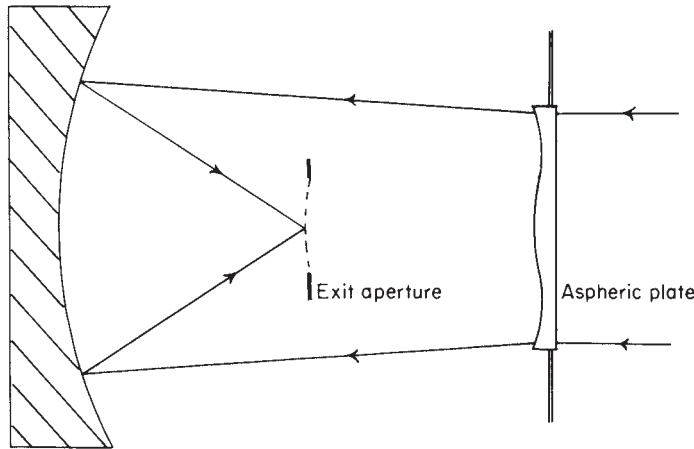
Figure 3.20   The Schmidt camera. This optical system has no spherical aberration or coma, so, in principle, it could be a good concentrator for small collecting angles. However, there are serious practical objections, such as cost and the central obstruction of the aperture.

It can be seen that this is less than 25% of the theoretical maximum concentration ratio and less than 50% of the ideal for the emergent angle used.

If, as is usual, the mirror is paraboloidal, the rays used for this calculation are actually the extreme rays—that is, the rays outside of the plane of the diagram all fall within the circle of radius a.

The large loss in concentration at high apertures is basically because the single concave mirror used in this way has large coma—in other words, it does not satisfy Abbe's sine condition [Eq. (3.2)]. The large amount of coma introduced into the image spreads the necessary size of the exit aperture and so lowers the concentration below the ideal value.

There are image-forming systems that satisfy the Abbe sine condition and have large relative apertures. The prototype of these is the Schmidt camera, which has an aspheric plate and a spherical concave mirror, as shown in Figure 3.20. The aspheric plate is at the center of curvature of the mirror, and thus the mirror must be larger than the collecting aperture. Such a system would have the ideal concentration ratio for a restricted exit angle apart from the central obstruction, but there would be practical difficulties in achieving the theoretical maximum. In any case a system of this complexity is clearly not to be considered seriously for solar work.

3.7   CONCLUSIONS ON CLASSICAL IMAGE-FORMING CONCENTRATORS

It must be quite clear by now that, whatever the theoretical possibilities, practical concentrators based on classical image-forming designs fall a long way short of the ideal. Our graphs of angular transmission will indicate this for some of the simpler designs. As to theoretical possibilities, it is certainly possible to have an ideal concentrator of theoretical maximum concentration ratio if we use a spheri-

cally symmetric geometry, a continuously varying refractive index, and quite unrealistic material properties (i.e., refractive index between 1 and 2 and no dispersion). This was proved by the example of the Luneburg lens, and Luneburg and others (e.g., Morgan, 1958, and Cornbleet, 1976) have shown how designs suitable for perfect imagery for other conjugates can be obtained.

Nevertheless, imaging and nonimaging are not opposite concepts. Not only the Luneburg lens and the other systems shown in this chapter proves it but also the new nonimaging designs found with the SMS method (Chapter 8), whose imaging properties are analyzed in Chapter 9, showing that the combination of imaging and nonimaging properties is possible with very simple devices, thus opening the field of new applications.

Perfect concentrators have not been obtained with plane apertures and axial symmetry only if we restrict ourselves to a finite number of elements. However, we will see in Chapter 6 that it is possible to have perfect concentrators with axial symmetry and plane apertures if we are allowed to use a continuously varying index or a multifoliate structure of thin waveguide surfaces.

REFERENCES

Born, M., and Wolf, E. (1975). "Principles of Optics," 5th Ed. Pergamon, Oxford, England.

Cornbleet, S. (1976). "Microwave Optics." Academic Press, New York.

Harper, D. A., Hildebrand, R. H., Pernic, R., and Platt, S. R. (1976). Heat trap: An optimised far infrared field optics system. *Appl. Opt.* **15**, 53–60.

Krenz, J. H. (1976). "Energy Conversion and Utilization." Allyn & Bacon, Rockleigh, New Jersey.

Luneburg, R. K. (1964). "Mathematical Theory of Optics." Univ. of California Press, Berkeley. This material was originally published in 1944 as loose sheets of mimeographed notes and the book is a word-for-word transcription.

Maxwell, J. C. (1958). On the general laws of optical instruments. *Q. J. Pure Appl. Math.* **2**, 233–247.

Morgan, S. P. (1958). General solution of the Luneburg lens problem. *J. Appl. Phys.* **29**, 1358–1368.

4

NONIMAGING OPTICAL SYSTEMS

4.1 LIMITS TO CONCENTRATION

The relationship between the concentration ratio and the angular field of view is a fundamental one that merits more than one demonstration. We shall give a thermodynamic argument in the context of solar energy concentration. Imagine the sun itself as a spherically symmetric source of radiant energy (Figure 4.1). The flux falls off as the inverse square of the distance R from the center, as follows from the conservation of power through successive spheres of area $4\pi R^2$. Therefore, the flux on the earth's surface, say, is smaller than the solar surface flux by a factor $(r/R)^2$, where r is the radius of the sun and R is the distance from the earth to the sun. By simple geometry, $r/R = \sin^2 \theta$ where θ is the angular subtense (half angle) of the sun. If we accept the premise that no terrestrial device can boost the flux above its solar surface value (which would lead to a variety of perpetual motion machines), then the limit to concentration is just $1/\sin^2 \theta$. We call this limit the sine law of concentration. This relationship may seem similar to the well-known Abbe's sine condition of optics, but the resemblance is only superficial. The Abbe's condition applies to well-corrected optical systems and is first order in the transverse dimensions of the image. There are no such limitations to the sine law of concentration, which is correct and rigorous for any size receiver. As already shown in Eqs. 2.14 and 2.15, there is an escape clause to this conclusion when the target is immersed in a medium with index of refraction (n) because then the limit is $n^2/\sin^2 \theta$. Of course, the limit we have derived is for concentration in both transverse dimensions, which we will refer to as 3-dimensional concentration (or 3D concentration for short). For concentration in one transverse dimension, which we'll call 2-dimensional (2D) concentration, the limit is clearly $1/\sin \theta$ While the concept of concentration in our demonstration refers to solar flux, implicit in our discussion is good energy throughput. We generally deal with concentrators that throw away as little energy as possible. Good energy conservation is an essential attribute of a useful concentrating optical system (see Figure 4.1).
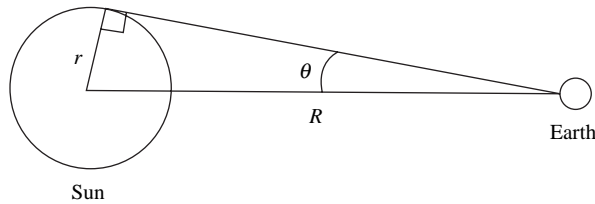
Figure 4.1   The flux from a spherically symmetric sun falls off $(r/R)^2 = (1/\sin\theta)^2$.

$$D = 2\, r \sin\phi$$

$$C = (D/d)^2 = (1/4) \sin^2 2\phi / \sin^2 \theta_s$$
$$\leq (1/4)1/\sin^2 \theta_s \leq (1/4)\, C_{max}$$

$$d = 2\, r \sin\theta / \cos\phi$$
$$D/d = \sin\phi \cos\phi / \sin\theta$$
$$= \sin 2\phi / 2 \sin\theta$$
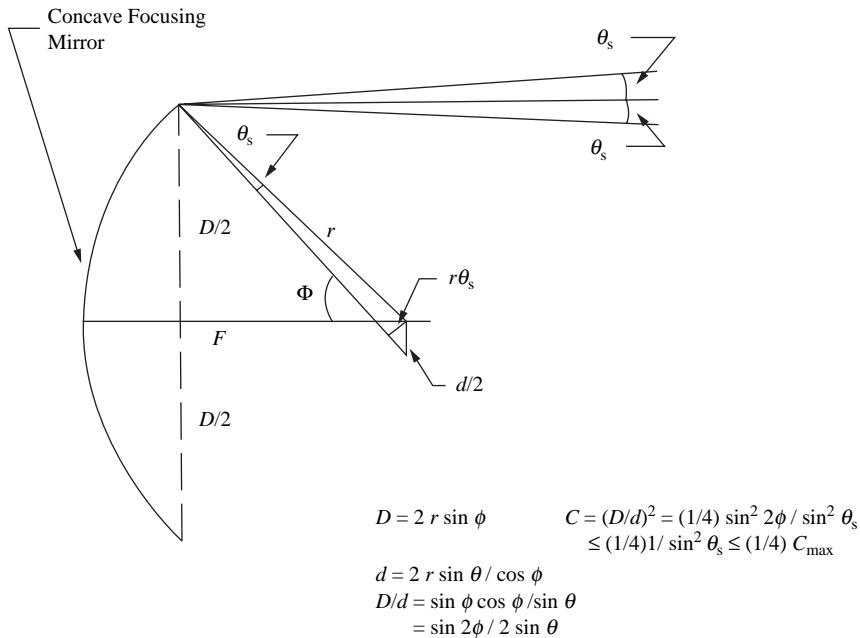
Figure 4.2   Image-forming solar concentrator for planar absorber.

4.2   IMAGING DEVICES AND THEIR LIMITATIONS

If one were to ask the proverbial man on the street for a suggestion of how one might attain the highest possible level of concentration of, say, solar flux, a plausible response would be to use a good astronomical telescope—perhaps the 200-inch telescope on Mt. Palomar or any preferred telescope. Of course, such an experiment had better remain in the realm of imagination only, since beginning astronomers are admonished never to point their telescopes at the sun at the risk of catastrophic damage to the instrument. But to continue this train of thought, the concentration limit of a telescope is readily shown to be $\sin^2(2\phi)/4\sin^2\theta$ after an elementary calculation (Figure 4.2). Here we have introduced a new parameter, ϕ, which is the rim angle of the telescope. The best one can do is make the numerator 1 for rim angle $\phi = 45°$, so the best concentration achieved is $1/4\sin^2\theta$, which falls short of the fundamental limit by a factor of 4! Now factors of 4 are significant in technology (and many other forms of human endeavor). It was the
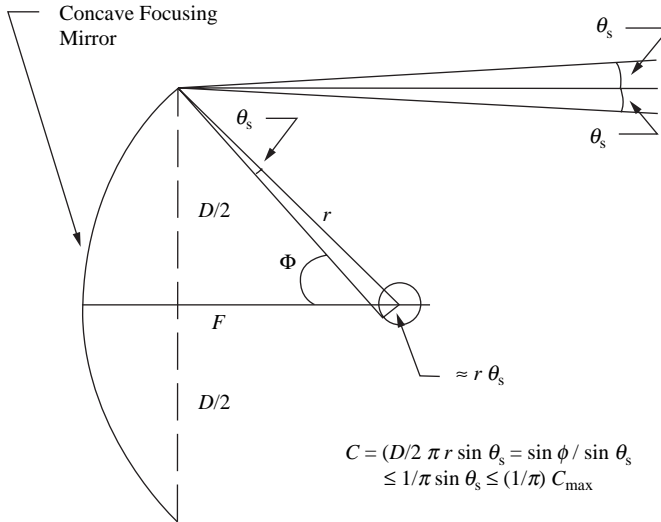
Figure 4.3   Image-forming solar concentrator for tubular absorber.

desire to bridge the gap between the levels of concentration achieved by common imaging devices and the sine law of concentration limit that motivated the invention of nonimaging optics. Entirely similar considerations can be applied to 2D or trough concentrators. A straightforward generalization to a strip absorber rather than a disk absorber gives a limit for, say, a parabolic trough of $\sin(2\phi)/2\sin\theta$, with a consequent upper limit of $1/2\sin\theta$ for rim angle $\phi = 90°$. This would be a useful configuration for a photovoltaic concentrator, with the strip consisting of solar cells. However, a more useful geometry for a parabolic trough thermal concentrator is a tubular receiver (Figure 4.3). In that case, the concentration relation becomes $\sin\phi/\pi\sin\theta$, which attains its maximum value of $1/\pi\sin\theta$ at 90° rim angle. This helps to explain why parabolic solar troughs tend to have large rim angles. In either case, we fall significantly short of the sine law of concentration limit, this time by a factor of 2 or π, depending on the configuration. As we shall see later, this factor of 2 or π makes the difference between the possibility and impossibility of fixed solar concentrators. The possibility of fixed concentrators opens up a broad vista of solar energy applications.

4.3   NONIMAGING CONCENTRATORS

These simple examples of imaging systems and their attendant shortfalls in concentrating performance (we could have examined lenses and reached similar conclusions) suggest that the requirement that an optical concentrator form an image is unduly restrictive. After all, we are after transport of radiant energy. Imaging the sun may be useful in solar astronomy or in the study of sun spots, but it has no obvious advantage in solar energy conversion systems. Thus, even taking a more empirical optimization approach, it is plausible that relaxing the imaging requirement has the potential of improving concentrating performance; in other words, one would expect to be able to trade off one against the other. Approaching

the subject this way does lead to incremental improvements over various classical imaging designs such as parabolic reflectors. Our approach in this book is to show methods that actually attain, or closely approach, the theoretical sine law limit to concentration while maintaining high throughput, methods that bear little resemblance to classical imaging approaches. An analogy with fluid dynamics may be useful to bring this point home. In fluid dynamics, as in optics, a useful representation is in "phase space." Phase space has twice the dimensions of ordinary space and consists of both the positions and momenta of elements of the fluid. In optics, the momenta are the directions of light rays multiplied by the index of refraction of the medium. In optics, as in fluid dynamics, the volume in this phase space is conserved, a sort of incompressible fluid flowing in this space of twice the number of physical dimensions. Now consider an imaging problem taking the simplest example of points on a line. An imaging system is required to map those points on another line, called the image, without scrambling the points (Figure 4.4a)—that is, to focus the rays issuing from every object point into their corresponding image points. All the rays issuing from a point are represented by a vertical line in the phase space, and the system is required to faithfully map line onto line (Figure 4.4b). That may appear quite demanding, but it is precisely what an
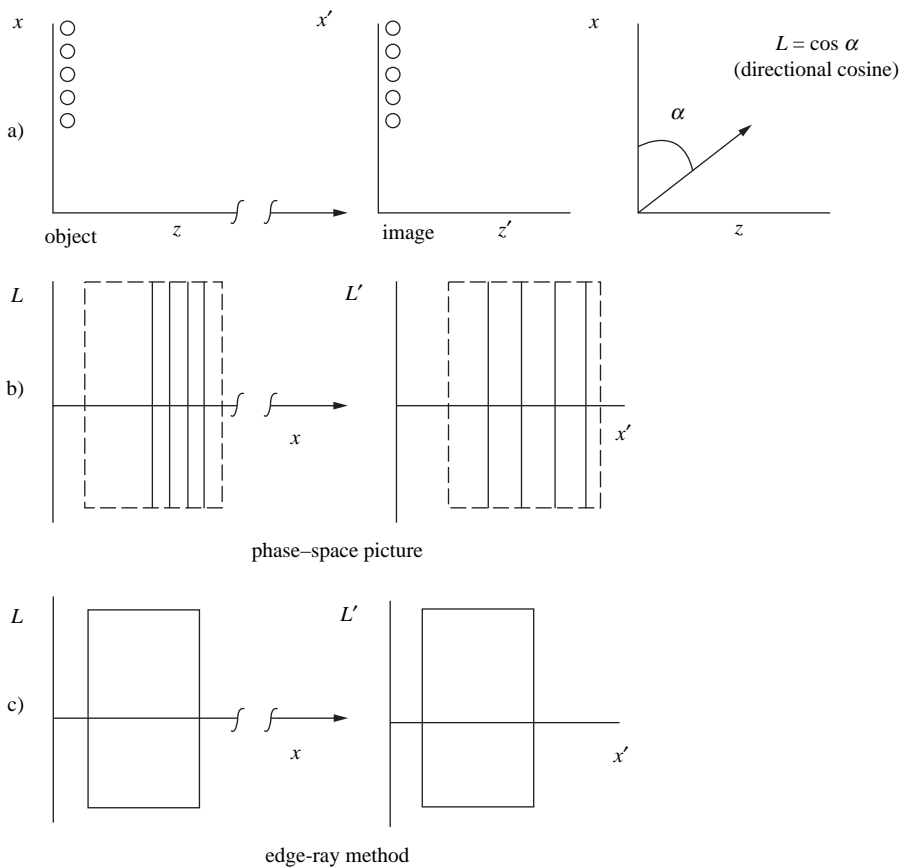
Figure 4.4   Comparing imaging with nonimaging in the phase-space representation.

imaging system is asked to perform. But suppose we consider only the boundary or edge of all the rays. Then all we require is that the boundary be transported from the source to the target. The interior rays will come along (Figure 4.4c). They cannot "leak out" because were they to cross the boundary, they would first become the boundary, and it is the boundary that is being transported. To complete the analogy, the volume of container of rays is unchanged in the process. This is the conservation of phase space volume. It is very much like transporting a container of an incompressible fluid—say, water. The fact that elements inside the container mix or the container itself is deformed is of no consequence. To carry the analogy a bit further, suppose one were faced with the task of transporting a vessel (the volume in phase space) filled with alphabet blocks spelling out a message. Then one would have to take care not to shake the container and thereby scramble the blocks. But if one merely needs to transport the blocks without regard to the message, the task is much easier. This is the key idea of nonimaging optics.

4.4   THE EDGE-RAY PRINCIPLE OR "STRING" METHOD

Figure 4.4c suggests the notion of transporting the boundary or edge of the container of rays in phase space. This leads to one of the most useful algorithms of nonimaging optics. We shall see that transporting the edges only, without regard to interior order, allows attainment of the sine law of concentration limit.

To motivate the method we begin with the statement that all imaging optical design derives from Fermat's principle. The optical path length between object and image points are the same for all rays (Figure 4.5a). When this same principle is applied to "strings" instead of rays, it gives the edge-ray algorithm of nonimaging optical design (Figure 4.5b). First we must explain what strings are, which is best done by example. We will proceed to solve the problem of attaining the sine law limit of concentration for the simplest case: that of a flat absorber. Referring to Figure 4.6, we loop one end of a "string" to a "rod" tilted at angle θ to the aperture AA' and tie the other end to the edge of the exit aperture B'. Holding the length fixed, we trace out a reflector profile as the string moves from C to A'. From simple geometry, the relation $BB' = AA' \sin\theta$ immediately follows. This construction gives the 2D compound parabolic concentrator, or CPC. Rotating the profile about the axis of symmetry gives the 3D CPC with radius (a) at the entrance and (b) at the exit. The 2D CPC is an ideal concentrator—that is, it works perfectly for all rays within the acceptance angle θ (in 2D geometry). As we will see, the 3D CPC is very close to ideal in 3D geometry. The flat absorber case is a natural candidate for rotating about the axis because the square of the ratio of diameters ($\sin^2\theta$) agrees with the maximum concentration. Other absorber shapes such as circular cross-sections (cylinders in 2D, spheres in 3D) do not have this correspondence because the area of the sphere is $4\pi b^2$, whereas the entrance aperture area is πa^2. Notice that we have kept the optical length of the string fixed. For media with varying index of refraction (n), the physical length is multiplied by n. Of course, we have not demonstrated that this construction actually works. One admittedly tongue-in-cheek approach is to state that anything this "neat"—in other words, that satisfies the conservation laws in a natural way—has to work. Perhaps a more serious "proof" is to notice that the 2D CPC rejects all stray radi-

Imaging Optics

$$\int_{P\ \text{ray}}^{P'} D\, dl = \text{constant [Fermat 1601-1665], where}$$

n = index of refraction

I = path length

a)

P P'

Nonimaging Optics:

b)

P

$$\int_{P\ \text{string}}^{P'} n\, dl = \text{constant}$$

P'
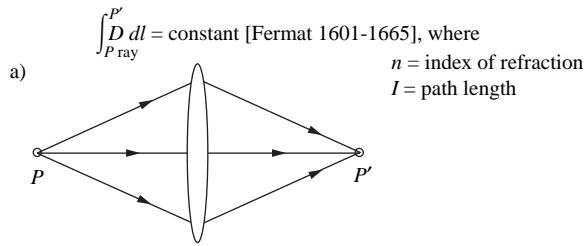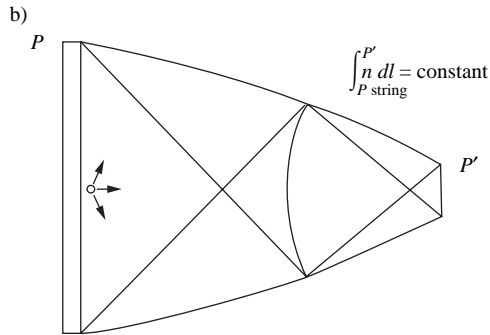
Figure 4.5   Fermat's principle for rays and strings.

String Method

Edge ray Wave front W

C

θ

A'

A

reflector profile

$$\int_{W}^{B'} n\, dl = \text{constant}$$

B' B

$$AC + AB' = A'B + BB'$$
$$AC = AA' \sin \Theta \implies AA' \sin \Theta = BB'$$
$$AB' = A'B$$
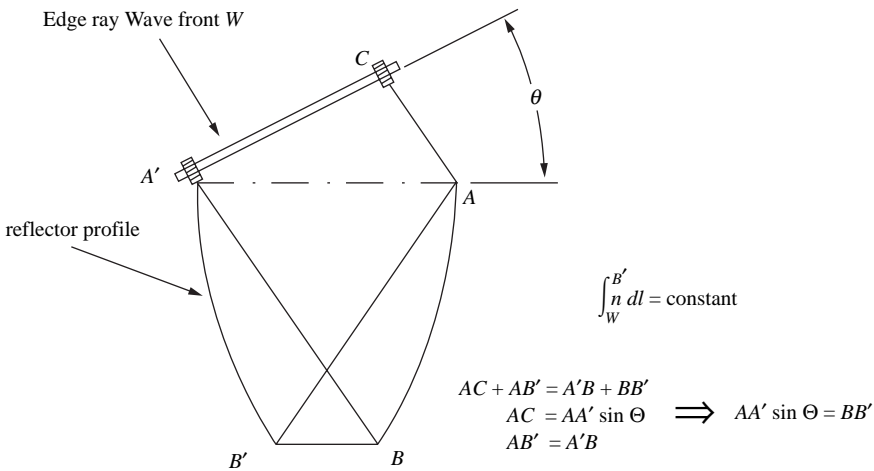
Figure 4.6   String construction for planar absorber.

ation and therefore must be ideal by conservation of phase space (Ries and Rabl, 1994; Winston, 1970). The string construction is very versatile and can be applied to any convex (or at least nonconcave) absorber. Figure 4.7 shows the string construction for a tubular absorber as would be appropriate for a solar thermal concentrator. This time $2\pi a = BB' \sin \theta$, where a is the radius of the cylindrical absorber (Winston, Hinterberger, 1975).
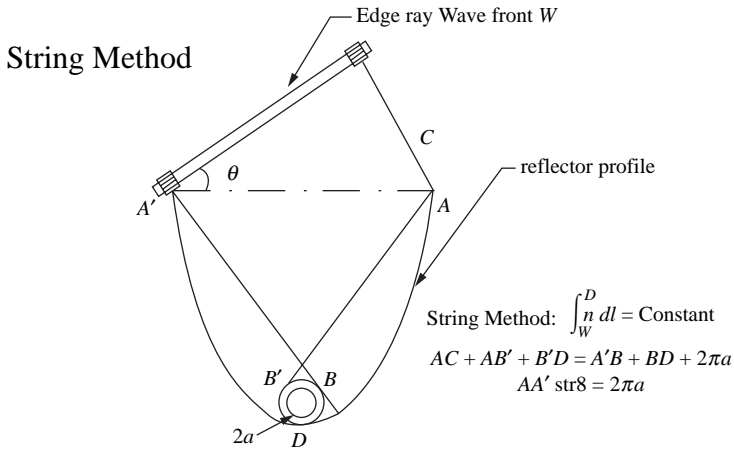
Figure 4.7   String construction for tubular absorber.
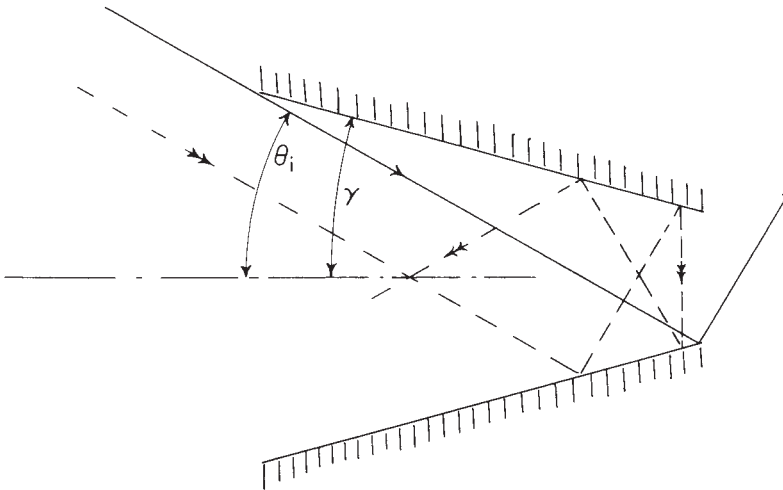
Figure 4.8   The cone concentrator.

4.5   LIGHT CONES

A primitive form of nonimaging concentrator, the light cone, has been used for many years (see, e.g., Holter et al., 1962). Figure 4.8 shows the principle. If the cone has semiangle γ and if θ_i is the extreme input angle, then the ray indicated will just pass after one reflection if $2\gamma = (\pi/2) - \theta_i$. It is easy to arrive at an expression for the length of the cone for a given entry aperture diameter. Also, it is easy to see that some other rays incident at angle θ_i, such as that indicated by the double arrow, will be turned back by the cone. If we use a longer cone with more reflections, we still find some rays at angle θ_i being turned back. Clearly, the cone is far from being an ideal concentrator. Williamson (1952) and Witte (1965) attempted some analysis of the cone concentrator but both restricted this treat-

ment to meridian rays. This unfortunately gives a very optimistic estimate of the concentration. Nevertheless, the cone is very simple compared to the image-forming concentrators described in Chapter 3 and its general form suggests a new direction in which to look for better concentrators.

4.6   THE COMPOUND PARABOLIC CONCENTRATOR

The flat absorber case occupies a special place because of its simplicity. Historically it was the first to be discovered. For these reasons its description and properties merit a separate discussion.

If we attempt to improve on the cone concentrator by applying the edge-ray principle, we arrive at the compound parabolic concentrator (CPC), the prototype of a series of nonimaging concentrators that approach very close to being ideal and having the maximum theoretical concentration ratio.

Descriptions of the CPC appeared in the literature in the mid-1960s in widely different contexts. The CPC was described as a collector for light from Cerenkov counters by Hinterberger and Winston (1976a,b). Almost simultaneously, Baranov (1965a), and Baranov and Mel'nikov (1966) described the same principle in 3D geometry, and Baranov (1966) suggested 3D CPCs for solar energy collection. Baranov (1965b; 1967) obtained Soviet patents on several CPC configurations. Axially symmetric CPCs were described by Ploke (1967), with generalizations to designs incorporating refracting elements in addition to the light-guiding reflecting wall. Ploke (1969) obtained a German patent for various photometric applications. In other applications to light collection for applications in high-energy physics, Hinterberger and Winston (1966a,b; 1968a,b) noted the limitation to $1/\sin^2\theta$ of the attainable concentration, but it was not until some time later that the theory was given explicitly (Winston, 1970). In the latter publication the author derived the generalized étendue (see appendix A) and showed how the CPC approaches closely to the theoretical maximum concentration.

The CPC in 2D geometry was described by Winston (1974). Further elaborations may be found in Winston and Hinterberger (1975) and Rabl and Winston (1976). Applications of the CPC in 3D form to infrared collection (Harper et al., 1976) and to retinal structure (Baylor and Fettiplace, 1975; Levi-Setti et al., 1975; Winston and Enoch, 1971) have also been described. The general principles of CPC design in 2D geometry are given in a number of U.S. patents (Winston, 1975; 1976a; 1977a,b).

Let us now apply the edge-ray principle to improve the cone concentrator. looking at Figure 4.9, we require that all rays entering at the extreme collecting angle θ_i shall emerge through the rim point P' of the exit aperture. If we restrict ourselves to rays in the meridian section, the solution is trivial, since it is well known that a parabolic shape with its axis parallel to the direction θ_i and its focus at P' will do this, as shown in Figure 4.10. The complete concentrator must have an axis of symmetry if it is to be a 3D system, so the reflecting surface is obtained by rotating the parabola about the concentrator axis (not about the axis of the parabola).

The symmetry determines the overall length. In the diagram the two rays are the extreme rays of the beam at θ_i, so the length of the concentrator must be such
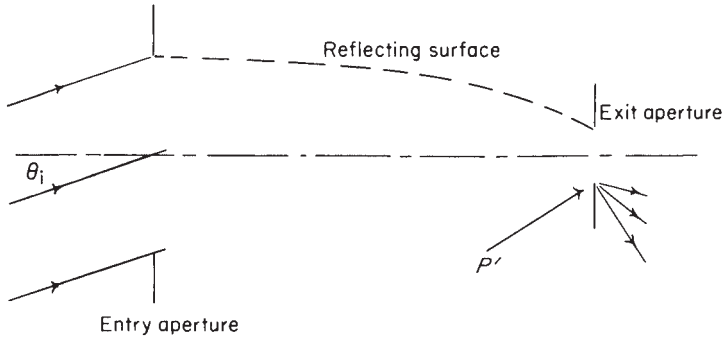
Figure 4.9   The edge-ray principle.
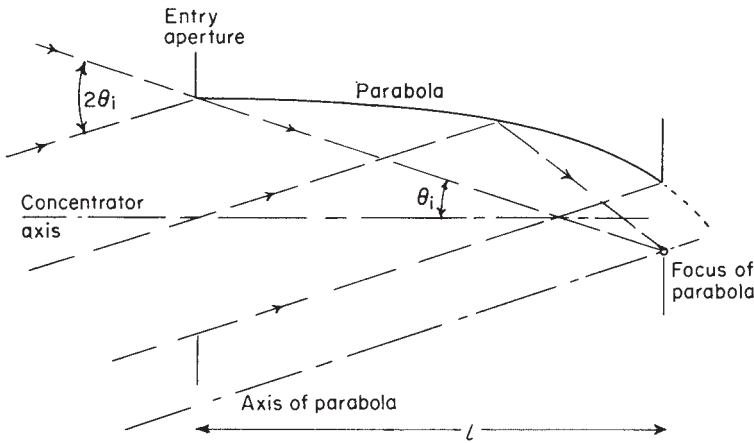
Figure 4.10   Construction of the CPC profile from the edge-ray principle.

as to just pass both these rays. These considerations determine the shape of the CPC completely in terms of the diameter of the exit aperture $2a'$ and the maximum input angle θ_i. It is a matter of simple coordinate geometry (Appendix G) to show that the focal length of the parabola is

$$f = \frac{a'}{1 + \sin\theta_i} \qquad (4.1)$$

the overall length is

$$L = \frac{a'(1 + \sin\theta_i)\cos\theta_i}{\sin^2\theta_i} \qquad (4.2)$$

and the diameter of the entry aperture is

$$a = \frac{a'}{\sin\theta_i} \qquad (4.3)$$

Also, from Eqs. (4.2) and (4.3) or directly from the figure,

$$L = (a + a')\cot\theta_i \qquad (4.4)$$
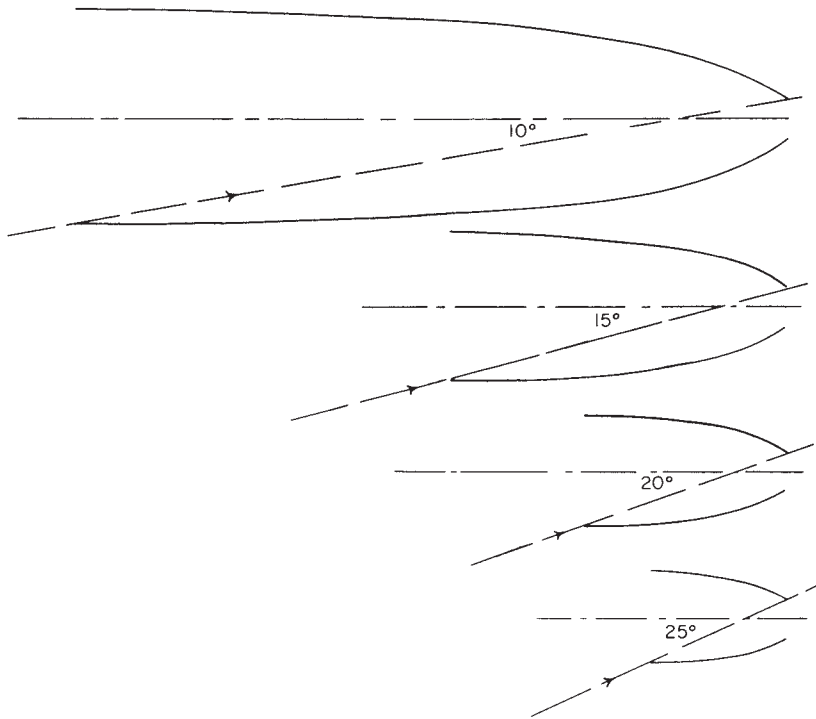
Figure 4.11   Some CPCs with different collecting angles. The drawings are to scale with the exit apertures all equal in diameter.

Figure 4.11 shows scale drawing of typical CPCs with a range of collecting angles. It is shown in Appendix G that the concentrator wall has zero slope at the entry aperture, as drawn.

The most remarkable result is Eq. (4.3). We see from this that the CPC would have the maximum theoretical concentration ratio (see Section 2.7)

$$\frac{a}{a'} = \frac{1}{\sin\theta_i} \tag{4.5}$$

provided all the rays inside the collecting angle θ_i actually emerge from the exit aperture. Our use of the edge-ray principle suggests that this ought to be the case, on the analogy with image-forming concentrators, but in fact this is not so. The 3D CPC, like the cone concentrator, has multiple reflections, and these can actually turn back the rays that enter inside the maximum collecting angle. Nevertheless, the transmission-angle curves for CPCs as calculated by ray tracing approach very closely the ideal square shape. Figure 4.12, after Winston (1970), shows a typical transmission-angle curve for a CPC with $\theta_i = 16°$.

It can be seen that the CPC comes very close to being an ideal concentrator. Also, it has the advantages of being a very practical design, easy to make for all wavelengths, since it depends on reflection rather than refraction, and of not requiring any extreme material properties. The only disadvantage is that it is very long compared to its diameter, as can be seen from Eq. (4.2). This can be overcome if we incorporate refracting elements into the basic design. In later sections of this
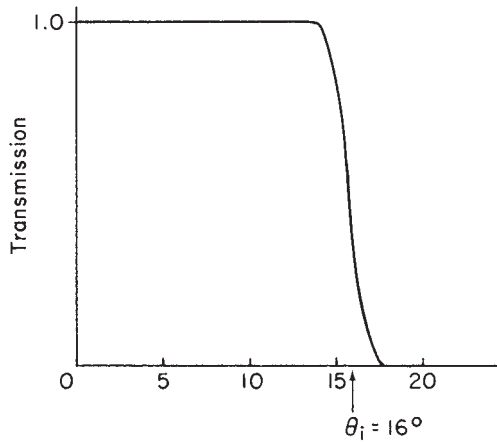
Figure 4.12   Transmission-angle curve for a CPC with acceptance angle $\theta_i = 16°$. The cutoff occurs over a range of about 1°.

chapter we shall study the optics of the CPC in detail. We shall elucidate the mechanism by rays inside the collecting angle which are turned back, give transmission-angle curves for several collecting angles, and give quantitative comparisons with some of the other concentrators, imaging and nonimaging, that have been proposed. In later chapters we shall discuss modifications of the basic CPC along various lines—for example, incorporating transparent refracting materials in the design and even making use of total internal reflection at the walls for all the accepted rays.

   We conclude this section by examining the special case of the 2D CPC or troughlike concentrator. This has great practical importance in solar energy applications, since, unlike other trough collectors, it does not require diurnal guiding to follow the sun. The surprising result is obtained that the 2D CPC is actually an ideal concentrator of maximum theoretical concentration ratio—that is, no rays inside the maximum collecting angle are turned back. To show this result we have to find a way of identifying rays that do get turned back after some number of internal reflections. The following procedure for identifying such rays actually applies not only to CPCs but to all axisymmetric conelike concentrators with internal reflections. It is a way of finding rays on the boundary between sets of rays that are turned back and rays that are transmitted. These extreme rays must just graze the edge of the exit aperture, as in Figure 4.13, so that if we trace rays in reverse from this point in all directions as indicated, these rays appear in the entry aperture on the boundary of the required region. Thus, we could choose a certain input direction, find the reverse traced rays having this direction, and plot their intersections with the plane of the input aperture. They could be sorted according to the number of reflections involved and the boundaries plotted out. Diagrams of this kind will be given for 3D CPCs in the next chapter.

   Returning to the 2D CPC, we note first that the ray tracing in any 2D troughlike reflector is simple even for rays not in a plane perpendicular to the length of the trough. This is because the normal to the surface has no component parallel to the length of the trough, and thus the law of reflection [Eq.(2.1)] can be applied
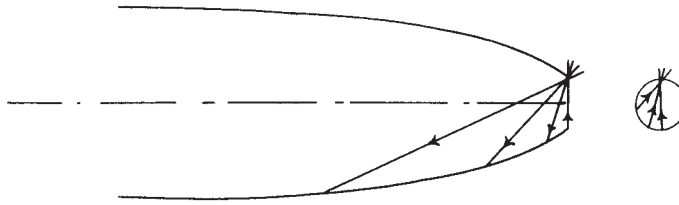
Figure 4.13   Identifying rays that are just turned back by a conelike concentrator. The rays shown are intended as projections of skew rays, since the meridional rays through the rim correspond exactly to θ_i by construction for a CPC.
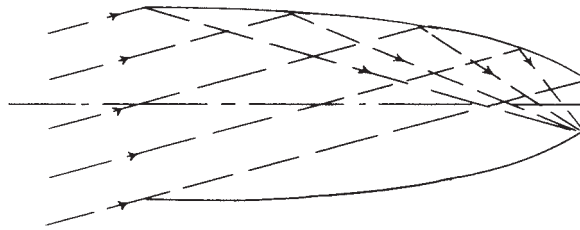
Figure 4.14   A 2D CPC. The rays drawn represent projections of rays out of the plane of the diagram.

in two dimensions only. The ray direction cosine in the third dimension is constant. Thus, if Figure 4.14 shows a 2D CPC with the length of the trough perpendicular to the plane of the diagram, all rays can be traced using only their projections on this plane. We can now apply our identification of rays that get turned back. Since, according to the design, all the rays shown appear in the entry aperture at θ_{max}, there can be no returned rays within this angle. The 2D CPC has maximum theoretical concentration ratio and its transmission-angle graph therefore has the ideal shape, as in Figure 4.15.[1]

Since this property is of prime importance, we shall examine the ray paths in more detail to strengthen the verification. Figure 4.16 shows a 2D CPC with a typical ray at the extreme entry angle θ_{max}. Say this ray meets the CPC surface at P. A neighboring ray at a smaller angle would be represented by the broken line. There are then two possibilities. Either this ray is transmitted as in the diagram, or else it meets the surface again at P_1. In the latter case we apply the same argument except using the extreme ray incident at P_1, and so on. Thus, although some rays have a very large number of reflections, eventually they emerge if they entered inside θ_{max}. Of course, in the preceding argument "ray" includes "projection" of a ray skew to the diagram.

This result shows a difference between 2D and 3D CPCs. The 2D CPC has maximum theoretical concentration, in the sense of Section 2.9. In extending it to

---

[1]   Strictly, this applies to 2D CPCs that are indefinitely extended along the length of the trough. In practice, this effect is achieved by closing the ends with plane mirrors perpendicular to the straight generators of the trough. This ensures that all rays entering the rectangular entry aperture within the acceptance angle emerge from the exit aperture.
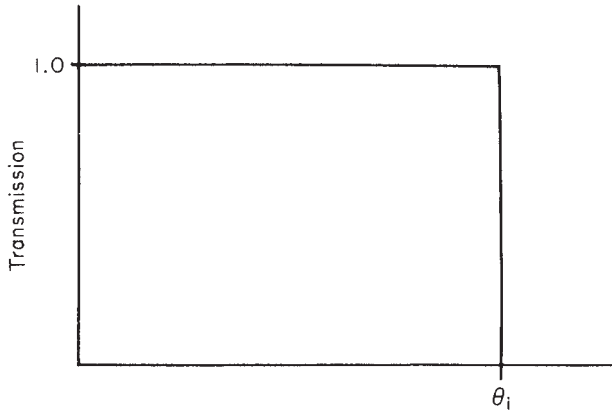
Figure 4.15   The transmission-angle curve for a 2D CPC.
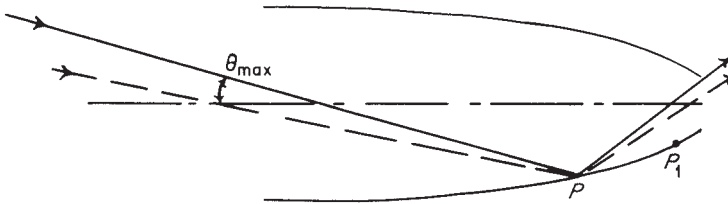
Figure 4.16   To prove that a 2D CPC has an ideal transmission-angle characteristic.

3D, however, we have included more rays (there is now a threefold infinity of rays, allowing for the axial symmetry, whereas in the 2D case we have to consider only a twofold infinity). We have no more degrees of freedom in the design, since the 3D concentrator is obtained from the 2D profile by rotation about the axis of symmetry. The 3D concentrator has to be a figure of revolution, and thus we can do nothing to ensure that rays outside the meridian sections are properly treated. We shall see in Section 4.7.3 that it is the rays in these regions that are turned back by multiple reflections inside the CPC.

This discussion also shows the different causes of nonideal performance of imaging and nonimaging systems. The rays in an image-forming concentrator such as a high-aperture lens all pass through each surface the same number of times (usually once), and the nonideal performance is caused by geometrical aberrations in the classical sense. In a CPC, on the other hand, different rays have different numbers of reflections before they emerge (or not) at the exit aperture. It is the effect of the reflections in turning back the rays that produces nonideal performance. Thus, there is an essential difference between a lens with large aberrations and a CPC or other nonimaging concentrator. A CPC is a system of rotational symmetry, and it would be possible to consider all rays having just, say, three reflections and discuss the aberrations (no doubt very large) of the image formation by these rays. But there seems no sense in which rays with different numbers of reflections could be said to form an image. It is for this reason that we continue to draw the distinction between image-forming and nonimaging concentrators.

4.7 PROPERTIES OF THE COMPOUND PARABOLIC CONCENTRATOR

In this section we examine the properties of the basic CPC of which the design was developed in the last section. We'll see how ray tracing can be done, the results of ray tracing in the form of transmission-angle curves, certain general properties of these curves, and the patterns of rays in the entry aperture that get turned back. This detailed examination will help in elucidating the mode of action of CPCs and their derivatives, to be described in later chapters.

4.7.1 The Equation of the CPC

By rotation of axes and translation of origin we can write down the equation of the meridian section of a CPC. In terms of the diameter $2a'$ of the exit aperture and the acceptance angle $\theta_{\max}$ this equation is

$$(r\cos\theta_{\max} + z\sin\theta_{\max})^2 + 2a'(1+\sin\theta_{\max})^2 r - 2a'\cos\theta_{\max}(2+\sin\theta_{\max})^2 z$$
$$- a'^2(1+\sin\theta_{\max})(3+\sin\theta_{\max}) = 0 \qquad\qquad (4.6)$$

where the coordinates are as in Figure 4.17. Recalling that the CPC is a surface of revolution about the z axis we see that in three dimensions, with $r^2 = x^2 + y^2$, Eq. (4.6) represents a fourth-degree surface.

A more compact parametric form can be found by making use of the polar equation of the parabola. Figure 4.18 shows how the angle ϕ is defined. In terms of this angle and the same coordinates (r, z) the meridian section is given by
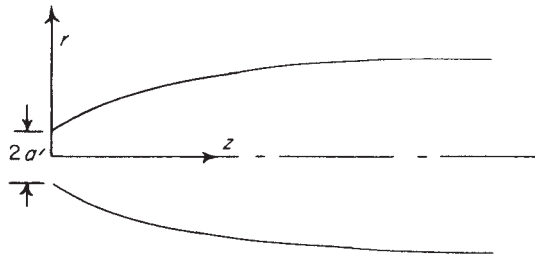
Figure 4.17   The coordinate system for the $r-z$ equation for the CPC.
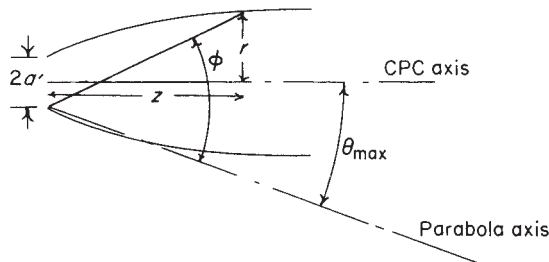
Figure 4.18   The angle ϕ used in the parameteric equations of the CPC.

$$r = \frac{2f \sin(\phi - \theta_{\max})}{1 - \cos\phi} - a', \quad z = \frac{2f \cos(\phi - \theta_{\max})}{1 - \cos\phi} \qquad (4.7)$$

$[f = a'(1 + \sin\theta_{\max})]$.

   If we introduce an azimuthal angle ψ we obtain the complete parametric equations of the surface:

$$x = \frac{2f \sin\psi \sin(\phi - \theta_{\max})}{1 - \cos\phi} - a' \sin\psi$$

$$y = \frac{2f \cos\psi \sin(\phi - \theta_{\max})}{1 - \cos\phi} - a' \cos\psi \qquad (4.8)$$

$$z = \frac{2f \cos(\phi - \theta_{\max})}{1 - \cos\phi}$$

The derivations of these equations are sketched in Appendix G.

4.7.2   The Normal to the Surface

We need the direction cosines of the normal to the surface of the CPC for ray-tracing purposes. There are well-known formulas of differential geometry that give these. If the explicit substitution $r = (x^2 + y^2)^{1/2}$ is made in Eq. (4.6), and the result is written in the form

$$F(x, y, z) = 0 \qquad (4.9)$$

the direction cosines are given by

$$n = (F_x, F_y, F_z)/(F_x^2 + F_y^2 + F_z^2)^{1/2} \qquad (4.10)$$

   The formulas for the normal are slightly more complicated for the parametric form. We first define the two vectors

$$a = (\partial x/\partial\phi, \partial y/\partial\phi, \partial z/\partial\phi), \quad b = (\partial x/\partial\psi, \partial y/\partial\psi, \partial z/\partial\psi) \qquad (4.11)$$

Then the normal is given by

$$n = a \times b / \{|a|^2 |b|^2 - |a \cdot b|^2\}^{1/2} \qquad (4.12)$$

These results are given in elementary texts such as Weatherburn (1931).

   Although the formulas for the normal are somewhat opaque, it can be seen from the construction for the CPC profile in Figure 4.10 that at the entry end the normal is perpendicular to the CPC axis—that is, the wall is tangent to a cylinder.

4.7.3   Transmission-Angle Curves for CPCs

In order to compute the transmission properties of a CPC, the entry aperture was divided into a grid with spacing equal to 1/100 of the diameter of the aperture and rays were traced at a chosen collecting angle θ at each grid point. The proportion of these rays that were transmitted by the CPC gave the transmission $T(\theta, \theta_{\max})$ for the CPC with maximum collecting angle $\theta_{\max}$. $T(\theta, \theta_{\max})$ was then plotted against θ to give the transmission-angle curve. Some of these curves are given in Figure 4.19. They all approach very closely the ideal rectangular cutoff that a concentrator with maximum theoretical concentration ratio should have. The transition
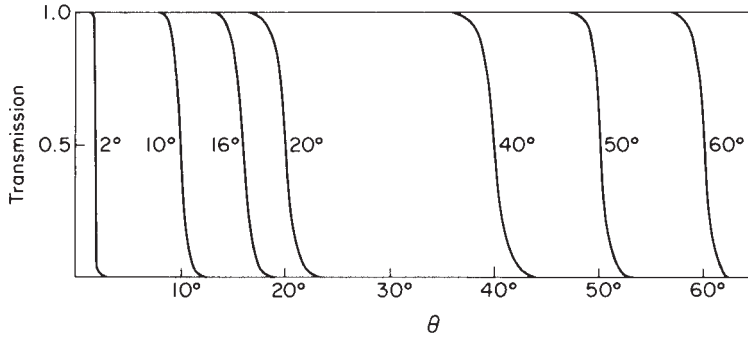
Figure 4.19   Transmission-angle curves for 3D CPCs with $\theta_{\max}$ from 2° to 60°.
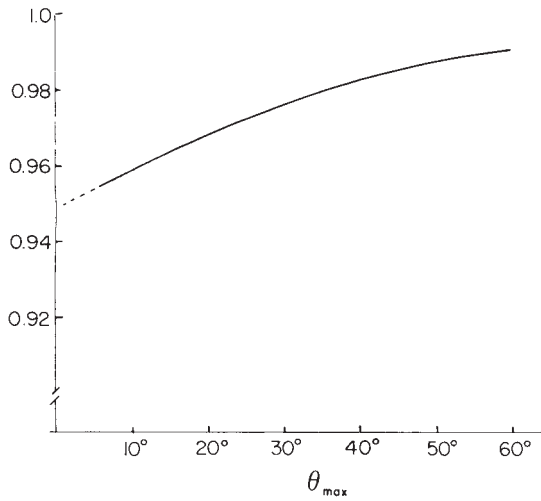
Figure 4.20   Total transmission within $\theta_{\max}$ for 3D CPCs.

from $T = 0.9$ to $T = 0.1$ takes place in $\Delta\theta$ less than 3° in all cases. Approximate values are

$\theta_{\max}$	2°	10°	16°	20°	40°	60°
$\Delta\theta$	0.4°	1.5°	2°	2.5°	2.7°	2.0°

We may also be interested in the total flux transmitted inside the design collecting angle $\theta_{\max}$. This is clearly proportional to

$$\int_{0}^{\theta_{\max}} T(\theta,\theta_{\max})\sin 2\theta \, d\theta \qquad (4.13)$$

and if we divide by $\int_{0}^{\theta_{\max}} \sin 2\theta \, d\theta$, we obtain the fraction transmitted of the flux incident inside a cone of semiangle $\theta_{\max}$. The result of such a calculation is shown in Figure 4.20. This gives the proportion by which the CPC fails to have the theoretical maximum concentration ratio. For example, the 10° CPC should have the theoretical concentration ratio $\csc^2 10° = 33.2$, but from the graph it will

actually have 32.1. The loss is, of course, because some of the skew rays have been
turned back by multiple reflections inside the CPC.

It is of some considerable theoretical interest to see how these failures occur.
By tracing rays at a fixed angle of incidence, regions could be plotted in the entry
aperture showing what happened to rays in each region. Thus, Figure 4.21 shows
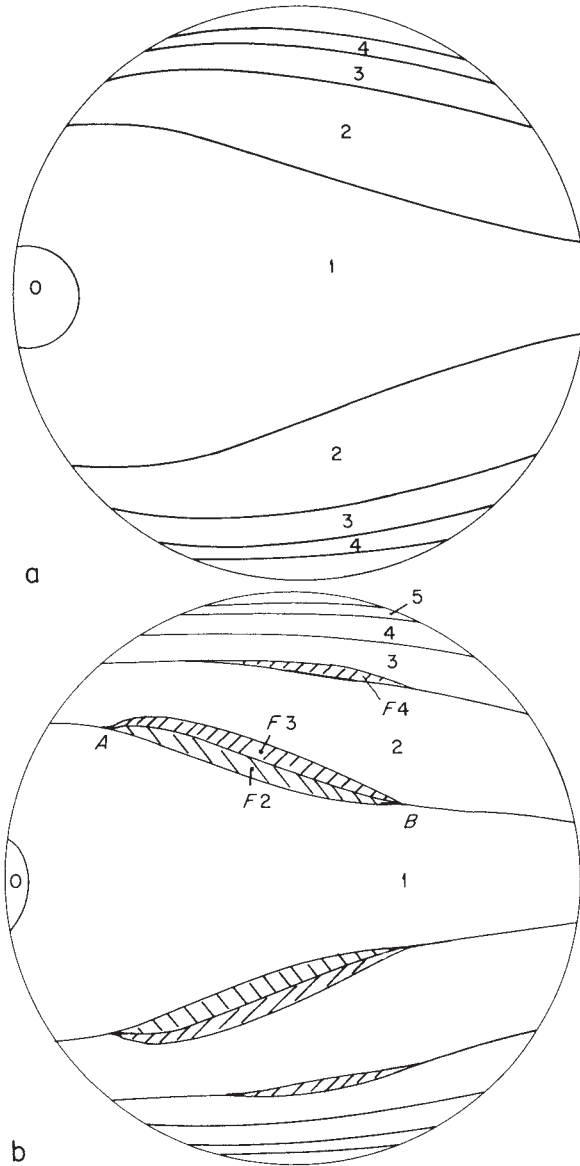
Figure 4.21   Patterns of accepted and rejected rays at the entry face of a 10° CPC. The
entry aperture is seen from above with incident rays sloping downward to the right. Rays
entering areas labeled n are transmitted after n reflections; those entering hatched areas
labeled Fm are turned back after m reflections. The ray trace was not carried to completion
in the unlabeled areas. (a) 8°, $\theta_{max} = 10°$; (b) 9°, $\theta_{max} = 10°$; (c) 9.5°, $\theta_{max} = 10°$; (d) 10°, $\theta_{max} =$
10°; (e) 10.5°, $\theta_{max} = 10°$; (f) 11°, $\theta_{max} = 10°$; (g) 11.5°, $\theta_{max} = 10°$.
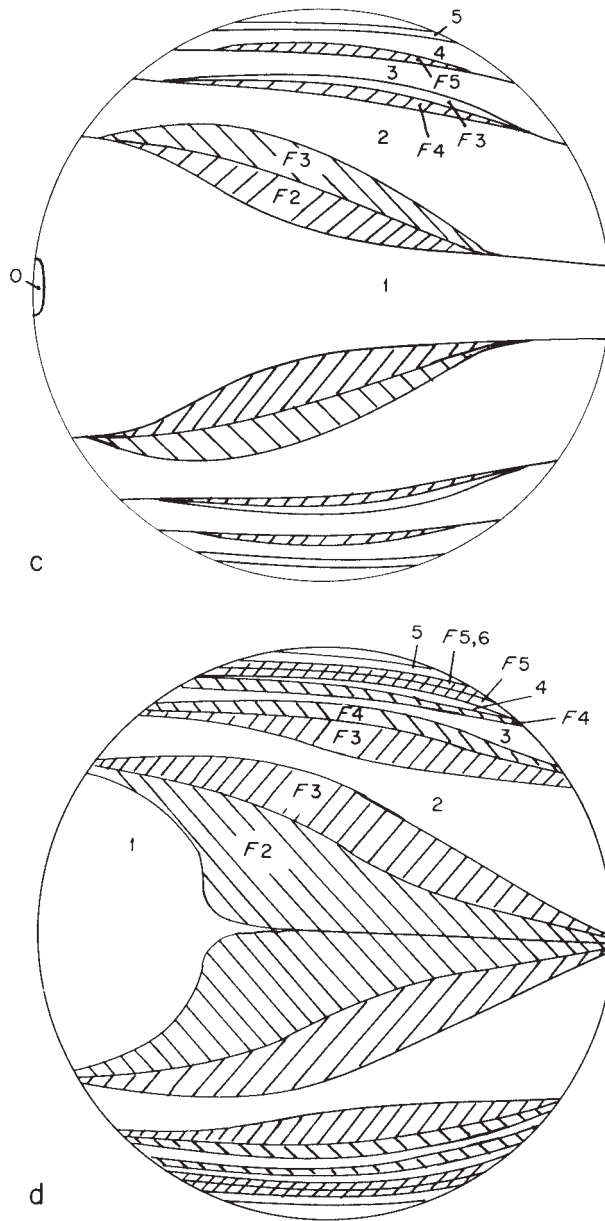
Figure 4.21   *Continued*

these plots for a CPC with $\theta_{\max} = 10°$ for rays at 8°, 9°, 9.5°, 10°, 20.5°, 11°, 11.5°. Rays incident in regions labeled 0, 1, 2, . . . are transmitted by the CPC after zero, one, two . . . reflections; F2, F3 . . . indicate that rays incident in those regions begin to turn back after two, three, . . . reflections. Rays in the blank regions will still be traveling toward the exit aperture after five reflections. The calculations were abandoned here to save computer time. In the computation of $T(\theta, \theta_{\max})$ where these rays were omitted for θ less than $\theta_{\max}$ it is most likely that all these rays are

e

f
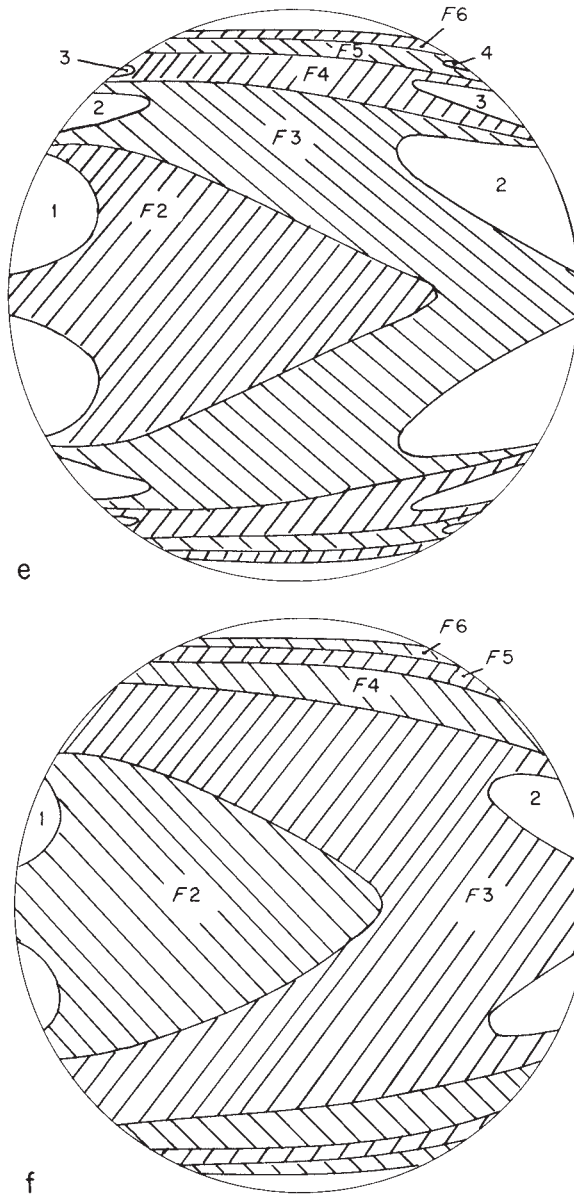
Figure 4.21   *Continued*

transmitted, as we shall show next, but for θ greater than the θ_{max}, this is prob-
ably not so. Thus, the transitions in the curves of Figure 4.19 are probably slightly
sharper than shown.

  The boundaries between regions in the diagrams of Figure 4.21 are, of course,
distorted images of the exit aperture seen after various numbers of reflections. It
can be seen that the failure regions—regions in which rays are turned back—
appear as a splitting between these boundary regions. For example, the regions
for failure after two and three reflections for 9° appear in the diagram as a split
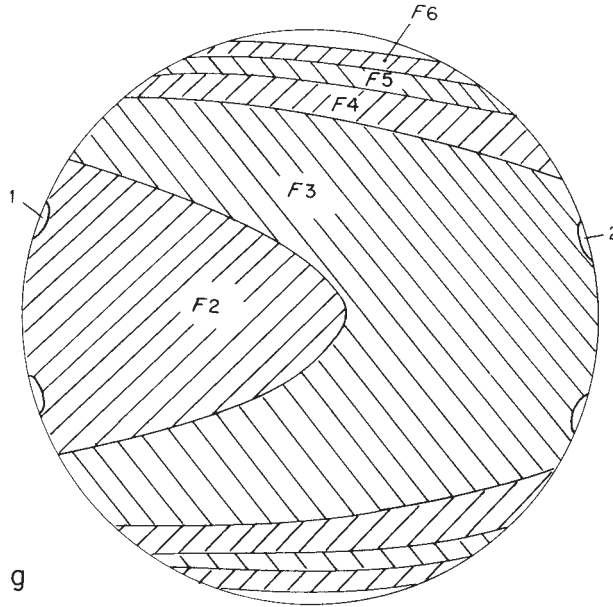
Figure 4.21   *Continued*
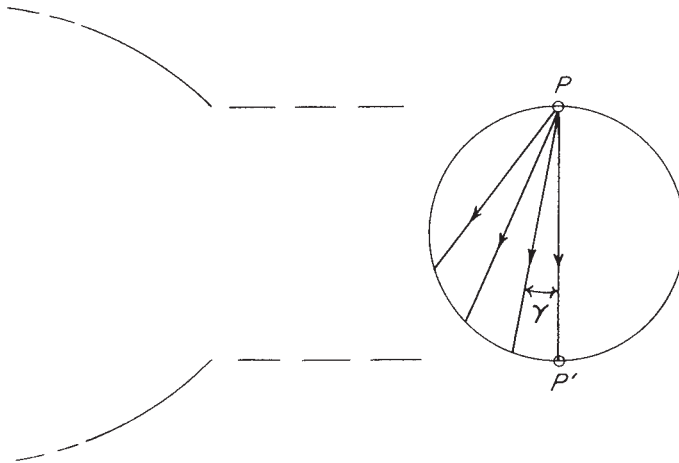
Figure 4.22   Rays at the exit aperture used to determine failure regions.

between the regions for transmission after one and two reflections. This confirms the principle stated in Section 4.3 that rays that meet the rim of the exit aperture are at the boundaries of failure regions. Naturally enough, each split between regions for transmission after n and $n + 1$ reflections produces two failure regions, for failure after $n + 1$ and $n + 2$ reflections.

We can delineate these regions in another way, by tracing rays in reverse from the exit aperture. Thus, in Figure 4.22 we can trace rays in the plane of the exit aperture from a point P at angles γ to the diameter P'. Each ray will eventually

emerge from the entry face at a certain angle $\theta(\gamma)$ to the axis and after n reflections. The point in the entry aperture from which this ray emerges is then the point in diagrams, such as those of Figure 4.21, at which the split between rays transmitted after $n - 1$ and n reflections begins. For example, to find the points A and B in the 9° diagram of Figure 4.21, we look for an angle γ that yields $\theta(\gamma) =$ 9° and find the coordinates of the ray emerging from the entry face after two reflections. There will, of course, be two such values of γ, corresponding to the two points A and B. This was verified by ray tracing.

Returning to the blank regions in Figure 4.21, the rays entering at these regions are almost tangential to the surface of the CPC. Thus, they will follow a spiral path down the CPC with many reflections, as indicated in Figure 4.23. We can use the skew invariant h explained in Section 2.8 to show that such rays must be transmitted if the incident angle is less than θ_{max}. For if we use the reversed rays and take a ray with $\gamma = \pi/2$ in Figure 4.22, this ray has $h = a'$. When it has spiraled back to the entry end (with an infinite number of reflections!), it must have the same $h = a' = a \sin \theta_{max}$—that is, it emerges tangent to the CPC surface at the maximum collecting angle. Any other ray in the blank regions closer to the axis or with smaller θ has a smaller skew invariant and is therefore transmitted.

The preceding argument holds for regions very close to the rim of the CPC. The reverse ray-tracing procedure shows that for angles below θ_{max} the failures begin well away from the blank region—in fact, at approximately half the radius at the entry aperture. Thus, we are justified in including the blank regions in the count of rays passed for $\theta < \theta_{max}$, as suggested above. This argument also shows that the transmission-angle curves (Figure 4.19) are precisely horizontal out to a few degrees below θ_{max}.

A converse argument shows, on the other hand, that rays incident in this region at angles above θ_{max} will not be transmitted. There seems to be no general argument to show whether the transmission goes precisely to zero at angles
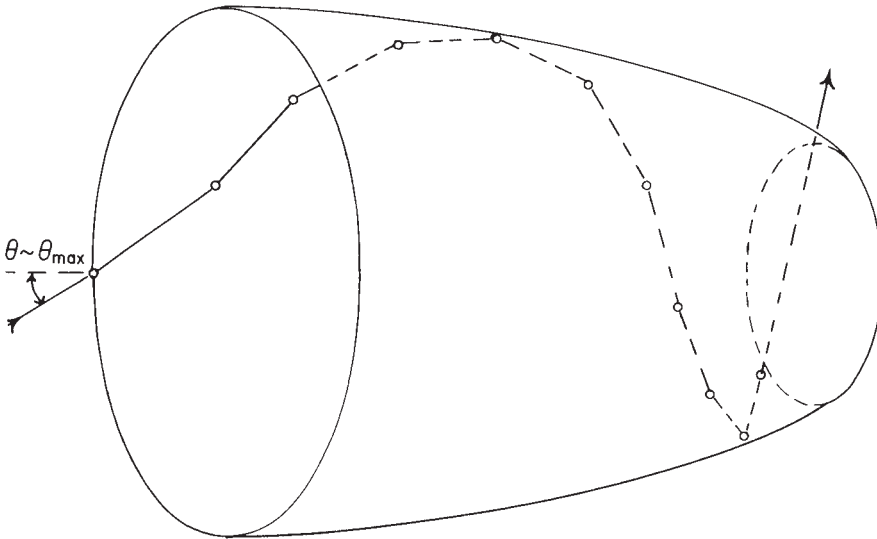
Figure 4.23   Path of a ray striking the surface of a CPC almost tangentially.

sufficiently greater than θ_{max}, but the ray tracing results suggest very strongly that it does.

4.8   CONES AND PARABOLOIDS AS CONCENTRATORS

Cones are much easier to manufacture than CPCs. Paraboloids of revolution (which of course CPCs are not) seem a more natural choice to conventional optical physicists as concentrators. We therefore provide some quantitative comparisons. It will appear from these that the CPC has very much greater efficiency as a concentrator than either of these other shapes.

In order to make a meaningful comparison, the concentration ratio as defined by the ratio of the entrance and exit aperture areas was made the same as for the CPC with $\theta_{max} = 10°$—that is, a ratio of 5.76 to 1 in diameter. The length of the cone was chosen so that the ray at θ_{max} was just cut off, as in Figure 4.24. For the paraboloid the exit aperture diameter and the concentration ratio completely determine the shape, as in Figure 4.25.
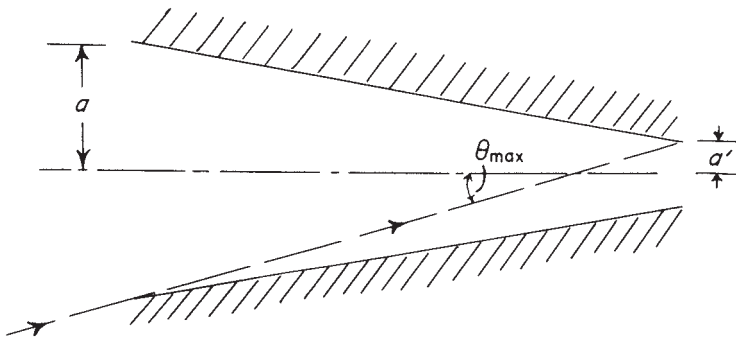
Figure 4.24   A cone concentrator, showing dimensions used to compare with a CPC.
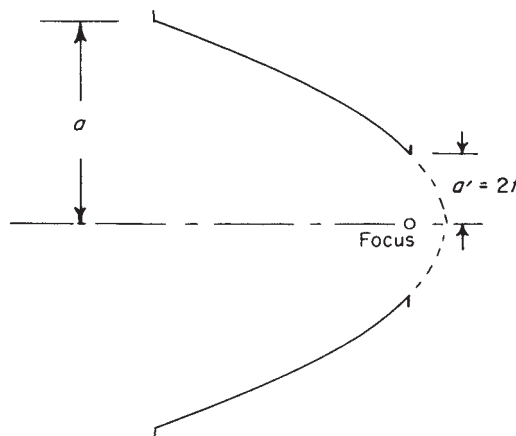
Figure 4.25   A paraboloid of revolution as a concentrator.
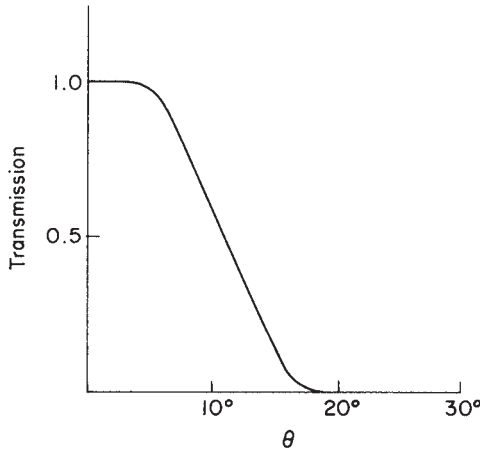
Figure 4.26   Transmission-angle curve for a cone; $\theta_{max} = 10°$.
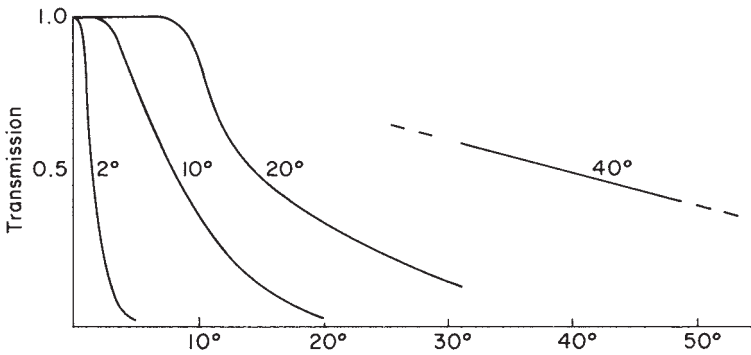
Figure 4.27   Transmission-angle curves for paraboloidal mirrors. The graphs are labeled with angles θ_{max} given by $\sin \theta_{max} = a'/a$ in Figure 4.25.

Figures 4.26 and 4.27 show the transmission-angle curves for cones and paraboloids, respectively. It is obvious that the characteristics of both these systems as concentrators are much worse than ideal. For example, the total transmission inside θ_{max} for the paraboloids, according to Eq. (4.13), is about 0.60 for all the angles shown. The cones clearly have definitely better characteristics than the paraboloids, with a total transmission inside θ_{max} of order 80%. This is perhaps a verification of our view that nonimaging systems can have better concentration than image-forming systems, since the paraboloid of revolution is an image-forming system, albeit with very large aberrations when used in the present way.[2]

Rabl (1976b) considered V-troughs—that is, 2D cones—and used a well-known construction (e.g., Williamson, 1952) shown in Figure 4.28 to estimate the angular

---

[2]   The section of a paraboloid of revolution in front of the focus is used for x-ray imaging, since most materials are good reflectors for x-rays at grazing incidence.
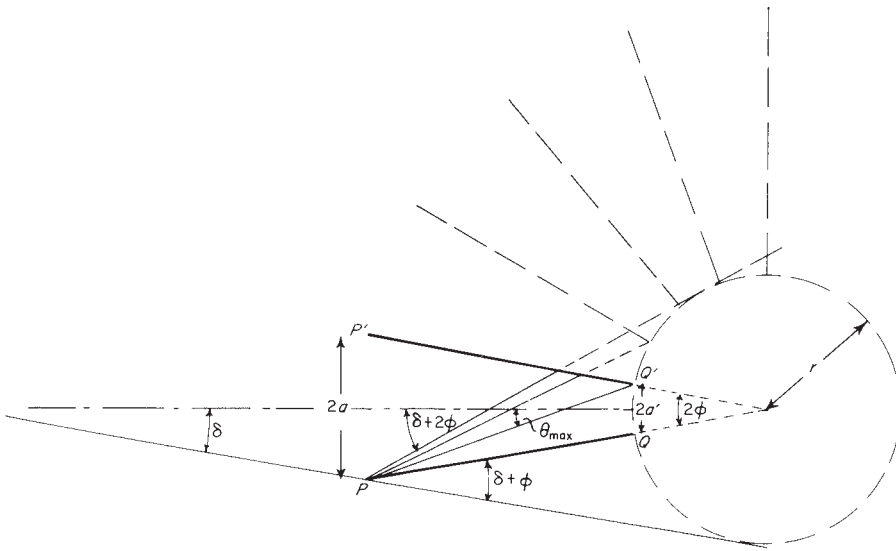
Figure 4.28   The pitch-circle construction for reflecting cones and V-troughs. A straight line through the entry aperture of the V-trough emerges from the exit aperture if it cuts the pitch circle. Otherwise, it is turned back. The construction is only valid for meridian rays of a cone.

width of the transition region in the transmission-angle curve. He showed that its width was equal to the angle 2ϕ of the V-trough and the center of the transition came at $\delta + \phi$, where δ is the largest angle of an incident pencil for which all rays are transmitted. If we assume the same holds for a 3D cone, it suggest that the transition in the transmission-angle curve becomes sharper as the cone angle decreases—in other words, smaller-angle cones are more nearly ideal concentrators. This accords with Garwin's result (Garwin, 1952), which may be said to imply that a very long cone with a very small angle is a nearly ideal concentrator.[3]

A different way to look at the performance of the cone is to note that for small ϕ the concentration ratio a/a' is approximately $1/\sin(\delta + \phi)$. Thus, as the cone length increases while a/a' is held fixed, ϕ/δ tends to zero, and from the diagram, the transition region in the transmission-angle diagram becomes sharper. Nevertheless, there is always a finite transition even for V-troughs and more so for cones so that the comparison with the CPC always shows that the cone is much less efficient and departs further from the ideal than the CPC.

Finally, Figure 4.29 shows the pattern of rays accepted and rejected by a 10° cone as seen at the entry aperture. This may be compared with Figure 4.21d, which shows the pattern for a 10° CPC.

---

[3]   Garwin showed that, in our terminology, a 3D concentrator can be designed to transform an area of any shape into any other while conserving étendue but that it is necessary to start the concentrator as a cylindrical surface and to change its shape adiabatically. In effect, this means that the concentrator would have to be infinitely long to achieve any concentration greater than unity!
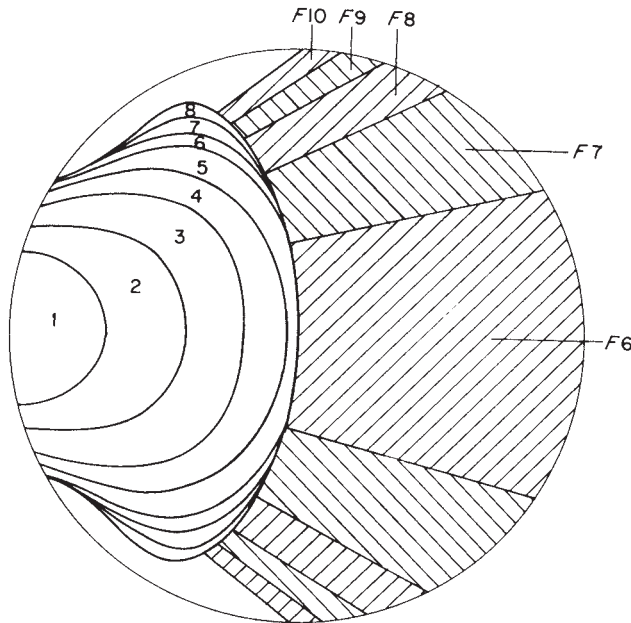
Figure 4.29   Patterns of accepted and rejected rays at the entry aperture of a 10° cone for rays at 10°. The notation is as for Figure 4.14, and this figure may be compared directly with Figure 4.14d.

REFERENCES

Baranov, V. K. (1965a). *Opt. Mekh. Prom*. **6**, 1–5.
  A paper in Russian that introduces certain properties of CPCs.
Baranov, V. K. (1966). *Geliotekhnika* **2**, 11–14 [*Eng transl*.: Parabolotoroidal mirrors as elements of solar energy concentrators. *Appl. Sol. Energy* **2**, 9–12.].
Baranov, V. K. (1967). "Device for Restricting in One Plane the Angular Aperture of a Pencil of Rays from a Light Source" (in Russian). Russian certificate of authorship 200530, specification published October 31, 1967.
  Describes certain illumination properties of the 2D CPC, called in other Russian publications a FOCLIN.
Baranov, V. K., and Melnikov, G. K. (1966). Study of the illumination characteristics of hollow focons. *Sov. J. Opt. Technol*. **33**, 408–411.
Brief description of the principle, with photographs to show illumination cut-off and transmission-angle curves plotted from measurements.
Baylor, D. A., and Fettiplace, R. (1975). Light and photon capture in turtle receptors. *J. Physiol. (London)* **248**, 433–464.
Harper, D. A., Hildebrand, R. H., Pernlic, R., and Platt, S. R. (1976). Heat trap: An optimized far infrared field optics system. *Appl. Opt*. **15**, 53–60.
Hinterberger, H., and Winston, R. (1966a). Efficient light coupler for threshold Čerenkov counters. *Rev. Sci. Instrum*. **37**, 1094–1095.
Hinterberger, H., and Winston, R. (1966b). Gas Čerenkov counter with optimized light-collecting efficiency. *Proc. Int. Conf. Instrum. High Energy Phys*. 205–206.

Holter, M. L., Nudelman, S., Suits, G. H., Wolfe, W. L., and Zissis, G. J. (1962). "Fundamentals of Infrared Technology." Macmillan, New York.

Levi-Setti, R., Park, D. A., and Winston, R. (1975). The corneal cones of *Limulus* as optimized light collectors. *Nature (London)* **253**, 115–116.

Ploke, M. (1967). Lichtführungseinrichtungen mit starker Konzentrationswirkung. *Optik* **25**, 31–43.

Ploke, M. (1969). "Axially Symmetrical Light Guide Arrangement." German patent application No. 14722679.

Rabl, A., and Winston, R. (1976). Ideal concentrators for finite sorces and restricted exit angles. *Appl. Opt.* **15**, 2880–2883.

Ries, H., and Rabl, A. (1994). Edge-ray principle of nonimaging optics, *J. Opt. Soc. Am. A* **11**, 2627–2632.

Weatherburn, C. E. (1931). "Differential Geometry of Three Dimensions." Cambridge Univ. Press, London.

Williamson, D. E. (1952). Cone channel condenser optics. *J. Opt. Soc. Am*. **42**, 712–715.

Winston, R. (1970). Light collection within the framework of geometrical optics. *J. Opt. Soc. Am.* **60**, 245–247.

Winston, R. (1974). Principles of solar concentrators of a novel design. *Sol. Energy* **16**, 89–95.

Winston, R. (1976a). Dielectric compound parabolic concentrators. *Appl. Opt*. **15**, 291–292.

Winston, R. (1976b). U.S. letters patent 3923 381, "Radiant Energy Concentration."

Winsion, R. (1977a). U.S. letters patent 4003 638, "Radiant Energy Concentration."

Winston, R. (1977b). U.S. letters patent 4002 499, "Cylindrical Concentrators for Solar Energy."

Winston, R., and Hinterberger, H. (1975). Principles of cylindrical concentrators for solar energy. *Sol. Energy* **17**, 255–258.

Witte, W. (1965). Cone channel optics. *Infrared Phys*. **5**, 179–185.

5

DEVELOPMENTS AND MODIFICATIONS OF THE COMPOUND PARABOLIC CONCENTRATOR

5.1   INTRODUCTION

There are several possible ways in which the basic CPC as described in Chapter 4 could be varied for specific purposes. Some of these have been mentioned already, such as a solid dielectric CPC using total internal reflection. Others spring to mind fairly readily for specific purposes, such as collecting from a source at a finite distance rather than at infinity. In this chapter we describe these developments and discuss their properties.

5.2   THE DIELECTRIC-FILLED CPC WITH TOTAL INTERNAL REFLECTION

Both 2D and 3D CPCs filled with dielectric and using total internal reflection were described by Winston (1976a). If we consider either the 2D case or meridian rays in the 3D case, we see that the minimum angle of incidence for rays inside the design collecting angle occurs at the rim of the exit aperture, as in Figure 5.1. If the dielectric has refractive index n, the CPC is, of course, designed with an acceptance angle θ' inside the dielectric, according to the law of refraction. It is then easy to show that the condition for total internal reflection to occur at all points is

$$\sin\theta_i' < 1 - (2/n^2) \quad or \quad \sin\theta_i < n - (2/n) \tag{5.1}$$

Since the sine function can only take values between 0 and 1, the useful values of n are greater than $\sqrt{2}$. This is in good agreement with the range of useful optical materials in the visible and infrared regions. In a 3D CPC, rays outside the meridian plane have a larger angle of incidence than meridian rays at the same inclination to the axis so Eq. (5.1) covers all CPCs. The expressions in Eq. (5.1) are plotted in Figure 5.2. For most purposes it is unlikely that collecting angles exceed-
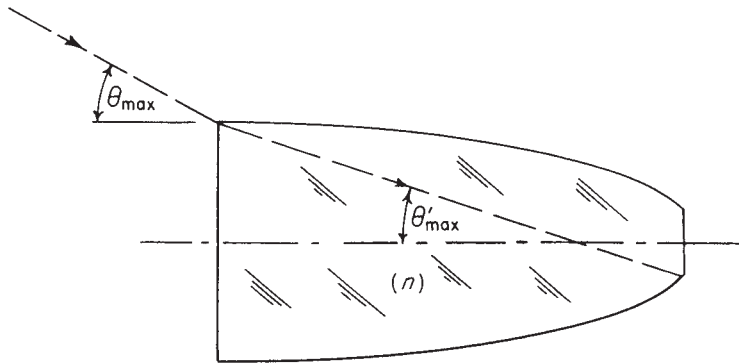
Figure 5.1   A dielectric-filled compound parabolic concentrator. The figure is drawn for an sentry angle of 18° and a refractive index of 1.5. The concentration ratio is thus 10.2 for a 3D concentrator.
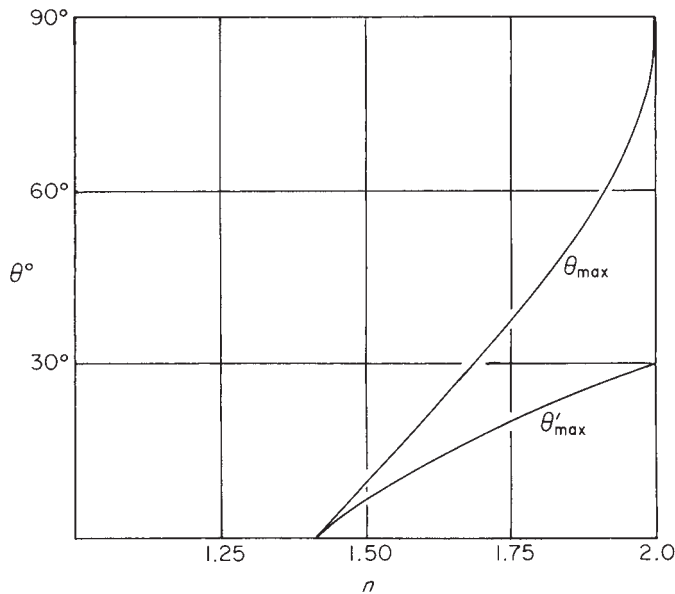
Figure 5.2   The maximum collecting angles for a dielectric-filled CPC with total internal reflection, as functions of the refractive index.

ing 40° would be needed so that the range of useful n coincides very well. For trough collectors there is always total internal reflection at the perpendicular end walls, since there is the same angle of incidence here as for the ray at θ' at the entry aperture on the curved surface. In fact, it could be shown that for $n > \sqrt{2}$ it is impossible for a ray to get into a trough CPC and not be totally internally reflected at the end face.

The angular acceptance of the dielectric-filled 2D CPC for nonmeridional rays is actually larger than a naive analogy with the $n = 1$ case would indicate. To see this, it is convenient to represent the angular acceptance by the direction cosine
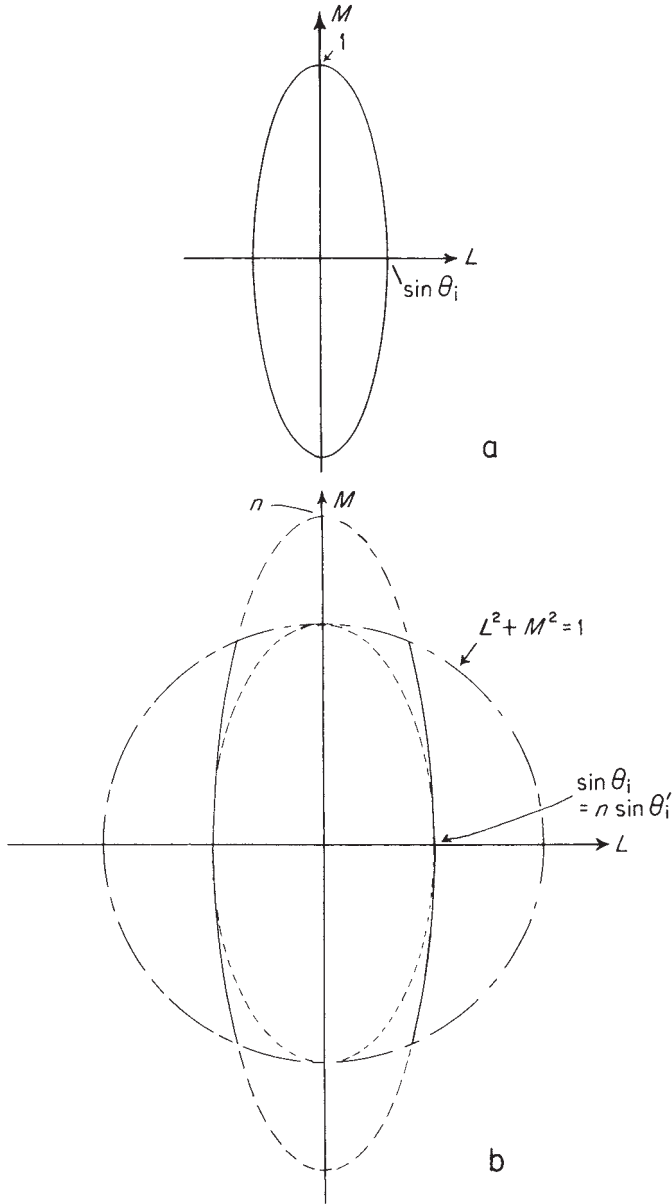
Figure 5.3 Angular acceptance for dielectric-filled concentrators, plotted in direction cosine space.

variables introduced earlier. Let x be transverse to the trough, let y lie along the trough, and let L and M be the corresponding direction cosines. We recall that the ordinary ($n = 1$) 2D CPC accepts all rays whose projected angles in the x, z plane are $\leq \theta_i$, the design cutoff angle. This condition is represented by an ellipse in the L, M plane with semidiameters $\sin \theta_i$ and 1, respectively (shown in Figure 5.3a). Therefore, the acceptance figure just inside the dielectric is such an ellipse with semidiameters $\sin \theta_i$ and a. In terms of direction cosines, Snell's law is simply

$$L = nL', \quad M = nM' \tag{5.2}$$

so one might expect the acceptance ellipse in the L, M plane to be scaled up by n. However, physical values of L and M lie inside the unit circle

$$L^2 + M^2 < 1 \tag{5.3}$$

It follows that the accepted rays lie inside the intersection of the scaled-up ellipse and the unit circle. This region, as seen in Figure 5.3b, is larger than an ellipse with semidiameters $\sin \theta_i'$ and 1, respectively. A quantitative measure of this enhancement is useful in discussing the acceptance of such systems for diffuse (i.e., totally isotropic) radiation. Diffuse radiation equipopulates phase space. Hence, it is uniformly distributed in the L, M plane. The area of the acceptance figure for an ordinary 2D CPC (corresponding to Figure 5.3a) is $\pi \sin \theta_i$. Therefore, the fraction of diffuse radiation accepted is just $\sin \theta_i$. However, for the dielectric-filled case, the area depicted by Figure 5.3b is found to be

$$2[n \sin \theta_i \tan^{-1}(\tan \theta_C \cos \theta_i) + \tan^{-1}(\tan \theta_i \cos \theta_C)] \tag{5.4}$$

where θ_C is the critical angle $\sin^{-1}(1/n)$. There exceeds $\pi \sin \theta_i$ by a factor

$$\frac{2}{n}[n \tan^{-1}(\tan \theta_C \cos \theta_i) + \tan^{-1}(\tan \theta_i \cos \theta_C)/\sin \theta_i] \tag{5.5}$$

This enhancement factor assumes the limiting value for small angles θ_i

$$\frac{2}{n}(n\theta_C + \cos \theta_C) \tag{5.6}$$

and slowly decreases to unity as θ_i increases to $\pi/2$. For example, for $n = 1.5$, a value typical for plastics, the enhancement is ≈ 1.17 for small θ_i ($\leq 10°$) and reduces to ≈ 1.13 for $\theta_i = 40°$. We shall return to this property in Chapter 13, where this extra angular acceptance is shown to be advantageous for solar energy collection.

The dielectric-filled CPC has certain practical advantages. Total internal reflection is 100% efficient, whereas it is difficult to get more than about 90% reflectivity from metallized surfaces. Also, for the same overall length the collecting angle in air is larger by the factor n, since Eqs. (4.1)–(4.4) for the shape of the CPC would be applied with the internal maximum angle θ_i', instead of θ_i. If the absorber can be placed into optical contact with the exit face and if it can utilize rays at all angles of incidence, the maximum theoretical concentration ratio becomes, from Eq. (2.13), $n^2/\sin^2\theta_i$—that is, it is increased by the factor n^2 (or n for a trough concentrator). However, if the rays had to emerge into air at the exit aperture, it is clear that many would get turned back at this face by total internal reflection. Thus, the CPC design needs to be modified for this case.

5.3   THE CPC WITH EXIT ANGLE LESS THAN $\pi/2$

There may well be instances such as that mentioned at the end of the last section where it is either impossible or inefficient to use rays emerging at up to $\pi/2$ from the normal to the exit aperture. The CPC designs can be easily modified to achieve this. It would then be close to being what we have called an ideal concentrator but without maximum theoretical concentration ratio.
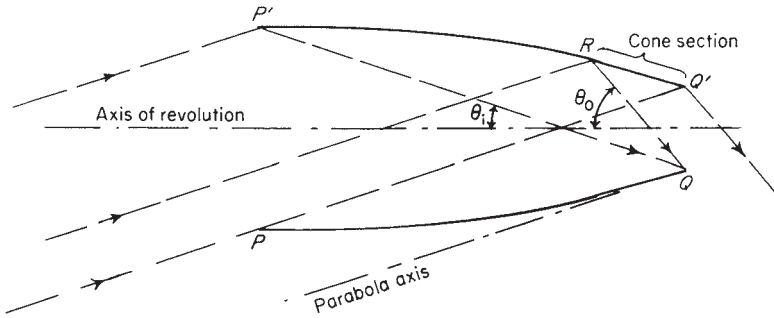
Figure 5.4   The $\theta_i\theta_0$ concentrator; as shown, $\theta_i = 18°$ and $\theta_0 = 50°$.

Let θ_i be the input collecting angle θ_0 and the maximum output angle. Then an ideal concentrator of this kind would, from Eq. (2.12), have the concentration ratio

$$C(\theta_i,\theta_0) = (n_0 \sin \theta_0)/(n_i \sin \theta_i) \tag{5.7}$$

for a 2D system or

$$C(\theta_i,\theta_0) = [(n_0 \sin \theta_0)/(n_i \sin \theta_i)]^2 \tag{5.8}$$

for a 3D system. Following Rabl and Winston (1976) we may call this device a θ_i/θ_0 transformer or concentrator. It is convenient to design the θ_i/θ_0 concentrator by starting at the exit aperture and tracing rays back. As for the basic CPC we start by considering the 2D case or the meridian rays for the 3D case. Let $QQ' = 2a'$ be the exit aperture in Figure 5.4. We make all reversed rays leaving any point on QQ' at angle θ_0, appear in the entrance aperture at angle θ_i to the axis. This is easily done by means of a cone section $Q'R$ making an angle $1/2(\theta_0 - \theta_i)$ with the axis. Next, we make all rays leaving Q at angles less than θ_0 appear at the entry aperture at angle θ_i; this is done in the same way as for a CPC by a parabola RP' with focus at Q and axis at angle θ_i to the concentrator axis. The parabola finishes as usual where it meets the extreme ray from Q at θ_i, so its surface is cylindrical at the entry end.

We have ensured by construction that in the meridian section all rays entering at θ_i emerge after one reflection at less than or equal to θ_0, and it is easily seen by examining a few special cases that all rays entering at angles less than θ_i emerge at less than θ_0. This must therefore be an ideal θ_i/θ_0 concentrator with

$$a/a' = \sin \theta_0 /\sin \theta_i \tag{5.9}$$

We can also prove this slightly more laboriously from the geometry of the design; this is done in Appendix H.

Clearly the 2D θ_i/θ_0 concentrator is an ideal concentrator, since the same reasoning as in Section 4.4 can be applied. In 3D form there will be some losses from skew rays inside θ_i being turned back. We show in Figure 5.5a the transmission-angle curve for a typical case. The transition is as sharp as for a full CPC, but when we compare Figure 5.5b with Figure 4.14, we see that the pattern of rejected rays is quite different.
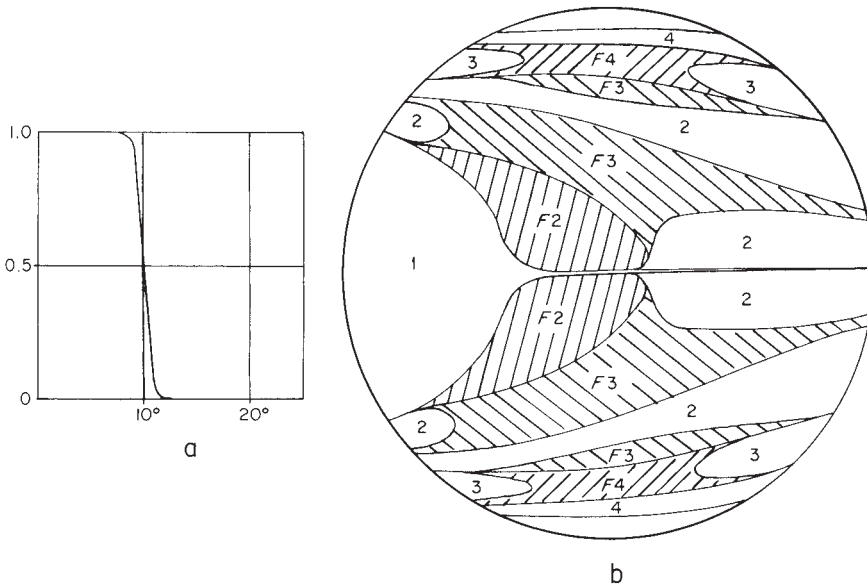
Figure 5.5   (a) The transmission-angle curve for a 10°/60° concentrator. (b) Rays transmitted and turned back in the 10°/60° concentrator for 10° input angle (see Figure 4.14 for full legend).

Although the 2D θ_i/θ_0 concentrator transmits within θ_0 all rays incident inside θ_i, it does not reject all rays incident outside θ_i but transmits some of them outside θ_0. The reason for this can be seen by noting that the straight cone section in Figure 5.4 lies outside the continuation of the parabolic section, and this parabolic section is the same as the section of an ordinary CPC with entry angle θ_i. Thus, some rays outside θ_i must be transmitted, and since the concentrator is ideal for rays inside θ_i, these extra rays must emerge outside θ_i. The same applies to the 3D version and the effect of these extra rays is not shown in Figure 5.5a.

5.4   THE CONCENTRATOR FOR A SOURCE AT A FINITE DISTANCE

So far we have assumed the source to be at infinity, as in the straightforward application to solar energy collection. There are, however, obviously cases where we should like to collect from a source at a finite distance, and the edge-ray principle enables us to derive the shape very simply (Winston, 1976c).

In Figure 5.6, let AA' be the finite source and let QQ' be the desired position of the absorber. Then if we apply the edge-ray principle, it is clear that the reflecting surface has the cross section of an ellipse, $P'Q$, with foci at A and Q. For a 3D system the complete surface would be obtained by rotating this ellipse about the axis of symmetry.

We can show that as a 2D system this has maximum theoretical concentration by noting that all rays from AA' that enter the concentrator do emerge (by the same reasoning as for the basic CPC), and the calculating the dimensions of
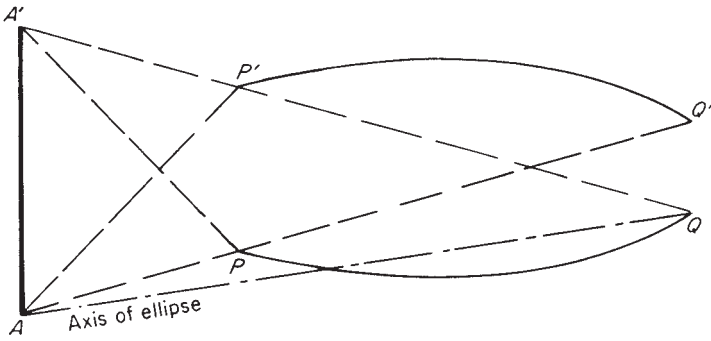
Figure 5.6   A concentrator for the object AA' at a finite distance.
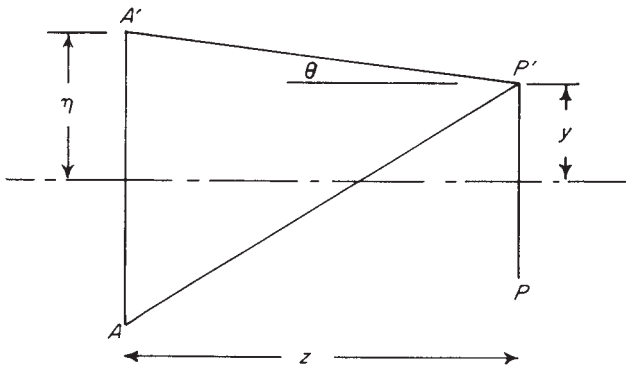
Figure 5.7   Calculating the étendue for a source AA' at a finite distance. The collecting aperture is at PP'.

the system by coordinate geometry. This approach is complicated, partly because of the geometry but also because it is not so easy to define the collecting angle for a source at a finite distance. It is better to use a more physical approach and calculate the étendue at either end of the system.

We take an object $AA' = 2\eta$ and an aperture PP' a distance z apart, as in Figure 5.7; if the coordinate y is measured from the center of the aperture, the étendue is

$$\iint d(\sin\theta)dy = \int_{-y_{max}}^{y_{max}} \left\{ \frac{\eta - y}{\sqrt{z^2 + (\eta - y)^2}} + \frac{\eta + y}{\sqrt{z^2 + (\eta - y)^2}} \right\} dy$$

$$= \left[-\sqrt{z^2 + (\eta - y)^2} + \sqrt{z^2 + (\eta + y)^2} \right]_{-y_{min}}^{y_{max}} \qquad (5.10)$$

$$= 2(AP' - AP)$$

It is useful to rewrite Eq. (5.10) in the form

$$(A'P + AP') - (A'P' + AP) \qquad (5.11)$$

because this version, a remarkably simple result due to Hottel (1954), actually gives the correct étendue formula even when there is no particular symmetry between the relative positions of aperture and source.

Now, returning to Figure 5.6, we have from the fundamental property of the ellipse that the sum of the distances from the two foci top any point on the curve is a constant—that is,

$$AP + PQ' + Q'Q = AP' + PQ;$$

or

$$AP' - AP = Q'Q \qquad (5.12)$$

so that the étendue measured at the output end is $2Q'Q$. Since $Q'Q$ is perpendicular to the axis, this must mean that rays emerge from all points of the exit aperture with their direction cosines distributed uniformly over ± 1—that is, this system has the maximum theoretical concentration ratio.

For the 3D case with rotational symmetry (Figure 5.6 now represents a meridional section) a straightforward calculation gives for the étendue (Winston, 1978)

$$\pi^2 (AP' - AP)^2 \big/ 4 \qquad (5.13)$$

while the maximum value assumed by the skew invariant is

$$h_{\max} = (AP' - AP)/2 \qquad (5.14)$$

Notice that, just as for the case of an infinitely distant source ($\theta_i = $ constant), both Eqs. (5.13) and (5.14) are consistent with maximum concentration onto an exit aperture of the diameter given by Eq. (5.12). Nevertheless, this system in 3D will turn back some rays, just as for the basic CPC, and so will not be quite ideal.

5.5   THE TWO-STAGE CPC

In Section 5.2 we discussed the dielectric-filled CPC using total internal reflection, and we noted that the use of a refractive index greater than unity at the exit aperture permits in principle a greater concentration ratio. In order to utilize this the absorber must be in optical contact with the dielectric, and preferably also the interface must be matched to minimize reflection losses by a suitable coating or a grading of the refractive index. This seems possible, but we then have the practical difficulty that CPCs are very long and therefore a large volume of possibly expensive dielectric is needed. This difficulty may be circumvented by a two-stage system like the one in Figure 5.8. The first stage in air is a θ_i/θ_0 concentrator to give whatever final output angle is needed.
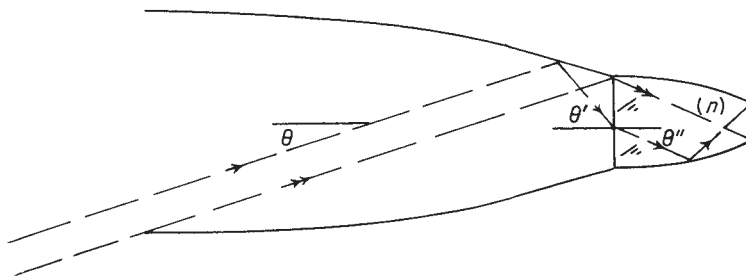
Figure 5.8   Two-stage system ending in a dielectric of index n. In the diagram, $\theta = 10°$, $\theta' = 60°$, $\theta'' = 24.5°$, and $n = 1.85$.
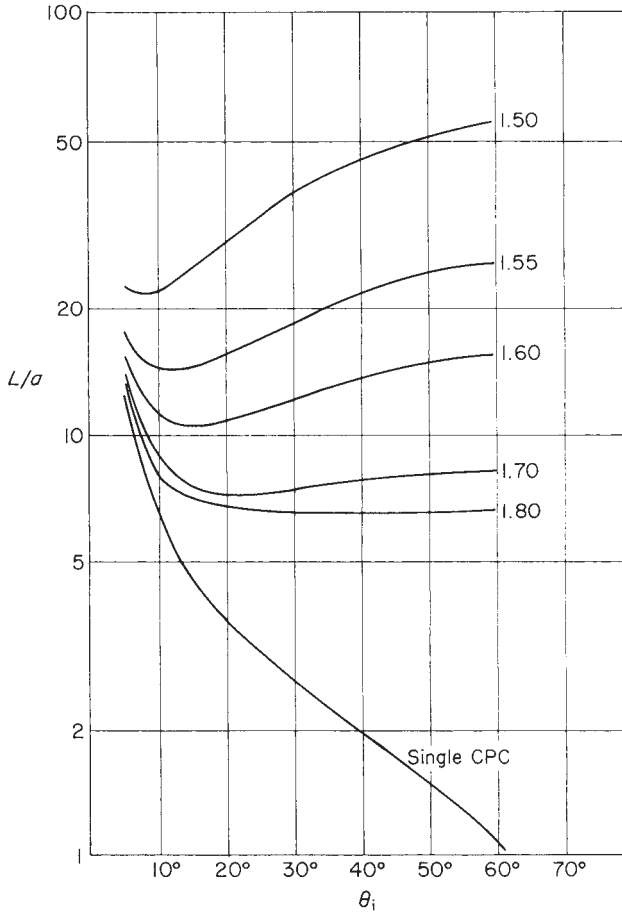
Figure 5.9 A comparison of the overall lengths of single- and two-stage concentrators according to Eqs. (5.15) and (5.16). The ordinate is the ratio of length to radius of collecting aperture, and the graphs are labeled with the refractive index of the second stage.

This two-stage system, if designed to give maximum concentration $n/\sin\theta$, where θ is the collecting angle in air, is always longer than a basic CPC designed for concentration $1/\sin\theta$ but with the same collecting aperture. If we make the dielectric part as short as possible—that is, if we maximize θ'' according to Eq. (5.1)—then it can be shown that the overall length of the two-stage system is

$$a\left\{\cot\theta + \frac{n}{n^2-2}\cos\theta + \frac{4(n^2-1)^{3/2}}{n(n^2-2)^2}\sin\theta\right\} \quad for \quad \sqrt{2} < n \le 2$$

$$a\left\{\cot\theta + \cos\theta + \frac{3}{2}\sqrt{3}\sin\theta\right\} \qquad for \quad n > 2 \qquad (5.15)$$

whereas the basic CPC has length

$$a(\cot\theta + \cos\theta) \qquad (5.16)$$

It is easily seen that (5.15) is greater than (5.16). Figure 5.9 shows some typical values for comparison.
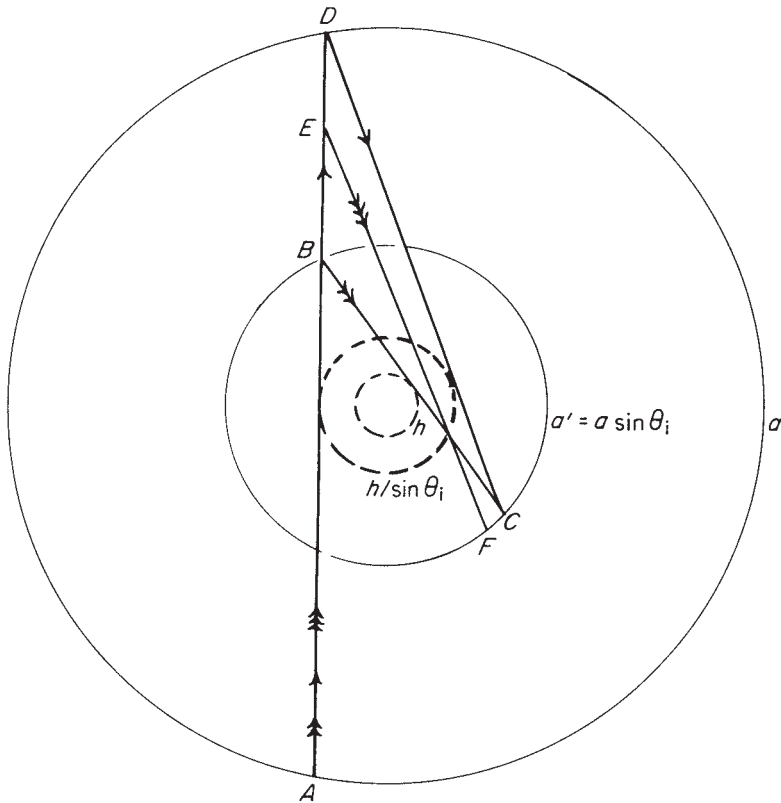
Figure 5.10   Design of a concentrator by applying the edge-ray principle to rays with nonzero skew invariant.

5.6   THE CPC DESIGNED FOR SKEW RAYS

As we saw in Chapter 4, the basic 3D CPC turns back some skew rays, and this makes its concentration efficiency slightly less than ideal, as shown in Figure 4.15. The basic CPC is designed by applying the edge-ray principle to meridian rays, and this uses up all the degrees of freedom available, so it is not surprising that some skew rays fail. This suggests that we should try designing a concentrator by applying the edge-ray principle to skew rays. The result might then have a slightly closer approach to maximum theoretical concentration ratio.

Figure 5.10 shows a view from the entry aperture of a concentrator. The entry aperture has diameter $2a$ and the exit aperture $2a' = 2a \sin \theta_i$. Let this system be designed to fulfill the edge-ray principle for skew rays with skew invariant h (see Section 2.8 and Appendix C for the skew invariant). In this projected view a ray that enters at θ_i and that grazes the entry aperture at A and the exit aperture at B is reflected across to C on the exit aperture (double arrows). The segment BC is thus tangent to a circle of radius $h/\sin \theta_i$. Another ray AD, also entering at θ_i, meets the opposite entry edge at D and is reflected there. In order to satisfy the edge-ray principle it must meet the edge of the exit aperture and therefore is like the
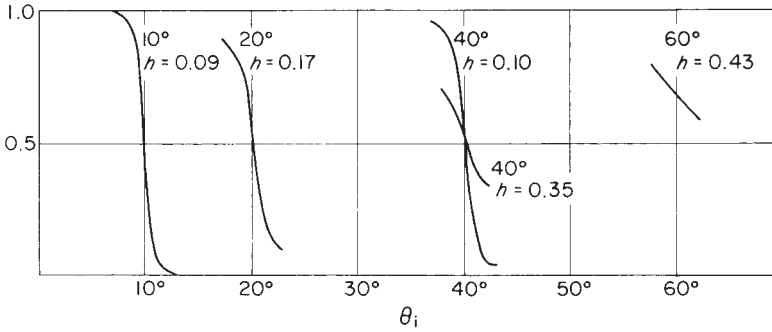
Figure 5.11   Transmission-angle curves for concentrators designed for nonzero skew invariant h. All the concentrators have exit apertures of diameter unity.

double-arrowed segment AB rotated around the axis so that A reaches D. Thus, the segment reflected from D touches the circle of radius $h/\sin\theta_i$. We can show that this segment meets the exit aperture at the same point C as the segment BC reflected at B by calculating the angles subtended at the center by the various ray segments; the proof is given in Appendix C.

The design is completed by requiring that all rays that leave the exit aperture at its rim shall have entered at the input angle θ_i. These rays in the projection of Figure 5.10 are typified by the three-arrowed segments—that is, reflection at some point E along the concentrator to emerge at F on the exit rim. It turns out that the rays do not in general emerge at the same point on the exit rim. Thus, we have a situation where the edge-ray principle is less restrictive than, say, a requirement for imaging at the exit rim. This process completely determines the concentrator as a surface of revolution, but it does not seem to be possible to represent it by any analytical expression. We show in Appendix C how the differential equation of the surface, which is first-order nonlinear, is obtained. But it can only be solved numerically. P. Greenman computed the solution for several values of the input angle θ_i and skew invariant h. The results are, briefly, that the shapes are very similar to those of the basic CPC for the same θ_i but that the overall lengths are less and the transitions in the transmission-angle curves are correspondingly more gradual. Figure 5.11 shows some of these curves.

The overall length of this concentrator is determined, as for the basic CPC, by the extreme rays, as in Figure 5.12. This figure shows the rays aDC and aBC of Figure 5.10, both inclined at the extreme input angle to the axis and both grazing the exit aperture after one and no reflections from the concentrator, respectively. Then in order to admit all rays at θ_i or less, the concentrator surface must finish at the point A determined by the intersection of the ray ABC with the surface. This geometry is obvious for the basic CPC, and also in that system the ratio of input to output diameters is set as part of the design data at the desired value $1/\sin\theta_i$. In the present system the design is developed from one end, and it does not follow that the ratio of input to output diameters will have any particular simple value. In fact, it can be shown (Appendix C) that this ratio is again $1/\sin\theta_i$—a result that is by no means obvious. In spite of this, the concentrator with nonzero h has even less than the maximum theoretical concentration ratio than the basic CPC, as can be seen by comparing Figures 5.11 and 4.19. This is
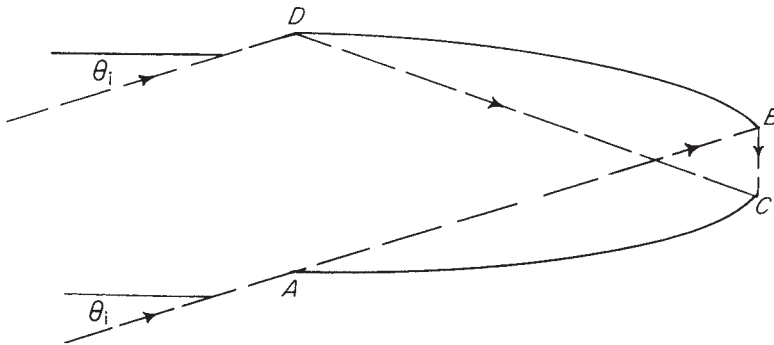
Figure 5.12   How the length of a concentrator is determined by the extreme rays.
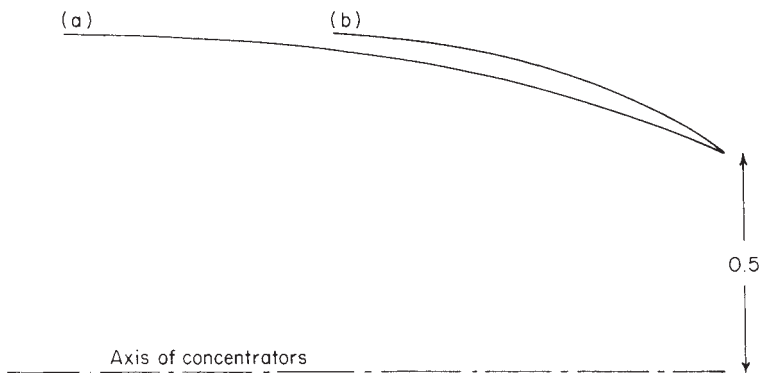
Figure 5.13   Comparison of concentrator profiles for $\theta_i = 40°$; (a) $h = 0$; (b) $h = 0.32$. It can be seen that meridian rays from the edge of (b) at angles greater than 40° are transmitted, since rays at 40° from the edge of (a) just get through the exit aperture.

because the shorter length permits meridian rays at angle greater than θ_i to reach the exit aperture directly. Thus, by volume conservation of phase space (Appendix A), more rays inside θ_i must be rejected.

Figure 5.13 shows a scale-drawing comparison of the basic 40° CPC with concentrators designed for nonzero h. It can be seen how meridian rays at angles greater than θ_i reach the exit aperture.

In Appendix C we show that in an ideal concentrator most transmitted rays have small values of the skew invariant h. In fact, we calculate the relative frequencies of occurrence of h and show that the greatest frequency is at $h = 0$. This tends to support our finding that the solution for the concentrator design with $h = 0$ is best.

5.7   THE TRUNCATED CPC

A disadvantage of the CPC compared to systems with smaller concentration is that it is very long compared to the diameter of the collecting aperture (or width for 2D systems). This is naturally important for economic reasons in large-scale appli-
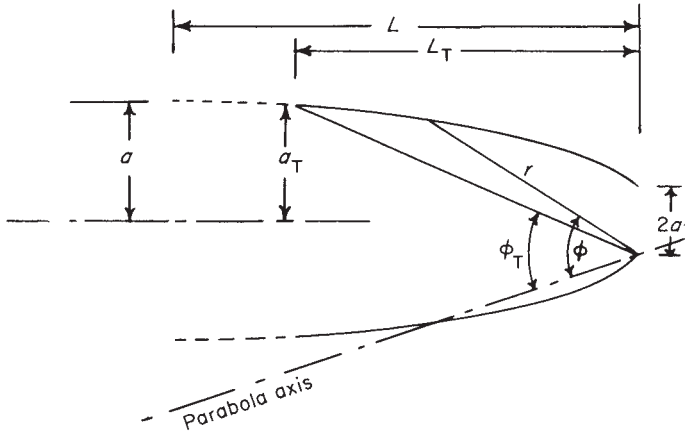
Figure 5.14   The polar coordinates used in computing truncation effects.

cations such as solar energy. From Eq. (4.2) the length L is approximately equal to the diameter of the collecting aperture divided by the full collecting angle—that is,

$$L \sim (2a/2\theta_i) \qquad (5.17)$$

If we truncate the CPC by removing part of the entrance aperture end, we find that a considerable reduction in length can be achieved with very little reduction in concentration, so this may be a useful economy.

It is convenient to express the desired relationships in terms of the (r, ϕ) polar coordinate system as in Figure 5.14. We denote truncated quantities by a subscript T. We are interested in the ratio of the length to the collecting apertures and also in the ratio of the area of the reflector to that of the collecting aperture. We find

$$a_T = \frac{f \sin(\phi_T - \theta_i)}{\sin^2(1/2)\phi_T} - a' \quad f = a'(1+\sin\theta_i)$$

$$a \frac{a'}{\sin\theta_i} \qquad (5.18)$$

$$L_T = \frac{f \cos(\phi_T - \theta_i)}{\sin^2(1/2)\phi_T} \qquad (5.19)$$

$$L = f \cos\theta_i / \sin^2\theta_i \qquad (5.20)$$

so that

$$\frac{L_T}{a_T} = \frac{(1+\sin\theta_i)\cos(\phi_T - \theta_i)}{\sin(\phi_T - \theta_i)(1+\sin\theta_i) - \sin^2\frac{1}{2}\phi_T} \qquad (5.21)$$

Plots of this quantity against the theoretical concentration ratio a_T/a' for 2D truncated CPCs were given by Rabl (1976c) and by Winston and Hinterberger (1975). Figure 5.15 shows some of these curves, and it can be seen that initially— for points near the broken line locus for full CPCs—the curves have a very large slope, so the loss in concentration ratio is quite small for useful truncations.
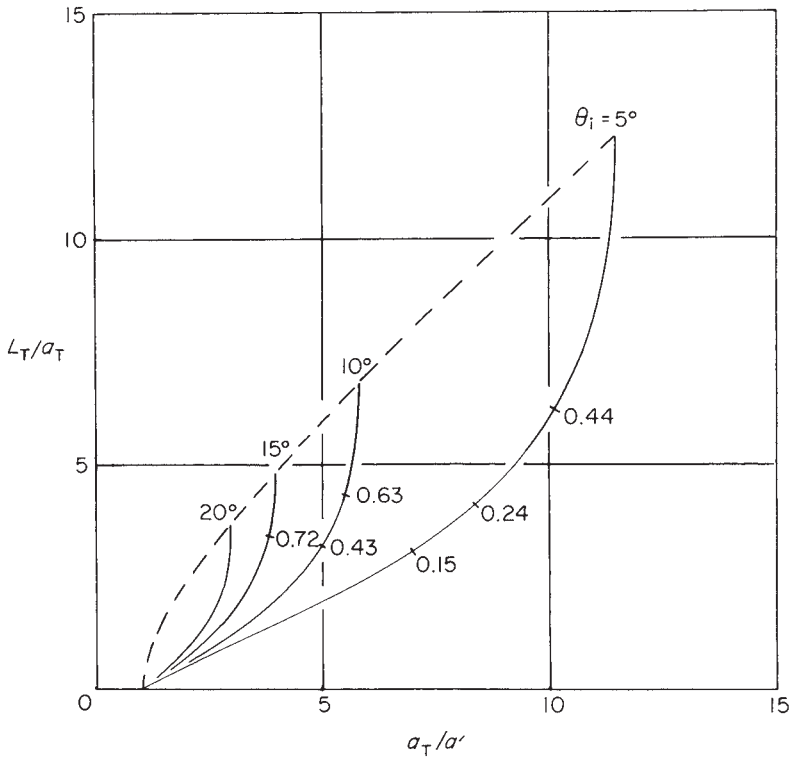
Figure 5.15   Length as a function of concentration ratio for 2D truncated concentrators. The numbers marked on the curves are the actual truncation ratios—that is, L_T/L.

In addition to the ratio of length to aperture diameter we may be interested in the ratio of surface area of reflector to aperture area, since this governs the cost of material for the reflector. The general forms of the curves would be similar to those in Figure 5.15 but with differences between 2D and 3D concentrators. The explicit formulae for reflector area divided by collector area are, for a 2D truncated CPC,

$$\frac{-f}{a_T}\left\{\frac{\cos\dfrac{\phi}{2}}{\sin^2\dfrac{\phi}{2}}-\ln\left(\tan\frac{\phi}{4}\right)\right\}\Bigg|_{\phi_T}^{\theta_i+\pi/2} \tag{5.22}$$

and for a 3D truncated concentrator

$$\frac{2f}{a_T^2}\int_{\phi_T}^{\theta_i+\pi/2}\left\{\frac{f\,\sin(\phi-\theta_i)}{\sin^5\dfrac{\phi}{2}}-\frac{a'}{\sin^3\dfrac{\phi}{2}}\right\}d\phi \tag{5.23}$$

Derivations of the preceding results and the explicit form of the integral in Eq. (5.23) are given in Appendix I. Some representative plots of these functions are given in Figures 5.16 and 5.17. It should be noted that in these figures the theoretical concentration ratios are respectively (a_T/a') and $(a_T/a')^2$. We conclude from
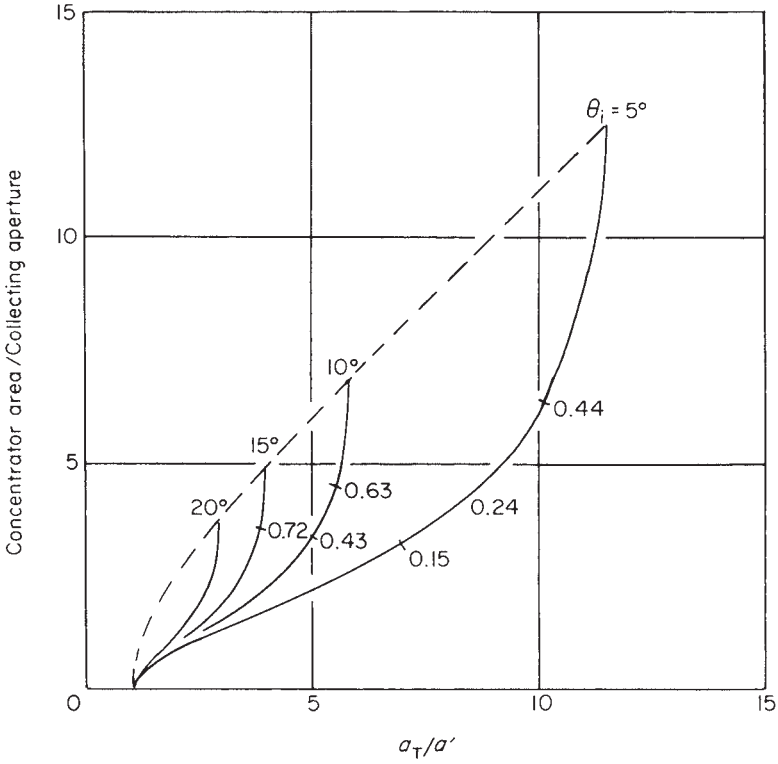
Figure 5.16   Concentrator surface area as a function of concentration ratio for 2D truncated concentrators. The numbers marked on the curves are the actual truncation ratios—that is, L_T/L.
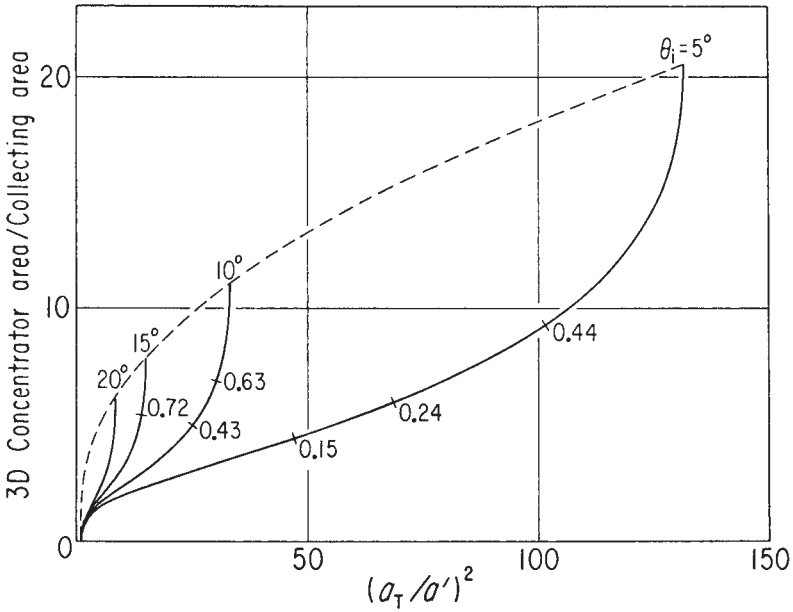
Figure 5.17   Concentrator surface area as a function of concentration ratio for 3D truncated concentrators. The numbers marked on the curves are the actual truncation ratios—that is, L_T/L.
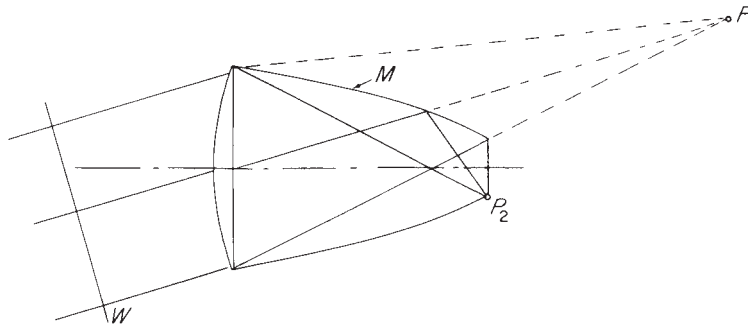
Figure 5.18   The lens-mirror CPC.

this that losses in performance due to moderate truncation would be acceptable in many instances on account of the economic gains.

5.8   THE LENS-MIRROR CPC

A more fundamental method for overcoming the disadvantage of excessive lengths incorporates refractive elements to converge the pencil of extreme rays. By consistent application of the edge-ray principle we leave the optical properties of the concentrator essentially identical to the all-reflecting counterpart while substantially reducing the length in many cases. The edge-ray principle requires that the extreme incident rays at the entrance aperture also be the extreme rays at the exit aperture. In the all-reflecting construction (Figure 4.10) this is accomplished by a parabolic mirror section that focuses the pencil of extreme rays from the wave front W onto the point P_2 on the edge of the exit aperture. To incorporate, say, a lens at the entrance aperture, the rays from W, after passage through the lens, are focused onto P_2 by an appropriately shaped mirror M (Figure 5.18). Therefore, the profile curve of M is then determined by the condition

$$\int_W^{P_2} n\,ds = \text{const.} \tag{5.24}$$

To comprehend the properties of the lens-mirror collector, it is useful to consider a hypothetical lens that focuses rays from P onto a point F. From Eq. (5.24) the appropriate profile curve for M is a hyperbola with conjugate foci at F and P_2 (Figure 5.18).[1] This example illustrates the principal advantage of this configuration. The overall length is greatly reduced from the all-reflecting case to $L = f$, the focal length of the lens. A real lens would have chromatic aberration, so M would no longer be hyperbolic but simply a solution to Eq. (5.24). A solution will be possible so long as the aberrations are not so severe as to form a caustic between the lens and the mirror. For the example in Figure 5.18, where the lens is plano-convex with index of refraction $n \sim 1.5$, this means we must not choose too small a value for the focal ratio of this simple lens (an $f/4$ choice works out nicely). Alternatively, we may say that the mirror surface corrects for lens aberrations, providing these

---

[1]   By suitable choice of parameters, the hyperbola can be a straight line.

are not too severe, to produce a sharp focus at P_2 for the extreme rays. Of course, this procedure can only be successful for monochromatic aberrations so that it is advantageous to employ a lens material of low dispersion over the wavelength interval of interest.

We may expect the response to skew rays in a rotationally symmetric 3D system to be nonideal just as in the all-reflecting case, and, in fact, ray tracing of some sample lens-mirror configurations shows angular cutoff characteristics indistinguishable from the simple CPC counterpart. We note that certain configurations of the lens-mirror type were proposed by Ploke (1967).

5.9   2D COLLECTION IN GENERAL

For moderate concentration ratios for solar energy collection, there is considerable interest in systems that do not need diurnal guiding for obvious reasons of economy and simplicity (see, e.g., Winston, 1974; Winston and Hinterberger, 1975). These naturally would have troughlike or 2D shapes and would be set pointing south[2] at a suitable elevation so as to collect flux efficiently over a good proportion of the daylight hours. So far our discussion has suggested that these might take the form of 2D CPCs, truncated CPCs, or compound systems with a dielectric-filled CPC as the second stage. In discussing all these it was tacitly assumed that the absorber would present a plane surface to the concentrator at the exit aperture, and this, of course, made the geometry particularly simple. In fact, when applications are considered in detail, it becomes apparent that other shapes of absorber would be useful. In particular, it is obvious that cylindrical absorbers—that is, tubes for heating fluids—suggest themselves. In this chapter we discuss the developments in design necessary to take account of such requirements.

5.10   EXTENSION OF THE EDGE-RAY PRINCIPLE

In Chapter 3 we proposed the edge-ray principle as a way of initiating the design of concentrators with concentration ratios approaching the maximum theoretical value. We found that for the 2D CPC this maximum theoretical value was actually attained by direct application of the principle.

We now propose a way of generalizing the principle to nonplane absorbers in 2D concentrators. Let the concentrator be as in Figure 5.19, which shows a generalized tubular absorber. We assume the section of the absorber is convex everywhere, and we also assume it is symmetric about the horizontal axis indicated. Then we assert that the required generalization of the edge-ray principle is that rays entering at the maximum angle θ_i shall be tangent to the absorber surface after one reflection, as indicated.

The generalization can easily be seen to reduce to the edge-ray principle for a plane absorber. In order to calculate the concentration we need to have a rule for constructing the concentrator surface beyond the point P'_0, at which the extreme reflected ray meets the surface. Here we choose to continue the reflector as an involute of the absorber surface, as indicated by the broken line. A reason for this

---
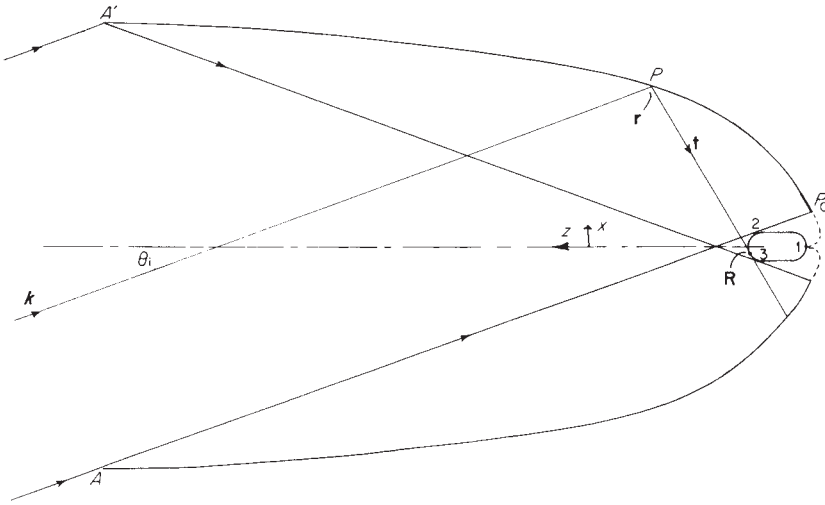
Figure 5.19   Generalizing the edge-ray principle for a nonplane absorber.

choice will be suggested following. We shall be able to show that this design for a 2D concentrator achieves the maximum possible concentration ratio, defined in this case as the entry aperture area divided by the area of the curved absorber surface.

Following Winston and Hinterberger (1975) we let r be the position vector of a current point P on the concentrator surface and take R as the position vector of the point of contact of the ray with the absorber. Then we have

$$r = R - lt \tag{5.25}$$

where t is the unit tangent to the absorber—that is, we have

$$t = dR/dS \tag{5.26}$$

where S is the arc length round the absorber. Let k be a unit vector along the direction of the extreme rays so that in the coordinate system shown $k = (\sin \theta_i, 0, -\cos \theta_i)$. Our condition that the tangent to the absorber be reflected into k takes the form, equating sines of the angles of incidence and reflection,

$$t \cdot dr = k \cdot dr$$

or

$$t \cdot dr/dS = k \cdot dr/dS \tag{5.27}$$

Now by differentiating Eq. (5.25) we obtain

$$dr/dS = dR/dS - dl/dSt - ldt/dS$$

and on scalar multiplication by t this gives

$$t \cdot dr/dS = 1 - dl/dS \tag{5.28}$$

On substituting in Eq. (5.27) and integrating we obtain

$$(S - l)|_2^3 = (r_3 - r_2) \cdot k \tag{5.29}$$

In this equation the points 2 and 3 would be those corresponding to the extreme reflected rays, as in the diagram.

Between points 1 and 2 we have postulated that the concentrator profile shall be an involute of the absorber, and the condition for this is

$$\boldsymbol{t} \cdot d\boldsymbol{r}/dS = 0 \tag{5.30}$$

Thus, for this section of the curve we have from Eq. (5.28)

$$S_2 - S_1 = l_2 - l_1$$

or, since our involute is chosen to be the one that starts at point 1

$$S_2 = l_2 \tag{5.31}$$

Thus, Eq. (5.29) gives

$$S_3 - l_3 = (\boldsymbol{r}_3 - \boldsymbol{r}_2) \cdot \boldsymbol{k} \tag{5.32}$$

From the figure it can be seen that $(r_3 - r_2) \cdot k$ is equal to the projection of $P_0'A'$ onto AP_0'. Thus

$$(\boldsymbol{r}_3 - \boldsymbol{r}_2) \cdot \boldsymbol{k} = -(l_3 + l_2) + 2a \sin \theta_i$$

and on substituting into Eq. (5.32) we find

$$S_3 + l_2 = 2a \sin \theta_i \tag{5.33}$$

Recalling that the second section of the concentrator is an involute, we see that $l_2 = S_2$. Thus,

$$S = S_3 + S_2 = 2a \sin \theta_i \tag{5.34}$$

We have proved that the concentrator profile generated in this way has the theoretical ratio of input area to absorber area—that is, it has the maximum theoretical concentration ratio if no rays are turned back.

If the property of the involute that its normal is tangent to the parent curve is remembered, it is easy to see that a concentrator designed in this way sends all rays inside the angle θ_i to the absorber, including those outside the plane of the diagram if it is a 2D system. Thus, from arguments based on étendue and on phase space conservation (see Section 2.7 and Appendix A) the system is optimal. For a further generalization of the the edge ray principle see Appendix B.

5.11   SOME EXAMPLES

It is easy to apply our generalization to plane absorbers. Figure 5.20 shows an edge-on fin absorber QQ' with extreme rays AQP_0' and $A'QP_0$. Clearly the section of the concentrator between P_0 and P_0' is an arc of a circle centered on Q and the section $A'P_0'$ is a parabola with focus at Q and axis AQP_0'.

The two-sided flat plate collector normal to the axis, as in Figure 5.21, is a slightly more complicated case. Following our rules, there are three sections to the profile. OP' receives no direct illumination and is thus an involute of the segment OQ'; that is, it is an arc of a circle centered on Q' and therefore part of a parabola with focus at Q and axis $AQ'P'$. $R'A'$ must focus extreme rays on Q and is therefore a parabola with focus Q and axis parallel to $AQ'P'$.
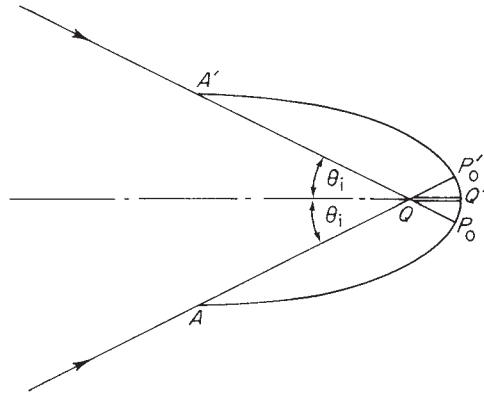
Figure 5.20   The optimum concentrator design for an edge-on fin.
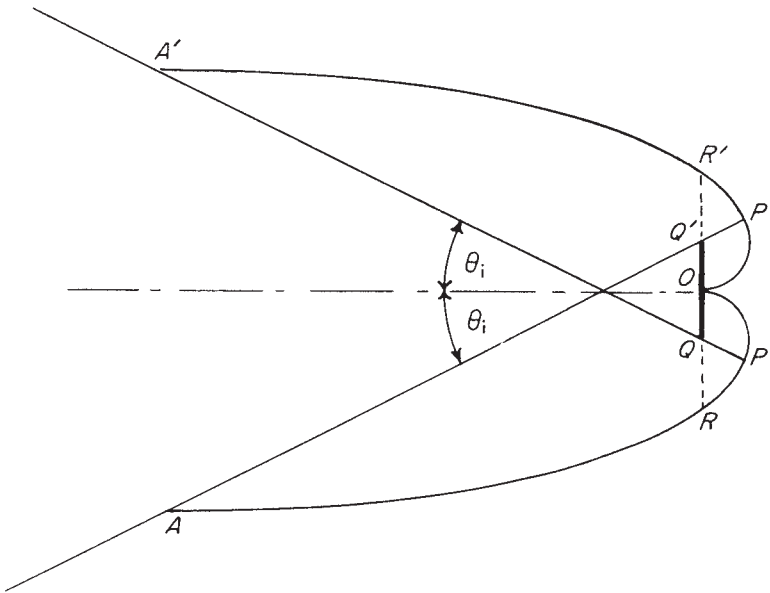
Figure 5.21   The optimum concentrator design for transverse fin.

In all cases it can easily be seen from the general mode of construction described in Section 5.10 that the segments of different curves have the same slope where they join; for example, in Figure 5.21 the normal at P' is a ray for the circular segment OP' and the parabolic segment $P'R'$, and at R' the incident ray at angle θ_i is required to be reflected to Q' by the segment $P'R'$ and to Q by the segment $R'A'$.

Figure 5.22 shows to scale the profile for a circular section absorber. Here the actual profile does not have a simple parabolic or circular shape, but we shall give the solution in Section 5.12. It is noteworthy, however, that in Section 5.10 we deduced the property of having maximum theoretical concentration ratio without explicit reference to the profile, just as we were able to do for the basic CPC (see
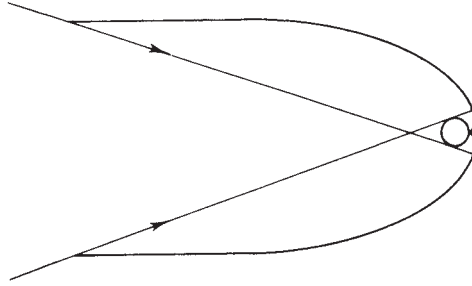
Figure 5.22   The optimum concentrator design for a cylindrical absorber.

Chapter 4). The case of the circular section absorber is important for solar energy application (Chapter 10) and has been actively pursued by a number of investigators. For example, Ortobasi (1974) independently developed the ideal mirror profile in an innovative collector program at Corning Glass Company.

5.12   THE DIFFERENTIAL EQUATION FOR THE CONCENTRATOR PROFILE

It is straightforward but laborious to set up a differential equation for the concentrator profile. The equation, given with its solution in Appendix H, is used for the region of the profile that sends extreme rays tangent to the absorber after one reflection—that is, the region between points 2 and 3 in Figure 5.19. The remaining region is an involute arranged to join the profile smoothly, and the equation for this is also given in the appendix. However, it is worth noting that for many practical applications the involute curve can be drawn accurately enough to scale by the draftsman's method of unwinding a taut thread from the absorber profile.

5.13   MECHANICAL CONSTRUCTION FOR 2D CONCENTRATOR PROFILES

In the discussion of Figure 5.19 in Section 5.20, it appeared that part of the concentrator surface was generated as an involute of the absorber section. It is possible to combine this result with the fact that optical path lengths from a wave front to a focus are constant to obtain a simple geometrical construction for the concentrator profile. Figure 5.23 shows a system similar to that shown in Figure 5.19, but we have drawn in a wave front *AB* of the incoming extreme pencil, and we assume the source is at a large but finite distance.

The construction is then as follows. We tie a string between the source and the point 1 at the rear of the absorber and pull the string taut with a pencil, as in the so-called gardener's method of drawing an ellipse. The length of the string must be such that it will be just taut when it is pulled right around the absorber to reach point 1 from the other side, as in Figure 5.24. It is then unwound, keeping the string taut, and the pencil describes the correct profile. To check this we simply have to show that the line drawn is at the correct angle to produce reflection. In
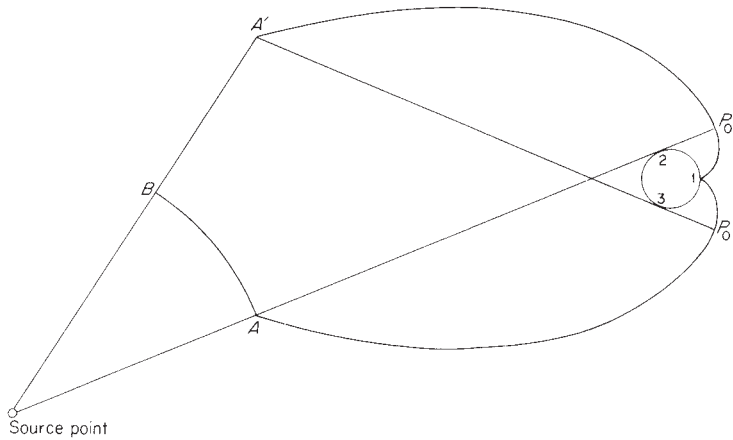
Figure 5.23   A concentrator for a source at a finite distance and a nonplane absorber.
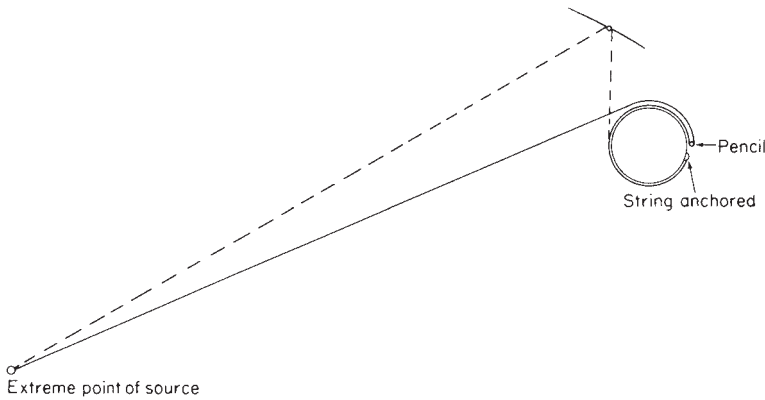
Figure 5.24   The string construction for the concentrator profile.

Figure 5.25 the string is tangent to the absorber at A, the source point is at C, and B is a typical position of the pencil. If the pencil is moved to B' where BCB' is a small angle ε, then we have

$$CBA = CB'A' - AA' + O(\varepsilon^2) \qquad (5.35)$$

so that

$$CBA = CB'A + 0(\varepsilon^2) \qquad (5.36)$$

Thus, by Fermat's principle BB' must be a portion of a reflecting surface that reflects CB into BA.

Figure 5.26 shows this method generalized further to a convex source and a convex absorber. The string is anchored at two suitably chosen points A and B and stretched with the pencil P. The length of the string is chosen so that it just reaches to the point Q when wound around the absorber. It is easily seen that this generates a 2D ideal concentrator. Here we are generating a concentrator that collects
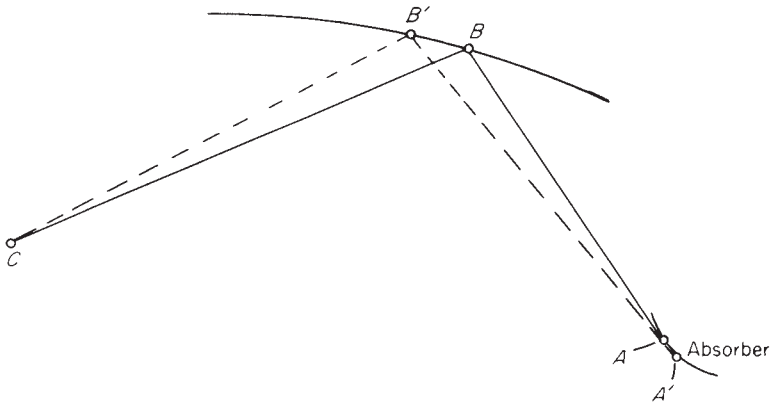
Figure 5.25   Proof that the string construction gives a concentrator surface that agrees with Fermat's principle.
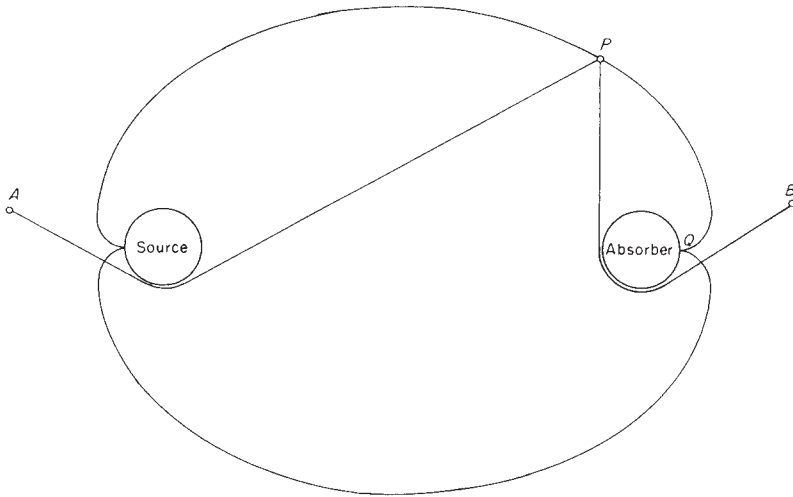
Figure 5.26   The case of a convex source and a convex absorber treated by the string construction.

all the flux from the source and sends it all to the absorber, and for this to be physically possible, we must make the perimeters of the source and absorber equal. If this were not so, the construction would not work because the string would not be the correct length to just close the curve, and this would mean we were trying to infringe the rules dictated by conservation of étendue.

Returning to Figure 5.26 we note that a solution is possible for any distance between source and absorber. We note also that the solution appears not to be unique, in the sense that we could break the reflector at its widest part and insert a straight parallel-sided section of any length, since such a section clearly transforms an étendue of width $2a$ and angle π. However, from obvious practical considerations, it is desirable to minimize the number of reflections of a given ray, and this is clearly done by not inserting such a straight section.

We could, of course, vary these solutions still further by adding in CPCs and θ_0/θ_i systems (Section 5.3) at the preceding break points instead of or in addition to the cylindrical sections. Again, this adds to the number of reflections, and it seems that there is always a unique "most economical" solution for which the extreme rays have at most one reflection.

We could generalize the string method to media with nonuniform refractive index by postulating that the string is elastic in such a way that it always assumes the optical path length, $\int n \, ds$, of the medium through which it passes. There is, of course, no way in which such a string could be realized, but the concept shows how ray trajectories—that is, geodesics—between the extremes of the apertures always define the correct mirror surface.

In another example of parallel development, the "string method" for generating ideal mirror profiles was discovered independently by Bassett and Derrick (1978) of the University of Sidney.

5.14   A GENERAL DESIGN METHOD FOR A 2D CONCENTRATOR WITH LATERAL REFLECTORS

In this section we formulate a more general treatment of the 2D concentrator with lateral reflectors, from which most other results in this chapter could be derived. A further generalization will be shown in Chapter 6. We shall describe a procedure in which an input surface and an output surface of given shapes are postulated enclosing and surrounded by regions of given refractive index distributions. On each of these surfaces a distribution of extreme rays is given. Then the procedure enables us to design a 2D concentrator that will ensure that all rays between the extreme incoming rays and none outside are transmitted so that the concentrator is optimal.

Suppose we have, as in Figure 5.27, two surfaces AB and $A'B'$, and let AB be illuminated in such a way that the extreme angle rays at each point form pencils belonging respectively to wave fronts Σ_α and Σ_β. Similarly, rays at intermediate angles belong to other wave fronts, so the whole ensemble of rays comes ultimately from a line of point sources and is transformed by a possibly inhomogeneous medium in such a way that the rays just fill the aperture AB. These rays then
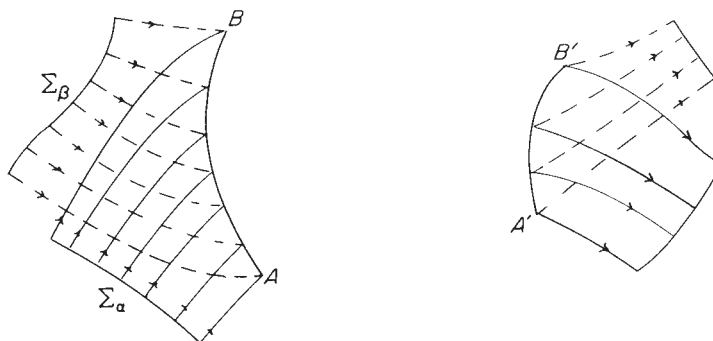
Figure 5.27   Beams of equal étendue to fit a concentrator.

have a certain étendue H, and we shall see following how to calculate it. Similarly, we draw rays and wavefronts emerging from $A'B'$ as indicated, and we postulate that these shall have the same étendue H.

Now we want to know how we can design a concentrator system between the surfaces AB and $A'B'$, possibly containing an inhomogeneous medium, that shall transform the incoming beam into the emergent beam without loss of étendue.

To solve this problem, we postulate a new principle (we shall see that our edge-ray principle of Chapter 4 can be regarded as derived from it): The optical system between AB and AB' must be such as to exactly image the pencil from the wave front Σ_α into one of the emergent wave fronts and Σ_β into the other. By "exactly" we mean that all rays from Σ_α as delimited by the aperture AB must just fill $A'B'$ so that none is lost and there is no unused space, and the same for the rays from Σ_β. This principle is relaxed in the formulation of Chapter 6, where it is only required that the optical system images the rays of $\Sigma_\alpha, \cup \Sigma_\beta$ into rays of any of the emergent wavefronts.

At this point it may be objected that the preceding seems to have little connection with our original edge-ray principle. But consider a system such as in Figure 5.28, which shows rays from one extreme wave front Σ_α in a CPC-like concentrator, and also a wave-front Σ'_α gradually moves into coincidence with the edge A' of the exit aperture. We then recover the CPC geometry and the original edge-ray principle. Thus, this new principle could be stated in the form that extreme points of the source must be imaged through the system by rays that just fill the exit and entry apertures.

To see how this principle leads to a solution of the general problem stated at the beginning of this section, we must first show how to calculate the étendue of an arbitrary beam of rays at a curved aperture, as in Figure 5.27. We use the Hilbert integral, a concept from the calculus of variations. In the optics context (Luneburg, 1964) the Hilbert integral for a path from P_1 to P_2 across a pencil of rays that originated in a single point is

$$I(1,2) = \int_{P_1}^{P_2} n\mathbf{k} \cdot ds \qquad (5.37)$$

where n is the local refractive index, k is a unit vector along the ray direction at the current point, and ds is an element along the path P_1P_2. Thus, $I(1, 2)$ is simply the optical path length along any ray between the wave fronts that pass through P_1 and P_2, so it is independent of the form of the path of integration. We can now use this to find the étendue of the beams in Figure 5.29. The Hilbert integral from A to B for the α pencil is seen from Eq. (6.14) to be
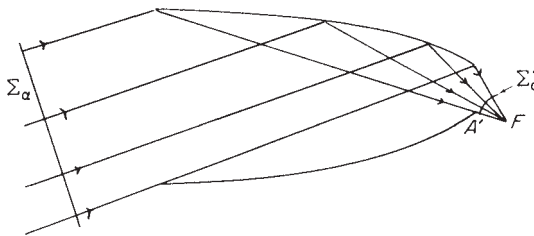
Figure 5.28   The edge-ray principle as a limiting case of matching wave fronts.
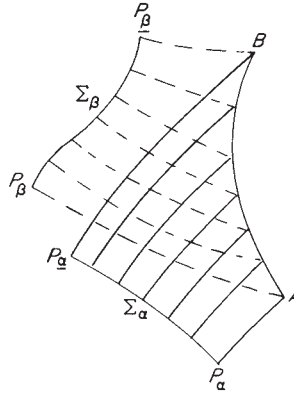
Figure 5.29   Calculating the étendue.

$$I_\alpha(AB) = \int_A^B n \sin \phi \, ds \qquad (5.38)$$

where ϕ is the angle of incidence of a ray on the line element ds. Thus,

$$I_\alpha(AB) = <n \sin \phi> L_{AB} \qquad (5.39)$$

where $<\ >$ denotes the average and L_{AB} is the length of the curve from A to B.

It follows that provided the aperture AB is filled with rays from the line of point sources just indicated, the étendue is simply

$$I_\alpha(AB) - I_\beta(AB)$$

But

$$I_\alpha(AB) = [P_\alpha B] - [P_\alpha A] \qquad (5.40)$$

where the square brackets denote optical path lengths along the rays so that we obtain for the étendue

$$H = [P_\alpha B] + [P_\beta A] - [P_\alpha A] - [P_\beta B] \qquad (5.41)$$

This result can be seen to be a simple generalization of the result of Eq. (5.11).

Now let there be some kind of system constructed that achieves the desired transformation of incident extreme pencils with emergent extreme pencils, as in Figure 5.30. The system takes Σ_α into Σ'_α and Σ_β into Σ'_β, and we want it to do so without loss of étendue. We write down the optical path length from $P_\alpha P'_\alpha$ and equate it to that from P_α to P'_α. Similarly, for the other pencil,

$$[P_\alpha B] + [BA']_\alpha + [A'P'_\alpha] = [P_\alpha A] + [AB']_\alpha + [B'P'_\alpha][P_\beta] + [AB']_\beta + [B'P'_\beta]$$
$$= [P_\beta B] + [BA']_\beta + [A'P'_\beta] \qquad (5.42)$$

where

$$\{[BA'] + [P_\beta] - [P_\alpha A] - [P_\beta B]\} - \{[A'P'_\beta][B'P'_\alpha] - [B'P'_{\beta'}]\}$$
$$= [AB']_\alpha - [AB']_\beta + [BA']_\beta - [BA']_\alpha \qquad (5.43)$$
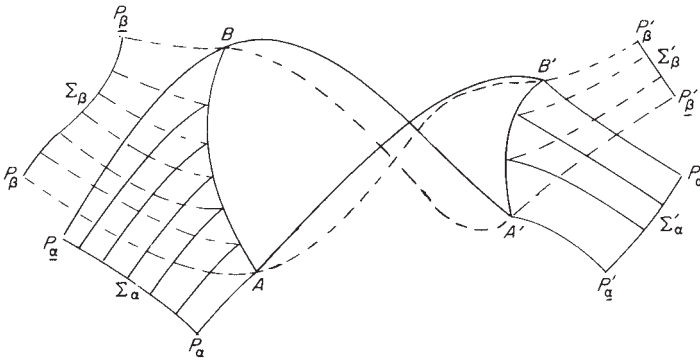
Figure 5.30   Rays inside the concentrator.

The left-hand side of this equation can be seen by comparison with Eq. (5.41) to be the difference between the étendues at the entry and exit apertures. Since we require this difference to vanish, we have to make the right-hand side of Eq. (6.19) vanish. A simple way to do this would be to ensure that the optical system of the concentrator is such that the α and β ray paths from A to B' coincide, and the same for those from B to A'. We can do this by starting segments of mirror surfaces at A and A' in such directions as to bisect the angles between the incoming α and β rays. We then continue the mirror surfaces in such a way as to make all β rays join up with the corresponding emerging β' rays; in other words, we image the β pencil exactly into the β' pencil, and similarly for the other mirror surface connecting B and B'. We have thus completed the construction and used up all degrees of freedom in doing so. For this construction we need to have one and only one α ray and one and only one β ray crossing the points of the mirror. This condition excludes wave fronts whose caustic is close to the mirror to design.

It appears from this theorem that it is always necessary to incorporate lateral mirrors in the design of ideal concentrators. Something like this emerges in considering simpler symmetric concentrators based on the edge-ray principle. The concentrator has to image the extreme ray pencils sharply at the rim of the exit aperture, and since any lens extends equally on either side of the axis, it must operate on both extreme ray pencils. This requires that a given region of the lens away from the axis shall produce stigmatic imagery of two pencils at different angles, which is the way in which the SMS design method of Chapter 8 solves the problem. The need of the lateral reflectors for this design method will be clearer in Chapter 6 with the introduction of the flow lines.

5.15   APPLICATION OF THE METHOD: TAILORED DESIGNS

Let us reconsider the case of Figure 5.18 where we have a lens at the entry of the concentrator to design. Assume that the lens can no longer be considered a thin lens, but just as a refracting surface. The objective is to design a reflector according to the rules of the preceding section. The first step of the design is to trace the

rays of the wave front W across the lens. After the refraction, the wave front W' will not be plane in general, and since we don't know anything about the shape of the lens, we cannot conclude that the wave front after refraction is spherical. The reflector must be designed to focus the rays of W' into the point P_2. This is a generalized Cartesian Oval problem (see Chapter 8). If the wave front W' were spherical, as in the case of Figure 5.18 (with center at F in this example), then the reflector obtained would be the branch of a hyperbola (with focuses at P_2 and F).

These types of designs, in which the incident wave fronts are not planes, have been called tailored designs. The first reference of one application of these designs (excluding the cases like the one of Figure 5.18, in which the refracted wave front is spherical, so the reflector has an analytic expression) is the Linear Integrated Tubular Concentrator[1] (see Appendix C) whose cross section is a 2D design with a refracting circle at the entry aperture. The plane wave front W is refracted by this circle.

In other tailored designs, the plane wave fronts are reflected at a prescribed mirror instead of being refracted. Reference[2] shows more details of this application. Many practical applications of nonimaging concentrators are within this case[3,4,5,6]. As we will see in Chapter 7, there is other type of tailored designs for which the calculation of the incident wave fronts is not so straightforward.

5.16   A CONSTRUCTIVE DESIGN PRINCIPLE FOR OPTIMAL CONCENTRATORS

We conclude this chapter with a discussion of a design prescription for optimal concentrators that historically has been a fertile source of new and useful solutions. In contrast to the edge-ray principle, which is allied to such abstract notions as étendue and the Hilbert integral theorem, this practical procedure directly instructs us how to draw the profile curve of the mirror. The statement is simply that we maximize the slope of the mirror profile curve consistent with reflecting the extreme entrance rays onto the absorber and subject to various subsidiary conditions that we may wish to impose. This generic principle for designing optimal concentrators is inherent in the earliest references (Hinterberger and Winston, 1966a,b).

To see how this rule operates we notice that the CPC designs for variously shaped absorbers are obtained when we don't impose any subsidiary conditions. On the other hand, the θ_i/θ_0 design results when we impose the condition that the maximum exit angle not exceed θ_0. This approach is particularly useful in situations where the subsidiary conditions preclude an ideal system—hence strict application of the edge-ray principle. For example, the various totally internally reflecting designs are obtained by specifying total internal reflection at the external wall as a condition. As discussed earlier in this chapter, when the input ray distribution and index of refraction fall outside a specified range, ideal solutions are not possible. However, efficient designs are still available.

Consider the example in Figure 5.31, where the front entrance face is curved and the index of refraction is not sufficiently high to totally reflect the extreme rays along the portion P_2, P_3, were we to follow the edge-ray principle. The maximum slope rule would suggest we maintain a constant angle (the critical
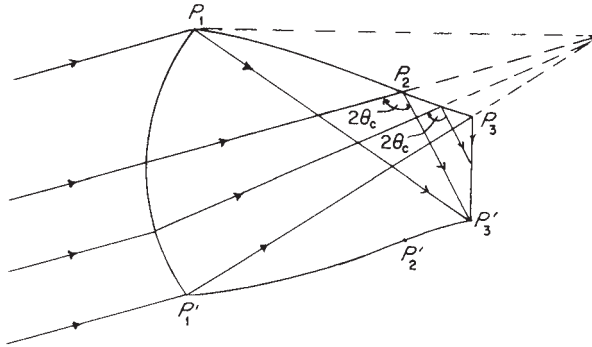
Figure 5.31   The solid concentrator that just satisfies the critical angle condition. The profile is hyperbolic from P_1 to P_2 (assuming the entry surface is aberration free). Then if the critical angle θ_c for total internal reflection is reached at P_2, it is an equiangular spiral.

angle) between the extreme ray and normal to the wall along this portion of the profile curve. In the approximation that the curve face images the extreme pencil onto a point, the portion between P_1, P_2 is a hyperbola, whereas the portion between P_2, P_3 is an arc of an equiangular spiral.

REFERENCES

Bassett, I. M., and Derrick, G. H. (1978). The collection of diffuse light onto an extended absorber. *Opt. Quantum Electron.* **10**, 61–82.

Gordon, J. M. (1996). Simple string construction method for tailored edge-ray concentrators in maximum-flux solar energy collectors. *Solar Energy* **56**, 279–284.

Hinterberger, H., and Winston, R. (1966a). Efficient light coupler for threshold Čerenkov counters. *Rev. Sci. Instrum.* **37**, 1094–1095.

Hinterberger, H., and Winston, R. (1966b). Gas Čerenkov counter with optimized light-collecting efficiency. *Proc. int. Conf. Instrum. High Energy Phys.* 205–206.

Hottel. H. (1954). Radiant heat transmission. *In* "Heat Transmission" (W. H. McAdams, ed.), 3rd Ed. McGraw-Hill, New York.

Luneburg, R. K. (1964). "Mathematical Theory of Optics." Univ. of California Press, Berkeley. This material was originally published in 1944 as loose sheets of mimeographed notes and the book is a word-for-word transcription.

Miñano, J. C., Ruiz, J. M., and Luque, A. (1983). Design of optimal and ideal 2D concentrators with the collector immersed in a dielectric tube. *Applied Optics* **22**, 3960–3965.

Ong, P. T., Gordon, J. M., and Rabl, A. (1995). Tailoring lighting reflectors to prescribed illuminance distributions: Compact partial-involute designs. *Applied Optics* **34**, 7877–7887.

Ong, P. T., Gordon, J. M., and Rabl, A. (1996). Tailored edge-ray designs for illumination with tubular sources. *Applied Optics* **35**, 4361–4371.

Ong, P. T., Gordon, J. M., Rabl, A., and Cai, W. (1995). Tailored edge-ray designs for uniform illumination of distant targets. *Optical Engineering* **34**, 1726–1737.

Ortobasi, U. (1974). "Proposal to Develop an Evacuated Tubular Solar Collector Utilizing a Heat Pipe." Proposal to National Science Foundation (unpublished).

Ploke, M. (1967). Lichtführungseinrichtungen mit starker Konzentrationswirkung. *Optik* **25**, 31–43.

Ries, H. R., and Wintson, R. (1994). Tailored edge-ray reflectors for illumination. *JOSA, A*, **11**, 1259–1264.

Winston, R. (1974). Principles of solar concentrators of a novel design. *Sol. Energy* **16**, 89–95.

Winston, R. (1978a). Cone collectors for finite sources. *Appl. Opt.* **17**, 688–689.

Winston, R. (1978b). Ideal light concentrations with reflector gaps. *Appl. Opt.* **17**, 1668.

Winston, R., and Hinterberger, H. (1975). Principles of cylindrical concentrators for solar energy. *Sol. Energy* **17**, 255–258.

6

THE FLOW-LINE METHOD FOR DESIGNING NONIMAGING OPTICAL SYSTEMS

6.1 THE CONCEPT OF THE FLOW LINE

We have seen that the edge-ray approach to nonimaging concentration has considerable flexibility in handling 2D problems. Its application in 3D is more limited, and, although useful designs are obtainable in 3D, these fall short of ideal concentration. This motivates the search for alternatives to edge-ray designs, especially in 3D applications (Winston and Welford, 1979a).

In the geometrical optics approximation, we describe the propagation of light rays as trajectories in a six-dimensional phase space (p, x), where the components of x are the generalized coordinates (x_1, x_2, x_3) and the components of p are the generalized momenta. These momenta give the optical path length when the light ray traverses length ds through a medium with an index of refraction n by the relation $n\, ds = p_1 dx_1 + p_2 dx_2 + p_3 dx_3$. For example, in cartesian coordinates (x, y, z) the components of p are just the optical-direction cosines.

It is useful to picture the propagation of the totality of light rays as if it were a fluid flow in this six-dimensional space. This flow is subject to conservation theorems that are essentially geometric and follow from the law of propagation of light rays. These are loosely referred to as Liouville's theorem, but they are more properly called the integral invariants of Poincaré. For our purpose, the most useful invariant is constructed out of a vector $\mathbf{J}$, whose components are

$$\mathbf{J}_1 = \int dp_2 dp_3, \quad \mathbf{J}_2 = \int dp_1 dp_3, \quad \mathbf{J}_3 = \int dp_1 dp_2, \tag{6.1}$$

Postulating as always a medium with no attenuation or radiation sources, the surface integral of $\mathbf{J} \approx \int \mathbf{J} \cdot dA$ is invariant so that $\mathbf{J}$ has zero divergence. We call $\mathbf{J}$ the geometrical vector flux and the direction of $\mathbf{J}$ the flow line. Quantities like $\mathbf{J}$ occur in radiative transfer where $\mathbf{J}\, dA$ is proportional to the net flux through the

surface element dA and in photometry where $\mathbf{J}_z$ is proportional to the illumination on the x, y plane. This is most readily seen by rewriting

$$\mathbf{J}_z = \int n^2 d\Omega \cos\theta_z \tag{6.2}$$

where θ_z, is the inclination of the elementary beam to the z axis.

6.2   LINES OF FLOW FROM LAMBERTIAN RADIATORS: 2D EXAMPLES

All Lambertian radiators that subtend the same solid angle at a point in the strict sense that they are bounded by the same cone give the same vector flux $\mathbf{J}$ at that point. This can be made clear by some simple examples (see Winston and Welford, 1979b).

Consider a strip AA' in the x, y plane, of width 2 in the y direction, as in Figure 6.1. A simple calculation shows that $\mathbf{J}$ points along the bisector of the angle APA', and its magnitude is $|\mathbf{J}| = 2\sin\theta$, where the angle $APA' = 2\theta$. From elementary geometry the lines of flow are confocal hyperbolae with A and A' as foci. Notice that the difference of distances from P to the foci $|AP| - |A'P|$ is constant along a flow line from the property of hyperbolae. As a second example, Figure 6.2 shows a semi-infinite strip. It is easily seen that the flow lines are confocal parabolas, with the focus at the end A of the strip. A final example is a wedge: Figure 6.3 shows an example where the wedge angle is 60°. In regions where only one face of the wedge is seen, the pattern is the same as for the semi-infinite strip of Figure 6.2. From any point where both faces are seen, the effect is of a uniform
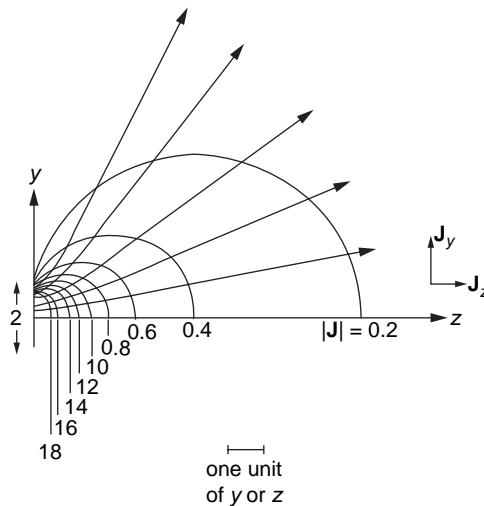
Figure 6.1   Lines of flow and loci of constant $|\mathbf{J}|$ for the two-dimensional strip: The width is taken as two units. The axes of circles are labeled with the values of $|\mathbf{J}|$. The two sets of lines are not orthogonal because $\mathbf{J}$ is not derived from a scalar potential.
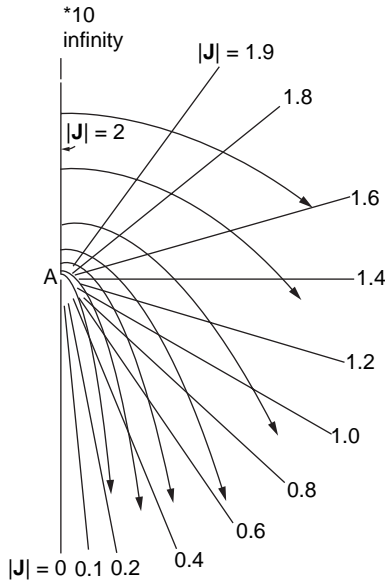
Figure 6.2 Lines of flow and loci of constant |**J**| for a semi-infinite strip.
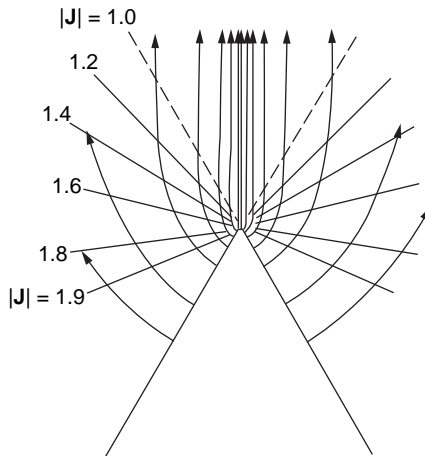
Figure 6.3 Lines of flow and loci of constant |**J**| for a wedge.

Lambertian source subtending an angle of $60°$. Thus |**J**| is constant in this region, and the lines of flow are parallel as indicated.

Figure 6.4 shows a truncated $60°$ wedge $RQQ'R'$, which is infinitely extended back from the apex with an included angle $2\theta_o, = 60°$. We have to consider four different regions in calculating vector flux fields; these are separated by the boundary lines of the wedge produced, as shown in broken lines, and they are labeled regions A (and A'), B, C (and C'), and D. A point P in region A receives flux only from the face RQ so that the **J** distribution is the same as for a semi-infinite strip; that is, the flow lines are confocal parabolas with focus at Q, and the loci of constant |**J**| are straight lines radiating from Q. If P is in region B, it receives flux
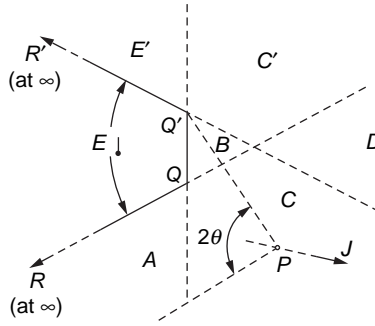
Figure 6.4   A truncated wedge of angle $2\theta_o$. The truncated QQ' is shown as forming equal angles with the two sides of the wedge, but in fact the argument is still valid if this is not so.

only from the strip QQ', and again we have the result that the lines of flow are confocal hyperbolas with Q and Q' as foci, and the loci of constant $|\mathbf{J}|$ are arcs of circles through QQ and Q' If P is anywhere in region D, the source has the constant angular subtense $2\theta_o$; thus $|\mathbf{J}|$ is constant at $2\sin\theta_o$ throughout this region, and the lines of flow are straight lines perpendicular to QQ'. Finally, in region C the point P receives radiation from RQ and QQ' and the angular subtense is 2θ as show in Figure 6.14. The direction of $\mathbf{J}$ thus bisects the angle between a constant direction (parallel to QR) and the radius from a fixed point Q'; thus, the lines of flow in region C are confocal parabolas with focus at Q' and axis through Q' and parallel to RQ. The loci of constant $|\mathbf{J}|$ are straight lines radiating from Q', since the angle 2θ as constructed is constant along these lines. The lines of flow for a truncated wedge are shown in Figure 6.5.

6.3   3D EXAMPLE

A disk as a Lambertian radiator—that is, the circular opening of a black body cavity—is the most important example to consider. The components of $\mathbf{J}$ can be calculated by integrating expressions such as Eq. (1.1) over the surface of the disk. These integrations have been carried out for radiative transfer purposes, and we shall simply quote the results. Sparrow and Cess (1978), for example, provide the details, and they also give a useful transformation of the surface integral over the disk into a line integral around its edge; this considerably simplifies the calculation.

Let c be the radius of the disk and take the coordinate axes as in Figure 6.6. Then the components of $\mathbf{J}$ in the y, z plane, which from symmetry are all that need be considered, are given by

$$
\mathbf{J}y = \frac{\pi}{2}\frac{z}{y}\left\{\frac{c^2 + y^2 + z^2}{\left[\left(c^2 + y^2 + z^2\right)^2 - 4c^2y^2\right]^{1/2}} - 1\right\}
$$

$$
\mathbf{J}z = \frac{\pi}{2}\left\{\frac{c^2 - y^2 - z^2}{\left[\left(c^2 + y^2 + z^2\right)^2 - 4c^2y^2\right]^{1/2}} + 1\right\}
$$

(6.3)
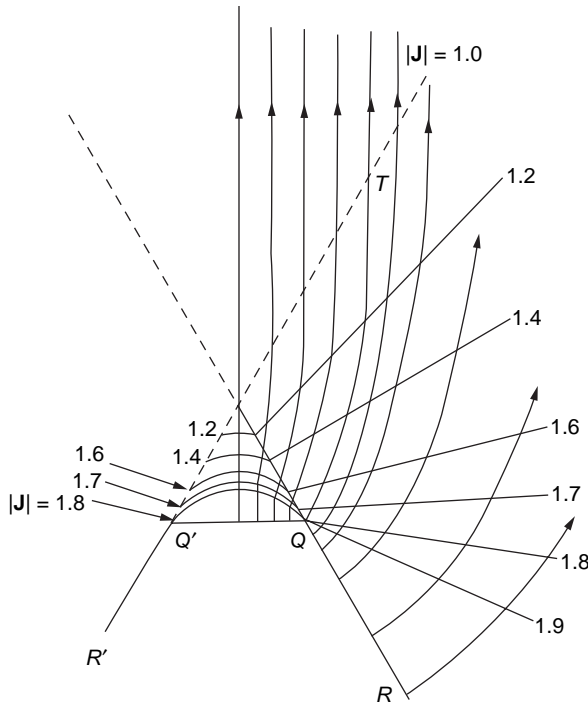
Figure 6.5   Lines of flow of **J** and labeled loci of constant |**J**| for the truncated wedge; only the loci below the line of symmetry are shown. The wedge angle is 60°.
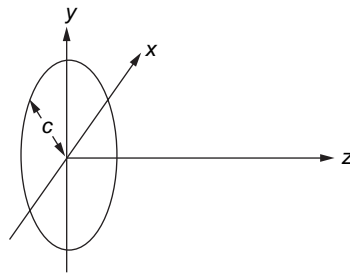
Figure 6.6   Geometry for calculating **J** for a luminous disk.

The slope of the lines of flow is given by $\arctan(\mathbf{J}_y/\mathbf{J}_z)$, and when this is calculated, it turns out that the lines of flow are the same hyperbolas as for the 2D case. The lines of flow in three dimensions are obtained by rotation about the z axis, as shown in Figure 6.7.

6.4   A SIMPLIFIED METHOD FOR CALCULATING LINES OF FLOW

In certain cases it is easier to calculate the étendue than to perform the preceding integrals. In case the étendue is known, we can use the solenoidal property of
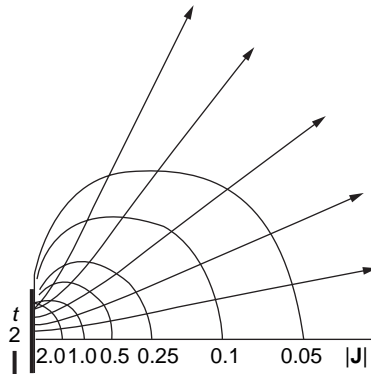
Figure 6.7   Lines of flow and loci of constant |**J**| for a luminous disk.
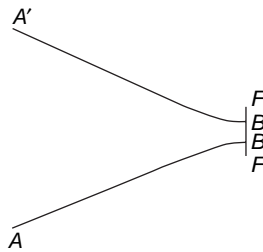
Figure 6.8   Étendue for a luminous disk.

J to find the lines of flow. For example, in the case described by Figure 6.8, the étendue H from F, F' to A, A' is equal to $[A'F] - [AF']$ in the 2D case and $\pi^2\{[AF] - [AT']\}^2/4$ in the 3D case (the Hottel strings!). Then holding H = constant means we intercept a constant flux of **J** and therefore move along a line of flow. But holding the difference of distances to two foci, F, F' constant is the geometric definition of a hyperbola. Hence we recover the hyperbolic flow lines with lighter computation. By similar reasoning we can also obtain the components of **J**. Thus, the relations $\mathbf{J}z = (\partial H/\partial y)/2\pi y$ and $Jy = -(\partial H/\partial z)/2\pi y$ give Eq. (3).

6.5   PROPERTIES OF THE LINES OF FLOW

Let $\delta\Sigma$ be a small surface element with a 100% reflecting mirror surface on one side. If we place $\delta\Sigma$ in a vector flux distribution at a point P (Figure 6.9) so the lines of flow lie in its surface, then the lines of flow locally on the mirror side will be undisturbed. To see this, we note that before the mirror is placed at position P, the resolved part of **J** normal to the surface is zero, since **J** lies along the lines of flow. Now this zero normal component is made up of equal and opposite components $\mathbf{J}_n^+$ and $\mathbf{J}_n^- = -\mathbf{J}_n^+$ from rays coming from either side of $\delta\Sigma$. But the effect of inserting the mirror is to remove $\mathbf{J}_n^-$ and replace it by $\mathbf{J}_n^+$ with reversed sign, according to the law of reflection, so there is no local disturbance.
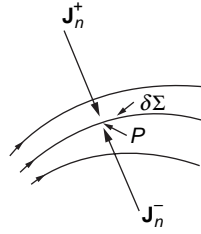
Figure 6.9   Placing a mirror in a vector flux field.

The insertion of a mirror can, of course, disturb the lines of flow ′ at some distance because it does not follow from $\mathbf{J}_n^- = -\mathbf{J}_n^+$ that each ray from one side is paired with a ray in the same plane of incidence and with the same angle of incidence on the other side. However, this simple situation does hold for a 2D vector flux field illuminated by a Lambertian radiator without obstructions and in the important case of a 3D field produced by a Lambertian disk.

6.6   APPLICATION TO CONCENTRATOR DESIGN

We have noted that a mirror surface placed along lines of flow of the vector flux will not disturb the overall flow pattern under a certain condition. This condition is that the zero normal component of $\mathbf{J}$ must result from the fact that, for each ray r_1 incident on one side of the surface, there must be a corresponding ray r_2 incident on the other side such that r_2 is in the direction of the reflection of r_1; we call this a condition of detailed balance of the rays. It is possible to have $\mathbf{J}_{\mathrm{normal}} = 0$ at a surface containing the lines of flow without the condition of detailed balance being fulfilled, and then $\mathbf{J}$ would remain undisturbed locally but not over the entire illuminated region. Clearly, the condition of detailed balance holds for a 2D field illuminated by a Lambertian source without obstructions.

We may therefore place mirrors along any of the flow lines of Figure 6.5 without disturbing the flux patterns. For example, suppose we place a mirror along the flow line from Q, ending at the point T where the flow line meets the continuation of the wedge boundary $R'Q'$, and similarly another mirror $Q'T'$ symmetrically on the other side. These mirrors take the Lambertian flux ($|\mathbf{J}| = 2$) from QQ' and convert it to a uniform flux with $|\mathbf{J}| = 1$ at TT'. The rays in the emergent beam must spread over an angle $\pm 30°$, since $|\mathbf{J}| = 1$. Clearly, we have constructed a concentrator in reverse, and our construction with parabolas with tilted axes is precisely that for the compound parabolic concentrator.

Since light rays, and therefore flow lines of $\mathbf{J}$, are reversible, we see that a way to construct a concentrator with maximum theoretical concentration is to place mirrors in the flow lines under conditions of detailed balance of the rays. This constitutes an entirely new perspective on concentrator design.
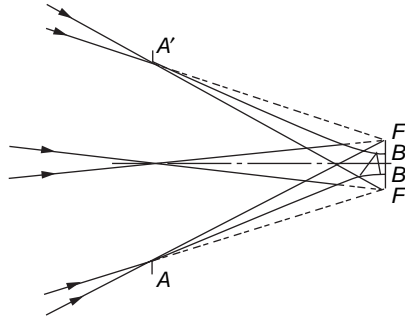
Figure 6.10   The hyperboloid of revolution as a concentrator.

6.7   THE HYPERBOLOID OF REVOLUTION AS A CONCENTRATOR

The results of the previous section lead to another interesting new perspective on the design of nonimaging concentrators. Suppose we start with the disk radiator and we place a mirror in the form of a hyperboloid of revolution coincident with a set of flow lines as in Figure 6.10. We truncate the mirror at some distance so the open end is a circle of diameter AA'. Considering the inside as a mirror, this forms a nonimaging concentrator with unusual properties. The foci of the hyperbolas in the section shown are at F and F', the ends of the diameter of the original disk. Then all rays entering the aperture AA' and pointing somewhere inside the disk will be reflected by the mirror so as to strike, eventually, the inner disk of diameter BB'. Thus, the concentrator takes all rays from the virtual source FF', which can pass the entry aperture AA' and concentrates them into an exit aperture BB'.

This result is easily proved for rays in a meridional section—that is, in the plane of the diagram. Consider a ray passing through A' and aimed at F, as indicated by the double arrow; after the first reflection, it will, by a basic property of conic sections, be aimed at F', and so on as indicated. Thus, the extreme angle rays emerge from the exit aperture but only after an infinite number of reflections. It is easily seen that rays at angles inside the extreme angles all emerge. Thus, in the meridional plane this is a concentrator of maximum theoretical concentration. This property also holds for skew rays, although this is not quite so obvious. We give in Appendix K proof that any skew ray incident on the mirror and pointing on the rim of the disk FF' will be reflected to another point on the rim and the result follows directly. When used in reverse, the same design produces a virtual ring that fills the space between a Lambertian source of diameter BB' and the larger diameter FF'. The visual effect produced is striking.

6.8   ELABORATIONS OF THE HYPERBOLOID: THE TRUNCATED HYPERBOLOID

The properties of the truncated hyperboloid are evident from the preceding discussion. We refer to Figure 6.11, where the hyperboloid stops short of the plane of the virtual source. For example, rays collected from the source of diameter FF' by
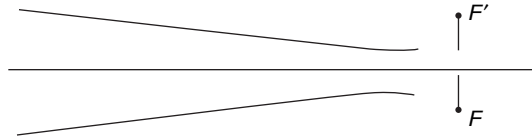
Figure 6.11   The truncated hyperboloid of revolution as a concentrator.
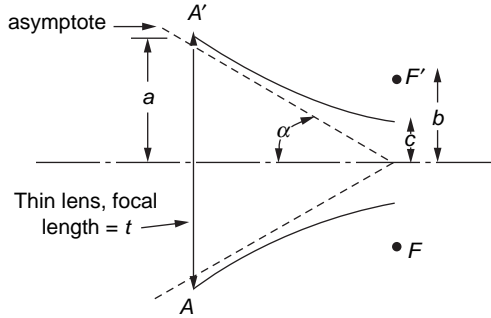
Figure 6.12   The lens/flow-line concentrator. The entry aperture is FF', and the exit aperture is at the waist of the hyperboloid. The semiangle of the asymptotic cone is arcsin c/b, and $(a/c)^2$ is the concentration ratio.

the small aperture will appear to originate from the same source upon emerging from the large aperture. Alternatively, rays directed toward the source of diameter FF' at the large aperture will illuminate the same source upon exiting the small aperture. The truncated hyperboloid may be useful on its own, but usually it is in combination with lenses, as discussed following.

6.9   THE HYPERBOLOID COMBINED WITH A LENS

We have seen that, for the bare hyperboloid, both real and virtual Lambertian sources are in the same plane. This is a limitation in certain applications; however, we can manipulate these positions with lenses to place the sources at more convenient locations. We show in Figure 6.12 the full hyperboloid with a positive lens of diameter $AA' = 2a$ at the large aperture, which takes the source plane to infinity—that is, the source is in the focal plane of the lens. Consider the properties of this design. For simplicity we neglect off-axis aberrations of the lens; a complete discussion is found in Welford, O'Gallagher, and Winston (1987). We only remark that, since the requirements on the lens need not be severe (e.g., the focal ratio may be chosen large, as will be shown following), the aberrations can be made negligible in practice. In this approximation, the cone of rays of angle θ is imaged by the lens onto the circle of diameter $FF' = 2b$. The hyperboloid collects the rays that fill this cone and concentrates them to a smaller circle of diameter $BB' = 2c$. We show in Appendix L that $c/a = \sin\theta$. Therefore, this is an ideal concentrator. This result could have been anticipated from the fact that we have postulated an ideal
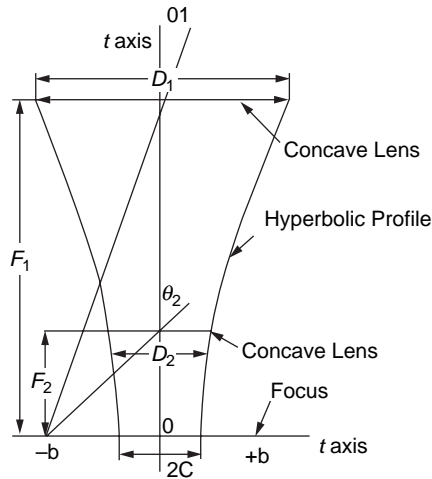
Figure 6.13 $\theta_i \neq \theta_o$ transformer using an FLC with two lenses.

lens while the hyperboloid on its own is an ideal concentrator, albeit operating on a virtual source.

The design considerations readily follow from the geometry of a hyperbola. With reference to Figure 6.12, the focal ratio of the lens can be estimated from the slope α of the asymptote to the hyperbola. The slope is given by $\sin \alpha = c/b$; the acceptance angle θ determines the focal length of the lens, completing the design.

6.10   THE HYPERBOLOID COMBINED WITH TWO LENSES

We have seen that we can construct a 3D concentrator approaching ideal properties by placing a suitable lens at the large aperture of the hyperboloid. Similarly, we can construct a 3D θ_1/θ_2 concentrator (angle transformer) by placing an additional negative lens at the small aperture of the truncated hyperboloid, which moves the source plane to infinity as shown by Ning (1988). The construction is shown in Figure 6.13, where it is apparent that the focal lengths of the two lenses are related by $F_1/F_2 = \tan \theta_1/\tan \theta_2$, whereas their focal ratios are related as $f_1/f_2 = \cos \theta_1/\cos \theta_2$. We see that this design lends itself to all dielectric versions, since the positive and negative lenses are readily molded into the shape of the transformer, whereas in practical cases the condition for total internal reflection is likely to be satisfied.

6.11   GENERALIZED FLOW LINE CONCENTRATORS WITH REFRACTIVE COMPONENTS

In 2D it is possible to generalize the flow line construction following the notions expressed in Section 6.4. For simplicity, we confine our discussion to systems that
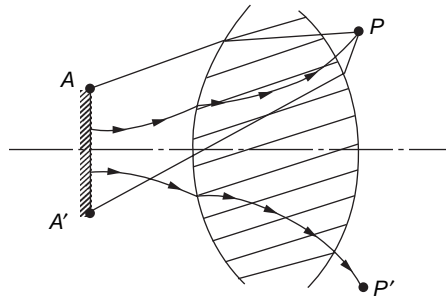
Figure 6.14   Flow lines with refractive components AA' are a Lambertian source. The arrows indicate row lines; the plain lines, rays.

are symmetrical transverse to the optic axis. As already noted in Section 6.4, the étendue H generalizes to the difference of optical path lengths (up to an overall constant). This remains true even in the presence of refractive media, provided the optical path lengths are measured along rays. These rays need not be straight lines. Thus, in Figure 6.14, the étendue H from Lambertian source AA' to section PP' is proportional to $[A'P] - [AP]$, where the brackets indicate optical path lengths. It follows that the lines of flow (indicated by arrows in the figure) lie along contours of $H =$ constant. Since the detailed balance condition holds in 2D, we may construct concentrators by placing mirrors along the flow lines. However, it does not follow that the 3D construction obtained by rotating the 2D flow line about the optic axis will automatically satisfy detailed balance. Specific cases will have to be checked with respect to detailed balance before the usefulness of the 3D designs can be evaluated.

6.12   HAMILTONIAN FORMULATION

6.12.1   Introduction

The principles of Geometrical Optics can be formulated in several ways, all of them being equivalent in the sense that they can provide the same information. Nevertheless, there are some particular problems for which one formulation is better than the others—for example, the problem is more easily stated and sometimes more easily solved using one of the formulations. This is common to disciplines having more than one mathematical model. Probably, the most well-known formulation of Geometrical Optics is the variational one (Fermat's principle). In Section 6.12.2 we will see another well-known formulation: the Hamiltonian equations. This formulation will be useful for stating and solving some nonimaging design problems both in 2D and 3D geometry with the Poisson Brackets method. This method is unique in the sense that it is able to give ideal designs in 3D geometry in some cases. Unfortunately, the 3D designs obtained with this method require graded refractive index materials, which limits its practical use.

The Hamiltonian formulation has been widely used in imaging optics. The most important results are the characteristic functions and the simplicity with which some optical invariants are recognized (see, for instance, Luneburg, 1964).

One of these invariants is a common tool in nonimaging optics, the conservation of étendue. Within the Hamiltonian formulation, this invariant is one of the Poincaré's invariants. Although Hamilton originally developed his equations for optics, their applications in mechanics developed faster, so some of the results of the theory may sound as if they belong to mechanics more than to optics. This may be the case of the Poisson brackets. In other cases, the same result has two different names: one for optics and one for mechanics. For instance, in mechanics Fermat's principle is known as the principle of least action or the principle of Maupertuis. The Hamiltonian formulation, when applied to nonimaging optics, makes little use of the results for imaging optics, and because of this, its results may appear more mechanic than optic.

6.12.2   Hamilton Equations and Poisson Bracket

As we will see, the Hamiltonian formulation is not unique. We start with the description of the Hamilton equations that we will use in the most general form we need. Let $x_1 = x_1(s)$, $x_2 = x_2(s)$, $x_3 = x_3(s)$, $t = t(s)$ be the equations of a ray trajectory in parametric form (s is the parameter) in the space $x_1 - x_2 - x_3 - t$ ($x_1 - x_2 - x_3$ are the Cartesian coordinates, and t is the time). For each point of the trajectory of a ray—that is, for each value of s, we have a value of the wave vector $k = (k_1, k_2, k_3)$ and a value of the angular frequency ω. Let $k_1 = k_1(s)$, $k_2 = k_2(s)$, $k_3 = k_3(s)$, $\omega = \omega(s)$ be the values of the three components of the wave vector and the angular frequency, respectively. The set of eight functions $x_1 = x_1(s)$, $x_2 = x_2(s)$, $x_3 = x_3(s)$, $t = t(s)$, $k_1 = k_1(s)$, $k_2 = k_2(s)$, $k_3 = k_3(s)$, $\omega = \omega(s)$ define a ray trajectory in the phase space $x_1 - x_2 - x_3 - t - k_1 - k_2 - k_3 - \omega$. In general we are only interested in the trajectory of the ray in the space $x_1 - x_2 - x_3$, sometimes also including t. The introduction of the other variables in this case is still interesting because they simplify the formulation of the equations. The variables $k_1, k_2, k_3, -\omega$ are called the conjugate variables of x_1, x_2, x_3, t in the Hamiltonian formulation.

A key point of the Hamiltonian formulation is the so-called Hamiltonian function. In the case of optics, $K(x_1, x_2, x_3, t, k_1, k_2, k_3, \omega)$ is a function such that $K = 0$ defines the surface of the wave vector k (Arnaud, 1974). The equation $K = 0$ is also called Fresnel's surface of wave normals, and it is directly related to the Fresnel's Differential Equation (Kline and Kay, 1965). The function K can be determined by the properties of the medium where the rays are evolving.

The Hamiltonian equations can be written as

$$\frac{dx_1}{ds} = \frac{\partial K}{\partial k_1} \quad \frac{dk_1}{ds} = \frac{\partial K}{\partial x_1}$$
$$\frac{dx_2}{ds} = \frac{\partial K}{\partial k_2} \quad \frac{dk_2}{ds} = \frac{\partial K}{\partial x_2}$$
$$\frac{dx_3}{ds} = \frac{\partial K}{\partial k_3} \quad \frac{dk_3}{ds} = \frac{\partial K}{\partial x_3} \qquad (6.4)$$
$$\frac{dt}{ds} = \frac{\partial K}{\partial \omega} \quad \frac{d\omega}{ds} = \frac{\partial K}{\partial t}$$

The solutions of this system of equations are sets of eight functions $x_1 = x_1(s)$, $x_2 = x_2(s)$, $x_3 = x_3(s)$, $t = t(s)$, $k_1 = k_1(s)$, $k_2 = k_2(s)$, $k_3 = k_3(s)$, $\omega = \omega(s)$. It can be proved that the solutions of this equation system are curves contained in the hyper-

surfaces K = constant of the phase space $x_1 - x_2 - x_3 - t - k_1 - k_2 - k_3 - \omega$—that is, the function K is a first integral of the system (Arnold, 1976). A function $F(x_1, x_2, x_3, t, k_1, k_2, k_3, \omega)$ is a first integral of the system of Eq. (6.4) if F is constant along any ray trajectory—that is, $dF/ds = 0$. F is said to be a "constant of motion" in mechanics (Abraham and Marsden, 1978; Leech, 1958). This can be written as

$$\frac{dF}{ds} = \frac{\partial F}{\partial x_1}\frac{dx_1}{ds} + \frac{\partial F}{\partial k_1}\frac{dk_1}{ds} + \frac{\partial F}{\partial x_2}\frac{dx_2}{ds} + \frac{\partial F}{\partial k_2}\frac{dk_2}{ds} + \frac{\partial F}{\partial x_3}\frac{dx_3}{ds}$$
$$+ \frac{\partial F}{\partial k_3}\frac{dk_3}{ds} + \frac{\partial F}{\partial t}\frac{dt}{ds} + \frac{\partial F}{\partial \omega}\frac{d\omega}{ds} = 0 \qquad (6.5)$$

The total derivative of F with respect to s can be written as (using Eq. (6.4))

$$\frac{dF}{ds} = \frac{\partial F}{\partial x_1}\frac{\partial K}{\partial k_1} - \frac{\partial F}{\partial k_1}\frac{\partial K}{\partial x_1} + \frac{\partial F}{\partial x_2}\frac{\partial K}{\partial k_2} - \frac{\partial F}{\partial k_2}\frac{\partial K}{\partial x_2} + \frac{\partial F}{\partial x_3}\frac{\partial K}{\partial k_3}$$
$$- \frac{\partial F}{\partial k_3}\frac{\partial K}{\partial x_3} - \frac{\partial F}{\partial t}\frac{\partial K}{\partial \omega} + \frac{\partial F}{\partial \omega}\frac{\partial K}{\partial t} \qquad (6.6)$$

The right-hand side of Eq. (6.6) is called the Poisson bracket of F and K, and it is noted by $\{F, K\}$. With this notation Eq. (6.6) can be written as $dF/ds = \{F, K\}$. Thus, a function F is a first integral of the Hamiltonian system when $\{F, K\} = 0$. It is easy to check that $\{K, K\} = 0$ and thus to conclude that the solutions of the system of Eq. (6.4) are contained in hypersurfaces K = constant.

Not all the solutions of Eq. (6.4) represent ray trajectories. The ray trajectories in the phase space $x_1 - x_2 - x_3 - t - k_1 - k_2 - k_3 - \omega$ are only the solutions of this equation system that are consistent with $K = 0$—that is, the curves contained in the hypersurface $K = 0$.

The equation of this hypersurface $K = 0$ can be expressed in different ways. For instance, let $f(x)$ be any function such that $f(x) = 0$ only if $x = 0$. Then $f(K) = 0$ represents the same surface as $K = 0$. It can be easily seen that if $f(K)$ is used as the Hamiltonian function instead of K in Eq. (6.4), then the same ray trajectories are obtained (with different parameterization) provided that $df/dx \neq 0$ when $x = 0$. In particular, if we multiply the Hamiltonian function by a nonzero function, the solutions of the Hamiltonian system remain the same but with different parameterization, that is, instead of getting $x_1 = x_1(s)$, $x_2 = x_2(s)$, $x_3 = x_3(s)$, $t = t(s)$, $k_1 = k_1(s)$, $k_2 = k_2(s)$, $k_3 = k_3(s)$, $\omega = \omega(s)$, we would get another set $x_1 = x_1(s')$, $x_2 = x_2(s')$, $x_3 = x_3(s')$, $t = t(s')$, $k_1 = k_1(s')$, $k_2 = k_2(s')$, $k_3 = k_3(s')$, $\omega = \omega(s')$ but still giving the same phase space trajectories.

One useful property fulfilled by the solutions of the Hamiltonian system is given by the Maupertius principle (in mechanics), also known as the least action principle, which corresponds to the Fermat's principle of optics (Arnold, 1974). In terms of the Hamiltonian system of Eq. (6.4) this principle says that the integral

$$\int_B^A k_1 dx_1 + k_2 dx_2 + k_3 dx_3 - \omega dt \qquad (6.7)$$

along the ray trajectories in the space x_1, x_2, x_3, t is an extremal among all the curves connecting point A and point B that also fulfill $K = 0$—that is, among the curves whose trajectory in the phase space $x_1 - x_2 - x_3 - t - k_1 - k_2 - k_3 - \omega$ is contained in the hypersurface $K = 0$. A and B are two points of the space x_1, x_2, x_3, t.

In other words, choose any curve of the space x_1, x_2, x_3, t connecting A and B. Now choose arbitrary functions $k_1 = k_1(x_1, x_2, x_3, t)$, $k_2 = k_2(x_1, x_2, x_3, t)$, $k_3 = k_3(x_1, x_2, x_3, t)$, $\omega = \omega(x_1, x_2, x_3, t)$ such that the Hamiltonian K vanishes along the curve. If these functions are compatible with the solution of the Hamiltonian system of Eq. (6.4), then the integral in Eq. (1.7) is an extremal among the other possible choices (Arnold, 1974). Observe that there is no restriction on the relationship of k_j and dx_j/dt in this way to establish Fermat's principle in contrast with the usual way to present it. Nevertheless, it can be proved that both ways to present the principle are equivalent (Arnold, 1974).

We shall restrict the analysis to time-invariant isotropic media. In this case, the surface of the wave vectors is a simple equation

$$K \equiv k_1^2 + k_2^2 + k_3^2 - \frac{\omega^2 n^2(x_1, x_2, x_3, \omega)}{c_o^2} = 0 \tag{6.8}$$

where c_o is the light velocity in vacuum and $n(x_1, x_2, x_3, \omega)$ is the refractive index at the point x_1, x_2, x_3 for the angular frequency ω (see Arnaud, 1976, and Kline and Kay, 1965, for obtaining the Hamiltonian function in other cases). Because the media is time-invariant, the Hamiltonian function does not depend on t and thus the last equation of the Hamiltonian system Eq. (6.4) expresses that ω is invariant along any ray trajectory ($d\omega/ds = 0$). Thus, ω is a first integral of the Hamiltonian system in this case.

If $\omega = $ constant and we are not interested in the dependence of t with the parameter s, then we only need the first six equations of the system. Furthermore, if we make the change of variables $p_j = k_j \cdot c_o/\omega$ a new Hamiltonian system is obtained

$$\begin{aligned}
\frac{dx_1}{ds} &= \frac{\partial P}{\partial p_1} & \frac{dp_1}{ds} &= -\frac{\partial P}{\partial x_1} \\
\frac{dx_2}{ds} &= \frac{\partial P}{\partial p_2} & \frac{dp_2}{ds} &= -\frac{\partial P}{\partial x_2} \\
\frac{dx_3}{ds} &= \frac{\partial P}{\partial p_3} & \frac{dp_3}{ds} &= -\frac{\partial P}{\partial x_3}
\end{aligned} \tag{6.9}$$

where the parameterization s now is not the same as before. The system of Eq. (6.9) is also a Hamiltonian system for the independent variables x_1, x_2, x_3, p_1, p_2, p_3 (the last three variables are the conjugate variables of the first three ones). Again, the ray trajectories are only the solutions of the system of Eq. (6.9) that are consistent with $P(x_1, x_2, x_3, p_1, p_2, p_3) = 0$, being the Hamiltonian function P

$$P \equiv p_1^2 + p_2^2 + p_3^2 - n^2(x_1, x_2, x_3, \omega) \tag{6.10}$$

Observe that now ω is a constant and thus an independent analysis can be done for each value of ω. The variables p_1, p_2, p_3 are called the optical direction cosines of a ray—that is, p_1 is $n(x_1, x_2, x_3)$ times the cosine of the angle formed by the tangent to the ray trajectory with respect to the x_1 axis (p_2 and p_3 are defined in a similar way with respect to the x_2 axis and the x_3 axis).

The Poisson bracket is defined in a similar way as before and the total derivative of a function $F(x_1, x_2, x_3, p_1, p_2, p_3)$ along the trajectories can also be written as

$$\frac{dF}{ds} = \{F,P\} = \frac{\partial F}{\partial x_1}\frac{\partial P}{\partial p_1} - \frac{\partial F}{\partial p_1}\frac{\partial P}{\partial x_1} + \frac{\partial F}{\partial x_2}\frac{\partial P}{\partial p_2} - \frac{\partial F}{\partial k_2}\frac{\partial P}{\partial p_2} + \frac{\partial F}{\partial x_3}\frac{\partial P}{\partial p_3} - \frac{\partial F}{\partial p_3}\frac{\partial P}{\partial x_3} \quad (6.11)$$

Thus, if F is a first integral of the Hamiltonian system, then it must fulfill $\{F, P\} = 0$.

When the Hamiltonian function is a first integral (and it is so in all the formulations that we have shown), then a new Hamiltonian system with two fewer variables can be built up, provided that the equation $P = 0$ can be solved for one variable (Arnold, 1974). Assume that this variable is p_3. Then, the new formulation of Hamilton equations is

$$\begin{aligned}
\frac{dx_1}{dx_3} &= \frac{\partial H}{\partial p_1} & \frac{dp_1}{dx_3} &= -\frac{\partial H}{\partial x_1} \\
\frac{dx_2}{dx_3} &= \frac{\partial H}{\partial p_2} & \frac{dp_2}{dx_3} &= -\frac{\partial H}{\partial x_2}
\end{aligned} \quad (6.12)$$

The ray trajectories are now the solutions of the system, without restriction to $H = 0$. The parameter of these ray trajectories is x_3—that is, the conjugate variable of p_3 in the system of Eq. (1.9). The function H is $H = -p_3$ when solved from the equation $P = 0$—that is,

$$H \equiv -\sqrt{n^2(x_1, x_2, x_3, \omega) - p_1^2 - p_2^2} \quad (6.13)$$

Eqs. (6.12) and (6.13) are the usual way in which Hamiltonian equations are introduced in optics (Luneburg, 1964). Nevertheless, we won't use it. For our purposes, Eq. (6.9) with the condition $P = 0$ is a more convenient way to set the basic equations of Geometrical Optics.

Before going further, we still need a last system of Hamilton equations. This is the one obtained when a change of variables from x_1, x_2, x_3 to a new set of orthogonal coordinates i_1, i_2, i_3 is done. This transformation belongs to a class of variable transformations called canonical (Leech, 1958), and owing to this fact, the Hamilton equations remain very similar (Leech, 1958; Miñano, 1986). Canonical transformations are characterized by a "generating function" G. For our purposes the expression of G is

$$G(x_1, x_2, x_3, u_1, u_2, u_3) = u_1 i_1(x_1, x_2, x_3) + u_2 i_2(x_1, x_2, x_3) + u_3 i_3(x_1, x_2, x_3) \quad (6.14)$$

where the functions i_1, i_2, i_3 in Eq. (6.14) give the values of the coordinates i_1, i_2, i_3 for a point x_1, x_2, $x_3 \cdot u_1$, u_2, u_3 are the conjugate variables of i_1, i_2, i_3. According to the canonical transformation theory, the new conjugate variables can be expressed as

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} \dfrac{\partial i_1}{\partial x_1} & \dfrac{\partial i_2}{\partial x_1} & \dfrac{\partial i_3}{\partial x_1} \\ \dfrac{\partial i_1}{\partial x_2} & \dfrac{\partial i_2}{\partial x_2} & \dfrac{\partial i_3}{\partial x_2} \\ \dfrac{\partial i_1}{\partial x_3} & \dfrac{\partial i_2}{\partial x_3} & \dfrac{\partial i_3}{\partial x_3} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \quad (6.15)$$

The resulting Hamiltonian system is
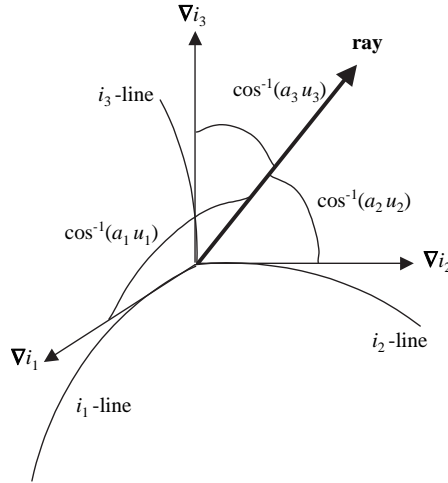
Figure 6.15   Physical meaning of the conjugate variables u_i.

$$\frac{di_1}{ds} = \frac{\partial H}{\partial u_1} \quad \frac{du_1}{ds} = -\frac{\partial H}{\partial i_1}$$
$$\frac{di_2}{ds} = \frac{\partial H}{\partial u_2} \quad \frac{du_2}{ds} = -\frac{\partial H}{\partial i_2} \qquad (6.16)$$
$$\frac{di_3}{ds} = \frac{\partial H}{\partial u_3} \quad \frac{du_3}{ds} = -\frac{\partial H}{\partial i_3}$$

and the Hamiltonian function is

$$H \equiv u_1^2 a_1^2(i_1, i_2, i_3) + u_2^2 a_2^2(i_1, i_2, i_3) + u_3^2 a_3^2(i_1, i_2, i_3) - 1 \qquad (6.17)$$

where a_1, a_2, and a_3 are, respectively, the modulus of the gradient of i_1, i_2, and i_3 over the refractive index n (i.e., $a_j = |\nabla i_j|/n$). Remembering the expressions of the scale factors h_j (Weisstein, 1999) of Differential Geometry, we can write $a_j = 1/(h_j\, n)$. The refractive index n is in general a function of i_1, i_2, i_3.

With the aid of Eq. (6.15) it is easy to find the physical meaning of the conjugate variables u_i: A point i_1, i_2, i_3, u_1, u_2, u_3 of the new phase space represents a ray passing by the point i_1, i_2, i_3 with optical direction cosines $a_1 u_1$, $a_2 u_2$, $a_3 u_3$ with respect to the three orthogonal vectors ∇i_1, ∇i_2, ∇i_3. Figure 6.15 shows these three orthogonal vectors and an arbitrary ray. The i_1 lines are given by equations $i_2 = $ constant, $i_3 = $ constant. The i_2, i_3 lines are defined in a similar way.

6.12.3   Optical Path Length

With the information provided in Figure 6.15 it is easy to see that the differential of path length dL can be written as

$$dL = \frac{di_1}{a_1^2 u_1} = \frac{di_2}{a_2^2 u_2} = \frac{di_3}{a_3^2 u_3} = \frac{u_1 di_1 + u_2 di_2 + u_3 di_3}{a_1^2 u_1^2 + a_2^2 u_2^2 + a_3^2 u_3^2} \qquad (6.18)$$

Taking into account Eq. (6.17), the optical path length L_{AB} of a ray is given by the integral of Eq. (6.7) applied to our problem—that is,

$$L_{AB} = \int_B^A u_1 di_1 + u_2 di_2 + u_3 di_3 \tag{6.19}$$

This integral is evaluated along a ray trajectory in the phase space. With Eq. (6.16) we get

$$L_{AB} = \int_B^A \left(u_1 \frac{di_1}{ds} + u_2 \frac{di_2}{ds} + u_3 \frac{di_3}{ds} \right) ds = \int_B^A \left(u_1 \frac{\partial H}{\partial u_1} + u_2 \frac{\partial H}{\partial u_2} + u_3 \frac{\partial H}{\partial u_3} \right) ds \tag{6.20}$$

Taking into account Eq. (6.17),

$$L_{AB} = \int_B^A 2(H+1)\, ds = 2 \int_B^A ds \tag{6.21}$$

Note that $H = 0$ for the ray trajectories. Eq. (6.21) provides the information we need to understand the physical meaning of ds: $\frac{1}{2}$ of the optical path length differential dL. It should be remembered that the parameterization of the ray trajectory, and thus the physical meaning of s, is associated with the Hamiltonian function we are using.

6.13   POISSON BRACKET DESIGN METHOD

The Poisson bracket design method is, as yet, one of the few known 3D nonimaging concentrator design methods. In general, this method provides concentrators requiring variable refractive index media, which is impractical in most of the cases. The main interest of the Poisson bracket method is that it provides ideal 3D concentrators, and thus it proved that such ideal concentrators exist. In particular, we will design a 3D maximal concentrator illuminated by a bundle of rays having an angular spread θ with respect the entry aperture's normal, that is, the set of rays that are concentrated are formed by all the rays that impinge a flat entry aperture forming an angle smaller than a certain value θ with the normal to this aperture. The concentrator has maximal concentration, and thus the ratio of entry to exit apertures areas is $n^2/\sin^2\theta$, where n is the refractive index of the points of the exit aperture, which is the same for all of them. Figure 6.16 shows a scheme of such a concentrator.

   The work presented here was developed some years ago (Miñano, 1985b; 1985c; Miñano, 1993a; 1993b; Miñano and Benítez, 1999). Some nontrivial ideal 3D nonimaging concentrators were already known when the Poisson brackets method was developed. Among these, the most important is the hyperboloid of revolution (Winston and Welford, 1979). Figure 6.17 shows one of these concentrators. A reflector whose cross-section is a hyperboloid forms it. The foci of this hyperboloid generate the circumference C when the cross-section is rotated around the axis of revolution symmetry. If the inner side of the hyperboloid of revolution is mirrored, then it becomes an ideal nonimaging concentrator with the following definitions of the input and output bundles: The input bundle is formed by all the rays crossing the entry aperture that would reach any point of the circle C (virtual receiver) if there was no mirror. The set of rays crossing the exit aperture forms the output bundle. The concentrator is ideal in the sense that any ray of the input
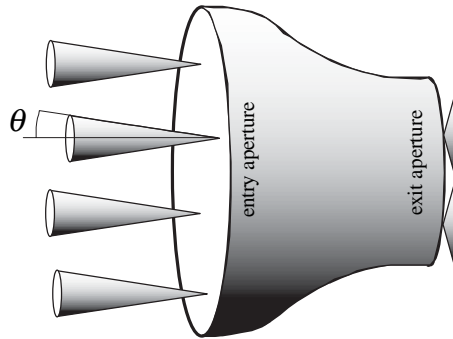
Figure 6.16   3D ideal concentrator designed to collect the rays impinging its entry aperture with directions within a cone of angle θ.
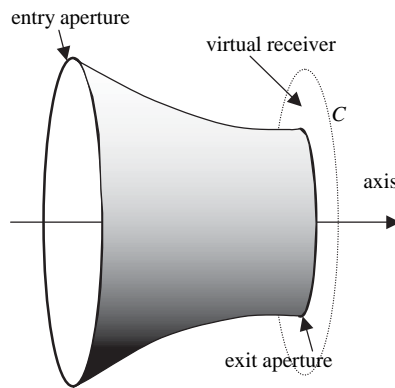
Figure 6.17   Hyperboloid of revolution as an ideal 3D concentrator.

bundle is transformed in a ray of the output bundle by the concentrator, and any ray of the output bundle comes from a ray of the input bundle. Thus, the same rays form both bundles. The only difference is that the input bundle describes the transmitted bundle at the entry aperture and the output bundle describes it at the exit aperture. Additionally, the concentrator has maximal concentration because the output bundle comprises all the rays crossing the exit aperture, and thus the exit aperture has the minimum possible area such that all the rays of the transmitted bundle cross it.

From the preceding definition of ideal concentrator we can conclude that any device may be an ideal concentrator with a proper definition of the input and output bundles. Nevertheless, the name "ideal" used to be restricted to cases in which both input and output bundles have a practical interest. There are two types of bundle that deserve special attention.

1. *Finite source.* The rays of this bundle are those linking any point of a given surface with any point of another given source (see Figure 6.18).
2. *Infinite source.* This bundle can be described as formed by all the rays that meet (real or virtually) a given surface forming an angle smaller than or equal
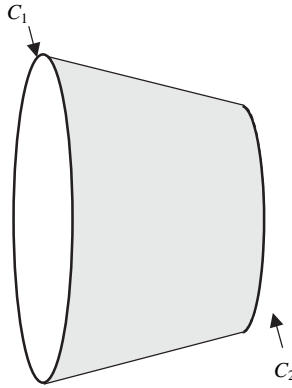
Figure 6.18 Example of finite source. The rays of this bundle are those linking any point of the circle C_1 with any point of the circle C_2.
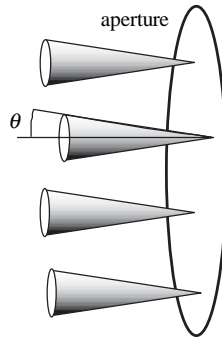
Figure 6.19   Example of infinite source of angle θ. It can also be considered as a particular case of the bundle shown in Figure 6.18 when one of the circles is infinitely far from the other and of infinite radius.

θ with a given reference direction. Then, this bundle is fully characterized by the surface (also called aperture), by the angle θ, and by the reference direction. This bundle is a typical input bundle for solar applications: The rays to be collected are those reaching the concentrator aperture forming an angle with the normal to this aperture smaller than the acceptance angle of the system (see Figure 6.19).

The input bundle of the hyperboloid of revolution of Figure 6.17 is a finite source where C_1 is the entry aperture and C_2 is the virtual receiver. The output bundle is an infinite source of the type shown in Figure 6.19 with $\theta = 90°$.

A thin lens with focal length f can be considered as a concentrator whose input bundle is an infinite source of angle θ and whose output bundle is a finite source of the type shown in Figure 6.18, C_1 being the lens aperture and C_2 being a circle located at the focal plane with radius equal to $f \cdot \tan(\theta)$. For a real lens this description is approximate. The approximation is better for smaller since θ is smaller. Therefore, a combination of a hyperboloid of revolution reflector and a thin lens
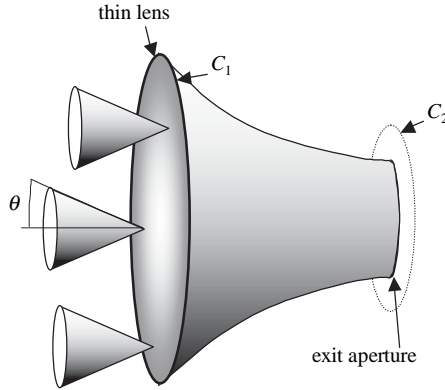
Figure 6.20   A thin lens combined with a hyperboloid of revolution behaves approximately like an ideal 3D concentrator with maximal concentration for an infinite source subtending an angle θ.
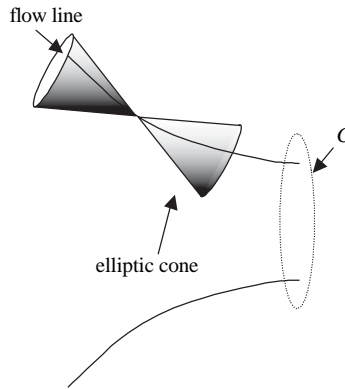
Figure 6.21   The edge rays of an elliptical bundle passing through a point form a cone with an elliptical basis.

is, approximately, an ideal concentrator of the type shown in Figure 6.16 (at least for small values of θ) (Welford, O'Gallagher, and Winston, 1987), if the combination is done in such a way that the output bundle of the thin lens, which is the finite source defined by the circles C_1 and C_2, is made to coincide with the input bundle of the hyperboloid (see Figure 6.20).

A characteristic of the hyperboloid of revolution as a nonimaging concentrator is that its transmitted bundle is what we call an elliptic bundle. An elliptic bundle is defined as one whose edge rays cross any point of the $x_1 - x_2 - x_3$ space form—in this space, a cone with an elliptic basis. Figure 6.21 shows one of these cones corresponding to the bundle of rays illuminating the circle C. This bundle is of the elliptic type, and thus the rays form an elliptic cone at any point of the space. The figure shows also two flow lines of this bundle.

If the elliptic bundle is such that its flow lines are the coordinate lines of a three-orthogonal coordinate system (i_1, i_2, i_3), then it can be easily proved that the edge rays conjugate variables u_1, u_2, u_3 fulfill an equation like

$$u_1^2\alpha_1^2(i_1,i_2,i_3) + u_2^2\alpha_2^2(i_1,i_2,i_3) + u_3^2\alpha_3^2(i_1,i_2,i_3) = 1 \qquad (6.22)$$

where the functions α_1, α_2, α_3 are arbitrary functions of i_1, i_2, i_3. This equation, together with Eq. (6.17) defines a conic curve (ellipse, parabola, or hyperbole) in a $u_n - u_m$ plane ($n, m = 1, 2, 3, n \neq m$). Note that it is necessary that Eq. (6.17) be fulfilled because the rays are the solutions of the Hamiltonian system that are consistent with $H = 0$, which is Eq. (6.17).

6.13.1   Statement of the Problem

The ray trajectories in the phase space $x_1 - x_2 - x_3 - p_1 - p_2 - p_3$ or ($i_1 - i_2 - i_3 - u_1 - u_2 - u_3$) do not cross between them. This property, which derives from the uniqueness of the solution of a system of first order differential equations passing through a given point of the phase space is particularly useful for describing visually the problem that we want to solve and comparing it with the typical synthesis problem in imaging optics. For the purpose of describing qualitatively both problems, we are going to consider a simplified case. This is when the rays are contained in a plane (for instance the $x_1 - x_2$). We call this case a 2D system, and it can be derived from the general case by establishing $\partial n/\partial x_3 = 0$ and $p_3 = 0$. In this case the phase space can be limited to four variables $x_1 - x_2 - p_1 - p_2$. Moreover, since the ray trajectories are restricted to $P = 0$ (P is defined in Eq. (6.10)), then $p_2 = \pm[n^2(x_1, x_2) - p_1^2]^{1/2}$—that is, for each point x_1, x_2, p_1 there are only two possible values of p_2 such that x_1, x_2, p_1, p_2 describes a ray. Both values of p_2 give the same ray path (x_2 increases with the parameter s for one value of p_2 and for the other value x_2 decreases with increasing s). Thus, if we forget one of the two possible directions of the ray, we can say that each ray can be fully characterized by a point x_1, x_2, p_1. Fortunately, these are only three variables, and the trajectories can be easily represented. For instance, Figure 6.22 shows the trajectory of a ray in the phase space $x_1 - x_2 - p_1$ and its projection on the $x_1 - x_2$ plane. This projection has the equation

$$x_1 = A\sin(Bx_2 + C) \qquad (6.23)$$

This ray trajectory is the one obtained for meridian rays in a fiber whose square of the refractive index has a parabolic profile versus x_1 (see Miñano, 1985b).

A one-parameter family of ray trajectories in the phase space forms, in general, a surface. For instance, consider the family of rays derived from Eq. (6.23) taken C as the parameter of the family (A and B are kept constants). The representation of this family in the phase space are the cylinders shown in Figure 6.23. The
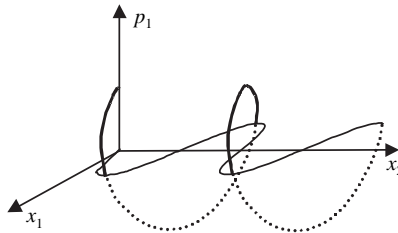
Figure 6.22   Ray trajectory in the phase space $x_1 - x_2 - p_1$ (heavy line) and in the $x_1 - x_2$ plane.
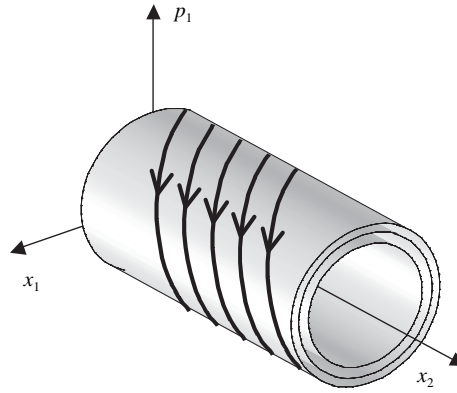
Figure 6.23   Surfaces of the phase space representing three one-parameter bundles of rays.
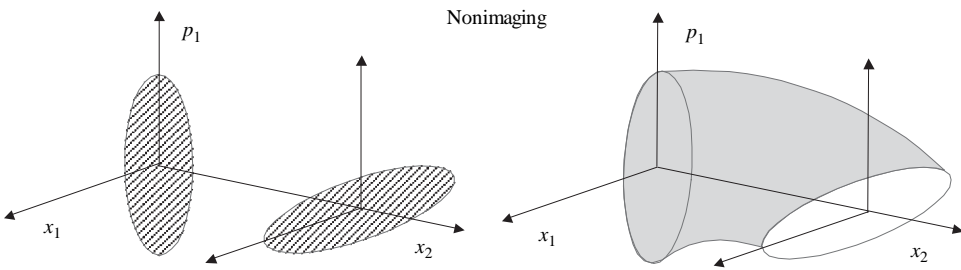
Figure 6.24   The nonimaging design problem: The bundles of rays at the entry aperture $(x_2 = 0)$ and at the exit aperture $(x_2 = 1)$ are prescribed (left side). An optical system has to be designed such that the edge rays at the entry are the same as the edge rays at the exit; that is, the edge ray trajectories in the phase space must form a surface connecting the edge rays' representations at both apertures.

ray trajectories in this phase space are wrapped around the cylinder, and they don't cross. Different cylinders correspond to different values of A and B.

Now let us consider the problem of designing a 2D nonimaging concentrator in the phase space. In general the problem involves determining the optical system such that a given bundle of rays described at a line called the entry aperture is transformed by the optical system in another prescribed bundle of rays at the exit aperture. Assume for simplicity that the entry aperture is at $x_2 = 0$ and that the exit aperture is at $x_2 = 1$. The bundle at the entry aperture can be defined by a region of the plane $x_2 = 0$, as well as the bundle at the exit aperture is defined by another region of the plane $x_2 = 1$ (see Figure 6.24 and Miñano, 1993a). Because of the conservation of étendue, both regions must have the same area.

The edge-ray principle simplifies the problem of design: To get the aforementioned goal we must design an optical system that transforms the edge rays of the bundle at the entry aperture in the edge rays of the bundle at the exit aperture. This is equivalent to stating that the edge rays' trajectories will form a tubelike surface in the phase space that cuts the $x_2 = 0$ plane and the $x_2 = 1$ plane at the contours of the regions defining the bundles at the entry and at the exit.
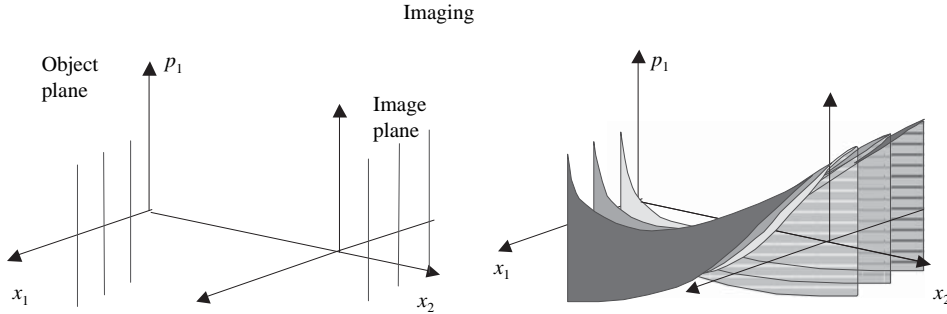
Imaging

Figure 6.25 The imaging problem: A family of one-parameter bundles of rays at the object plane ($x_2 = 0$) and another one at the image plane ($x_2 = 1$) are prescribed (left side). An optical system has to be designed such that each bundle at the object plane is imaged to its corresponding bundle at the image plane.

In general the imaging problem has less degrees of freedom. Figure 6.25 shows the phase space representation for this case. At the object and at the image plane there is a prescribed family of one-parameter bundles of rays to be coupled. The rays issuing or reaching a point of the object or imaging planes form each one of these one-parameter bundles. From the mathematical point of view, in the non-imaging problem we have to find an optical system that admits a given particular integral of the Hamilton equations, whereas in the image problem we have to find an optical system that admits a given first integral of the Hamilton equations.

Now let us go back to the 3D case. Let us call restricted entry phase space to the points of the phase space whose spatial coordinates x_1, x_2, x_3 belong to the entry aperture. The restricted exit phase space is defined in a similar way. In the non-imaging design the edge rays have prescribed descriptions in the restricted entry and exit phase spaces. The edge rays in the restricted phase spaces form a curve that encloses the set of points representing the rays of the transmitted bundle. Note that in the 3D case, the edge rays form a three-parameter bundle of rays, and thus their trajectories in the six-dimensional phase space ($x_1 - x_2 - x_3 - p_1 - p_2 - p_3$) form a four-dimensional subset that must be contained in the subset $P(x_1, x_2, x_3, p_1, p_2, p_3) = 0$. The subset $P = 0$ is then five-dimensional, and thus a four-dimensional subset can be characterized by an additional equation of the type $\omega(x_1, x_2, x_3, p_1, p_2, p_3) = 0$ (ω here has no relation with the angular frequency, which is not considered in this analysis). The surface of the edge rays' trajectories is then defined by $P = 0$ together with $\omega = 0$. The function $\omega(x_1, x_2, x_3, p_1, p_2, p_3)$ is not uniquely determined except when $\omega = 0$.

The question now is to find the conditions on the function ω so ω is a surface formed by ray trajectories. The answer is that the Poisson bracket of ω and P should be zero, when $P = 0$ and when $\omega = 0$—that is,

$$\{\omega, P\} = 0, \quad \text{when} \quad (\omega = 0 \quad \text{and} \quad P = 0) \tag{6.24}$$

Remember that the Poisson bracket is defined as

$$\{\omega, P\} = \sum_{j=1}^{j=3} \left(\frac{\partial \omega}{\partial x_j} \frac{\partial P}{\partial p_j} - \frac{\partial \omega}{\partial p_j} \frac{\partial P}{\partial x_j} \right) \tag{6.25}$$

Since the variable transformation from $x_1 - x_2 - x_3 - p_1 - p_2 - p_3$ to $i_1 - i_2 - i_3 - u_1 - u_2 - u_3$ is canonical, the problem can be easily established in the new variables: Eq. (6.24) becomes (now ω is a function of the variables $i_1, i_2, i_3, u_1, u_2, u_3$ obtained with the preceding transformation from the function $\omega(x_1, x_2, x_3, p_1, p_2, p_3)$)

$$\{\omega, H\} = 0, \quad \text{when} \quad (\omega = 0 \quad \text{and} \quad H = 0) \tag{6.26}$$

The Poisson bracket of the functions ω and H is expressed with the new variables as

$$\{\omega, H\} = \sum_{j=1}^{j=3} \left(\frac{\partial \omega}{\partial i_j} \frac{\partial H}{\partial u_j} - \frac{\partial \omega}{\partial u_j} \frac{\partial H}{\partial i_j} \right) \tag{6.27}$$

We have now new tools to proceed with the design problem. For instance, we can propose a function ω that fulfills the contour conditions at the restricted entry and exit phase spaces and apply Eq. (6.26) and find out the conditions on the Hamiltonian function H (see Eq. (6.17)) and thus the conditions on the refractive index distribution and on the modulus of the gradients of the coordinate variables (more precisely, the conditions on the functions $a_j = |\nabla i_j|/n$).

Because we are not completely free to choose the Hamiltonian function (because, for instance, its dependence with the squares of the conjugate variables must be linear if a three-orthogonal coordinate system is used), then choosing the function ω is not completely free either. In order to find the restrictions we must impose to the function ω, we must expand Eq. (6.27).

The problem is further simplified if we restrict the analysis to elliptic edge ray bundles that can be defined by a couple of equations $H = 0$ (H is defined in Eq. (6.17)) together with

$$\omega \equiv u_1^2 \alpha_1^2(i_1, i_2, i_3) + u_2^2 \alpha_2^2(i_1, i_2, i_3) + u_3^2 \alpha_3^2(i_1, i_2, i_3) - 1 = 0 \tag{6.28}$$

Observe that with this restriction, ω and H are both linear functions of the squares of the conjugate variables u_1, u_2, u_3. The definition of Eq. (6.28) implies that the cone formed by the rays of the bundle passing through any point i_1, i_2, i_3 has three planes of symmetry. Therefore, the flow lines of the bundle are one of the three coordinate lines. Thus, we are restricting elliptic bundles to those whose flow lines may be coordinate lines of a three-orthogonal system. This restriction implies, for instance, that the flow lines are orthogonal to a family of surfaces and thus that $\mathbf{J} \cdot \nabla \times \mathbf{J} = 0$ ($\mathbf{J}$ is the geometrical vector flux) which is not a necessary condition for $\mathbf{J}$. This restriction is not imposed in the analysis done further in this chapter using the Lorenz geometry tools, where, nevertheless, the refractive index is assumed to be constant.

The type of bundle given by Eq. (6.28) can be used as an edge ray bundle in the flow-line design method where the existence of a reflector surface connecting entry and exit apertures borders permits the edge ray bundle to be unbounded in the spatial variables i_1, i_2, i_3 (see Appendix B). This won't to be the case of the Simultaneous Multiple Surface design method described in Chapter 8.

Owing to the symmetries of the elliptic bundles, if we find a refractive index distribution that has as a solution a prescribed elliptic bundle, then we will easily be able to design a concentrator with the flow-line method. All we will have to do is choose a surface formed by flow lines of the bundle as reflector. In the definition of elliptic bundles given by Eq. (6.28) it is implicitly established that one of the

three coordinate lines are flow lines of the bundle and that the plane tangents to the surfaces i_j = constant (j = 1, 2, 3) at each point i_1, i_2, i_3 are planes of symmetry of the bundle at this point. Assume, for instance, that the coordinate lines i_1 (i.e., the lines i_2 = constant i_3 = constant) are the flow lines of interest. Then, any surface i_2 = constant will contain flow lines and its tangents are planes of symmetry of the bundle. Then, if the surface i_2 = constant is mirrored, the bundle result is unaffected (from the collection point of view), and thus we will obtain the desired concentrator provided the surface i_2 = constant defines the entry and exit aperture of the concentrator according to our requirements. The concentrator will be formed by a mirror with the i_2 = constant surface shape and the refractive index distribution. The mirror edges will define the shape of the entry and exit apertures. Additionally we know that the rotational hyperbolic concentrator is a solution of this type, and thus we know that there is at least one solution for this problem.

The analysis of elliptic bundles may appear too restrictive, and that may be so. Nevertheless, we have to take into account that the edge-ray bundle of any nontrivial ideal 3D concentrator known at present is an elliptic bundle. Moreover, the concept of an elliptic bundle can also be applied in 2D geometry (see Section 6.16), and in this case any edge-ray bundle can be viewed as an elliptic bundle. Since the majority of design methods of nonimaging concentrators (other than the Poisson brackets and the numerical methods) are actually 2D methods, we can conclude that the concept of elliptic bundle is not so restrictive.

6.13.2 Elliptic Edge-Ray Bundle Analysis

Let us define the following vectors.

$$\mathbf{u} \equiv \left[u_1^2, u_2^2, u_3^2\right] \tag{6.29}$$

$$\mathbf{a} \equiv \left[a_1^2, a_2^2, a_3^2\right] \tag{6.30}$$

$$\boldsymbol{\alpha} \equiv \left[\alpha_1^2, \alpha_2^2, \alpha_3^2\right] \tag{6.31}$$

The vectors a and α depend solely on i_1, i_2, i_3. Using this notation, equations $H = 0$, $\omega = 0$ remain as

$$H \equiv \mathbf{u} \cdot \mathbf{a} - 1 = 0 \tag{6.32}$$

$$\omega \equiv \mathbf{u} \cdot \boldsymbol{\alpha} - 1 = 0 \tag{6.33}$$

Note that for a given edge-ray bundle—that is, for a given surface ($H = 0$, $\omega = 0$)—the vector α is not uniquely determined: Any vector

$$\alpha_1 = \alpha + \mu(\mathbf{a} - \alpha) \tag{6.34}$$

where μ is an arbitrary parameter $\mu \neq 1$ defines the same edge-ray bundle. Eqs. (6.32) and (6.33) must be independent—that is, $a \times \alpha \neq 0$.

The equation $\{\omega, H\} = 0$ can be written as

$$\{\omega, H\} \equiv \mathbf{u} \cdot \sum_{j=1}^{j=3} \left(\frac{\partial \boldsymbol{\alpha}}{\partial i_j} \left(\frac{\partial \mathbf{u}}{\partial u_j} \cdot \mathbf{a} \right) - \frac{\partial \mathbf{a}}{\partial i_j} \left(\frac{\partial \mathbf{u}}{\partial u_j} \cdot \boldsymbol{\alpha} \right) \right) = 0 \tag{6.35}$$

Because only the power 2 of u_1, u_2, u_3 appears in the equations $\omega = 0$ and $H = 0$, then if there is a point (i_1, i_2, i_3, u_1, u_2, u_3) belonging to $\omega = 0$ and $H = 0$ it is nec-

essary that the points $(i_1, i_2, i_3, \pm u_1, \pm u_2, \pm u_3)$ also belong to $\omega = 0$ and $H = 0$. Thus, if $\{\omega, H\} = 0$ at $(i_1, i_2, i_3, u_1, u_2, u_3)$, it is necessary that $\{\omega, H\} = 0$ also at $(i_1, i_2, i_3, \pm u_1, \pm u_2, \pm u_3)$. This means that the factors of odd powers of u_1, u_2, and u_3 in Eq. (6.35) must be zero. From this result we obtain three equations instead of Eq. (6.35).

$$\mathbf{u} \cdot \left(a_j^2 \frac{\partial \boldsymbol{\alpha}}{\partial i_j} - \alpha_j^2 \frac{\partial \mathbf{a}}{\partial i_j} \right) \equiv \mathbf{u} \cdot \mathbf{T}_j = 0 \quad j = 1, 2, 3 \tag{6.36}$$

These three equations together with Eqs. (6.32) and (6.33) should be satisfied by a one-parametric set of vectors u at each point i_1, i_2, i_3. Eqs. (6.32) and (6.33) are assumed to be independent—that is, $a \times \boldsymbol{\alpha} \neq 0$. Thus, any of Eq. (6.36) must be a linear combination of Eqs. (6.32) and (6.33)—that is, any $\boldsymbol{T}_j$ must be parallel to $a - \boldsymbol{\alpha}$:

$$\mathbf{T}_j \times (\mathbf{a} - \boldsymbol{\alpha}) = 0 \quad j = 1, 2, 3 \tag{6.37}$$

From this result we get six differential equations involving the functions a_j and α_j, which, surprisingly, can be combined so they become easily integrated:

$$\frac{\partial}{\partial i_j} \left(\frac{a_j^2 \alpha_k^2 - \alpha_j^2 a_k^2}{a_j^2 - \alpha_j^2} \right) = 0 \quad j, k = 1, 2, 3 \quad k \neq j \tag{6.38}$$

$$\frac{a_j^2 - \alpha_j^2}{a_j^2 \alpha_k^2 - \alpha_j^2 a_k^2} = f_{j,k} \quad j, k = 1, 2, 3 \quad k \neq j \tag{6.39}$$

Where the functions $f_{j,k}$ fulfill that $\partial f_{j,k} / \partial i_j = 0$.

Eq. (6.47) can also be written as

$$[\mathbf{M}] \cdot \mathbf{a} = 1 \quad [\mathbf{M}] \cdot \boldsymbol{\alpha} = 1 \tag{6.40}$$

Where 1 is the column vector $(1, 1, 1)$ and $[\mathbf{M}]$ is the matrix

$$[\mathbf{M}] = \begin{bmatrix} 0 & f_{3,2} & f_{2,3} \\ f_{3,1} & 0 & f_{1,3} \\ f_{2,1} & f_{1,2} & 0 \end{bmatrix} \tag{6.41}$$

Since the matrix Eq. (45) must be satisfied by the vectors a and $\boldsymbol{\alpha}$, and these vectors fulfill $a \times \boldsymbol{\alpha} \neq 0$, then the determinant of $[\mathbf{M}]$ must be zero (this equation could be got directly from Eq. (6.39))

$$f_{1,3}(i_2, i_3) f_{2,1}(i_1, i_3) f_{3,2}(i_1, i_2) + f_{1,2}(i_2, i_3) f_{2,3}(i_1, i_3) f_{3,1}(i_1, i_2) = 0 \tag{6.42}$$

where the dependence on the variables i_1, i_2, i_3 has been explicitly written. Let us rearrange this equation as follows:

$$-\frac{f_{3,1}(i_1, i_2)}{f_{3,2}(i_1, i_2)} = \frac{f_{1,3}(i_2, i_3)}{f_{1,2}(i_2, i_3)} \frac{f_{2,1}(i_1, i_3)}{f_{2,3}(i_1, i_3)} \tag{6.43}$$

Eq. (6.43) is a functional equation in which the left-hand side does not depend on i_3. If we fix a value of i_3 in the right-hand side, we get an equation that establishes that the left-hand side is a product of a function of i_1 times a function of i_2 (see Castillo, 1996; and Castillo and Ruiz-Cobo, 1992). Therefore, we can write the following equations:

$$\frac{n_1(i_1)}{n_2(i_2)} = -\frac{f_{3,1}(i_1,i_2)}{f_{3,2}(i_1,i_2)} \quad \frac{n_2(i_2)}{n_3(i_3)} = -\frac{f_{1,2}(i_2,i_3)}{f_{1,3}(i_2,i_3)} \quad \frac{n_3(i_3)}{n_1(i_1)} = -\frac{f_{2,3}(i_1,i_3)}{f_{2,1}(i_1,i_3)} \tag{6.44}$$

where n_1, n_2, n_3 are three functions that only depend on one variable as specified; that is,

$$\frac{\partial n_j}{\partial i_k} = 0 \quad j,k = 1, 2, 3 \quad j \neq k \tag{6.45}$$

Let us introduce the functions s_1, s_2, s_3 as

$$\begin{aligned} s_1(i_2,i_3) &= -f_{1,2}(i_2,i_3)n_3(i_3) \\ s_2(i_1,i_3) &= -f_{2,3}(i_1,i_3)n_1(i_1) \\ s_3(i_1,i_2) &= -f_{3,1}(i_1,i_2)n_2(i_2) \end{aligned} \tag{6.46}$$

that is, s_j fulfills

$$\frac{\partial s_j}{\partial i_j} = 0 \quad j = 1, 2, 3 \tag{6.47}$$

With these new functions Eq. (6.40) can be written as

$$\mathbf{a} \times \mathbf{S} = \boldsymbol{\alpha} \times \mathbf{S} = \mathbf{N} \tag{6.48}$$

where the vectors $\mathbf{S}$ and $\mathbf{N}$ are

$$\mathbf{S} = [s_1(i_2,i_3), s_2(i_3,i_1), s_3(i_1,i_2)] \tag{6.49}$$

$$\mathbf{N} = [n_1(i_1), n_2(i_2), n_3(i_3)] \tag{6.50}$$

From Eq. (6.48) it is concluded that

$$\mathbf{S} \cdot \mathbf{N} = 0 \tag{6.51}$$

Expanding Eq. (6.51) with the definitions in Eqs. (6.49) and (6.50) and dividing over $n_1(i_1) \, n_2(i_2) \, n_3(i_3)$ we get a functional equation called the generalized Sincov's equation (Castillo, 1996; Castillo and Ruiz-Cobo, 1992). Its solution is

$$\mathbf{S} = \mathbf{N} \times \mathbf{V} \tag{6.52}$$

where $\mathbf{V}$ is a vector whose components fulfill

$$\mathbf{V} = [v_1(i_1), v_2(i_2), v_3(i_3)] \tag{6.53}$$

Introducing the solution in Eq. (6.48) leads to

$$\mathbf{a} \cdot \mathbf{N} = \boldsymbol{\alpha} \cdot \mathbf{N} = 0 \tag{6.54}$$

$$\mathbf{a} \cdot \mathbf{V} = \boldsymbol{\alpha} \cdot \mathbf{V} = 1 \tag{6.55}$$

Thus, if an elliptic bundle defined by Eq. (6.33) exists in a medium characterized by the Hamiltonian of Eq. (6.32), then there must be vectors $\mathbf{V}$ and $\mathbf{N}$ of the type shown in Eq. (6.50) and Eq. (6.53) fulfilling Eqs. (6.54) and (6.55).

From these equations it is concluded that the vector $a - \boldsymbol{\alpha}$ is parallel to $\mathbf{V} \times \mathbf{N}$. Therefore, taking into account Eq. (6.34), we get that different edge-ray bundles should have different directions of the vector $\mathbf{V} \times \mathbf{N}$.

Eqs. (6.32) and (6.33) together with Eqs. (6.54) and (6.55) give the following result concerning the values of u_1, u_2, u_3 of the edge rays

$$\mathbf{u} = \mathbf{V} + \lambda \mathbf{N} \tag{6.56}$$

where λ is a parameter. Note that we can choose any new vector V_n such that $V_n = V_o + k_1 N$, where k_1 is an arbitrary constant. We can also choose a new vector N_n such that $N_n = k_2 N_o$, where k_2 is another arbitrary constant ($k_2 \neq 0$). These changes only affect to the range of values of λ giving positive components of u.

Keeping constant the coordinates of the point i_1, i_2, i_3 and varying λ in Eq. (6.56), we can get the different values of the edge rays passing through i_1, i_2, i_3. Eq. (6.56) can also be written as

$$\frac{u_1^2 - v_1(i_1)}{n_1(i_1)} = \frac{u_2^2 - v_2(i_2)}{n_2(i_2)} = \frac{u_3^2 - v_3(i_3)}{n_3(i_3)} = \lambda \tag{6.57}$$

From Eq. (6.57) we can obtain the equation of the edge rays at a given point (i_1, i_2, i_3) and verify the conic shape in the $u_1 - u_2$, $u_1 - u_3$, or $u_2 - u_3$ planes. For instance, we can get:

$$\frac{u_1^2}{\left[\dfrac{n_2(i_2)v_1(i_1) - n_1(i_1)v_2(i_2)}{n_2(i_2)}\right]} + \frac{u_2^2}{\left[\dfrac{n_1(i_1)v_2(i_2) - n_2(i_2)v_1(i_1)}{n_2(i_2)}\right]} = 1 \tag{6.58}$$

Eq. (6.58) shows that the expression of the edge rays in terms of $u_1 - u_2$ is invariant when we move along an i_3-line—that is, a line where i_1 = constant, i_2 = constant. A similar result can be obtained for the other couples of variables.

6.13.3   Decomposition of the Edge-Ray Bundle in Normal Congruences

This section shows some properties of the parameter λ appearing in Eq. (6.56). These properties are not necessary to understand the design procedure. Using Eq. (6.57) we can obtain λ as a function of i_1, u_1 (or as a function of i_2, u_2 or as a function of i_3, u_3). Expressing i_1, u_1 as functions of the parameter s along the ray trajectories—that is, $i_1 = i_1(s)$ and $u_1 = u_1(s)$—we can easily evaluate the variation of λ along the trajectory; that is, we can evaluate $d\lambda/ds$

$$\frac{d\lambda}{ds} = \frac{\partial \lambda}{\partial i_1}\frac{di_1}{ds} + \frac{\partial \lambda}{\partial u_1}\frac{du_1}{ds} = \frac{\partial \lambda}{\partial i_1}\frac{\partial H}{\partial u_1} - \frac{\partial \lambda}{\partial u_1}\frac{\partial H}{\partial i_1} \tag{6.59}$$

Let's take the first component of Eq. (6.56)

$$u_1^2 = v_1(i_1) + \lambda n_1(i_1) \tag{6.60}$$

Derivation of this expression with respect to i_1 and u_1 gives

$$2u_1 = \frac{\partial \lambda}{\partial u_1} n_1 \tag{6.61}$$

$$0 = \frac{dv_1}{di_1} + \frac{\partial \lambda}{\partial i_1} n_1 = \lambda \frac{dn_1}{di_1} \tag{6.62}$$

Derivation of Eq. (6.32) with respect to i_1 and u_1 gives

$$\frac{\partial H}{\partial u_1} = 2u_1 a_1^2 \tag{6.63}$$

$$\frac{\partial H}{\partial i_1} = \mathbf{u} \cdot \frac{\partial \mathbf{a}}{\partial i_1} \tag{6.64}$$

Derivation of Eqs. (6.54) and (6.55) gives

$$\mathbf{N} \cdot \frac{\partial \mathbf{a}}{\partial i_1} = -a_1^2 \frac{dn_1}{di_1} \tag{6.65}$$

$$\mathbf{V} \cdot \frac{\partial \mathbf{a}}{\partial i_1} = -a_1^2 \frac{dv_1}{di_1} \tag{6.66}$$

Restricting the values of u in Eq. (6.64) to those of the edge-ray bundle—that is, introducing Eq. (6.56) in Eq. (6.64) and using Eqs. (6.65) and (6.66), we get

$$\left. \frac{\partial H}{\partial i_1} \right|_{\mathbf{u}=\mathbf{V}+\lambda\mathbf{N}} = -a_1^2 \left(\frac{dv_1}{di_1} + \lambda \frac{dn_1}{di_1} \right) \tag{6.67}$$

Introducing Eqs. (6.61), (6.62), (6.63), and (6.67) in Eq. (6.59), we obtain $d\lambda/ds = 0$ for the edge rays, so the value of λ along each edge ray is constant. λ is an invariant (a constant of motion) for the edge rays, but it is not for the remaining rays. This means that there is a function $\omega_2(i_1, i_2, i_3, u_1, u_2, u_3)$ such that it is a first integral of the Hamiltonian system (a constant of motion) and such that $\omega_2 = \lambda$ for the points of the phase space region $H = 0$, $\omega = 0$ (this is the phase space hypersurface formed by the trajectories of the edge rays).

Moreover, each value of λ defines a normal congruence (or orthotomic system) of rays. A normal congruence of rays is a two-parametric set of rays for which there is a family of surfaces (the wavefronts) normal to the trajectories of the rays (in the space i_1, i_2, i_3) (Starroudis, 1972)—that is, there is a function $\Phi(i_1, i_2, i_3)$ such that its gradient $\nabla\Phi$ is a vector giving the optical direction cosines of the ray with respect to the three unit vectors i_1, i_2, i_3. These three optical direction cosines are a_1u_1, a_2u_2, and a_3u_3. Thus, what we have to prove is that the curl of $n(a_1u_1, a_2u_2, a_3u_3)$ is zero—that is, $\nabla \times n(a_1u_1, a_2u_2, a_3u_3) = 0$. It can be easily checked that this expression is zero (see Weisstein, 1999, for the expression of the curl in three orthogonal curvilinear coordinates, remembering that a_j is the inverse of the corresponding scale factor times the refractive index n) when u_1, u_2, and u_3 are given by Eq. (6.56) with λ = constant. Another way to prove that the subset of edge rays defined by λ = constant is a normal congruence is to check that the differential of 2D étendue dE_{2P} is zero for this bundle (see Appendix D). dE_{2P} can be expressed as $dE_{2P} = di_1du_1 + di_2du_2 + di_3du_3$, which is clearly zero when each u_j is a function of i_j solely, as it is in Eq. (6.56) when λ = constant. Thus, we have got a decomposition of the edge rays in normal congruencies (or orthotomic systems) of rays.

The phase space trajectory of each of these normal congruences can be defined by the equations $H = 0$, $\omega = 0$, $\lambda = \lambda_0$. It is thus the intersection of the three integrals in a six-dimensional phase space. The Liouville theorem can then be applied, and the angle-action variables can be introduced. We will not pursue this subject further. The reader can find more details in Abraham and Marsden, 1978, and Arnold, 1974.

6.14   APPLICATION OF THE POISSON BRACKET METHOD

As an example of application of the preceding results we are going to solve two problems: First, we are going to design a concentrator with rotational symmetry and graded refractive index. Second, we are going to find the elliptic bundles associated to an elliptic system of coordinates with rotational symmetry. Because there is rotational symmetry in both examples, we will analyze the conditions for rotational symmetry first.

6.14.1   Rotational Symmetry

Let us restrict our problem to systems having rotational symmetry around an axis. Assume that i_2 is the angular coordinate around this axis, $i_2 = \theta$. The second component of vector a is $a_2 = |\nabla i_2|/n = 1/(n\,\rho)$, where ρ is the shortest distance from a point to the symmetry axis. Both $\rho(i_1, i_3)$ and the refractive index distribution $n(i_1, i_3)$ are not functions of θ. The conjugate variable of θ is the skew invariant h; $u_2 = h$. This can be verified with the definitions given in Figure 6.15. i_1 and i_3 are coordinates on a meridian plane. Then, neither a_1 nor a_3 depend on θ, and thus vector a does not depend on θ. We are interested in solutions having rotational symmetry, too, so vector $\boldsymbol{\alpha}$ does not depend on θ either. With Eqs. (6.50), (6.53), (6.54), and (6.55) it is concluded that the second components of V and N, (v_2 and n_2) are constants. Summarizing,

$$i_2 = \theta \quad u_2 = h \quad a_2 = \frac{1}{n\rho} \tag{6.68}$$

$$\frac{\partial \mathbf{a}}{\partial \theta} = \frac{\partial \boldsymbol{\alpha}}{\partial \theta} = \frac{\partial \mathbf{V}}{\partial \theta} = \frac{\partial \mathbf{N}}{\partial \theta} = 0 \tag{6.69}$$

Application of the preceding results to the second and the last terms of Eq. (6.57) establishes that λ is a function of h and thus that the invariance of λ along the edge rays is just a particular case of the skew invariant for rotational symmetric systems. Besides this, Eq. (6.57) contains two more equalities.

Because we can choose a new V as the sum of the old V and N, and we can choose a scale factor for N, then we can set without loss of generality that

$$v_2 = 0 \quad n_2 = -1 \tag{6.70}$$

6.14.2   Design of a Nonimaging Concentrator with Graded Refractive Index

The goal of this section is to design an ideal 3D nonimaging concentrator with flat entry and exit apertures, collecting the rays that impinge its entry aperture forming an angle smaller than β with the aperture normal, with maximal concentration, and such that the refractive index of the exit aperture points is n_x. The medium outside the concentrator has refractive index 1.
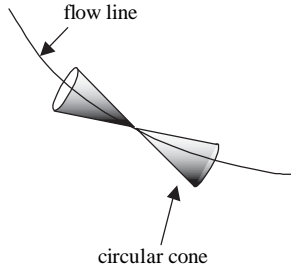
Figure 6.26 Circular cone of a circular bundle.

6.14.2.1 Circular Bundles

Because the possibility of designing the refractive index distribution provides enough degrees of freedom, we can impose additional requirements to the design. In particular we will require that the edge-ray bundle be circular. A circular bundle is a particular case of elliptic bundle in which the ray cones at any point have a circular base. Observe that any plane containing the tangent to a flow line at a given point is a plane of symmetry for the circular bundle at this point (see Figure 6.26). This property will allow us to design nonrotational symmetric concentrators once we have obtained the refractive index distribution.

Assume that the i_3 lines are the flow lines of the bundle (we have seen before that the flow lines are one of the coordinate lines). Then a circular edge-ray bundle can be characterized by an equation like this

$$\omega \equiv u_3^2 \alpha_3^2(i_1, i_2, i_3) - 1 = 0 \qquad (6.71)$$

Note that Eq. (6.71) expresses that the optical direction cosine of the edge rays with respect to the i_3 line is the constant at any point i_1, i_2, i_3—that is, they form a circular cone. Therefore, a circular bundle can be characterized by a vector $\boldsymbol{\alpha}$ in which $\alpha_1 = \alpha_2 = 0$.

Application of these results to Eqs. (6.54) and (6.55) gives

$$v_3(i_3) = \frac{1}{\alpha_3^2} \quad n_3 = 0 \qquad (6.72)$$

Note that the preceding result implies that α_3 depends only on i_3. Thus, $\boldsymbol{\alpha}$ is

$$\alpha \equiv [0, 0, \alpha_3^2(i_3)] \qquad (6.73)$$

6.14.2.2 Basic Equations

Combination of the circular bundle equations and those of rotational symmetric systems with the equalities in Eqs. (6.54) and (6.55) regarding vector a gives

$$a_1^2 n_1(i_1) - a_2^2 = 0 \qquad (6.74)$$

$$a_1^2 v_1(i_1) + a_3^2 v_3(i_3) = 1 \qquad (6.75)$$

Let's now express a_i as functions of the refractive index n and of the gradients of the coordinates ($a_j = |\nabla i_j|/n$) using the results in Eq. (6.68).

$$(\nabla i_1)^2 n_1(i_1) = \frac{1}{\rho^2} \tag{6.76}$$

$$(\nabla i_1)^2 v_1(i_1) + (\nabla i_3)^2 v_3(i_3) = n^2 \tag{6.77}$$

These equations do not contain any dependence with θ. The first one is an eikonal type equation, and it expresses that the lines $i_1 = $ constant (these are the i_3 lines because θ is also constant) have the shape of wave fronts in a media with refractive index $1/\rho$. The lines orthogonal to them—that is, the i_1 lines are thus rays in a medium with refractive index $1/\rho$. The rays in such media are circles with centers at the symmetry axis—that is, at $\rho = 0$ (Miñano, 1986). Since i_3 is constant in a i_1 line, we can define a function $R(i_3)$ giving the radius of the circle of each of these lines. The second equation, Eq. (6.77), will be used to calculate the refractive index distribution.

6.14.2.3   Contour Conditions

The contour conditions are the edge-ray bundle descriptions at the entry and exit apertures given before. From these descriptions the geometrical vector flux at both surfaces can be calculated. Since the flow lines are tangent to the geometrical vector flux, it can be concluded that the flow lines—the i_3 lines—must be normal to both surfaces. Therefore, two i_1 lines must coincide with the entry and exit apertures. We can choose the value of i_3 for each of these surfaces as $i_3 = 0$ for the entry aperture and $i_3 = 1$ for the exit aperture. Thus, we need the surfaces $i_3 = 0$ and $i_3 = 1$ to be flat. This result is compatible with the spherical shape of the $i_3 = $ constant surfaces imposed by Eq. (6.76). All we have to require is that

$$R(i_3 = 0) = R(i_3 = 1) = \infty \tag{6.78}$$

Using Eq. (6.56) and the previous results for the components of the vectors V and N of our problem we get

$$\begin{aligned} u_1^2 &= v_1(i_1) + \lambda n_1(i_1) \\ h^2 &= -\lambda \\ u_3^2 &= v_3(i_3) \end{aligned} \tag{6.79}$$

For maximal concentration, the direction cosine with respect to the i_3 lines must be 0 at the exit aperture, $i_3 = 1$ (see Figure 6.15 for the relationship between the conjugate variables and the direction cosines). Thus,

$$(\nabla i_3(i_1, i_3 = 1))^2 v_3(i_3 = 1) = 0 \tag{6.80}$$

Using Eq. (6.77) and remembering that the refractive index of the exit aperture points must be n_x, we get

$$(\nabla i_1(i_1, i_3 = 1))^2 v_1(i_1) = n_x^2 \tag{6.81}$$

With a similar reasoning at the entry aperture and taking into account that the medium outside the concentrator has refractive index 1 and so that there is a refraction at the entry aperture, we get

$$(\nabla i_3(i_1, i_3 = 0))^2 v_3(i_3 = 0) = n^2(i_3 = 0) - \sin^2 \beta \tag{6.82}$$

$$(\nabla i_1(i_1, i_3 = 0))^2 v_1(i_1) = \sin^2 \beta \tag{6.83}$$

6.14.2.4   Refractive Index Distribution Calculation

For the following calculations it is useful to introduce now the last cylindrical co-ordinate z. This coordinate z, as well as ρ, can be expressed as a function of i_1, i_3. Without loss of generality we can choose the coordinate i_3 so $i_3(\rho = 0) = z$. In the same way we can choose i_1 so $i_1(z = 0) = \rho$. With this selection, the equation of the spherical surfaces $i_3 = $ constant can be given as

$$z = i_3 + R(i_3) - \sqrt{R^2(i_3) - \rho^2} \tag{6.84}$$

This equation, which mixes cylindrical and $i_1 - i_3$ coordinates for simplicity, con-siders only the half sphere down the center. With it we can express $(\nabla i_3)^2$ as a func-tion of $R(i_3)$ and ρ:

$$\frac{\partial i_3}{\partial z} = \left(1 + R' - \frac{RR'}{\sqrt{R^2 - \rho^2}} \right)^{-1}$$

$$\frac{\partial i_3}{\partial \rho} = \frac{-\rho}{\sqrt{R^2 - \rho^2}} \left(1 + R' - \frac{RR'}{\sqrt{R^2 - \rho^2}} \right)^{-1} \tag{6.85}$$

$$(\nabla i_3)^2 = \left((1 + R') \sqrt{1 - \frac{\rho^2}{R^2}} - R' \right)^{-2}$$

We already know that the i_3 lines have the shape of wave fronts in a medium of refractive index $1/\rho$ and that $i_1(z = 0) = \rho$. The entry aperture is at $z = 0$, which coincides with $i_3 = 0$. With these data we can calculate i_1 as function of $R(i_3)$ and ρ:

$$i_1 = \frac{2R\rho \exp\left(\int_0^{i_3} \frac{di_3}{R} \right)}{R + \sqrt{R^2 - \rho^2}} \tag{6.86}$$

We can easily check that $i_1(z = 0) = i_1(i_3 = 0) = \rho$. Note that $R(i_3 = 0) = \infty$.

$(\nabla i_1)^2$ can be calculated by derivation of Eq. (6.86) with respect to ρ and z (using Eq. (6.85) for the partial derivatives with respect to i_3):

$$(\nabla i_1)^2 = \frac{i_1^2}{\rho^2} \tag{6.87}$$

When $i_3 = 1$, such as at the exit aperture (which also coincides with $z = 1$), $R(i_3 = 1) = \infty$ and thus

$$i_1|_{\text{exit aperture}} = \rho(i_1, i_3 = 1) \exp\left(\int_0^1 \frac{di_3}{R} \right) \tag{6.88}$$

Let us call $\sqrt{C_g}$ to the ratio of the entry and exit aperture diameters. Since $\rho \, i_1 \, i_3 = 0 = i_1$ (at the entry aperture) and the flow lines are $i_1 = $ constant, we conclude that $\sqrt{C_g}$ is the ratio of ρ at the entry and exit apertures for any $i_1 = $ constant line.

$$\frac{\rho(i_1, i_3 = 0)}{\rho(i_1, i_3 = 0)} = \exp\left(\int_0^1 \frac{di_3}{R} \right) = \sqrt{C_g} \tag{6.89}$$

From the étendue conservation we know that $\sqrt{C_g} = n_x/\sin\beta$. We shall check later that this equation is fulfilled.

Using Eqs. (6.76) and (6.87), we see

$$n_1(i_1) = \frac{1}{i_1^2} \tag{6.90}$$

Using Eqs. (6.81), (6.83), (6.86), and (6.87) and noting that $R \to \infty$ for the points of the entry and exit apertures, we get

$$v_1(i_1) = \frac{n_x^2}{\exp\left(2\int_0^1 \frac{di_3}{R}\right)} \tag{6.91}$$

$$v_1(i_1) = \sin^2 \beta \tag{6.92}$$

These two equations imply that $\sqrt{C_g} = n_x/\sin\beta$.

All we need now to obtain the refractive index distribution is to choose the functions $R(i_3)$ and $v_3(i_3)$. $R(i_3)$ has to fulfill Eqs. (6.78) and (6.89). We choose

$$R(i_3) = \frac{m}{i_3^2(1-i_3)^2} \tag{6.93}$$

where m is a constant ($m = 4/(15 \, ln(C_g))$), satisfying Eq. (6.89).

With $R(i_3)$ and with the aid of Eq. (6.85), we can calculate $(\nabla i_3)^2$ at the entry and exit apertures. The result is $(\nabla i_3)^2 = 1$ in both cases. The conditions on the function $v_3(i_3)$ can now be obtained with Eqs. (6.80) and (6.82)

$$v_3(i_3 = 1) = 0 \tag{6.94}$$

$$v_3(i_3 = 0) = n^2(i_3 = 0) - \sin^2 \beta \tag{6.95}$$

There are no more conditions on $v_3(i_3)$, so we still have degrees of freedom. We shall use these degrees of freedom to choose a refractive index equal to n_x for the points of the entry aperture ($i_3 = 0$ or $z = 0$) and for the points of the axis ($i_1 = 0$ or $\rho = 0$). We could impose this condition in any other line crossing all the range of values of i_3.

$$v_3(i_3) = n_x^2\left(1 - \frac{\sin^2 \beta}{n_x^2}\exp\left(2\int_0^{i_3} \frac{di_3}{R}\right)\right) \tag{6.96}$$

Finally, we can express the refractive index distribution as a function of R, ρ and i_3:

$$n^2 = \frac{(2R)^2 \sin^2 \beta \exp\left(2\int_0^{i_3} \frac{di_3}{R}\right)}{\left(R + \sqrt{R^2 - \rho^2}\right)} + \frac{n_x^2 - \sin^2 \beta \exp\left(2\int_0^{i_3} \frac{di_3}{R}\right)}{\left((1+R')\sqrt{1-\frac{\rho^2}{R^2}} - R'\right)^2} \tag{6.97}$$

Giving values to ρ and i_3, we can calculate z with Eq. (6.84), and thus we can calculate $n(\rho, z)$ with the preceding equation.

Figure 6.27 shows the resulting refractive index distribution for $n_x = 1.5$, $\beta = 30°$, a function $R(i_3)$ given by Eq. (6.93) and a function $v_3(i_3)$ such that the refractive index is equal to n_x at the points of the axis. Only values in the range $0 \le i_1 \le 0.3$ and $0 \le i_3 \le 1$ are shown. The refractive index is symmetric with respect to the z axis—that is, $n(i_1, i_3) = n(-i_1, i_3)$. A wider range of i_1 requires unrealistic values of the refractive index.
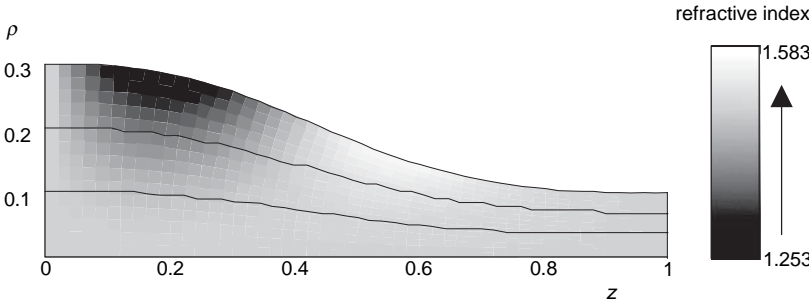
Figure 6.27   Refractive index distribution with rotational symmetry such that the rays impinging its entry aperture ($z = 0$) with an angle smaller than 30° find the exit aperture ($z = 1$). And any ray hitting the exit aperture finds the entry aperture at an angle smaller than 30°. ($n_x = 1.5$, $\beta = 30°$.) The refractive index at the points of both apertures ($z = 0$ and $z = 1$) and at the z-axis is equal to 1.5.

6.14.2.5   Reflector Surface

The last and easiest step to build up the concentrator is to define the reflector surface. The elliptic cones formed by the rays of an elliptic bundle at any point have two planes of symmetry that contains the cone axis. A reflector surface that has one of these two planes as a tangent plane at the referred point is a surface that does not disturb the flow of the bundle at this point. If the reflector surface fulfills this condition for all the points, then the bundle is not disturbed by the introduction of the reflector. With the definitions that we have used in the preceding sections, these two planes of symmetry are the tangent planes to the surfaces $i_1 = $ constant and $\theta = $ constant. Thus, a surface $i_1 = $ constant or a surface $\theta = $ constant or a combination of a finite number of $i_1 = $ constant portions and $\theta = $ constant portions can all be reflectors that do not disturb the flow. Figure 6.28 shows the lines $i_1 = $ constant and the lines $i_3 = $ constant in a meridian plane. We can choose any line $i_1 = $ constant and its symmetric $i_1 = -$constant as the cross section of the reflector. With this reflector we can get a circular entry and exit apertures. The smallest is the value of i_1, the lowest the variation of the refractive index within the concentrator and the highest the aspect ratio (depth to entry aperture diameter ratio). In the limit case $i_1 \to 0$, an ideal 3D concentrator infinitely deep with constant refractive index is found. This limit case was already found by Garwin (1952). As we said before, to get realistic values of the refractive index, i_1 should be limited to $|i_1| \leq 0.3$.

We can also choose two nonsymmetric $i_1 = $ constant lines. In this case we will obtain ring-shaped apertures.

Combinations of $i_1 = $ constant and $\theta = $ constant surfaces would give us entry and exit apertures with shapes formed by circular arcs and straight radial portions. Note that according to Eq. (6.89) the ratio of ρ at the entry aperture to the ρ at the exit aperture is a constant for the i_3 lines (these are the flow lines, which are contained in the reflector surface). Since $\theta = $ constant along these lines we can conclude that the shape of the entry and exit apertures is the same except a scale factor that is $\sqrt{C_g}$.

As we have seen before, the rays of a circular bundle passing through any point form a circular cone that is symmetric with respect to any plane containing
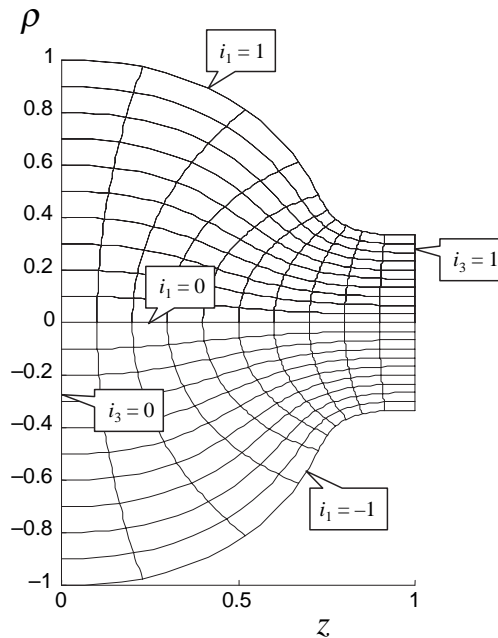
Figure 6.28   Reflector surface formed by a i_1 = constant surface. Since the flow is not disturbed with respect to the one existing without the reflector (the one corresponding to the refractive index distribution of Figure 6.27), we can state that a concentrator formed by this reflector and this refractive index distribution will collect all the rays impinging its entry aperture with an angle smaller than 30° with maximal concentration, and thus it is an ideal 3D concentrator.

the cone axis. This property implies that any reflector whose surface is formed by flow lines will be a reflector that won't disturb the flow—that is, for circular bundles—the reflectors are not limited to surfaces (or portions of surfaces) i_1 = constant or θ = constant. Any surface defined by an equation $\Gamma(i_1, \theta) = 0$, where Γ is an arbitrary function, can be a reflector surface that does not disturb the flow. Figure 6.29 shows the case forming squared entry and exit apertures. Note that the lines z = constant are squares only at the apertures; in other words, this reflector is not formed by crossing two troughlike ones. Of course, the refractive index distribution is rotationally symmetric.

6.14.3   Elliptic Bundles in an Elliptic Coordinate System with Constant Refractive Index

Manufacturing a dielectric material with a prescribed variation of the refractive index is a difficult task except in some well-known cases. Understanding this, it is of major practical interest to find concentrators using a small number of different homogeneous materials. The subject of this section is to use the preceding results to find elliptic bundles in a given coordinate system when n is fixed to a constant ($n = 1$). This is another way to work with Eqs. (6.54) and (6.55). Vector a is now prescribed, and we have to find the vectors N and V and then find the elliptic bundles.
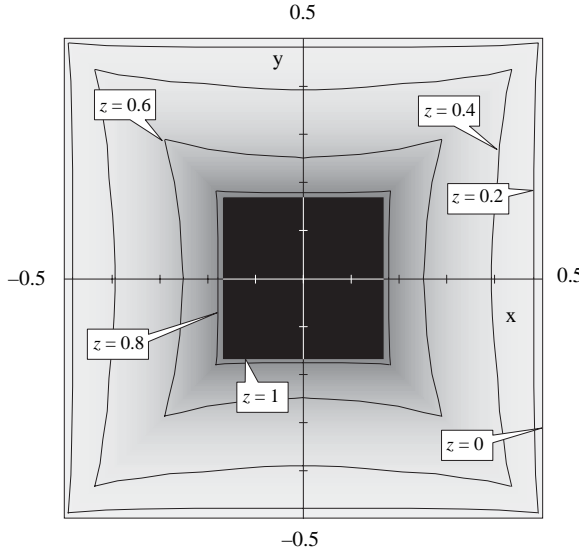
Figure 6.29 z = constant lines of a reflector surface formed by a $\Gamma(i_1, \theta) = 0$ surface defining a square at the entry and exit apertures. The reflector surface is not formed by two crossed linear surfaces, although it is close to this case. A concentrator formed by this reflector and the refractive index distribution of Figure 6.27 collects all the rays impinging its entry aperture with an angle smaller than 30° with maximal concentration, and thus it is again an ideal 3D concentrator.

The hyperboloid of revolution concentrator, which is the first ideal 3D concentrator known excluding trivial cases, can be derived from the elliptic bundle formed by all the rays emitted by a disk in media with $n = 1$ (Welford and Winston, 1989). The flow lines of this bundle are coordinate lines of a elliptic coordinate system with symmetry of revolution. Not only concentrators but also confining cavities have been generated from this bundle (Miñano, 1990). Elliptic bundles in media with $n = 1$ have been analyzed successfully with other tools (Gutiérrez et al., 1996), and from this study we know that there are other elliptic bundles in this type of coordinate system (see Section 6.17). Summarizing, we know that we will find solutions if we study the elliptic coordinate system. For simplicity we restrict ourselves to the case of rotational symmetry—we'll say, one of the coordinates, i_2, is the angular coordinate θ. Consequently, its conjugate variable is the skew invariant $u_2 = h$. There are also solutions in the general elliptic coordinate system (Benítez, 1999).

The equations defining the coordinate change from Cartesian to elliptic coordinates with rotational symmetry is

$$i_1 = \frac{\sqrt{\left(\sqrt{x_1^2 + x_2^2} + 1\right)^2 + x_3^2} - \sqrt{\left(\sqrt{x_1^2 + x_2^2} - 1\right)^2 + x_3^2}}{2} \tag{6.98}$$

$$\theta = \arctan\frac{x_2}{x_1} \tag{6.99}$$

$$i_3 = \frac{\sqrt{\left(\sqrt{x_1^2 + x_2^2} + 1\right)^2 + x_3^2} + \sqrt{\left(\sqrt{x_1^2 + x_2^2} - 1\right)^2 + x_3^2}}{2} \tag{6.100}$$
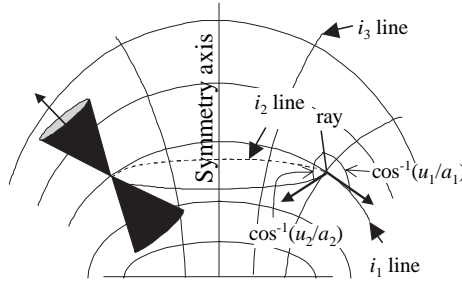
Figure 6.30   Description of i_1, i_2, and i_3 lines. The coordinate i_2 is the angular coordinate θ ($i_2 = \theta$), and u_2 is the skew invariant h ($u_2 = h$). The cone of edge rays passing through a point is also represented.
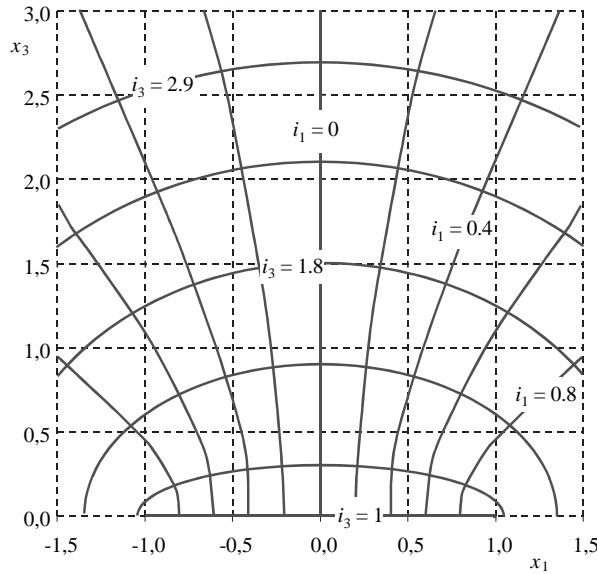
Figure 6.31   Lines i_1 = constant and i_3 = constant in a θ = constant plane given by Eqs. (6.98), (6.99), and (6.100).

The i_1 lines (θ = constant and i_3 = constant) are confocal ellipses (with foci at the circumference in the plane $x_3 = 0$ of radius 1 and centered at $x_1 = x_2 = 0$). The i_3-lines are confocal hyperbolas with foci in the same locus (see Figures 6.30 and 6.31).

The vector a can be calculated from the preceding equations (note that $n = 1$).

$$\mathbf{a} \equiv \frac{1}{n^2}\left((\nabla i_1)^2,(\nabla i_2)^2,(\nabla i_2)^2\right) = \left(\frac{1-i_1^2}{i_3^2-i_1^2},\frac{1}{i_1^2 i_3^2},\frac{i_3^2-1}{i_3^2-i_1^2}\right) \tag{6.101}$$

Eqs. (6.54) and (6.55) applied to vector a can be considered as functional equations where the unknowns are the vectors N and V. As an example, we are going to find the vector N.

$$\mathbf{N} \cdot \mathbf{a} = n_1(i_1)\frac{1-i_1^2}{i_3^2-i_1^2} + n_2(i_2)\frac{1}{i_1^2 i_3^2} + n_3(i_3)\frac{i_3^2-1}{i_3^2-i_1^2} = 0 \qquad (6.102)$$

Solving for $n_2(i_2)$, we get

$$-n_1(i_1)\frac{1-i_1^2}{i_3^2-i_1^2}i_1^2 i_3^2 - n_3(i_3)\frac{i_3^2-1}{i_3^2-i_1^2}i_1^2 i_3^2 = n_2(i_2) \qquad (6.103)$$

The right-hand side of this equation only depends on i_2, whereas the other side only depends on i_1 and i_3. The only solution is that both sides are equal to a constant that we will call $-n_a$. Now let us group Eq. (6.103) as follows:

$$n_1(i_1)(1-i_1^2) - \frac{n_a}{i_1^2} = n_3(i_3)(1-i_3^2) - \frac{n_a}{i_3^2} \qquad (6.104)$$

Again, we find that the left-hand side of Eq. (6.104) only depends on i_1, whereas the other side only depends on i_2. Both sides must be equal to a constant that we will call $n_b - n_a$. Thus, $\mathbf{N}$ can be written as

$$\mathbf{N} = n_a\left(\frac{1}{i_1^2}, -1, \frac{1}{i_3^2}\right) + n_b\left(\frac{1}{1-i_1^2}, 0, \frac{1}{1-i_3^2}\right) \qquad (6.105)$$

With a similar reasoning, using Eq. (6.55), we find

$$\mathbf{V} = \left(\frac{-i_1^2}{1-i_1^2}, 0, \frac{-i_3^2}{1-i_3^2}\right) + v_a\left(\frac{1}{i_1^2}, -1, \frac{1}{i_3^2}\right) + v_b\left(\frac{1}{1-i_1^2}, 0, \frac{1}{1-i_3^2}\right) \qquad (6.106)$$

Where v_a and v_b are arbitrary constants. As said before, we can choose the scale factor for the vector N and we can choose a new vector V as the sum of the old vector V plus the vector N times any factor, and still Eq. (6.56) represent the same edge-ray bundle. Therefore, no set of four parameters n_a, n_b, v_a, and v_b provides a different edge-ray bundle. The different edge-ray bundles depend only on two parameters. We can write the expressions for $\mathbf{N}$ and $\mathbf{V}$ using two parameters (γ and k_v) so different pairs represent different edge-ray bundles:

$$\mathbf{N} = \left(\frac{1}{i_1^2}, -1, \frac{1}{i_3^2}\right)\cos\gamma + \left(\frac{1}{1-i_1^2}, 0, \frac{1}{1-i_3^2}\right)\sin\gamma \qquad (6.107)$$

$$\mathbf{V} = \left(\frac{-i_1^2}{1-i_1^2}, 0, \frac{-i_3^2}{1-i_3^2}\right) + \left(\frac{1}{i_1^2}, -1, \frac{1}{i_3^2}\right)k_v \sin\gamma - \left(\frac{1}{1-i_1^2}, 0, \frac{1}{1-i_3^2}\right)k_v \cos\gamma \qquad (6.108)$$

With these equations and Eq. (6.56) we can find the expression of any elliptic bundle in $n = 1$ such that the i_1, i_2, or i_3-lines are flow lines of the bundle. For instance, the bundle of rays emitted by the disk of radius 1 (this is the bundle with which the hyperbolic rotational concentrator is generated) corresponds to the case $\gamma = 0$, $k_v = -1$. The disk is defined by the equation $i_3 = 1$. The equation of the edge rays with variables u_1 and u_2 for this bundle—that is, Eq. (6.58) applied to this bundle is simply

$$u_1^2 + \frac{u_2^2}{i_1^2} = 1 \qquad (6.109)$$

This equation can be easily formulated with the variables x_1, x_3 (using Eqs. (6.98) and (6.100)) and with the direction cosines with respect the tangents to the i_1 lines and θ lines (using Eq. (6.101) and the expressions given in Figure 6.15).

Any concentrator derived from these bundles is formed by reflector surfaces shaped with one or more coordinate surfaces. The hyperbolic concentrator is one case. The elliptic cavity (Miñano, 1990) (which is not a concentrator but is generated in a similar way) also derives from one of these elliptic bundles.

6.15 MULTIFOLIATE-REFLECTOR-BASED CONCENTRATORS

When a mirror is positioned such that it intercepts the rays of a bundle, in general it will disturb the flow lines. The flow lines are not disturbed when the mirror surface contains these lines and the rays of the bundle have some symmetry. This property is used to fix a spatial limitation of the concentrator's designs based in the flow-line method. This spatial limitation is needed because the definitions of the input and output bundle do not contain such spatial limitation. For instance, the edge-ray bundle defined by Eqs. (6.71), (6.72) and (6.73) does not contain any limitation in the i_1, i_2 spatial variables. This edge-ray bundle exists in the refractive index distribution of Figure 6.27. This refractive index distribution has only realistic values if i_1 is less than $\cong 0.25$. The reflector placed in a surface $i_1 =$ constant avoids this problem. If we don't want to use these mirrors, we need a definition of the edge-ray bundle that is bounded in its phase-space representation—in other words, that is also spatially bounded. It is done this way in the SMS design method (Chapter 8).

The way flow lines of a bundle can be disturbed by a mirror depends on both mirror and flow lines geometry. In some cases, the flow lines are disturbed in such a way that the new flow lines are contained in the reflector surface. In the general case, the new flow lines are "reflected" by the reflector. For instance, Figure 6.32 shows the flow lines of the bundle of rays collected by a CPC before and after the insertion of the CPC (2D geometry). In this case, the new flow lines are either contained in the reflector surface or never touch it. Figure 6.33 shows the other

before
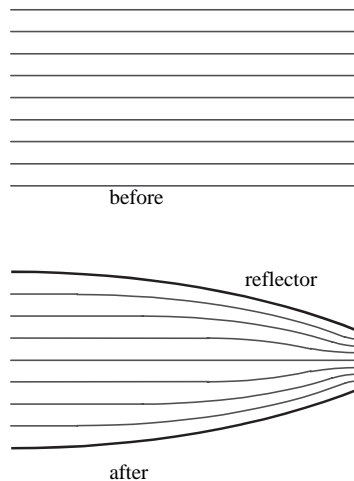
reflector

after

Figure 6.32   Flow lines before and after the insertion of a CPC reflector (2D geometry).
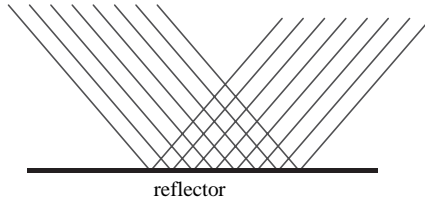
Figure 6.33   Flow lines reflected by a flat mirror.
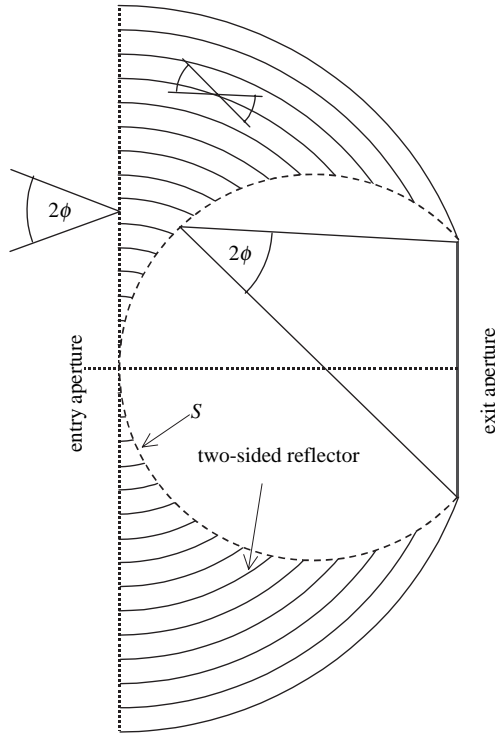
Figure 6.34   Multifoliate ideal 3D nonimaging concentrator that collects with maximal concentration all the rays impinging its entry aperture within an angular cone $\alpha = 20°$.

case in which the flow lines are "reflected" by the flat mirror. For our purposes, we do not need to go further in these considerations (which also would require a more detailed definition of the flow lines).

When the flow lines find the reflector surface such that the geometrical vector flux is contained in the tangent to the surface, then it may happen that the new flow lines are contained in the reflector surface. This property has been used to design a multifoliate ideal 3D concentrator by Forbes and Basset (1982), which is shown in insert Figure 6.34. This concentrator is formed by many (infinite in the theoretical limit case) two-sided mirrors with a spherical shape. The center of the spheres is at the center of the entry aperture. The reflectors' surfaces are extended in the right hemisphere defined by the entry aperture plane, excluding the points
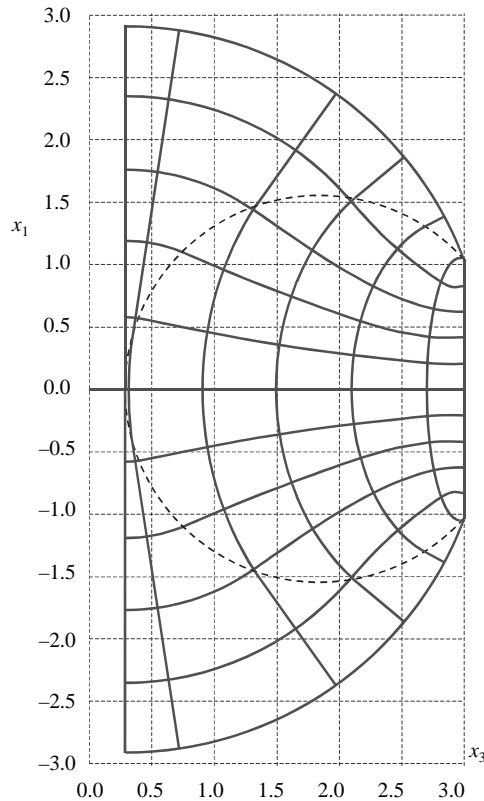
Figure 6.35   Flow lines in the multifoliate concentrator of Figure 6.34.

inside the sphere S. In this region, the meridian rays form a constant angle equal to ϕ with respect to the tangents to the circumferences centered at the entry aperture center. At the entry aperture, not only the meridian rays but the remaining ones form an angle ϕ with these tangents. The meridian section of S is the circumference whose points subtend an angle 2ϕ with the exit aperture diameter.

Because of the multiple reflections suffered by the rays in their passage through the multifoliate mirrors, Forbes and Basset suggested total internal reflection. Simply using a set of transparent solid thin spherical shells can get the desired performance. This 3D ideal concentrator can be very well approximated in the real world using nonzero thickness dielectric layers. Both articles by Bassett and Forbes (1982) examine other cases in which the desired output bundle is not the one illuminating the exit aperture isotropically but illuminating it with a restricted angle, called angle transformers. Again, the design of Forbes and Basset gives ideal concentrators in 3D geometry.

It can be proved that the bundle of collected rays of the multifoliate concentrator is also an elliptic bundle. For this reason, we can use the Poisson bracket method to find out the refractive index distribution that can substitute the multifoliated region. For this, it is convenient to choose a triorthogonal coordinate system so the flow lines are one of the coordinate lines. These coordinates are

spherical. The coordinate change from cylindrical coordinates $(\rho,\ \theta,\ z)$ to the new ones is given by

$$
\begin{aligned}
i_1 &= \sqrt{\rho^2 + z^2}\\
i_2 &= \theta\\
i_3 &= -\tan^{-1}\left(\frac{z}{\rho}\right)
\end{aligned}
\tag{6.110}
$$

Vector a is thus

$$
a = \frac{1}{n^2}\left((\nabla i_1)^2,(\nabla i_2)^2,(\nabla i_3)^2\right) = \frac{1}{n^2}\left(1,(i_1\cos i_3)^{-2},i_1^{-2}\right)
\tag{6.111}
$$

The rotational symmetry of the bundle we are considering implies that the second component of the vectors $\mathbf{V}$ and $\mathbf{N}$ should be a constant. Application of Eqs. (6.54) and (6.55) to vector a, taking into account this consideration, gives

$$
\mathbf{N} = \left(-\frac{C_1}{i_1^2},n_2,C_1-\frac{n_2}{\cos^2 i_3}\right)
\tag{6.112}
$$

$$
\mathbf{V} = \left(n^2-\frac{C_2}{i_1^2},v_2,C_2-\frac{v_2}{\cos^2 i_3}\right)
\tag{6.113}
$$

We are looking for a refractive index distribution such that the meridian rays form an angle ϕ with the tangents to the circumferences centered at $\rho = 0$ and the bundle is circular in the plane $z = 0$—that is, when $i_3 = 0$. With this description we can calculate vector $\boldsymbol{\alpha}$.

We can choose vector $\boldsymbol{\alpha}$ so its third component is zero because the third component of vector a is zero only when $i_1 \rightarrow \infty$ (see Section 6.13.2, Eq. (6.34)). Then $\alpha_1 = 1/(u_1)^2$ when $u_2 = 0$, (Eq. (6.33))—and u_1 is the optical direction cosine of meridian edge rays with respect the i_1 lines. Thus, we can write

$$
\boldsymbol{\alpha} = (\alpha_1,\alpha_2,0) = \left(\frac{1}{n^2\sin^2\phi},\alpha_2,0\right)
\tag{6.114}
$$

From Eq. (6.54), and taking into account that the edge rays form a circular bundle at the plane $i_3 = 0$, we derive

$$
\boldsymbol{\alpha} = \frac{1}{n^2\sin^2\phi}\left(1,\frac{1}{i_1^2},0\right)
\tag{6.115}
$$

From Eq. (6.55) it is concluded that

$$
n^2 = \frac{C^2}{i_1^2}
\tag{6.116}
$$

Where C is an arbitrary constant $(C^2 = (C_2 - v_2)/\cos^2\phi)$.

What is important in this result for our purposes is (1) there is an arbitrary multiplicative constant, and (2) the flow lines are contained in surfaces of the same refractive index (surfaces $i_1 = $ constant).

Let us take a thin layer between the surfaces $i_1 = $ const. and $i_1 + \Delta i_1 = $ const. Since the flow lines are contained in these surfaces, we can isolate this region by mirroring the two surfaces. The bundle of edge rays is not disturbed by this action, and, therefore, it is still the same one we used for the preceding calculations.

Assume that Δi_1 is thin enough to consider that the refractive index is constant in this layer. We can still choose C so the refractive index of this layer is a given one—for instance, $n = 1$.

This reasoning is independent of the value of i_1, and thus it can be applied to any thin layer in between two i_1 = constant. In this way we get that multifoliate region of the concentrator of Figure 6.34 can also be derived from the Poisson bracket method. In the Forbes and Bassett (1988) articles it is explained in detail that the bundle exiting the multifoliate region is coincident, at the sphere S, with the bundle of rays reaching the receiver disk.

As also noted in these articles, a practical way to manufacture an efficient multifoliate concentrator should rely on total internal reflection. The geometry of the concentrator of Figure 6.34 permits the total internal reflection condition with practical materials of refractive index around 1.5.

6.16   THE POISSON BRACKET METHOD IN 2D GEOMETRY

When the spatial variables are only 2, i_1, and i_2, it is called a 2D geometry problem. Consequently, their conjugate variables are also 2, u_1, and u_2. The formulation of the problem is identical to the 3D case, except i_3 and u_3 do not appear. The equations $H = 0$ and $\omega = 0$ can be written as in Eqs. (6.32) and (6.33), but now, $\mathbf{u}$, $\mathbf{a}$, and $\boldsymbol{\alpha}$ are two-dimensional vectors.

$$H \equiv \mathbf{u} \cdot \mathbf{a} - 1 = 0 \tag{6.117}$$

$$\omega \equiv \mathbf{u} \cdot \boldsymbol{\alpha} - 1 = 0 \tag{6.118}$$

Since it is assumed that $\mathbf{a}$ and $\boldsymbol{\alpha}$ are independent, then these two equations define uniquely a vector $\mathbf{u}$. This means that for any point i_1, i_2 of the concentrator there is a single vector $\mathbf{u}$ associated to it—that is, there are four possible combinations of values of the conjugate variables ($\pm u_1$, $\pm u_2$), which represent the two possible directions of the edge rays passing through the point i_1, i_2.

The result of the elliptic bundle analysis in 2D geometry is the equivalent to Eq. (6.55).

$$\mathbf{a} \cdot \mathbf{V} = \boldsymbol{\alpha} \cdot \mathbf{V} = 1 \tag{6.119}$$

Where now $\mathbf{V}$ is a two-dimensional vector that fulfills

$$\mathbf{V} = [v_1(i_1), v_2(i_2)] \tag{6.120}$$

Vector $\mathbf{u}$ can be written as (equivalent to Eq. (6.56))

$$\mathbf{u} = \mathbf{V} \tag{6.121}$$

In the 3D problem we restricted the analysis to a certain type of bundles (the elliptic bundles). In the 2D problem, the description of the edge-ray bundle given by Eqs. (6.117) and (6.118) is more general. In fact, it covers all the cases in which there are a couple of edge rays at every point i_1, i_2 of the concentrator. This case includes all 2D concentrators considered in this book.

The design procedure is much simpler in the 2D case, and the number of solutions is much broader than in the 3D case. One possible design procedure is to fix

the refractive index distribution and then find the vector $\mathbf{V}$ and the coordinates i_1 and i_2 that fulfill the contour conditions. This procedure is essentially the string method (see Section 4.4). The way in which we fix the string is what we call here the contour conditions. The fact that the string is straight between fixing points is equivalent to establish the refractive index distribution (rays are straight lines in a constant refractive index distribution). The movement of the pencil along the string is equivalent here to the calculation of the flow lines—one of the coordinate lines. In order to clarify this equivalence, let us introduce Eq. (6.121) in Eq. (6.119) using Eq. (6.120), and using the expression of a_j in terms of $|\nabla i_j|$ and n.

$$v_1(i_1)(\nabla i_1)^2 + v_2(i_2)(\nabla i_2)^2 = n^2 \qquad (6.122)$$

This equation together with the condition of orthogonality of the coordinate lines $(\nabla i_1 - \nabla i_2 = 0)$ and noting that the components of vector u—that is, v_1 and v_2 must be positive lead to

$$(\nabla m_1)^2 = (\nabla m_2)^2 = n^2 \qquad (6.123)$$

where

$$
\begin{aligned}
m_1 &= \int \sqrt{v_1(i_1)}\,di_1 + \int \sqrt{v_2(i_2)}\,di_2 \\
m_2 &= \int \sqrt{v_1(i_1)}\,di_1 - \int \sqrt{v_2(i_2)}\,di_2
\end{aligned}
\qquad (6.124)
$$

The functions m_1 and m_2 are the optical path lengths of the two sets of edge rays (remember that there were two edge rays at every point) and Eq. (6.123) is simply their eikonal equation.

6.16.1   Example: The Compound Triangular Concentrator

Another possible design procedure is to propose a vector $\mathbf{V}$ that fulfills the contour condition, select a set of orthogonal coordinates i_1 and i_2, and then calculate the refractive index distribution in a way quite similar to the one shown in Section 6.14.2. For instance, let's choose the set of orthogonal coordinates show in Figure 6.36. These coordinates can be chosen such that $|\nabla i_j|$ has a constant value in each of the four regions defined by the straight-lines AE and $A'E'$. Let $v_1(i_1)$ and $v_2(i_2)$ be linear functions of i_1 and i_2, respectively. These functions and this coordinate system define what is called the Compound Triangular Concentrator (CTC) (Miñano, 1985a, 1985c).

Figure 6.36 shows the i_1 and i_2 lines of a CTC concentrator. These lines have three straight segments. Consequently, there is a discontinuity on ∇i_1 and ∇i_2 at the points where these straight segments connect (lines AE' and $A'E$). The design is done in such a way that the refractive index is the same for all the points out of the lines where the discontinuities are (n_b). The refractive index at the discontinuities was studied as if the CTC was a limited case of concentrators in which ∇i_1 and ∇i_2 vary sharply but continuously at lines AE' and $A'E$. It was concluded that the refractive index has a minimum in these lines, whose value is called n_l. This refractive index minimum causes the lines AE' and $A'E_1$ behave as non-existent layers for rays forming an angle with the normal smaller than $\sin^{-1}(n_1/n_b)$ and behaves like a mirror (by total internal reflection) otherwise. Figure 6.37
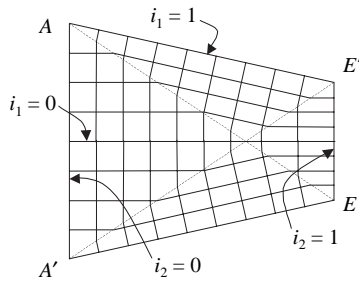
Figure 6.36 i_1 and i_2 lines of a CTC concentrator.
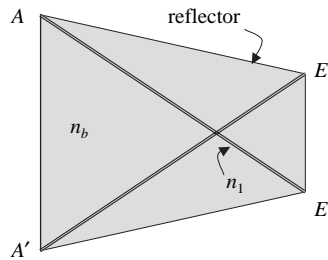
Figure 6.37   Refractive index distribution of a CTC. In this example there are two differ-
ent refractive indices for the materials within the concentrator, n_b, which is the refractive
index of the bulk material, and n_1, which is the refractive index of the lines AE' and $A'E$.

shows the refractive index distribution. As in any other concentrator designed with
the flow line method, the final step is to choose two flow lines as reflectors. In the
case of Figure 6.37 the lines chosen are $i_1 = \pm 1$. These flow lines define the edges
of the entry (AA') and exit (EE') apertures.

It should be emphasized that with our definitions, a 2D concentrator is not
the same as a 3D concentrator with linear symmetry. An ideal 2D concentrator
will not in general generate an ideal 3D concentrator by linear symmetry. The con-
fusion arises from the fact that when the refractive index is constant (for instance,
in the CPC), then the linear concentrator derived from the 2D concentrator is ideal
with respect to the input and output bundles that are generated from the 2D case
by the same linear symmetry. This is not the case of the CTC because its refrac-
tive index is not constant.

6.17   ELLIPTIC BUNDLES IN HOMOGENEOUS MEDIA

Gutiérrez et al. (1996) applied the Lorentz geometry to the problem of finding ellip-
tic bundles in a medium of constant refractive index. The rapid development of
Lorentz geometry is due to Einstein's general relativity theory of gravitation. It
provides a cone structure on an open subset $\Re$ of the space—that is, a cone at each
point X of $\Re$ (formally, in the tangent space $T_x \Re$ of X, which is isomorphic to $\mathbb{R}^3$).
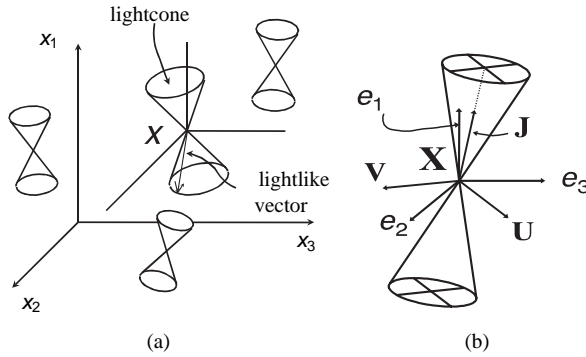
Figure 6.38   (a) The Lorentz geometry provides a cone structure in $\Re$, formed by the set of lightlike vectors at each point X. (b) The vectors {$\mathbf{J}$, $\mathbf{U}$, $\mathbf{V}$} define the three planes of symmetry of the cone of edge rays of the elliptical bundle at the point $\mathbf{X}$.

This cone is called the light cone in Einstein's theory because it is formed by the directions of the light rays in the continuous space-time. In nonimaging optics, Lorentz metric is used in a quite different interpretation. This cone is formed by the directions of the edge rays passing through a point in the conventional 3D space (see Figure 6.38a). We'll briefly discuss Lorentz geometry here, but the interested reader can consult O'Neill (1983) for some background on differential geometry.

Let $B = \{e_1, e_2, e_3\}$ be the canonical global basis vector field on $\Re$, which is orthonormal with respect to the usual Euclidean metric. For each point X of $\Re$, B_x is the canonical basis of the tangent space $T_x\Re$. A Lorentz metric g on $\Re$ is a map that assigns to every point X of $\Re$ a bilinear map g_x: $T_x\Re \times T_x\Re \to \mathbb{R}$, with some additional properties. These properties may be expressed through the matrix G associated with the map g in the basis B. Using this matrix, one can formulate the map as $g_x(Y, Y') = Y^tG(X)Y'$, where Y and Y' are two vectors of $T_x\Re$ (the superscript t denotes transposition). The matrix G must verify the following three properties:

1. G is symmetric on every point X of $\Re$.
2. $|G| \neq 0$ on every point of $\Re$.
3. G has one negative and two positive eigenvalues on every point of $\Re$. This is expressed by saying that g has the signature $(-, +, +)$.

The elements of $G(X)$ are functions on $\Re$ defined by $g_{if}(X) = g_x(e_i, e_j)$. If they are differentiable of class C^∞, the above properties characterize a C^∞ Lorentz metric on $\Re$. Similarly, a C^r Lorentz metric can be introduced ($r \in \mathbb{N}$).

Three different types of vector Y are considered: spacelike if $Y^tGY > 0$ or $Y = 0$, timelike if $Y^tGY < 0$, and lightlike if $Y^tGY = 0$ and $Y \neq 0$. The set of all lightlike vectors forms the light cone, whose equation is $Y^tGY = 0$. If $\gamma: \mathbb{R} \to \Re$ is a C^1 curve and $\dot{\gamma}$ is its derivative, then γ is called spacelike if $\dot{\gamma}(t)$ is spacelike for all t. The timelike and the lightlike curves are defined in a similar way.

Let us look at the role of the eigenvalues and eigenvectors of the matrix G of a Lorentz metric g. Because G is a symmetric matrix, it can be diagonalized,

leading to a matrix $G' = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$, where $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$. This transformation from G to G' is just a change of basis, and thus G' may be expressed as $G' = A^{-1}GA$, where A is the matrix changing between the orthonormal canonical basis B_x and a new one B_x', which is also orthonormal (and thus $A^{-1} = A^t$). The new basis B_x' is formed by three eigenvectors of G. The columns of A are the components of these eigenvectors in the canonical basis B_x, and the elements $\lambda_1, \lambda_2, \lambda_3$ of the diagonal of G' are the eigenvalues of G associated with the corresponding eigenvectors. Because g is Lorentzian, the signature is $(-, +, +)$; thus only one eigenvalue is negative, and the other two are positive. There is no loss of generality if we assume that $\lambda_1 < 0$ and $|A| = 1$ (changing the order of the columns of A if necessary). Thus, the basis B_x' in which G' is a diagonal matrix has a nice geometric interpretation, since it can be obtained from the canonical basis B_x by a rotation.

Now the study of the cone structure in the basis B_x' is trivial. $B_x' = \{\mathbf{J}, \mathbf{U}, \mathbf{V}\}$, where $\mathbf{J}$, $\mathbf{U}$, and $\mathbf{V}$ are the eigenvectors of G associated with λ_1, λ_2, and λ_3, respectively. Let $Y \neq 0$ be a lightlike vector. If the components of Y in the basis B_x' are (y_1, y_2, y_3), the equation of the lightlike vectors in this basis is $Y^tG'Y = 0$—that is,

$$\lambda_1 y_1^2 + \lambda_2 y_2^2 + \lambda_3 y_3^2 = 0 \qquad (6.125)$$

which is a cone whose axis is parallel to the vector $\mathbf{J}$ (remember that $\lambda_1 < 0$, and $\lambda_2, \lambda_3 > 0$). The intersection of the cone and the plane $y_1 = 1$ is the ellipse

$$\lambda_2 y_2^2 + \lambda_3 y_3^2 = -\lambda_1 \qquad (6.126)$$

The ellipse axes are parallel to the vectors $\mathbf{U}$ and $\mathbf{V}$ (see Figure 6.38b), and their respective semiaxis lengths are $(-\lambda_1/\lambda_2)^{1/2}$ and $(-\lambda_1/\lambda_3)^{1/2}$.

To summarize:

1. The light cone at X is elliptic, $\forall X \in \mathfrak{R}$.
2. The direction of the elliptic cone axis is the direction of J, the eigenvector of G associated with the negative eigenvalue of G.
3. The directions of the principal axis of the elliptic cone are the directions of U and V, the eigenvectors of G associated with the positive eigenvalues of G.

Sections 6.17.1 and 6.17.2 are devoted to the "optical condition," which is the condition that the Lorentz geometry must fulfill to make the lightlike curves define the edge rays of the bundle. In Section 6.17.2 we examine the system of partial differential equations, which are derived from the general optical condition. We also discuss a specific set of solutions of said system, which coincide with the already identified bundles found by other methods. Section 6.17.4 describes some examples of elliptic bundles, and the conclusions are given in Section 6.18.

6.17.1   The Restricted Optical Condition

Let us consider a medium with constant refractive index $n = 1$. Thus, the ray trajectories will be straight lines. In the Lorentz geometry approach of Gutierrez et al. (1996), a "restricted optical condition" was used, which stated that every lightlike geodesic of the Lorentz metric must be a geodesic of the Euclidean metric. The application of this condition forces the lightlike curves to be geodesics of both the Lorentz metric and the Euclidean metric but provided tools to find new elliptic bundles.

Let us eliminate requirement of the lightlike curves being geodesics of the Euclidean metric from restricted optical condition. Then every lightlike geodesic of the Lorentz metric must be a straight line. This condition will be called "reformulated restricted optical condition".

Thus, we can reformulate the restricted optical condition of Gutiérrez et al. (1996) as follows: If γ is a C^2 curve in $\mathfrak{R}$ such that $\dot{\gamma}(t) \neq 0$, $\forall\ t \in \mathbb{R}$, then, for $k \in \{1, 2, 3\}$,

$$\left.\begin{array}{r}\ddot{y}^k(t) + [\dot{\gamma}(t)]^t\,\Gamma^k[\gamma(t)]\dot{\gamma}(t) = 0 \\ [\dot{\gamma}(t)]^t\,\mathbf{G}[\gamma(t)]\dot{\gamma}(t) = 0\end{array}\right\} \Rightarrow \ddot{\gamma}(t) \wedge \dot{\gamma}(t) = 0 \tag{6.127}$$

where the symbol $\wedge$ denotes the vector product.

The upper left equation establishes that γ is a geodesic curve in Lorentz geometry. The lower left one implies that γ is a lightlike curve. The equation on the right-hand side states that $\ddot{\gamma}(t)$ is parallel to $\dot{\gamma}(t)$, which is the general condition for γ to be a straight line.

Γ^k is a matrix formed by the elements Γ_{ij}^k, which are the Christoffel symbols of the Lorentz metric (O'Neill, 1983). The expression of these symbols using Einstein's notation for the summation (in which the summation symbol is omitted) is

$$\Gamma_{ij}^k = \frac{1}{2} g^{km}(g_{jm,i} + g_{mi,j} - g_{ij,m}) \tag{6.128}$$

where $G = (g_{ij})$; $G^{-1} = (g^{ij})$.

Let us see that the restricted optical condition Eq. (6.127) is fulfilled if

$$\Gamma^k = f^k \mathbf{G} + \mathbf{A}^k, \quad k \in \{1, 2, 3\} \tag{6.129}$$

where f^k are functions on $\mathfrak{R}$, and the matrices A^k are formed by the elements $A_{ij}^k = \dfrac{a^i}{2}\delta_{kj} + \dfrac{a^j}{2}\delta_{ki}$, a^k are also functions on $\mathfrak{R}$ and δ_{ki} is the Kronecker delta function. By substitution of Γ^k in the upper-left equation in Eq. (6.127), and taking into account the lower-left equation in Eq. (6.127), we obtain:

$$\gamma^k + \dot{\gamma}^t \mathbf{A}^k \dot{\gamma} = 0 \Rightarrow \gamma^k + (a^i\dot{\gamma}^i)\dot{\gamma}^k = 0 \Rightarrow \frac{\ddot{\gamma}^k}{\dot{\gamma}^k} = -\mathbf{a}\cdot\dot{\gamma} \Rightarrow \ddot{\gamma}\wedge\dot{\gamma} = 0 \tag{6.130}$$

where $\mathbf{a}$ denotes the vector $\mathbf{a} = (a^1, a^2, a^3)$.

Equations Eq. (6.129) are the reformulated restricted optical equations. Let us prove that Eq. (6.129) is not only sufficient but necessary for the restricted optical condition Eq. (6.127) to be fulfilled. Consider a point X in $\mathfrak{R}$ and a lightlike vector Y at X, which must fulfill

$$\mathbf{Y}^t\mathbf{G}(X)\mathbf{Y} = 0 \tag{6.131}$$

Let the straight line $\gamma(t)$ be the lightlike curve fulfilling $\gamma(t_o) = X$ and $\dot{\gamma}(t_o) = \mathbf{Y}$. If Γ denotes the vector whose components are the matrices Γ^k, from (1):

$$\left.\begin{array}{r}\ddot{\gamma}(t) = -[\dot{\gamma}(t)]^t\,\Gamma[\gamma(t)]\dot{\gamma}(t) \\ \ddot{\gamma}(t) \wedge \dot{\gamma}(t) = 0\end{array}\right\} \Rightarrow \dot{\gamma}(t) \wedge [\dot{\gamma}(t)]^t\,\Gamma[\gamma(t)]\dot{\gamma}(t) = 0 \tag{6.132}$$

where the right-hand side equation is obtained by substitution of the upper left equation in the lower left one. Setting $t = t_o$ in the equation on the right, the following two independent equations are obtained:

$$Y_2[\mathbf{Y}^t\Gamma^3(X)\mathbf{Y}] - Y_3[\mathbf{Y}^1\Gamma^2(X)\mathbf{Y}] = 0$$
$$Y_3[\mathbf{Y}^t\Gamma^1(X)\mathbf{Y}] - Y_1[\mathbf{Y}^1\Gamma^3(X)\mathbf{Y}] = 0$$

(6.133)

In Eq. (6.132) we indicate explicitly that Γ^k is not a function of Y.

As the equations Eq. (6.132) are third-order polynomials in the components of vector Y and must vanish for all the vectors Y defined by the quadric Eq. (6.131), then the third-order polynomials must be decomposed as:

$$Y_2[\mathbf{Y}^t\Gamma^3\mathbf{Y}] - Y_3[\mathbf{Y}^t\Gamma^2\mathbf{Y}] \equiv [\mathbf{Y}^t G\mathbf{Y}](\mathbf{DY})$$
$$Y_3[\mathbf{Y}^t\Gamma^1\mathbf{Y}] - Y_1[\mathbf{Y}^t\Gamma^3\mathbf{Y}] \equiv [\mathbf{Y}^t G\mathbf{Y}](\mathbf{EY})$$

(6.134)

where $D = (d^1(X), d^2(X), d^3(X))$ and $E = (e^1(X), e^2(X), e^3(X))$ are vectors whose coefficients are functions on $\mathfrak{R}$. Developing Eq. (6.134) and identifying the coefficients of the powers of Y, it is found that D and E fulfill

$$\mathbf{D} = (0, f^3, -f^2)$$
$$\mathbf{E} = (-f^3, 0, f^1)$$

(6.135)

being f^k functions on $\mathfrak{R}$, and the following system of differential equations is obtained:

$$\Gamma_{ij}^k = f^k g_{ii} \quad i \neq k$$
$$\Gamma_{ij}^k = f^k g_{ij} \quad i \neq j, \quad i \neq k, \quad j \neq k$$
$$\Gamma_{ii}^i - 2\Gamma_{ik}^k = f^i g_{ii} - 2f^k g_{ik} \quad i \neq k$$

(6.136)

It is straightforward to see that the 15 equations Eq. (6.136) and the 18 equations Eq. (6.129) are equivalent: Eq. (6.136) can be obtained from Eq. (6.129) by elimination of the 3 unknown functions a^k.

Let us solve the restricted optical equations Eq. (6.129). Using the expressions for the Christoffel symbols Γ_{ij}^k, we can write the restricted optical equations as

$$g_{im,j} = g_{mk}f^k g_{ij} + g_{ik}f^k g_{jm} + \frac{a^i}{2}g_{mj} + \frac{a^m}{2}g_{ij} + a^j g_{im}, \quad i, j, m \in \{1, 2, 3\} \quad (6.137)$$

Calling $F^k = f^k + g^{mk}a^m/2$, this system of equations can be rewritten as

$$g_{im,j} = g_{mk}F^k g_{ij} + g_{ik}F^k g_{jm} + a^j g_{im}, \quad i, j, m \in \{1, 2, 3\}$$

(6.138)

Assume that the metric is C^2 and that the functional coefficients F^k and a^k are C^1. The integrability conditions of system Eq. (6.138), given by forcing the equality of the second-order cross derivatives ($g_{ij,mk} = g_{ij,km}$), lead to the following expressions for the functions F^k and a^k:

$$a_j^i = a_i^j$$
$$F_i^j - a_i F^j = 0 \quad i \neq j, \quad i, j \in \{1, 2, 3\}$$

(6.139)

The upper equations say that the vector $a = (a^1, a^2, a^3)$ is irrotational, and thus the must be a potential function $V(X)$ on $\mathfrak{R}$ fulfilling:

$$a = \nabla V$$

(6.140)

Taking into account Eq. (6.140), the lower equations in Eq. (6.139) can be easily integrated, and the solution can be written in terms of V as:

$$F^k = e^{-V}(\theta x_k + \mu_k) \quad k \in \{1, 2, 3\}$$

(6.141)

where θ, μ_1, μ_2, $\mu_3 \in \mathbb{R}$. Eqs. (6.140) and (6.141) are the integrability conditions of the system.

Let us show that, without loss of generality, we can set $V = 0$ in expression Eq. (6.141). This implies that $a^k = 0$, $F^k = f^k = \theta x_k + \mu_k$ and will simplify the integration of the restricted optical equations Eq. (6.137).

Consider an arbitrary C^2 function $h(X)$ on $\mathfrak{R}$ fulfilling $h > 0$ and a Lorentz metric G that fulfills the restricted optical equations for certain functions f^k and a^k. The matrix $G' = hG$ also defines a Lorentz metric with the same cone structure. By direct calculation using Eq. (6.128) it can be obtained that

$$\Gamma_{ij}'^k = \Gamma_{ij}^k - \frac{1}{2}\left[\frac{h_j}{h}\delta_{ik} + \frac{h_i}{h}\delta_{jk} - \left(\frac{h_m}{h}g^{mk}\right)g_{ij}\right], \quad k \in \{1, 2, 3\} \qquad (6.142)$$

where $\Gamma_{ij}'^k$ denotes the Chistoffel symbols for the matrix G'. Taking into account that the restricted optical equations are fulfilled for G (Eq. (6.129)), then from Eq. (6.142) we get that the restricted optical equations are also fulfilled for G', for the functions f'^k and a'^k given by

$$f'^k = \frac{1}{h}\left(f^k - \frac{1}{2}\frac{h_m}{h}g^{mk}\right) \qquad k \in \{1, 2, 3\} \qquad (6.143)$$
$$a'^k = a^k + \frac{h_k}{h}$$

Let us select the function $h = e^{-V}$, where V is the potential function associated to the vector $a = (a^1, a^2, a^3)$. By the definition of V in Eq. (6.140) and from Eq. (6.143), we get that for this function h, G' fulfills the restricted optical equations Eq. (6.129) with $f'^k = e^V F^k$ and $a'^k = 0$. This implies, again by Eq. (6.140) that the potential function for G' is $V' = V_0$ (= constant). Therefore, by Eq. (6.141), $F'^k = e^{-V_o}(\theta'x_k + \mu'_k)$ $= \eta x_k + \varepsilon_k$ where η and ε_k are also real numbers that group the exponential factor in the unknown constants θ' and μ'_k associated to G'. Obviously, with no loss of generality we can set $V_0 = 0$.

Therefore, we have proved that for any Lorentz metric G fulfilling the restricted optical equations Eq. (6.129), there exist an equivalent Lorentz metric G', which fulfills the more restrictive equations

$$\Gamma'^k = f'^k \mathbf{G}' \quad k \in \{1, 2, 3\} \qquad (6.144)$$

which differs from Eq. (6.129) in that $a'^k = 0$. This condition, when introduced in Eq. (6.130), provides that $\gamma(t) = 0$, which means that the edge rays of G' are not only geodesics of the Lorentz metric but also geodesics of the Euclidean metric. These equations Eq. (6.144) are the restricted optical equations defined in Gutiérrez et al. (1996), where the restricted optical condition was stated forcing also the lightlike curves to be geodesics of the Euclidean metric. Therefore, here we have shown that eliminating the additional constrain set does not lead to new elliptic bundles.

6.17.2   The General Optical Condition

The "general optical condition" that we are interested in can be stated this way: For every lightlike vector Y at X, Y must be also a lighlike vector at $X' = X + Yt$, for all t. This can be expressed as

$$\mathbf{Y}^t \mathbf{G}(X)\mathbf{Y} = 0 \Rightarrow \mathbf{Y}^t \mathbf{G}(X + t\mathbf{Y})\mathbf{Y} = 0, \quad \forall t \tag{6.145}$$

This general optical condition can equivalently be stated as follows: For every lightlike vector Y at X, there must be a lightlike curve with $\dot{\gamma}(0) = \mathbf{Y}$ that is a straight line.

In order to express the general optical condition Eq. (6.145) by a system of partial differential equations in the elements $g_{ij}(X)$ of G, let us consider the Poisson bracket method (Miñano, 1986) in Cartesian coordinates. In this framework, a ray is described in the phase space by the Hamilton equations, which constitutes a system of ordinary differential equations on the functions $(x_1, x_2, x_3; p_1, p_2, p_3)$ with respect to a parameter s that progress along the ray. The solutions of this system of equations must vanish the Hamiltonian $H \equiv (p_1^2 + p_2^2 + p_3^2 - n^2)/2$—that is, it must be consistent with the additional equation $p_1^2 + p_2^2 + p_3^2 = n^2$. Since we are considering the case $n = 1$ and the Hamilton equations state that $\dot{x}_i = p_i$, then $|\dot{X}| = 1$. Therefore, in this case $n = 1$, s is the arc length along the ray and the vector $P = (p_1, p_2, p_3)$ is a vector with unit modulus.

Let us consider the cone structure of equation $P^t G(X)P = 0$ defined in $\Re$. The general optical condition states that this equation must define a manifold $\omega \equiv P^t G(X)P = 0$ in the phase space that is composed by ray trajectories. As indicated in Miñano (1986), the necessary and sufficient condition for a manifold to be composed by ray trajectories is given by the vanishing of the Poisson bracket of the Hamiltonian and the manifold. In Cartesian coordinates, that is given by

$$0 = \{H, \omega\} = \sum H_{xi}\omega_{pi} - H_{pi}\omega_{xi} \tag{6.146}$$

Taking into account the expressions of H and ω, we get

$$0 = \{(\mathbf{P}^2 - 1)/2, \mathbf{P}^t \mathbf{G P}\} = \mathbf{P}[\mathbf{P}^t \nabla \mathbf{G P}] \tag{6.147}$$

for the values of X and P vanishing the manifold equation and the Hamiltonian. ∇G denotes the vector whose components are the matrices $(g_{ij,1})$, $(g_{ij,2})$, and $(g_{ij,3})$.

Let us consider now a lightlike vector Y, in general, not unitary. This means that $Y = aP$ for certain $a \in \mathbb{R}$. From the cone structure equation and Eq. (6.147), the lightlike vectors Y fulfill

$$\mathbf{Y}^t \mathbf{G Y} = 0$$
$$\mathbf{Y}[\mathbf{Y}^t \nabla \mathbf{G Y}] = 0 \tag{6.148}$$

Note that the second equation in Eq. (6.148) can also be obtained by deriving the left-hand side of Eq. (6.145) with respect to the parameter t and setting $t = 0$ for all $X \in \Re$. As the second equation in Eq. (6.148) is a third-order polynomial in the components of vector Y and must vanish for all the vectors Y defined by the first equation in Eq. (6.148), which is a quadric, then the third-order polynomial must be decomposed as

$$\mathbf{Y}[\mathbf{Y}^t \nabla \mathbf{G Y}] = [\mathbf{Y}^t \mathbf{G Y}](\mathbf{C Y}) \tag{6.149}$$

where $C = (c^1(X), c^2(X), c^3(X))$ is a vector whose coefficients are functions on $\Re$. Developing Eq. (6.149) and identifying the coefficients of the powers of Y, the following system of differential equations is obtained:

$$g_{ij,k} + g_{ki,j} + g_{jk,i} = g_{ij}c^k + g_{ki}c^j + g_{jk}c^i \tag{6.150}$$

where i, j, $k \in \{1, 2, 3\}$. We will refer to these equations as the general optical equations.

6.17.3 Results

Let us consider the particular case in which the functions c^k fulfill following conditions:

$$c^i_j = c^j_i \quad i \neq j, \quad i, j \in \{1, 2, 3\} \tag{6.151}$$

The upper equations say that the vector $C = (c^1, c^2, c^3)$ is irrotational, and thus there must be a potential function $U(X)$ on $\Re$ fulfilling:

$$\mathbf{C} = \nabla U \tag{6.152}$$

where ∇ denotes the gradient operator. Let us show that we can set additionally without loss of generality that:

$$c^1 = c^2 = c^3 = 0 \tag{6.153}$$

Consider an arbitrary function $h(X)$ on $\Re$ fulfilling $h > 0$, and a Lorentz metric G fulfilling the general optical equations Eq. (6.150) for certain functions c^k. As can be easily checked, the matrix $G' = hG$ also defines a Lorentz metric. Moreover, both metrics are equivalent for our purpose because they define the same cone structure:

$$\mathbf{Y}^t\mathbf{G}\mathbf{Y} = 0 \Leftrightarrow \mathbf{Y}^t\mathbf{G}'\mathbf{Y} = \mathbf{Y}^t[h\mathbf{G}]\mathbf{Y} = h[\mathbf{Y}^t\mathbf{G}\mathbf{Y}] = 0 \tag{6.154}$$

Let U be the corresponding potential function of C, and thus $c^k = U_k$. Taking $h = e^U$, and since the metric $G' = (g'_{ij}) = e^U G$ defines the same cone structure as G, obviously G' also fulfills the general optical condition. By substitution of $G = e^{-U}G'$ in Eq. (6.150), we immediately find that the functions c'^k associated to G' are of the form given by Eq. (6.153).

The integration of Eq. (6.150), taking into account Eq. (6.153), is straightforward. It leads to the following matrix G:

$$\begin{aligned}
g_{11} &= ax_2^2 + bx_3^2 + cx_2x_3 + dx_2 + ex_3 + f \\
g_{12} &= (rx_3 + s)x_2 - tx_3^2 + \gamma x_3 + \delta - \frac{1}{2}(2ax_2 + cx_3 + d)x_1 \\
g_{13} &= (tx_2 + u)x_3 - rx_2^2 + ax_2 + \beta - \frac{1}{2}(2bx_3 + cx_2 + e)x_1 \\
g_{22} &= mx_3^2 - 2Bx_3 + p - 2(rx_3 + s)x_1 + ax_1^2 \\
g_{23} &= -mx_2x_3 + Bx_2 + Dx_3 + E - (-tx_3 + \gamma - rx_2 + \alpha)x_1 + \frac{1}{2}cx_1^2 \\
g_{33} &= mx_2^2 - 2Dx_2 + q - 2(tx_2 + u)x_1 + bx_1^2
\end{aligned} \tag{6.155}$$

where a, b, c, d, e, f, p, q, r, s, t, u, m, B, D, E, α, β, γ, and δ are real numbers that appear as constants of integration. There are other conditions, apart from the optical conditions stated in Section 6.17.1, that we have to impose on the matrix G in order to describe the elliptic bundles. The values of the parameters in Eq. (6.155) must be such that G defines a cone—that is, that the determinant of G does not vanish, and its signature is $(-, +, +)$, as indicated in the Introduction. These conditions define the region $\Re$ where the elliptic bundle exists.

This matrix G was first discussed in Benítez (1999), with an approach based on the application of a series expansion from the bundle defined at a reference surface.

On the other hand, the elliptic bundles described recently using the Poisson bracket design method in curvilinear coordinates (Miñano and Benítez, 1999) coincide with the subset of elliptic bundles given by Eq. (6.155) with rotational symmetry (see the next section on imposing this condition) and when $\gamma = 0$. With this approach, the coordinate curves of the curvilinear coordinate system coincide with the curves tangent to the eigenvectors of the matrix G. The bundles with $\gamma \neq 0$ are not found with this approach because no curvilinear coordinate system can be defined in this case. This is due to the nonintegrability of the eigenvector field, as shown in Benítez (1999) and in the next section.

With respect to the solution of the restricted optical condition (Gutiérrez et al., 1996), it is equivalent to Eq. (6.155), except only 10 of the 20 parameters in Eq. (6.155) are independent. Therefore, for the elliptic bundles given by Eq. (6.155) that do not fulfill the restricted optical condition, the edge rays are not geodesics of the Lorentz metric.

Therefore, all the known elliptic bundles are the solutions of the general optical equations when the vector C is irrotational. Finding new elliptic bundles when the vector C is not irrotational is an open problem and deserves further research. We have checked that not all vectors C provide elliptic bundles. For instance, let consider the vectors C of the form

$$\mathbf{C} = \nabla\Phi + x_i e_k \quad i \neq k \tag{6.156}$$

where Φ is an arbitrary C^1 function in $\Re$, and e_k is one of the vectors of the canonical basis. It is straightforward to check that $\nabla \times C \neq 0$ (where $\nabla \times$ denotes the rotational operator). By direct calculations we have proven that the only matrix G fulfilling the general optical equations for this vector C is the null matrix (which is the trivial degenerate solution). What are the conditions on the vector C for the existence of solutions of the system Eq. (6.150)? This is a problem to work on. Necessary conditions may come out from the integrability conditions of the system, which are given by the equality of the second-order cross derivatives of the matrix coefficients ($g_{ij,mk} = g_{ij,km}$).

6.17.4   Examples

Let us consider the subset of the elliptic bundles given by Eq. (6.155) with rotational symmetry with respect to the x_1 axis. Imposing this symmetry (Benítez, 1999) lead to the following conditions for the parameters in the matrix G in Eq. (6.155)

$$a = b, \quad p = q, \quad s = u, \quad \alpha = -\gamma, \quad c = d = e = r = t = B = D = E = \beta = \delta = 0 \tag{6.157}$$

Therefore, the elliptic bundles in homogeneous media with rotational symmetry when that $\nabla \times C = 0$ are described by the matrix

$$\mathbf{G} = \begin{bmatrix} a(x_2^2 + x_3^2) + f & (-ax_1 + s)x_2 + \gamma x_3 & (-ax_1 + s)x_3 - \gamma x_2 \\ (-ax_1 + s)x_2 + \gamma x_3 & mx_3^2 + ax_1^2 - 2sx_1 + p & -mx_2 x_3 \\ (-ax_1 + s)x_3 - \gamma x_2 & -mx_2 x_3 & mx_2^2 + ax_1^2 - 2sx_1 + p \end{bmatrix} \tag{6.158}$$

The subset of the rotational-symmetric solutions Eq. (6.158) that fulfill the restricted optical equations are the ones given by Eq. (6.158) setting $\gamma = 0$ and $ap-s^2 = mf$. Although the bundles not fulfilling these two conditions are different from those fulfilling it, the concentrators that are designed by the application of the flow line method are the same ones. The flow line concentrator design method is based on the fact that if the reflector surface is tangent to one of the symmetry planes of the elliptic cones at any point, the reflection does not disturb the bundle, in other words, the reflector exchanges ray trajectories among rays of the bundle, but the cones formed by the rays of the bundle remain the same.

Therefore, Eq. (6.158) contains, for instance, apart from the elliptic bundle defined by a disk, which is ideally concentrated in 3D by the FLC, other bundles that are also ideally transmitted in 3D by this well-known concentrator. Let us examine just two representative examples of solutions that do not fulfill the restricted optical equations.

A. Case $\gamma = 0$, $s = 0$, $a = 0$, $p < 0$, $f > 0$, $m > 0$

For $m = 0$, the trivial bundle was obtained, for which the matrix $G = \mathrm{diag}(f, p, p)$, independent of X, represents an elliptic bundle (for $f < 0$, $p > 0$) of vertical and circular cones that are equal at all points in space. Again, the additional degree of freedom $m \neq 0$ provides four new nontrivial types of elliptic bundles. Let us consider the one obtained for the parameters $p < 0$, $f > 0$, $m > 0$ (that we will call bundle A). Introducing the values of the parameters in the matrix Eq. (6.158), and particularizing in the plane $x_3 = 0$

$$G|_{x_3=0} = \begin{bmatrix} f & 0 & 0 \\ 0 & p & 0 \\ 0 & 0 & mx_2^2 + p \end{bmatrix} \tag{6.159}$$

which is also a diagonal matrix. The eigenvalues of G are, obviously, the elements of the principal diagonal, and their eigenvectors are the canonic basis. Through revolution symmetry it is deduced that the eigenvectors corresponding to $\lambda_1 = f$ and $\lambda_2 = p$ are parallel to the meridian planes—specifically, $(1, 0, 0)$ is an eigenvector of $\lambda_1 = f$ and $(0, x_2, x_3)$ of $\lambda_2 = p$. The eigenvectors corresponding to $\lambda_3 = mx_2^2 + p$ are perpendicular to the meridian planes. The region $\mathfrak{R}$ where the bundle exists (i.e., where the signature $(-, +, +)$ is obtained) is $x_2^2 + x_3^2 > -p/m$—that is, the exterior of the cylinder defined by the equation $|G| = 0$. The eigenvalue corresponding to vector J (which is tangent to the flow line) is λ_2.

Figure 6.39a shows some edge-ray cones of this bundle. The cone axes are horizontal pointing toward the axis X_1. It can be easily proven that the edge rays contained in horizontal planes are tangent to the surface $|G| = 0$. The vertical semiaxis of the ellipse has a constant value, and the horizontal axis increases as ρ decreases, becoming infinite on $|G| = 0$.

As the flow lines are horizontal straight lines and are contained in meridian planes, it is deduced that the CC with linear symmetry, shown in Figure 6.39b, transmits this bundle ideally (independently of the constant semiaxis value), as well as the conventional linear bundle (into which this bundle A degenerates when $f = 0$). This means that the linear CC collects ideally any elliptic bundle tangent to the cylinder that defines it, maintaining the value of the semiaxis of the cones of edge rays in the direction of the cylinder axis.
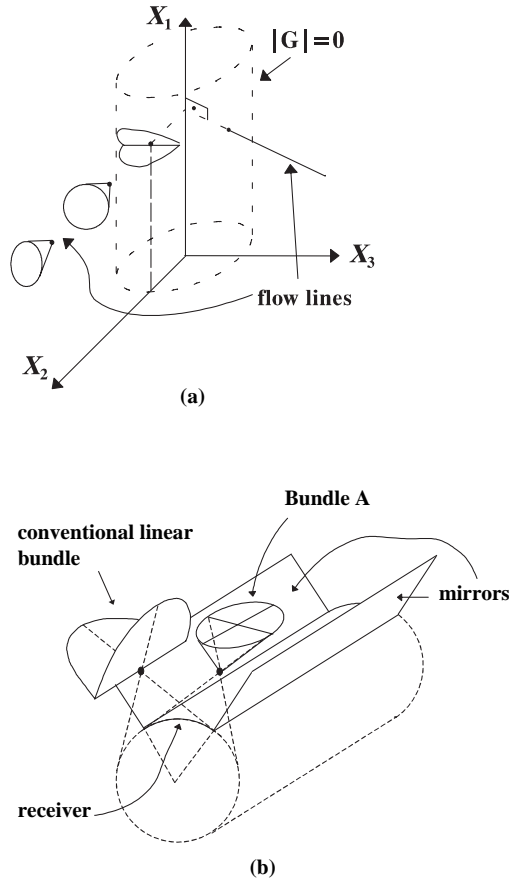
Figure 6.39   Bundle A: (a) Edge rays and flow lines. (b) This bundle is ideally transmitted by the cone concentrator with linear symmetry.

The preceding example is a particular case of the general solution in which $\gamma = 0$. Let us consider now that $\gamma \neq 0$. In this situation the vector $(0, 0, 1)$ is not an eigenvector of the matrix G in $x_3 = 0$, as occurred in the bundles with $\gamma = 0$ (including the solutions of the restricted optical equations). This implies that the meridian planes are not planes of symmetry of the cones of edge rays of the bundle. To illustrate this new situation, let us consider the following example.

B.   Case $s = 0$, $a = 0$, $m = 0$, $fp \neq 0$, $\gamma \neq 0$

Once again, if $\gamma = 0$, we obtained the trivial bundle. The condition $\gamma \neq 0$ leads to the appearance of two new nontrivial bundles, which will be called $B1$ and $B2$. The matrix G in $x_3 = 0$, given by Eq. (6.158), for the parameter values of this case, takes the form

$$G|_{x_3=0} = \begin{bmatrix} f & 0 & \gamma x_2 \\ 0 & p & 0 \\ \gamma x_2 & 0 & p \end{bmatrix} \qquad (6.160)$$

whose eigenvalues are

Table 6.1. Some characteristics of the elliptic bundles with $\gamma \neq 0$, s = 0, a = 0, m = 0, pf $\neq$ 0 $\left(\rho = \sqrt{x_2^2 + x_3^3}\right)$.

Case	p	f	$\mathfrak{R}$	Negative Eigenvalue
$B1$	>0	<0	R^3	λ_3
$B2$	>0	>0	$\rho > \dfrac{\sqrt{fp}}{|\gamma|}$	λ_3

$$\lambda_1 = \frac{(f + p) + \sqrt{(f - p)^2 + 4\gamma^2 + x_2^2}}{2}$$

$$\lambda_2 = p \tag{6.161}$$

$$\lambda_3 = \frac{(f + p) - \sqrt{(f - p)^2 + 4\gamma^2 x_2^2}}{2}$$

The range $p > 0$ leads to the two elliptic bundles $B1$ and $B2$ described in Table 6.1. Both bundles have the negative eigenvalue λ_3. The eigenvalue $\lambda_2 = p$ has as its eigenvector (0, 1, 0) in $x_3 = 0$, as is easily confirmed. Through revolution symmetry it is deduced that the eigenvector of λ_2 has always $y_1 = 0$ and is parallel to the meridian planes, and that the eigenvectors of λ_1 and λ_3 are normal to this direction and therefore tangent to cylinders of equation $\rho = cte$. In the case of the eigenvector J (corresponding to λ_3), from its equation $GJ = \lambda_3 J$ it is found that the angle it forms with the vertical is

$$\phi = \text{arctag}\left(\frac{2\gamma\rho}{-(p - f) + \sqrt{(p - f)^2 + 4\gamma^2\rho^2}}\right) \tag{6.162}$$

Figure 6.40a-b represent schematically the bundles $B1$ and $B2$ for the case $\gamma > 0$. In elliptic bundle $B1$, the value of the angle ϕ is zero (i.e., the vector J is vertical) on the axis X_1 and increases as ρ increases, tending toward the value $\phi = 45°$. Bundle $B2$ only exists outside the cylinder $\rho > (fp)^{1/2}/|\gamma|$, on whose surface the two semiaxes of the elliptic cone vanish. A particular case of bundle $B2$, $p = f$, $\gamma > 0$, is interesting because the angle $\phi = 45°$ for all points in $\mathfrak{R}$.

The flow lines of the bundles $B1$ and $B2$ are helices with constant pitch (given by the angle ϕ of Eq. (6.162)) and centered with respect to axis X_1 (see Figure 6.40a–b). The corresponding concentrator is the vertical cylinder, which also transmits the trivial bundle. An important difference with the bundles A (and with all bundle when $\gamma = 0$) is that bundles $B1$ and $B2$ verify that the vector fields corresponding to the eigenvectors parallel to the cylinder $\rho = cte$ (U and J) cannot be integrated—that is, there does not exist a continuous surface normal to each vector field. This is deduced from the integrability condition of the vector field, which they do not fulfill (i.e., $U \cdot \nabla \times U \neq 0$, and $J \cdot \nabla \times J \neq 0$), as can be easily seen by direct calculations. Figure 6.40c shows this for the particular case $p = f$ of bundle $B2$.

6.18 CONCLUSION

Using poisson brackets is a smart way to establish the general nonimaging design problem. Nevertheless it is not by itself a useful tool unless the bundle to work
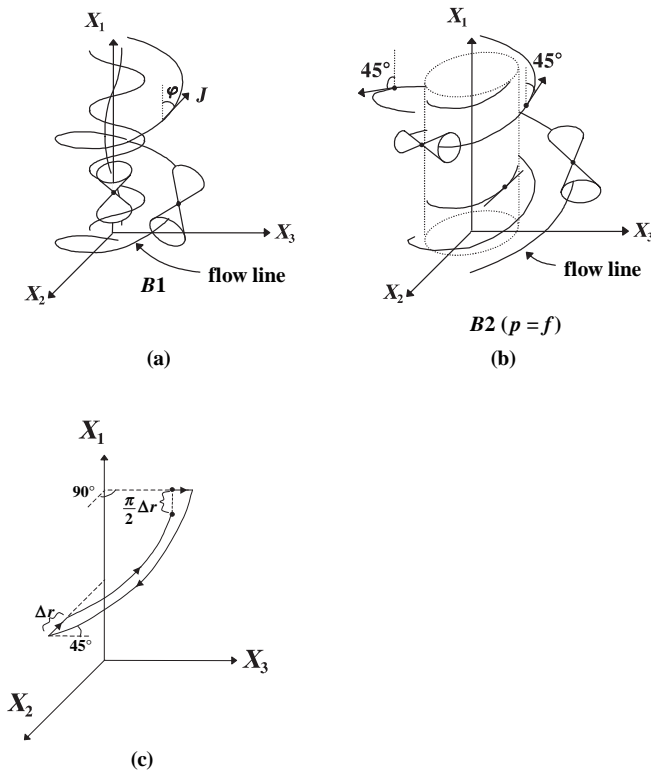
Figure 6.40 In elliptic bundles $B1$ (a) and $B2$ (b), the cones of edge rays are not symmetrical with respect to the meridian planes. The flow lines are helices with constant pitch. (c) In bundle $B2$ with $p = f$, the path shown is normal at every point to U if $\gamma > 0$ and to J if $\gamma < 0$, but it is not closed. This is equivalent to the nonintegrability of these vector fields.

with has remarkable characteristics. This is the case of elliptic bundles whose flow lines are coincident with one of the coordinate lines of a three-orthogonal system. The equation defining the elliptic bundle in this case (Eq. (6.28)) is an expression of the same type as the Hamiltonian function H (Eq. (6.17)) in the Hamiltonian formulation that we use. The condition for $\omega = 0$ to be a particular integral of the Hamiltonian system (Eq. (6.26)) results in a system of three linear equations where the independent variables are the squares of the conjugate variables. This allows us to reach relatively simple solutions. In fact, these solutions are connected with the general solutions of the Eikonal equation with arbitrary orthogonal coordinates and separation of variables.

The condition for ω to be a particular integral of the Hamiltonian system (Eq. (6.26)) is symmetrical with respect to ω and H. Therefore, if a bundle defined by ω is a particular integral of a Hamiltonian system with function H, then H is also a particular integral of a Hamiltonian system with function ω. This reasoning would lead us to the concept of dual concentrators if ω, as a Hamiltonian function, represents an optical system. In the case of elliptic bundles whose flow lines are coincident with one of the coordinate lines of a three-orthogonal system, the expression $\omega = 0$ has the same form as $H = 0$, and so the function ω may be the Hamiltonian function of an optical system. All we need is that the functions $|\alpha_1|$,

$|\alpha_2|$, $|\alpha_3|$ be the modulus of the gradients of the coordinates of a three-orthogonal system over an arbitrary positive function (this positive function would be the square of the refractive index distribution). The conditions that three functions must fulfill to be the modulus of the gradients of the coordinates of a three-orthogonal system are not simple. These conditions are a system of six differential equations called the Lamé equations (Benítez, 1999; Gutiérrez et al., 1996). Owing to this difficulty, the problem of finding the dual concentrator (or more exactly, the dual bundle) remains unsolved.

One of the reasons to study elliptic bundles is that at any point these bundles are symmetrical with respect to three planes, and this allows applying the flow-line method to build up the reflectors. Only one plane of symmetry is required for the application of the flow-line design method. Moreover, the existence of such symmetry is only needed at the surface where the reflector is going to be placed. Therefore, we are clearly reducing the degrees of freedom with respect what is strictly needed. Fortunately, enough degrees of freedom still remain to find a solution. These considerations suggest that there may be a much wider set of solutions (i.e., of ideal 3D concentrators) than the ones found here and that it is very likely that ideal 3D concentrators with prescribed apertures and with constant refractive index exist.

In fact, the multifoliate concentrator can be considered as one of these cases. In this chapter we have analyzed the multifoliate concentrator with the Poisson brackets tools, and this analysis has opened new possibilities for designing ideal concentrators. If a concentrator is found such that the flow lines are contained in surfaces of constant refractive index, then a multifoliate solution using constant refractive index may exist.

For the case of elliptic bundles in a homogeneous medium, we have presented both the Lorentz geometry and the Poisson bracket approaches (applied in Cartesian coordinates). The latter provides the general partial differential equations that must be accomplished by all possible elliptic bundles in a homogeneous medium. As a result, all the known elliptic bundles, which were found by other methods, are obtained as a particular solution of the system, and the question about the existence of other bundles remains open.

REFERENCES

Abraham, R., and Marsden, J. E. (1978). *Foundations of Mechanics*. Benjamin Cummings, Reading, MA.

Arnaud, J. A. (1976). *Beam and Fiber Optics*. Academic Press, New York.

Arnold, V. I. (1974). *Les Méthodes Mathematiques de la Mécanique Classique* (French translation). Mir, Moscow.

Bassett, I. M., and Forbes, G. W. (1982). A new class of ideal non-imaging transformers. *Optica Acta* **29**, 1271–1282.

Benítez, P. (1999). Elliptic bundles in three-dimensional geometry for nonimaging optics: A new approach. *J. Opt. Soc. Am. A.* **16**(9), 2245–2252.

Castillo, E. (1996). *Algunas aplicaciones de las ecuaciones funcionales*. Servicio de Publicaiones de la Universidad de Cantabria, Santander, Spain.

Castillo, M. R., and Ruiz-Cobo, M. R. (1992). Functional Equations and Modelling in Science and Engineering. Marcel Dekker, NY.

Darboux, G. (1910). *Leçons sur les systèmes ortogonaux et les Coordeneés*. Currilignes, Gauthier-Villar, Paris, France.

Eisenhart, L. P. (1925). Reimannian Geometry. Princeton University Press, Princeton, NJ.

Forbes, G. W., and Bassett, I. M. (1982). An axially symmetric variable-angle nonimaging transformer. *Optica Acta* **29**, 1283–1297.

Garwin, R. L. (1952). The design of liquid scintillation cells. *Rev. Sci. Instrum*. **23**, 755–757.

Gutiérrez, M., Miñano, J. C., Vega, C., and Benítez, P. (1996). Applications of the Lorentz geometry to nonimaging optics: New three-dimensional ideal concentrators. *J. Opt. Soc. Am. A*. **13**(3), 532–540.

Kline, M., and Kay, I. W. (1965). *Electromagnetic Theory and Geometrical Optics*. Interscience Publishers, New York.

Leech, J. W. (1958). *Classical Mechanics*. Methuen, London.

Luneburg, R. K. (1964). *Mathematical Theory of Optics*. University of California Press, Berkeley.

Miñano, J. C. (1985a). New family of 2-D non-imaging concentrators: The Compound Triangular Concentrators. *Applied Optics* **24**, 3872–3876.

Miñano, J. C. (1985b). Refractive index distribution in two-dimensional geometry for a given one-parameter manifold of rays. *J. Opt. Soc. Am. A*. **2**, 1821–1825.

Miñano, J. C. (1985c). Two-dimensional non-imaging concentrators with inhomogeneous media. A new look. *J. Opt. Soc. Am. A*. **2**, 1826–1831.

Miñano, J. C. (1986). Design of three-dimensional nonimaging concentrators with inhomogeneous media. *J. Opt. Soc. Am. A*. **3**(9), 1345–1353.

Miñano, J. C. (1990). Optical confinement in photovoltaics. In *Physical Limitations to Photovoltaic Energy Conversion* (Luque, A., and Araujo, G. L., eds.). Adam Hilgher, Bristol, UK.

Miñano, J. C. (1993a). Poisson brackets method of design of nonimaging concentrators: A review. In *Nonimaging Optics: Maximum-Efficiency Light Transfer II*. (Winston, R., and Holman, R. L., eds.). Proc. SPIE 2016, 98–108.

Miñano, J. C. (1993b). The Poisson Brackets method of design of nonimaging concentrators. Optical Society of America Annual Meeting. Optical Society of America, Washington, DC. Postconference Edition. Vol. 16, 315.

Miñano, J. C., and Benítez, P. (1999). Poisson bracket design method review. Application to the elliptic bundles. In *Nonimaging Optics: Maximum-Efficiency Light Transfer V*. (Winston, R., ed.). Proceedings of SPIE. Vol. 3781, 2–11.

O'Neill, B. (1983). *Semi-Reimannian Geometry with Applications to Relativity*. Academic, New York.

Stavroudis, O. N. (1972). *The Optics of Rays, Wavefronts, and Caustics*. Academic, New York.

Weisstein, E. W. (1999). *CRC Concise Encyclopedia of Mathematics*. CRC Press, Boca Raton.

Welford, W. T., and Winston, R. (1989). *High Collection Nonimaging Optics*. Academic Press, New York.

Welford, W. T., O'Gallagher, J., and Winston, R. (1987). Axially symmetric nonimaging flux concentrators with the maximum theoretical concentration ratio. *J. Opt. Soc. Am. A* **4**, 66–68.

Winston, R., and Welford, W. T. (1979). Geometrical Vector Flux and some new imaging concentrators. *J. Opt. Soc. Am. A*. **69**, 532–536.

7

CONCENTRATORS FOR PRESCRIBED IRRADIANCE

7.1 INTRODUCTION

Generally speaking, designs in Nonimaging Optics act as transformers of extended ray bundles. The bundle of rays impinging on the surface of the entry aperture of the nonimaging device is called the input bundle and is denoted by M_i. The bundle of rays that exits the surface of the exit aperture of the concentrator is the exit bundle M_o. Collected bundle M_c is the name given to the set made up of the rays common to M_i and M_o, connected to one another by means of the concentrator.

There are two main groups of design problems in Nonimaging Optics. Although both groups have usually been treated separately in the nonimaging literature, there are examples of design problems, such as photovoltaic concentrators, that belong to both groups.

For the first group, which can be referred to as "bundle-coupling" problems, the design problem consists in specifying the bundles M_i and M_o, and the objective is to design the concentrator to couple the two bundles—that is, making $M_i = M_o = M_c$. This design problem is the one that has been considered in previous chapters.

This chapter is devoted to the second group of design problems, usually referred to as "prescribed irradiance" or "prescribed intensity" type. In this case, it is only specified that one bundle must be included in the other—for example, M_i in M_o (so that M_i and M_c will coincide)—with the additional condition that the bundle M_c produces a certain prescribed irradiance distribution on one target surface at the output side. Since M_c is not fully specified, this problem is less restrictive than the bundle-coupling one (this states, for instance, that some design problems can be exactly solved with a single optical surface).

The prescribed irradiance/intensity problems appear in most of the illumination applications. For instance, in automotive lighting, the light source is a lightbulb or an LED, and the target surface is the far-field sphere, where the intensity distribution is prescribed. It is also interesting in nonillumination applications. For example, it is used for wide-angle ceiling IR receivers in indoor wireless communications, where the receiver sensitivity is prescribed to compensate the different link distances for multiple emitters on the desks (in this case, M_o is included

in M_i, and the irradiance distribution is prescribed at the input side, at the plane of the desks). The receiver angular sensitivity plays the same role as the intensity pattern in a luminary design.

Although there are more degrees of freedom in a "prescribed intensity" design, designing in 3D geometry is still quite complex. The last section of this chapter is devoted to an overview of some of these design methods, in particular those that simplify the problem considering the source as a point (nonextended) source.

7.2   REFLECTOR PRODUCING A PRESCRIBED FUNCTIONAL TRANSFORMATION

For illustrative and pedagogical reasons, we will first consider the simpler case in which the sources are small compared to the other parameters of the problem. This problem was studied by Boldirev (1932), Komissarov (1941), and Elmer (1980). The size of the source may be much smaller than the closest distance of approach to any reflective or refractive component, or the angle subtended by the source at any reflective or refractive component may be much smaller than the angular divergence of the beam (point source approximation). Then it is useful to consider the limiting case where the source has no extent. We will idealize the source by a line or point with negligible diameter and seek a one-reflection solution.

Using polar coordinates R, ϕ from the source as origin and θ for the angle of the reflected ray, the geometry in Figure 7.1 shows that the following relation between source angle and reflected angle holds:

$$\frac{d(\log R)}{d\phi} = \tan\alpha \qquad (7.1)$$

where α is the angle of incidence with respect to the normal. Clearly,

$$\alpha = \frac{\phi - \theta}{2} \qquad (7.2)$$

Eq. (7.1) is readily integrated,

$$\log R = \int \tan\alpha \, d\phi + \text{const} \qquad (7.3)$$
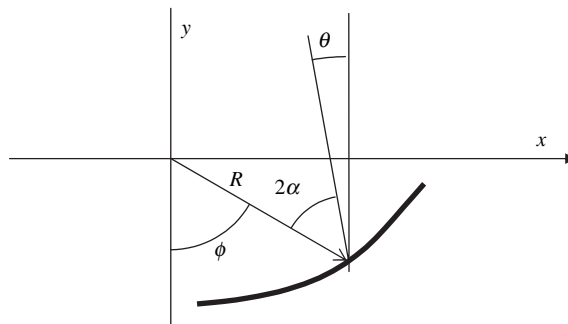
Figure 7.1 Reflector design for a point source producing a prescribed intensity distribution.

so that,

$$R = \text{const} \cdot \exp\left(\int \tan\alpha \, d\phi\right) \qquad (7.4)$$

Eqs. (7.4) and (7.2) determine the reflector profile $R(\phi)$ for any desired functional dependence $\theta(\phi)$.

7.3   SOME POINT SOURCE EXAMPLES WITH CYLINDRICAL AND ROTATIONAL OPTICS

Suppose we want to radiate power with a particular angular distribution $P_o(\theta)$; which we assume to be axially symmetric. For example, $P_o(\theta) = \text{const.}$ from $\theta = 0$ to θ_1 and $P_o(\theta) \approx 0$ outside this angular range. If the line source radiates an angular distribution $P_i(\phi)$, then, since by conservation of energy $P_o(\theta)\,d\theta = \pm P_i(\phi)d\phi$ (neglecting material reflection loss), we need only to ensure that

$$\frac{d\theta}{d\phi} = \pm \frac{P_i(\phi)}{P_o(\theta)} \qquad (7.5)$$

to obtain the desired radiated beam profile. The alternative in the sign corresponds to two possible families of solutions, depending on the sign of the derivative $d\theta/d\phi$. The positive sign is usually called converging or elliptic type surface, and the negative is called diverging or hyperbolic type. The difference between these two cases can also be seen because the rays after the reflection cross in the elliptic case and do not cross in the hyperbolic case (i.e., the caustic of the exit-ray bundle is real in the elliptic case and virtual in the hyperbolic case).

To illustrate the method, consider the preceding example of a constant $P_o(\theta)$ for a rotational symmetric line source. By this rotational symmetry, $dP_i(\phi)/d\phi = \text{constant}$ so that, according to Eq. (7.5) we want θ to be a linear function of ϕ—say, $\theta = a \cdot \phi$ (with the boundary condition that $\theta = 0$ at $\phi = 0$). Then the solution of Eq. (7.4) is

$$R = \frac{R_o}{\cos^k\left(\dfrac{\phi}{k}\right)} \qquad (7.6)$$

where,

$$k = \frac{2}{1-a} \qquad (7.7)$$

and R_o is the value of R at $\phi = 0$.

We note that the case $a = 0$ ($k = 2$) gives the parabola in polar form,

$$R = \frac{R_o}{\cos^2\left(\dfrac{\phi}{2}\right)} \qquad (7.8)$$

while the case $\theta = \text{constant} = \theta_1$ gives the off-axis parabola,

$$R = R_o \, \frac{\cos^2\left(\dfrac{\theta_1}{2}\right)}{\cos^2\left(\dfrac{\phi - \theta_o}{2}\right)} \qquad (7.9)$$

Suppose we desire instead to illuminate a plane with a particular intensity distribution. Then we correlate position on the plane with angle θ and proceed as before.

Turning next to a spherically symmetric point source, we consider the case of an exit constant $P_o(\Omega_o)$, where Ω_o is the radiated solid angle. Now we have by energy conservation,

$$P_i(\Omega_i)d\Omega_i = P_o(\Omega_o)d\Omega_o \tag{7.10}$$

where Ω_o is the solid angle radiated by the source. By spherical symmetry of the point source, $P_o(\Omega_o) = $ constant. Moreover, we have $d\Omega_o = 2\pi\, d(\cos\theta)$ and $d\Omega_i = 2\pi d(\cos\phi)$, so we need to make $\cos\theta$ a linear function of $\cos\phi$,

$$\cos\theta = a\cos\phi + b \tag{7.11}$$

With the boundary conditions that $\theta = 0$ at $\phi = 0$, $\theta = \theta_1$ at $\phi = \phi_o$, we obtain,

$$a = \frac{1 - \cos\theta_1}{1 - \cos\phi_0} \tag{7.12}$$

$$b = \frac{\cos\theta_1 - \cos\phi_0}{1 - \cos\phi_0} \tag{7.13}$$

This functional dependence is applied to Eq. (7.6), which is then integrated, perhaps numerically.

7.4   THE FINITE STRIP SOURCE WITH CYLINDRICAL OPTICS

7.4.1   Flow Line Mirror

We have now developed the analytical tools to solve the real problems that involve finite size sources. We do this combining the previous technique with the edge ray method that has proved so effective in nonimaging designs—that is, we apply the preceding methods to rays from the periphery. As an example, we apply the preceding methods to a planar Lambertian source in order to achieve a far-field illumination of predetermined shape. Looking at Figure 7.2 we design the edge rays of the nonimaging reflector to satisfy the condition.
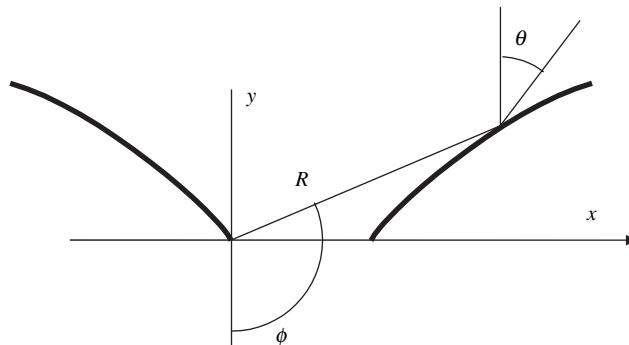
Figure 7.2   Flow line type reflector design for a strip emitter and a prescribed exit intensity distribution.

$$r \sin(\phi + \theta) \cos^n \theta = f(\theta) \tag{7.14}$$

The exponent n together with $f(\theta)$ determine the far-field illuminance. For example, for uniform far-field illuminance from a two-dimensional (strip) source, $f(\theta) = $ constant and $n = 2$ are appropriate, since $\sin(\phi + \theta) \cos^2 \theta$ just cancels the $\cos^3(\theta)$ fall-off in illuminance of a bare source. On the other hand, for rotationally symmetrical sources whose far-field illuminance would fall off as $\cos^4(\theta)$, $f(\theta) = $ constant and a value of n between 2 and 3 would be appropriate. After performing some algebra and using the identity $\tan 1/2(\phi + \theta) + \cot(\phi + \theta) = 1/\sin(\phi + \theta)$, we obtain

$$\frac{1}{\sin(\phi + \theta)} + \cot(\phi + \theta) \frac{d\theta}{d\phi} - n \tan \theta \frac{d\theta}{d\phi} = \frac{d(\log f)}{d\theta} \frac{d\theta}{d\phi} \tag{7.15}$$

This can be simplified by the substitutions $\psi = \phi + \theta - \pi/2$ so that $d\psi/d\theta = 1 + d\psi/d\theta$, $\sin(\phi + \theta) = \cos \psi$, $\cos(\phi + \theta) = -\sin \psi$. Then,

$$\frac{d\psi}{d\phi} = 1 + \sin \psi + n \tan \theta \cos \psi + \frac{d(\log f)}{d\theta} \tag{7.16}$$

This equation would be solved by standard numerical methods with starting values $\psi = 0$ at $\theta = 0$. Many other cases have been worked out along the lines illustrated in the example. See, for example, Winston and Ries (1993).

7.4.2   Sequential Optical Surface

It is also possible to produce in 2D geometry (i.e., for the meridian rays of a strip source) a prescribed intensity for an emitter or prescribed sensitivity for a receiver by a means of a sequential optical surface. This design procedure was first proposed by Ries and Winston (1994) for reflectors and tubular source (see Section 7.6). In the example considered in this section (Hernández, 2003), a refractive surface is used, which corresponds to the upper surface of a lens that encapsulates a flat receiver (see Figure 7.3). This specific example will be useful for the comparison with the design in 3D geometry, shown in Section 7.5.

Let $L_{ap}(\theta)$ be the prescribed sensitivity of the device, which is assumed for this example to be symmetric with respect to $\theta = 0$ and to extend in the angular range $-\theta_M \le \theta \le \theta_M$, and to be null outside that range. $L_{ap}(\theta)$ has units of length and
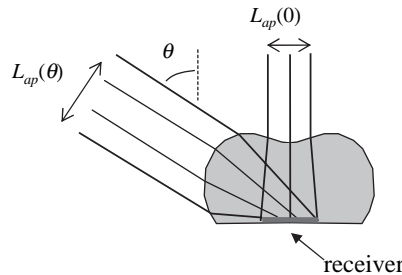
Figure 7.3   Lens to be calculated to provide a prescribed sensitivity $L_{ap}(\theta)$ to the flat receiver in two dimensions.
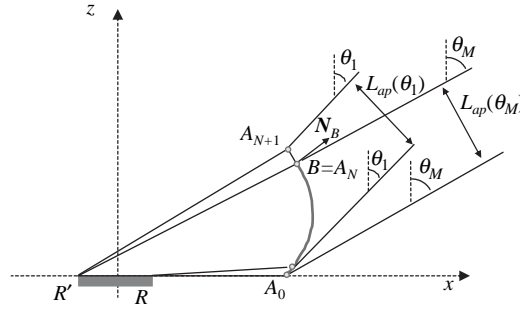
Figure 7.4 First steps of the design procedure for a refractive surface that produced a pre-scribed sensitivity in two dimensions.

(neglecting the optical losses) can be understood as the apparent projected length of the receiver as seen through the lens (see Figure 7.3).

Maximum concentration is required, and thus the étendue conservation theorem states an integral condition over the specification $L_{ap}(\theta)$, given by

$$2nL = \int_0^{\theta_M} L_{ap}(\theta) \cdot d\theta \qquad (7.17)$$

where L is the receiver length and n is the lens refractive index.

The input parameters of this design are $L_{ap}(\theta)$, d, n, and the x coordinate of the point A_0 (shown in Figure 7.4) of the lens along the receiver line. The design procedure has in general two types of solutions, converging and diverging (Ries and Winston, 1994). Due to the limited power of the refractive surface to deflect the ray trajectories, the converging solution is not practical, and we will refer only to the diverging one. These are the steps to follow.

1. Select point A along the receiver line.
2. Calculate the portion of the Cartesian oval passing through A_0 that focus the parallel rays impinging from direction $\theta = \theta_M$ onto point R' (See Figure 7.4). The edges of this portion of the Cartesian oval are A_0, and the point B, which is the one fulfilling that the projection of the segment $A_0 - B$ normal to the direction $\theta = \theta_M$ toward equals the prescribed value $L_{ap}(\theta_M)$. Take N points of the oval in point A_j, $j = 0, \ldots N$ (where $A_0 = A$ and $A_N = B$) and their corresponding normal vectors $\mathbf{N}_j$ to the surface.
3. Refract (inversely) at point $(A_1, \mathbf{N}_1)$ the ray r_1 emitted from R, and take the resulting direction angle $\theta = \theta_1$ (note that $\theta_1 < \theta_M$).
4. Calculate the new point A_{N+1} of the refractive surface as the intersection point between the straight line passing through the last point A_N with direction perpendicular to the normal vector $\mathbf{N}_N$ and the ray r_1' impinging from direction $\theta = \theta_1$, whose distance to ray r_1 is $L_{ap}(\theta_1)$.
5. Calculate the normal vector $\mathbf{N}_{N+1}$ by (inverse) application of the Snell's law considering that the ray r_1' is refracted toward R'.
6. Repeat steps 3 to 5 for tracing (inversely) through each successive point A_j (with $j = 2$ and so on) the ray r_j from R, calculating the direction angle θ_j (note that $\theta_j < \theta_M$), and new point A_{j+N} using the straight line through the last point A_{j-1+N} perpendicular to $\mathbf{N}_{j-1+N}$ and the ray r_j' (distant $L_{ap}(\theta_j)$ to r_j).
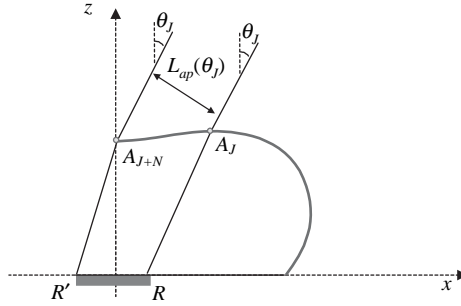
Figure 7.5   The design procedure from edge to center has not enough degrees of freedom to fulfill the prescribed sensitivity in the range $0 < \theta < \theta_J$.
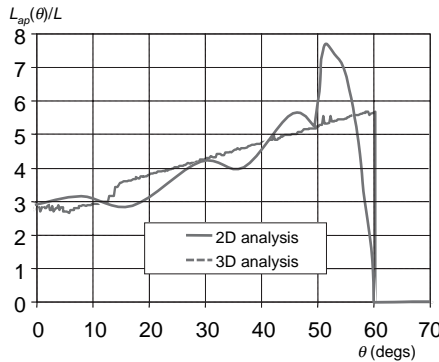
Figure 7.6   In continuous line, the 2D sensitivity ray-trace of a refractive lens analyzed by ray-trace designed for a prescribed linear sensitivity in $0 < \theta < \theta_M = 60°$ (and null outside) and design parameters are $L = 2\,$mm, $n = 1.49$, $|A_0| = 15\,$mm. As can be seen, the design does not follow the prescription for $\theta < \theta_J = 15.1°$. In dashed line, the 3D sensitivity of a rotational lens generated from the profile of the 2D design.

7. The design finishes when $j = J$ such that the point A_{J+N} reaches the z axis (see Figure 7.5).

In general, the resulting normal vector at the z-axis point, $\mathbf{N}_{J+N}$, is not parallel to the z-axis. Moreover, in general, the last designed direction $\theta_J > 0$ strictly, and thus there are no degrees of freedom to guarantee the prescribed sensitivity for $0 < \theta < \theta_J$.

Figure 7.6 shows the ray-tracing result (labeled as "2D analysis") over a lens designed using this procedure. The sensitivity was prescribed with the linear law

$$L_{ap}(\theta) = L_0 \cdot \left[\frac{\theta}{60°} + 1 \right] \tag{7.18}$$

from $0 < \theta(\text{degs}) < \theta_M = 60°$ and null outside that range. The design parameters are $L = 2\,$mm, $n = 1.49$, $A_0| = 15\,$mm, and it gets $|A_{J+N}| = 12.4\,$mm. As can be seen, the design does not follow the prescription for $\theta < \theta_J = 15.1°$.

The impossibility to adjust the prescribed sensitivity for $0 < \theta < \theta_J$ is intrinsic to the use of a single optical surface. The design can also be done from border to center, and then the prescription can be adjusted in the whole range $0 < \theta < \theta_M$, but the stepped transition to zero at $\theta = \theta_M$ cannot be achieved (i.e., the prescription for $\theta > \theta_M$ cannot be adjusted) because there are no degrees of freedom to make the outer portion of the lens be a Cartesian oval (as illustrated in Section 7.5 for the rotational case). However, it is possible to achieve the full-prescription (including the ranges $0 < \theta < \theta_J$ and $\theta > \theta_M$) if two surfaces are used, and the SMS design method (see Chapter 8) is applied. (This will also be illustrated in Section 7.5.)

In the bundle-coupling type problems, when the prescribed input and output ray bundles have rotational symmetry with respect to z-axis, it is usual to make a 2D design symmetric with respect to the z-axis, generate the 3D device by applying the rotation and analyze its 3D performance by ray-tracing. This leads to excellent results in many cases. However, if the same strategy is applied to prescribed-intensity problems, the results are not that good. This calculation has been done for the aforementioned 2D example, and the result is also shown in Figure 7.6 labeled "3D analysis." The expected linear sensitivity in 3D is given as apparent area by

$$A_{ap}(\theta) = A_0 \cdot \left[\frac{\theta}{60\,°} + 1 \right] \tag{7.19}$$

for $0 < \theta(\text{degs.}) < \theta_M = 60°$, and null outside this range. The low performance of the 3D lens generated from the 2D design suggests the need to design the rotational lens directly in 3D (using not only meridian but also skew rays). This is the topic of next section.

7.5   THE FINITE DISK SOURCE WITH ROTATIONAL OPTICS

A single rotational optical surface can also provide a prescribed rotational intensity for an extended source (Benítez et al., 2000). We will consider here again the example of a sequential refractive surface, as in Section 7.4.2.

The design procedure described can provide the desired angular response (the function $A_{ap}(\theta)$) for an optoelectronic receiver with a single rotational-symmetric refracting surface (see Figure 7.7). When the receiver is a photodiode or phototransistor, the material surrounding the receiving area may just substitute for the conventional optoelectronic encapsulation.
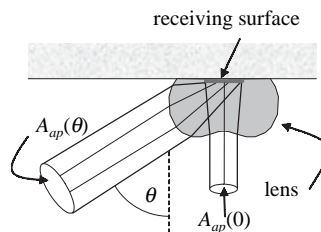
Figure 7.7   Receiver lens showing the trajectories of some rays defining $A_{ap}(\theta)$ and $A_{ap}(0)$.

Although the explanation in this section has points in common with the 2D case, it will be self-contained for clarity. The procedure can be applied (as the 2D procedure) for designing from rim to center, or the opposite. Since the 2D description was done from rim to center, in this section the center to the rim design will be used.

The angular response to be provided by the receptor or emitter is specified as input data for the design procedure. In the case, Figure 7.7, as an example, the lens provides the receiver set with an angular sensitivity inversely proportional to the Nth power of the cosine of the angle of incidence of the ray—that is, proportional to $1/\cos^N\theta$, N being an integer, for angle of incidence in a range Θ. This type of sensitivity for $N = 3$ is the one necessary if we wish, for example, to place the receiver on the ceiling of a room in which there is an isotropic emitter (that is, with an intensity independent of the angle of emission) situated at a certain height above the floor, and for the output signal from the receiver to be independent of the horizontal displacement of the emitter.

Other cases of note are obtained when the angular sensitivity provided by a lens to either receiver or the emitter is constant and when it is linear with θ. The latter case may be considered as an interesting approximation of these angular responses of the type $1/\cos^N\theta$, due to its simplicity.

In order to describe the design procedure we shall consider as an example the specific case in which the lens is composed of an axisymmetric aspherical refractive surface that illuminates a receiver, such as a photodiode with circular active area.

Let $I(\theta)$ be the prescribed angular dependence of the current photogenerated by the photodiode of the receptor when being illuminated by parallel rays of constant radiance and appropriate wavelength at angles in the range Θ. This specification is relative—that is, the photogenerated current will be $kI(\theta)$, where k is a constant. This is the same as specifying $k'A_{ap}(\theta)$, where k' is another constant. As an additional objective to that of obtaining this relative angular response $A_{ap}(\theta)$, we shall aim to make the value of the constant k' as high as possible. In order to get the highest possible k', no rays outside the range Θ must be collected and, additionally, the optical losses (Fresnel reflections, absorption, etc.) have to be reduced as much as possible.

A good approach consists in considering that the photogenerated current is proportional to the photon flux that penetrates the photodiode bulk. Thus, given an arbitrary optical system for illuminating the photodiode it is possible to evaluate this current within the framework of Geometric Optics by means of a conventional ray-tracing. If desired, losses of the optical system and those due to reflection on the photodiode surface can both be included in the design procedure so the desired $A_{ap}(\theta)$ is obtained taking into account these losses. The latter of the two types of losses may vary greatly according to whether the photodiode has an antireflective coating or not.

Figure 7.8 shows the coordinate system x-z that will be used to describe the design procedure of the lens. The y-axis, corresponding to the third dimension of the space, is perpendicular to the plane of that figure. R and R' are the points of intersection of the edge of the active area of the photodiode with the plane $y = 0$. As the system will have rotation symmetry with respect to the z-axis, without loss of generality, we can consider that the impinging rays will be parallel to the plane x-z with direction vector $\mathbf{U}(\theta) = (-\sin\theta, 0, -\cos\theta)$.
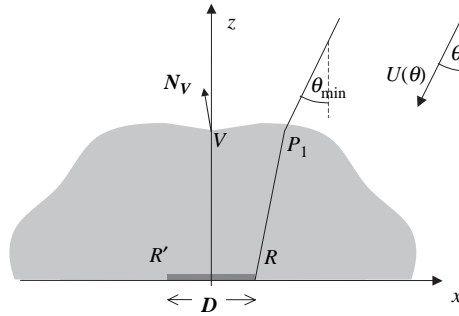
Figure 7.8   Coordinate system and contour conditions for the lens design. The receiver has an area $A = \pi(D/2)^2$.

To simplify the problem, the range Θ will be considered limited by two values—that is, $\theta \in \Theta$ if $\theta_{\text{MIN}} \leq \theta \leq \theta_{\text{MAX}}$. In addition to the function $A_{ap}(\theta)$ and the angles θ_{MIN} and θ_{MAX}, other parameters to be chosen in the design are the diameter D of the active area of the photodiode, the position of the vertex V of the refractive surface on the z-axis, the normal $\mathbf{N_V}$ to the refractive surface at that point (if $\mathbf{N_V}$ is not vertical, then the surface will have a peak at that vertex), the refractive index n of the lens, and the angular sensitivity of the photodiode active surface.

The design procedure consists in the calculation of isolated points of the refractive surface by means of an iterative process based on ray-tracing techniques and involves the following steps.

1. Choose a point P_1 of the refractive surface in the first quadrant of the plane x-z of Figure 7.8 (thus verifying that its y coordinate is null). Calculate the normal $\mathbf{N}_1$ of the refractive surface at P_1 with the condition that the ray impinging at P_1 with direction $\mathbf{U}(\theta_{\text{MIN}})$ is refracted toward R. This calculation is carried out by inverse application of Snell's law. From the normal calculate the tangent $\mathbf{T}_1$ to the refractive surface at P_1 contained in the plane x-z and with the direction that verifies that the y component of $P_1 \times \mathbf{T}_1 > 0$. Calculate the central portion of the refractive surface by interpolating between P_1 and V in the plane x-z a gentle curve (e.g., a polynomial of third order), in such a way that the normal to the curve at both points coincides with $\mathbf{N}_1$ and $\mathbf{N_V}$, respectively. Carry out the 3D ray-tracing on the portion of lens between P_1 and V to obtain the function $A_{ap1}(\theta)$, which is the angular response of that portion. Normalize the function $A_{ap}(\theta)$ by multiplying it by the value $A_{ap1}(\theta_{\text{MIN}})/A_{ap}(\theta_{\text{MIN}})$ (note that the specification $A_{ap}(\theta)$ was relative). In what follows, $A_{ap}(\theta)$ should be understood as the normalized function. Set a small $\Delta\theta$ value—for example, $(\theta_{\text{MAX}} - \theta_{\text{MIN}})/1000$—identify $\theta_1 \equiv \theta_{\text{MIN}}$ and calculate $\theta_2 = \theta_1 + \Delta\theta$.

2. Choose an initial value for a distance $s > 0$ and calculate the point $P_2 = P_1 + s\mathbf{T}_1$ as the next point of the refractive surface. Find the normal $\mathbf{N}_2$ at P_2 with the condition that the ray impinging at P_2 with direction $\mathbf{U}(\theta_2)$ is refracted toward R. From the normal calculate the tangent $\mathbf{T}_2$ to the refractive surface at P_2 contained in the plane x-z and with the direction that verifies that the

y component of $P_2 \times \mathbf{T}_2 > 0$. Estimate an intermediate point $P_{12} = (P_1 + P_2)/2$ and its normal $\mathbf{N}_{12} = (\mathbf{N}_1 + \mathbf{N}_2)/2$.

3. Carry out the 3D ray-tracing on P_{12}, associating it with the area of refractive surface between P_1 and P_2 with the direction of impingement $\mathbf{U}(\theta_2)$ to obtain the value ΔA_{ap12} and thus calculate $A_{ap2}(\theta_2) = A_{ap1}(\theta_2) + \Delta A_{ap12}$, estimated angular response for the angle θ_2 of the portion of lens up to P_2.

4. Repeat the steps from 2 to 3, iterating on the value of the parameter s (note that ΔA_{ap12} increases with s), until obtaining that $|1 - A_{ap}(\theta_2)/A_{ap2}(\theta_2)| < \varepsilon$, ε being a preset margin of error.

5. Carry out the 3D ray-tracing on the point P_{12} resulting from the iteration 4, associating it with the area of the refractive surface between P_1 and the point P_2 resulting from the iteration 4, in order to obtain the function $\Delta A_{ap12}(\theta)$ and thus calculate $A_{ap2}(\theta) = A_{ap1}(\theta) + \Delta A_{ap12}(\theta)$, estimated angular response of the portion of lens up to P_2.

6. Increase the value $\theta_3 = \theta_2 + \Delta\theta$ and repeat the steps from 2 to 5, increasing the subindices by one unit, until for a given angle θ_n the coordinate z of point P_n is negative.

7. Repeat the steps 1–6, iterating on the abscissa of point P_1 until $|1 - \theta_n/\theta_{\text{MAX}}| < \varepsilon'$, ε' being another preset margin of error.

The design is finished. The refractive surface is defined by the set of points calculated in the process. If required, it is possible to fit these points by a spline or a polynomial curve, which facilitates handling of the data.

The design guarantees that the prescription is adjusted in the whole range $0 < \theta < \theta_M$, but the stepped transition to zero at $\theta = \theta_M$ is not (i.e., the prescription for $\theta > \theta_M$ cannot be adjusted) because there are no degrees of freedom to make the outer portion of the lens perform as a Cartesian oval (as done in Section 7.4.2 for the linear case).

However, it should be emphasized that the design procedure uses rays impinging on the receiver from nearly all possible directions (the whole field of view of the photodiode is covered). This situation is close to optimum in terms of maximizing sensitivity—that is, making the constant A_0 (and k') as large as possible. The optimum is equivalent to get isotropic illumination of all the points of the receiver with rays from the specified range $\theta_{\text{MIN}} < \theta < \theta_{\text{MAX}}$.

In the case that the active surface is not flat, as is common in the case of the surface of LED or IRED emitters, the procedure described is applicable simply by considering the corresponding geometry of the active surface for the ray tracings and directing the refracted ray tangent to that surface (as a generalized concept of the point R) for the calculation of the normal at P_k.

The method can be easily generalized to include preset rotational sequential surfaces (either refractive or reflective), which deflect the ray trajectories. This is the case shown in Figure 7.9 shows the cross section of a lens designed for a circular photodiode active area of silicon without antireflection coating. The lens has $n = 1.49$, and an encapsulating material of $n' = 1.56$ is assumed. The surface separating both media is preset to a sphere. The prescribed $A_{ap}(\theta)$ function is the linear function of Eq. (7.19).

The actual function $A_{ap}(\theta)$ function is finally calculated by ray tracing (including Fresnel losses at the air-lens and lens-silicon interfaces) is shown along with the specifications in Figure 7.10a. The procedure, if applied to other angular sen-
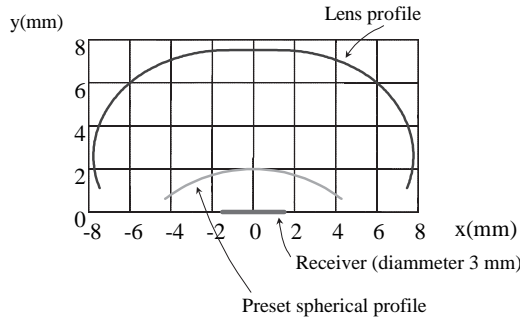
Figure 7.9   Cross-section of a lens designed to get a linear angular sensitivity function in the range $0 \leq \theta \leq 60°$. (lens refractive index $n = 1.49$; encapsulant refractive index $n' = 1.56$)
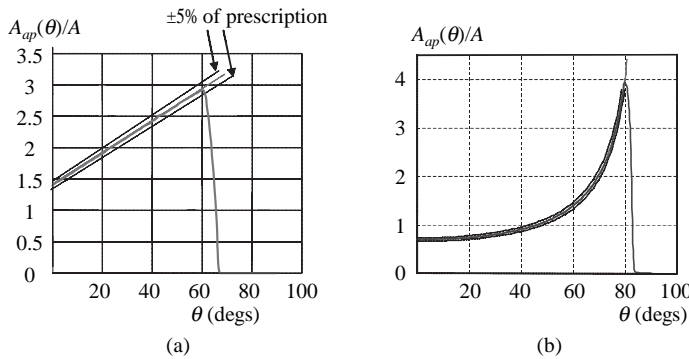
Figure 7.10   (a) Angular sensitivity $A_{ap}(\theta)$ of the lens of Figure 7.9 (bold line) obtained by ray tracing and specified curve and 5% tolerance curves (dashed lines). (b) Angular sensitivity $A_{ap}(\theta)$ of the another lens (profile not shown here) for producing $1/\cos(\theta)$ dependence for $0 < \theta < 80°$.

sitivity functions as $A_{ap}(\theta) = 1/\cos(\theta)$ inside for $0 < \theta < 80°$ (and null outside that range) lead to another lens (whose profile is not shown here) that also produces the specified sensitivity accurately, as shown in Figure 7.10b.

Figure 7.11 shows the results from M. Hernández (2003) of ray trace on several lenses designed for a linear prescribed relative sensitivity with different values of the z-coordinate of the on-axis point of the lens $\mathbf{V}$ (obtained from different selections of the z-coordinate of the point $\mathbf{A}_1$). All show very good agreement with the linear prescription (perhaps except near $\theta = 0$, especially for small V_z values). The very noticeable difference is that the smaller V_z value, the smoother the transition of the sensitivity at $\theta = \theta_M = 60°$. Note that, since by étendue conservation all curves in Figure 7.11 fulfill the integral condition

$$\int_{\theta=0}^{\pi/2} A_{ap}(\theta)\sin(\theta)d\theta = \frac{n^2 A}{2} \tag{7.20}$$

the useless sensitivity for $\theta > 60°$ makes that the constant A_0 in Eq. (7.19) increases with V_z.
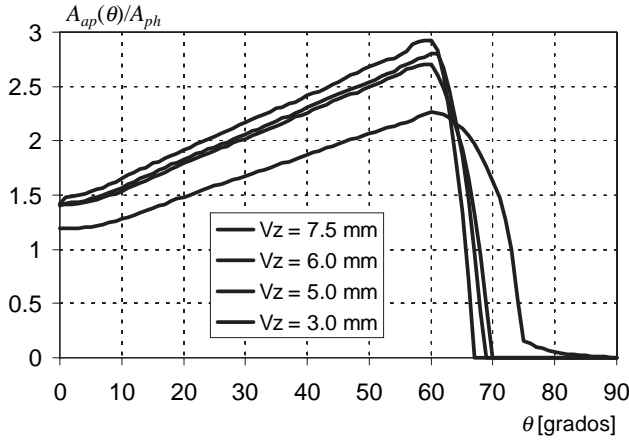
Figure 7.11   Effect of the lens size in the optical performance for the linear prescription of rotational lenses (receiver diameter $D = 3\,$mm).
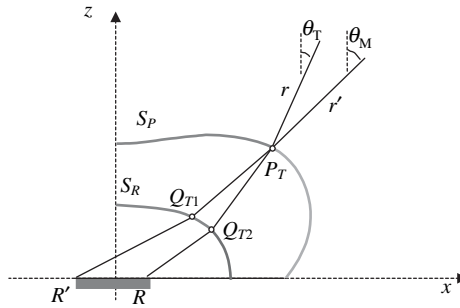
Figure 7.12   The null prescription for $\theta > \theta_M$ can be achieved if two surfaces are used (see SMS method, Chapter 8). This is the condition for maximum absolute sensitivity.

In order to get the maximum possible value of the constant A_0, the stepped transition to zero at $\theta = \theta_M$ is needed (i.e., the null prescription for $\theta > \theta_M$ must also be adjusted). This cannot be done with a single sequential optical surface, as already discussed, but it is possible if two surfaces are used. This is done with the SMS design method presented in Chapter 8. Two complete surfaces are not needed to solve this design problem. For both refractive surfaces, one possible design is indicated in the next steps, which will refer to the points indicated in Figure 7.12:

1. Preset surface S_Q from the center. The last point Q_{T2} of the present portion of surface S_Q will be calculated in step 3.
2. Apply the procedure just described to achieve the prescribed intensity for the calculation of refractive surface S_P through the present surface S_Q up to point P_T, which is the point such that the ray r' traced (inversely) from R' passing through P_T (after the refraction on S_R at point Q_{T1}) exits the lens toward direction $\theta = \theta_M$.
3. Calculate the point Q_{T2} as the point of S_Q on which the ray r from R is refracted toward P_T. Note that, up to this point, the intensity prescription has been designed for $0 < \theta < \theta_T$, which is the exit direction of ray r.
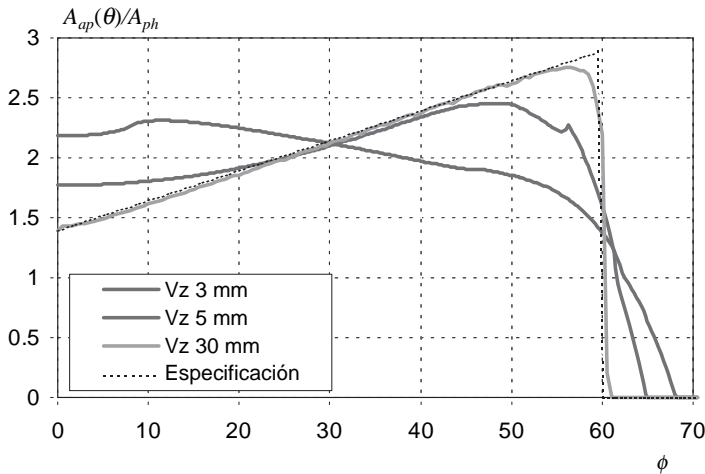
Figure 7.13   Effect of the lens size on the optical performance for the linear prescription of the design obtained with the point source approximation (receiver diameter $D = 3$ mm).

4. Calculate a new portion of surface S_P as the Cartesian oval that makes that the rays r' traced (inversely) from R' and refracted at S_Q at the portion between Q_{T1} and Q_{T2} are refracted on the new points of S_P toward direction $\theta = \theta_M$.
5. Apply the procedure just described to achieve the prescribed intensity for the calculation of refractive surface S_Q through the already known portion of surface S_P calculated in step 4.
6. Repeat steps 4 and 5 up to convergence onto the line R–R'.

7.5.1   Comparison with Point Source Designs

The point source approximation of Section 7.3 can also be applied to the design problem of a refractive sequential rotational surface for prescribed sensitivity. A comparison shows how important the finite dimension of the source is in a specific example (Hernández, 2003).

   The comparison of the performance for different lens sizes designed with the point source approximation but ray-traced with the receiver of diameter $D = 3$ mm is shown in Figure 7.13. The linear prescription is well achieved only for large sizes ($V_Z = 10D$). For the size of this lens with practical interest, which is about $V_z = 3$ mm, the point size model leads to a lens profile that performs far from the specification, in contrast with the result for $V_z = 3$ mm already presented in Figure 7.11.

7.6   THE FINITE TUBULAR SOURCE WITH CYLINDRICAL OPTICS

Another particularly useful case is producing a constant irradiance on a distant plane from a cylindrical source of uniform brightness, such as a Lambertian source. As already mentioned, this was worked out by Ries and Winston (1994). In fact, Ong, Gordon, and Rabl (1996) showed that there are four basic types of

solutions for this type of problem. Two classes derive from the fact that the reflector curve can be diverging or converging—that is, the caustics formed can fall behind or in front of the reflector. These types have been referred to as compound hyperbolic concentrator (CHC) or compound elliptical concentrator (CEC). The possibilities are then doubled because the design can be done with the near edge or far edge of the source being always illuminated. The interested reader can find further information in the cited reference.

7.7   FREEFORM OPTICAL DESIGNS FOR POINT SOURCES IN 3D

Freeform (without any prescribed symmetry) designs in 3D are not a simple extension from the 2D case. These designs become much more difficult, and consequently, they are less developed than their 2D equivalents. In this section we examine overview 3D freeform design methods for point sources—that is, methods that use the point source approximation. This means that the designs will perform as the theory foresees if the optical surfaces are far enough from the source (in terms of source diameter) so it can be considered as a point. At present only one method, which is currently being developed, is able to manage extended sources in 3D geometry. This method is the extension to 3D of the SMS method of Chapter 8 (Benítez et al., 2003).

A basic problem in illumination design is that of designing a single surface (reflective or refractive) that transforms a spherical wave front (point source) with a given intensity pattern into an output wave front with a prescribed intensity pattern. Variations of this basic problem are to have a prescribed irradiance pattern at a given surface instead of the output intensity pattern or to have a plane wave front at the input instead of the spherical one.

The basic equation governing the solution of this problem is a second order nonlinear partial differential equation of Monge-Ampere type. This was found in 1941 by Komissarov and Boldyrev (1994). Schruben (1972) created the equation governing the design of a luminary reflector that provides a prescribed irradiance pattern on a given plane when the reflector is illuminated by a nonisotropic punctual source.

During the 1980s and 1990s a strong development of the method was encouraged by reflector antennas designers. Wescott, Galindo, Graham, Zaporozhets, Mitra, Jervase, and (see References) others contributed to this field of antenna reflector design. The method starts with a procedure purely based in Geometrical Optics. This is the part in which we are more interested for illumination applications. After the Geometrical Optics design, a Physical Optics analysis and synthesis procedure is necessary for a fine-tuning of the design. At present there is commercial software for designing these antenna reflectors based on this method (see, for instance, http://www.ticra.dk/).

The method is particularly useful for satellite applications. Satellite reflector antennas must provide a given far-field (or intensity) pattern to fit, for instance, a continent contour, in satellite-to-earth broadcasting applications. And this should be done efficiently. In this case a single-shaped reflector is enough to solve the problem. The requirement is equivalent to saying that the amplitude of the field at the aperture is prescribed. In other cases it is required to achieve a prescribed

irradiance pattern at the antenna aperture (in general, this is required to reduce the side-lobes emissions) besides the specified far-field pattern. In these cases, two shaped reflectors are enough to solve the problem, and not only the output amplitude is controlled, but also the phase distribution at this aperture. This second problem is very similar to the first, although it may look different.

The single reflect or designs for satellite applications do not differ strongly from a parabola shape because the desired intensity pattern is highly collimated in general. This fact has allowed developing several approximate methods to solve the Monge Ampere that worked well within these conditions.

Beginning in the 1980s until the present, the subject has been of interest to mathematicians like Oliker, Caffarelli, Kochengin, Guan, Glimm, and Newman (see References). Conditions of existence and uniqueness of the solutions have been found as well as new design procedures have been proposed. For instance, Glimm and Oliker (2003) have shown recently that the problem can also be solved as a variational problem in the framework of a Monge-Kantorovich mass transfer problem, which allows solving the problem numerically by techniques from linear programming. The designs are not limited to reflectors but extend also to refractive surfaces. Already in the present decade, the subject has come back to the illumination field by Ries and Muschaweck (2002). In this reference, multigrid numerical techniques are efficiently used to solve the Monge-Ampere equation. The solutions are classified into four types depending on the location of the centers of curvature of the output wavefronts to design: In two of these types, the surfaces of curvature centers (each one corresponding to one of the two families of curvature lines) are at one "side" of the optical surface, whereas in the remaining types the surfaces of curvature centers are at both sides of the optical surface.

7.7.1   Formulation of the Problem

We shall restrict the explanations to the problem of designing a single optical surface (reflective or refractive) that transforms a given intensity pattern of the source into another prescribed intensity pattern (Miñano and Benítez, 2002). Let $\hat{r}$ be a unit vector characterizing an emitting direction of the source. This unit vector can be determined with two parameters, u and v. These two parameters can be, for instance, the two angular coordinates (θ, ϕ) of the spherical coordinates. In this case, $\hat{r}$ is given by (see Figure 7.14)

$$\hat{r} = (\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta) \qquad (7.21)$$
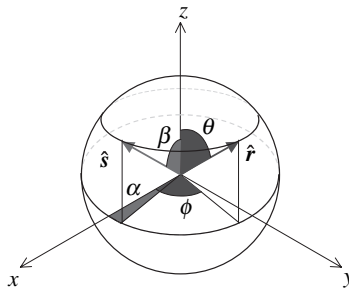
Figure 7.14   Definition of the unit vectors $\hat{r}$ and s.

Let the unit vector $\hat{s}$ define an outgoing direction of the rays after deflection on the optical surface. Using spherical coordinates (α, β), $\hat{s}$ can be written as

$$\hat{s} = (\cos\alpha \sin\beta, \sin\alpha \sin\beta, \cos\beta) \tag{7.22}$$

Because $\hat{r}$ and $\hat{s}$ are unit vectors, then

$$\begin{aligned}
|\hat{r}|^2 &= \hat{r} \cdot \hat{r} = 1 \Rightarrow \hat{f}\hat{f}_u = \hat{r} \cdot \hat{r}_v = 0 \\
|\hat{s}|^2 &= \hat{s} \cdot \hat{s} = 1 \Rightarrow \hat{s} \cdot \hat{s}_u = \hat{s} \cdot \hat{s}_v = 0
\end{aligned} \tag{7.23}$$

Nevertheless $\hat{r}_u$ is not necessarily normal to $\hat{r}_v$ and s_u is not necessarily normal to s_v.

The differential of solid angle $d\Omega_r$ subtended by the rays in a differential $du\,dv$ can be written as

$$d\Omega_r = |\hat{r}_u \times \hat{r}_v| du\,dv = (\hat{r}_u \times \hat{r}_v) \cdot \hat{r}\, du\,dv \tag{7.24}$$

The second equality of Eq. (7.24) assumes that we have chosen the parameters u and v such that the vector $\hat{r}_u \times \hat{r}_v$ points in the same direction as $\hat{r}$. A similar equation applies for the vector $\hat{s}$ and the solid angle $d\Omega_s$.

7.7.2   Basic Equations

7.7.2.1   Laws of Reflection and Refraction

According to Herzberger (1958), if we have two surfaces defined by the vectors $\bar{a}$ and $\bar{a}'$ that are crossed by a one-parameter beam of rays and such that u is the parameter, we have

$$\bar{a}_u \cdot \hat{s}n - \bar{a}_u' \cdot \hat{s}'n' = E_u \tag{7.25}$$

where the unit vectors $\hat{s}'$ and $\hat{s}$ are pointing in the ray directions at each one of the surfaces, E is the optical path length from the surface defined by $\bar{a}'$ to the surface defined by $\bar{a}$, and n', n are, respectively, the refractive indices at each one of the surfaces. We can obtain both the equation of reflection and the equation from Eq. (7.25).

Assume that the one-parameter bundle of rays is passing through the coordinate origin. The surface defined by the vector $\bar{a}'$ is just a point and thus $\bar{a}_u' = 0$. Let $\bar{r}$ be a vector defining a reflective surface. $\bar{r}$ is the vector $\bar{a}$ of Eq. (7.25).

Now consider a two-parameter bundle of rays passing through the coordinate origin. The two parameters are u and v. Then, application of Eq. (7.25) gives (see Figure 7.15)
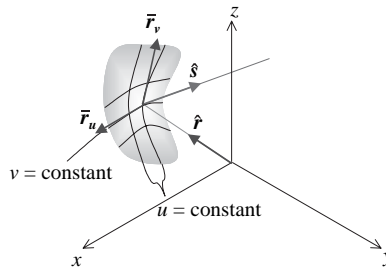
Figure 7.15   Vector $\hat{r}$ is impinging on the reflector where it is reflected as vector $\hat{s}$.

$$\bar{r}_u \cdot \hat{s} = r_u$$
$$\bar{r}_v \cdot \hat{s} = \hat{r}_v$$
(7.26)

where it has been taken into account that $E_u = r_u$; r is the modulus of $\bar{r}$—that is, $r = \bar{r}$. Eq. (7.27) is derived from this definition of the modulus.

$$\bar{r} = r\hat{r} \quad \bar{r}_u = r_u\hat{r} + r\hat{r}_u \quad \bar{r}_v = r_v\hat{r} + r\hat{r}_v$$
(7.27)

Combining Eqs. (7.26) and (7.27) we get the reflection law in the form that we are going to use (note that the vectors $\hat{r}_u$ and $\hat{r}_v$ are not a unit vectors)

$$\frac{r_u}{r} = \frac{\hat{r}_u \cdot \hat{s}}{1 - \hat{r} \cdot \hat{s}} \quad \frac{r_v}{r} = \frac{\hat{r}_v \cdot \hat{s}}{1 - \hat{r} \cdot \hat{s}}$$
(7.28)

We can apply Eq. (7.25) to a refractive surface to obtain the refraction law in a similar way as we got the reflection law. The result is

$$\frac{r_u}{r} = \frac{\hat{r}_u \cdot \hat{s}}{n - \hat{r} \cdot \hat{s}} \quad \frac{r_v}{r} = \frac{\hat{r}_v \cdot \hat{s}}{n - \hat{r} \cdot \hat{s}}$$
(7.29)

that is, for our purposes, both laws can be summarized in Eq. (7.29) taking $n = 1$ in the case of reflection.

7.7.2.2   Power Conservation

Let $E(\hat{s})$ be the desired output radiant intensity (for instance, in Watt/stereoradian), and let $I(\hat{r})$ be the intensity emitted by the source. Energy conservation can be written as

$$E(\hat{s})d\Omega_s = I(\hat{r})d\Omega r$$
(7.30)

Expressing the vectors $\hat{r}$ y $\hat{s}$ as functions of the two parameters (u, v), then we have that

$$E(\hat{s})|\hat{s}_u \times \hat{s}_v| = I(\hat{r})|\hat{r}_u \times \hat{r}_v|$$
(7.31)

Eq. (7.31) is the form that we will use for the energy conservation.

Using Eq. (7.24), Eq. (7.31) can be written as

$$E(\hat{s})(\hat{s}_u \times \hat{s}_v) \cdot \hat{s} = \pm I(\hat{r})(\hat{r}_u \times \hat{r}_v) \cdot \hat{r}$$
(7.32)

Where the sign $\pm$ takes into account that the trihedron $\hat{s} - \hat{s}_u - \hat{s}_v$ may have two possible orientations. We have chosen $\hat{r}, \hat{r}_u, \hat{r}_v$ to be in the positive orientation $(\hat{r} \cdot \hat{r}_u \times \hat{r}_v > 0)$, but we don't have the freedom to choose the orientation of $\hat{s} - \hat{s}_u - \hat{s}_v$.

7.7.2.3   Malus-Dupin Theorem

The dependence of $\hat{s}$ with (u, v) is not totally free. This is due to the Malus-Dupin theorem, which states that a normal congruence remains like this after being deflected by a mirror or a lens surface. For our particular case (a single reflective or refractive surface and a punctual source) the Malus-Dupin theorem is nothing else than the equality of the crossed derivatives of the function describing the optical surface—that is, $r_{uv} = r_{vu}$ (see Eq. (7.26)).

$$\bar{r}_u \cdot \hat{s}_v - \bar{r}_v \cdot \hat{s}_u = 0$$
(7.33)

which can also be written as

$$r_u \hat{r} \cdot \hat{s}_v + r \hat{r}_u \cdot \hat{s}_v - r_v \hat{r} \cdot \hat{s}_u - r \hat{r}_v \cdot \hat{s}_u = 0 \qquad (7.34)$$

7.7.3 Mathematical Statement of the Problem

Eqs. (7.29), (7.32), and (7.34) form a system of equations with unknown variables $r(u, v)$ and $\hat{s}(u, v)$. Variable r can be eliminated with Eqs. (7.29) and (7.34), resulting

$$(r \times r_u) \cdot (\hat{s} \times \hat{s}_v) - (\hat{r} \times \hat{r}_v) \cdot (\hat{s} \times \hat{s}_u) = n(\hat{r}_u \cdot \hat{s}_v - \hat{r}_v \cdot \hat{s}_u) \qquad (7.35)$$

The system is now formed by Eqs. (7.35) and (7.32), where the unknown function is $\hat{s}(u, v)$—that is, we have to find a mapping of the unit sphere into itself satisfying Eqs. (7.35) and (7.32). We can take (θ, ϕ) in Eq. (7.21) as the parameters u, v and take $\alpha(u, v)$, $\beta(u, v)$ (Eq. (7.22)) as the unknown functions of this equation system.

Eliminating $\hat{s}$ and its derivatives from the equation system Eqs. (7.29), (7.32), and (7.41) leads to a single, second order partial differential equation of the Monge Ampere type, which can be found, for instance, in Schruben (1972). In this case the unknown is the function $r(u, v)$.

7.7.4 Dual Optical Surfaces

The previous development allows us to introduce easily the concept of dual optical surfaces (Miñano and Benítez, 2002). As seen in the previous section, the mathematical problem can be summarized in

$$(\hat{r} \times \hat{r}_u) \cdot (\hat{s} \times \hat{s}_v) - (\hat{r} \times \hat{r}_v) \cdot (\hat{s} \times \hat{s}_u) = n(\hat{r}_u \cdot \hat{s}_v - \hat{r}_v \cdot \hat{s}_u)$$
$$E(\hat{s})(\hat{s}_u \times \hat{s}_v) \cdot \hat{s} = \pm I(\hat{r})(\hat{r}_u \times \hat{r}_v) \cdot \hat{r} \qquad (7.36)$$

assume that this equation system is solved—that we know the function $\hat{s}(u, v)$ satisfying Eq. (7.36) with the contour conditions. The calculation of the optical surface can be done with Eq. (7.28)—that is, by integration of $r(u, v)$.

Note that the system of Eq. (7.36) is the same that we would have if

(a) $\hat{r}$ is the output unit vector.
(b) $\hat{s}$ is a unit vector departing from the source.
(c) $I(\hat{r})$ is the required intensity distribution and $E(\hat{s})$ is the source intensity distribution.

In this case, the optical surface would be given by the function $s(u, v)$ by means of

$$\frac{s_u}{s} = \frac{\hat{s}_u \cdot \hat{r}}{n - \hat{r} \cdot \hat{s}} \quad \frac{s_v}{s} = \frac{\hat{s}_v \cdot \hat{r}}{n - \hat{r} \cdot \hat{s}} \qquad (7.37)$$

Assume that we have two functions $\hat{s}(u, v)$ y $\hat{r}(u, v)$ satisfying the system of Eq. (7.36), then we have two functions $r(u, v)$ ys(u, v) fulfilling Eq. (7.28) y Eq. (7.37) and generating two optical systems that we call duals. Functions $r(u, v)$ and $s(u, v)$ fulfill

$$\nabla_{u,v} \ln(n - \hat{r} \cdot \hat{s}) = -\left(\frac{\hat{r}_u \cdot \hat{s} + \hat{s}_u \cdot \hat{r}}{n - \hat{r} \cdot \hat{s}} : \frac{\hat{r}_v \cdot \hat{s} + \hat{s}_v \cdot \hat{r}}{n - \hat{r} \cdot \hat{s}} \right) = -\nabla_{u,v} \ln(r\,s) \qquad (7.38)$$

and so

$$(n - \hat{r} \cdot \hat{s}) \, r \, s = \text{const} \qquad (7.39)$$

If one of the systems is known, the other can be easily calculated with Eq. (7.39).

One of the systems produces a pattern $E(\hat{s})$ when the point source radiates as given by $I(\hat{r})$, and the other (dual) system produces the pattern $I(\hat{r})$ when the source is radiating as $E(\hat{s})$.

REFERENCES

Aoki, K., Miyahara, N., Makino, S., Urasaki, S., and Katagi, T. (1999). Design method for offset shaped dual-reflector antennas with an elliptical aperture of low cross-polarisation characteristics. *IEE Proc. Microw. Antennas Propag*. **146**, 60–64.

Benítez, P., Miñano, J. C., Blen, J., Mohedano, R., Chaves, J., Dross, O., Hernández, M., and Falicoff, W. (2004). "Simultaneous multiple surface optical design method in three demensions", *Opt. Eng.*, vol. 43, no. 7.

Benítez, P., Miñano, J. C., Hernández, M., Hirohashi, K., Toguchi, S., and Sakai, M. (2000). "Novel nonimaging lens for photodiode receivers with a prescribed angular response and maximum integrated sensitivity", *Optical Wireless Communications III*, Eric J. Korevaar, Editor Vol. 4214 pp. 94–103. Boston MA.

Boldyrev, N. G. (1932). About calculation of Asymmetrical Specular reflectors. *Svetotekhnika* **7**, 7–8.

Brown, K. A., and Prata, A. (1994). A design procedure for classical offset dual reflector antennas with circular apertures. *IEEE Trans. on Antennas and Propagation*, Vol. 42, **8**, 1145–1153.

Caffarelli, L., Kochengin, S., and Oliker, V. I. (1999). On the numerical solution of the problem of reflector design with given far-field scattering data. *Contemporary Mathematics*, Vol. 226, 13–32.

Elmer, W. B. (1980). *The Optical Design of Reflectors*, 2nd ed. Wiley, New York.

Galindo, V. (1964). Design of dual-reflector antennas with arbitrary phase and amplitude distributions. *IEEE Trans. Antennas Propagat*, 403–408.

Galindo Israel, V., Imbriale, W. A., and Mittra, R. (1987). On the theory of the synthesis of single and dual offset shaped reflector antennas. *IEEE Trans. Antennas Propagat.*, Vol. AP-35, 887–896.

Galindo-Israel, V., Imbriale, W. A., Mittra, R., and Shogen, K. (1991). *IEEE Trans. Antennas Propagat.*, Vol. 39, 620–626.

Glimm, T., and Oliker, V. I. (2003). Optical design of single reflector systems and the Monge-Kantorovich mass transfer problem. *Journal of Mathematical Sciences*, Vol. 117, **3**, 4096–4108.

Hernández, M. (2003). PhD. dissertation, UPM, Madrid.

Herzberger, M. (1958). *Modern Geometrical Optics*. Interscience, New York.

Jervase, J. A., Mittra, R., Galindo-Israel, V., and Imbriale, W. (1989). Interpolation solutions for the problem of synthesis of dual-shaped offset reflector antennas. *Microwave Opt. Technol. Lett.*, Vol. 2, 43–47.

Kildal, P. S. (1984). Comments on "Synthesis of offset dual shaped subreflector antennas for control of cassegrain aperture distributions." *IEEE Trans. Antennas Propagat.*, Vol. AP-32, 1142–1145.

Kochengin, S., and Oliker, V. I. (2003). Computational algorithms for constructing reflectors. *Computing and Visualization in Science* **6**, 15–21.

Kochengin, S. A., Oliker, V. I., and von Tempski, O. (1998). On the design of reflectors with prespecified distribution of virtual sources and intensities. *Inverse Problems* **14**, 661–678.

Komissarov, V. D. (1941). The foundations of calculating specular prismatic fittings. *Trudy VEI* **43**, 6–61.

Lee, J. J., Parad, L. I., and Chu, R. S. (1979). A shaped offset-fed dual reflector antenna. *IEEE Trans. Antennas Propagat.*, Vol. 27, **2**, 165–171.

Miñano, J. C., and Benítez, P. (2002). Design of reflectors and dioptrics for prescribed intensity and irradiance pattern. *Light Prescrptions LLC*, internal report.

Newman, E., and Oliker, V. I. (1994). Differential-geometric methods in design of reflector antennas. *Symposia Mathematica*, Vol. 35, 205–223.

Oliker, V. I. (2002). On the geometry of convex reflectors. *Banach Center Publications*, Vol. 57, 155–169.

Oliker, V. I. (2003). Mathematical aspects of design of beam shaping surfaces in geometrical optics. In *Trends in Nonlinear Analysis* (Kirkilionis, M., Kromker, S., Rannacher, R., and Tomi, F., eds.). Springer-Verlag, New York, pp. 192–224.

Ong, P. T., Gordon, J. M., and Rabl, A. (1996). Tailored edge-ray designs for illumination with tubular sources. *Applied Optics* **35**, 4361–4371.

Pengfei Guan and Xu-Jia Wang. (1998). On a Monge-Ampere equation arising in geometrical optics. *J. Differential Geometry* **48**, 205–223.

Ries, H., and Muschaweck, J. (2002). Tailored freeform optical surfaces. *J. Opt. Soc. Am. A.* **19**, 590–595.

Ries, H., and Winston, R. (1994). Tailored edge-ray reflectors for illumination. *J. Opt. Soc. Am. A.*, Vol. 11, **4** 1260–1264.

Rubiños-López, J. O., and García-Pino, A. (1997). A ray-by-ray algorithm for shaping dual-offset reflector antennas. *Microwave Opt. Technol. Lett.*, Vol. 15, 20–26.

Rubiños-López, J. O., Landesa-Porras, L., and García-Pino, A. (1998). Algorithm for shaping dual offset reflector antennas based on ray tracing. *Automatika* **39**, 39–46.

Schruben, J. S. (1972). Formulation of a reflector-design problem for a lighting fixture. *J. Opt. Soc. Am. A.*, Vol. 62, 1498–1501.

Westcott, B. S. (1983). *Shaped Reflector Antenna Design.* Wiley, New York.

Westcott, B. S., Graham, R. K., and Wolton, I. C. (1986). Synthesis of dual-offset, shaped reflectors for arbitrary aperture shapes using continuous domain deformation. *IEE Proc.*, Vol. 133, 57–64.

Westcott, B. S., Stevens, F. A., and Brickell, F. (1981). GO synthesis of offset dual reflectors. *IEE Proc.*, Vol. 128, 11–18.

Westcott, B. S., and Zaporozhets, A. A. (1993). Fast synthesis of aperture distributions for contoured beam reflector antennas. *Electronic Letters*, Vol. 29, **20**, 1735–1737.

Westcott, B. S., and Zaporozhets, A. A. (1994). Single reflector synthesis using an analytical gradient procedure. *Electronics Letters*, Vol. 30, **18**, 1462–1463.

Westcott, B. S., and Zaporozhets, A. A. (1995). Dual-reflector synthesis based on analytical gradient-iteration procedures. *IEE Proc.*, Vol. 142, **2**, 129–135.

Westcott, B. S., Zaporozhets, A. A., and Searle, A. D. (1993). Smooth aperture dis-
      tribution synthesis for shaped beam reflector antennas. *Electronics Letters*,
      Vol. 29, **14**, 1275–1276.

Winston, R., and Ries, H. (1993). Nonimaging reflectors as functionals of the
      desired irradiance. JOSA A, Vol. 10(9), 1902–1908.

Xu-Jia Wang. (1996). On the design of reflector antenna. *Inverse Problems* **12**(2),
      351–375.

8

SIMULTANEOUS MULTIPLE SURFACE DESIGN METHOD

8.1 INTRODUCTION

In this chapter we examine the Simultaneous Multiple Surface (SMS) (U.S. Letters Patent, 6,639,733, "High Efficiency Non-Imaging Optics") design method in 2D geometry. As with the 2D flow-line method, actual concentrators are generated by symmetries from the 2D designs. Typical symmetries are linear and rotational, but none are excluded. This procedure does not ensure the ideality of the actual 3D concentrators. Only the subsets of edge rays that have been used in the 2D design are fulfilling the hypothesis of the edge-ray principle. There is an infinite possible set of edge rays that contain this subset. In some special cases (linear symmetry with constant refractive index distribution, for instance) the invariant imposed by the symmetry allows one to calculate the trajectories of the rays from the trajectories of their projections (on a plane normal to the axis of symmetry, in the linear case). In this way, a full set of edge rays can be derived from the 2D subset of edge rays.

In general, ray tracing is necessary to calculate the bundle of rays transmitted by the 3D concentrator. This ray tracing will be the final step of the 3D design if its result is sufficiently satisfactory.

The reflectors joining entry and exit apertures' edges are an essential part of all concentrators designed with the flow-line design method. Sometimes these reflectors are inconvenient. For instance, in optoelectronic applications, the exit aperture is the semiconductor surface. A reflector close to this surface complicates the routing of the electrical contact. In solar thermal applications the reflector may be a source of thermal losses. To avoid the reflector being close to the receiver, incorporation of cavities in the design of the reflector has been proposed. This solution allows a sizable gap between the receiver and the reflector, with a small reduction in the concentration (Winston, 1980). For nonmaximal concentration, the exit aperture does not coincide with the receiver, and thus the reflectors do not touch it. If the receiver is circular, it is possible to design a set of nonmaximal concentrators that together give maximal concentration with reflectors not touching the receiver but with complex reflector structure (Chaves and Collares-Pereira, 1999). In the SMS method there are no reflectors that join entry and exit apertures. This will require handling the edge rays in a slightly different way, as with the flow-

line method. In the latter, some of the edge rays passing through the borders of the entry or exit apertures are not considered in the design procedure (see Appendix B). In the SMS method every edge ray must be considered.

8.2   DEFINITIONS

Let Σ_i be the entry aperture of the optical system, and let Σ_o be the exit aperture. These apertures may be real or virtual. Assume that both the entry and exit apertures are on a z = constant plane. Assume also that the rays coming from a source and impinging on the entry aperture form the input bundle and that the rays illuminating any point of a receiver from the exit aperture form the output bundle. These assumptions simplify the following reasoning. Figure 8.1 shows a 3D concentrator with its entry and exit apertures, a source, and a receiver. In our 2D problem we will be restricted to the plane x-z (see Figure 8.2).
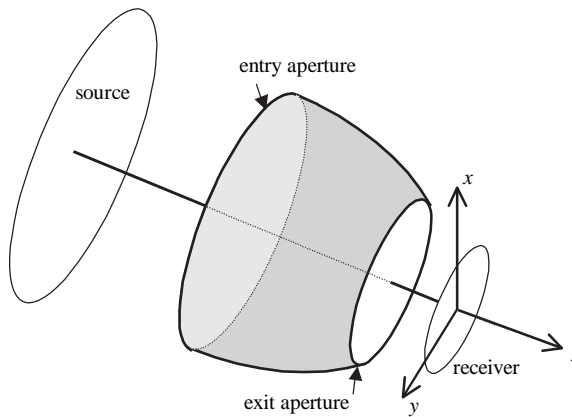
Figure 8.1   The rays of the source impinging on the entry aperture form the input bundle, and the rays illuminating the receiver from the exit aperture form the output bundle.
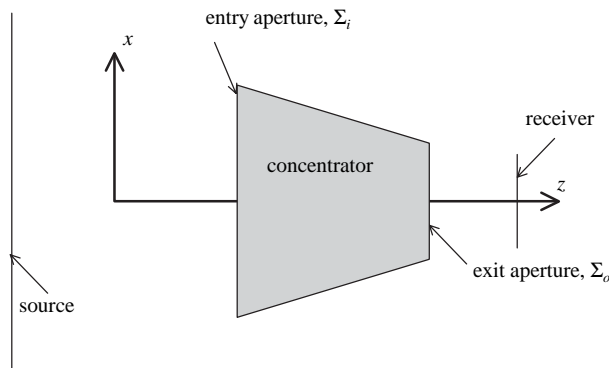
Figure 8.2   2D geometry definitions of source, entry aperture, receiver, and exit aperture.

Let n_i be the index of refraction of the medium between the source and the entry aperture, and let n_o be the index of refraction of the medium between the exit aperture and the receiver. Typically, $n_i = 1$ and $n_o = 1$, or it is in the range ≈ 1.4 to 1.6. Let p be the optical direction cosine of a ray with respect to the x-axis. For instance, p is n_i times $\cos(\alpha)$ (α is the angle formed between the ray and the x-axis) when p is calculated at a point of the ray-trajectory between the source and the entry aperture, and p is $n_o \cos(\alpha)$ at the point of the trajectory where the calculations are done between the exit aperture and the receiver. Let r be the optical direction cosine with respect to the z-axis. Then $p^2 + r^2 = n^2$, n being the index of refraction of the point where p and r are calculated.

In 2D geometry, every ray reaching the entry aperture Σ_i can be characterized by two parameters. These parameters can be, for instance, the coordinate x of the point of interception of the ray with the entry aperture and the coordinate p of the ray direction at this point of interception. Similarly, the rays issuing from the exit aperture Σ_o can be characterized by another pair of parameters—for instance, the coordinate x of the point of interception of the ray with the receiver and the optical direction cosine p of the ray at this point of interception. A region of the phase space x-p represents the set of rays linking the source with the entry aperture. We call this set of rays the input bundle, M_i. Similarly, the output bundle M_o is a region of the phase space x-p whose points represent the rays linking Σ_o with the receiver.

The purpose of this section is to design an optical system such that the rays of M_i leave the system as rays of M_o and the rays of M_o, if reversed, leave the system as rays of M_i. If this is the case, then the rays of M_i and M_o are the same ($M_i = M_o$), the only difference being that M_i is the representation at Σ_i and M_o is the representation at Σ_o. We consider as a particular case when M_o includes all possible rays reaching the receiver. This case is called maximal concentration.

The requirement for the optical system is that $M_i = M_o$, regardless of the particular transformation of each one of the rays, just that M_i and M_o represent the same bundle of rays at two different surfaces (at Σ_i and at Σ_o). In general, an actual optical system does not achieve this condition perfectly. The bundle of rays M_c connecting the source with the receiver through the optical system does not coincide with M_i, or with M_o in the general case. Obviously M_c must be a subset of M_i because the definition of M_c includes the rays connecting source and receiver through the optical system—that is, the rays of M_c should cross the entry aperture (and also the exit aperture). Similarly, M_c is a subset of M_o.

If M_i and M_o have to be the same set of rays (i.e., $M_i = M_o$), then necessarily the étendue E of M_i and M_o must be the same—that is,

$$E(M_i) = \int_{M_i} dx\,dp = \int_{M_o} dx\,dp \equiv E(M_o) \qquad (8.1)$$

As any other design method for nonimaging concentrators, a key part of the procedure is the edge-ray theorem (see Appendix B), which establishes that for $M_i = M_o$ it is enough that $\partial M_i = \partial M_o$, where ∂M_i and ∂M_o are, respectively, the sets of edge rays of M_i and M_o, which are represented by the points of the borders of the regions M_i and M_o in the phase space. In other words, the optical system to be designed must transform the rays of ∂M_i into the rays of ∂M_o and vice versa. Again, there are no requirements about which ray of ∂M_i has to be linked with a given ray of ∂M_o.

In our 2D geometry problem, the regions M_i and M_o are two-parametric (i.e., to distinguish one ray of M_i from another, it is necessary to give two parameters), whereas the regions ∂M_i and ∂M_o are one-parametric: This reduction of the number of parameters clearly simplifies the design problem.

8.3   DESIGN OF A NONIMAGING LENS: THE RR CONCENTRATOR

The simplest example to start with is the design of a nonimaging lens, also called RR concentrator. Figure 8.3 shows an example of these lenses. The source extends from S to S' and the receiver from R to R'. The lens entry aperture is a curve extending from N to N', and the exit aperture is a curve from X to X'. The purpose of the design is find two refractive curves such that every ray from the source hitting the entry aperture is refracted at these curves in such a way that it exits the lens as a ray linking the exit aperture and the receiver. We use the same explanation as in Miñano and González (1991; 1992).

In order to fix the conditions of the design, let us assume that the receiver width is 2 (the receiver edges R and R' are at $x = -1$ and $x = 1$). Figure 8.4 shows the representation of M_i and M_o in the phase space x-p at the entry aperture (left) and at the exit aperture (right). Each point (x, p) represents a ray. The points corresponding to the rays drawn in Figure 8.3 are represented in Figure 8.4.
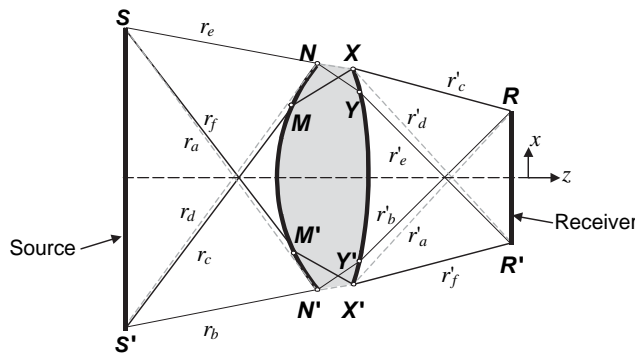
Figure 8.3   Location of the source and receiver and representation of some edge rays. Rays having the same subscript are the same ray.
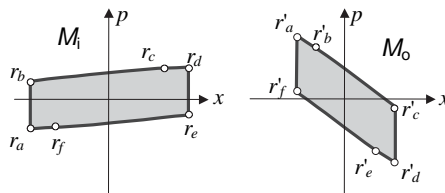
Figure 8.4   Representation in the phase space of M_i and M_o. Some special edge rays are marked with a dot, and their trajectories are shown in Figure 8.3.
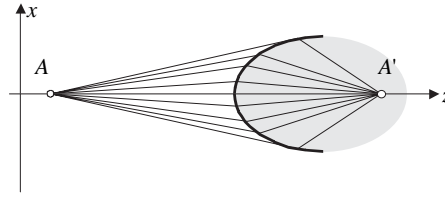
Figure 8.5   Refractive Cartesian oval focusing the rays issuing from A on A'. The refracting surface is rotationally symmetric about the z-axis.

It is well known that a single refractive or reflective surface can sharply image a bundle of rays into a point if no more than one ray passes through each point of the surface. In general, a single surface can transform a given bundle of rays into another predetermined one if there is no more than one ray crossing each point of this surface. We call these surfaces generalized Cartesian ovals (see, for instance, Luneburg, 1964, and Stavroudis, 1972). The problem of determining a generalized Cartesian can be solved simply requiring the constant path length between the incident and the emergent wavefront (see Wolf, 1948, and Wolf and Preddy, 1947, for an example of a generalized Cartesian Oval of refraction). A Cartesian oval problem is that of finding an optical surface (refractive or reflective) that couples two spherical wave fronts (including the case of infinite radius sphere—that is, the plane). We call it a generalized Cartesian oval problem when we don't require the wave fronts to be spherical. Figure 8.5 shows the cross-section of a refractive Cartesian oval.

Of course, Cartesian ovals also apply in 2D geometry: The rays issuing from one point can be sharply focused onto another by a single refractive curve. The problem here is slightly different: There are two curves to be designed (the two refractive curves) and there are two edge rays passing through every point of these two curves (except the extreme points of these surfaces, which are crossed by a bundle of edge rays). A solution to this problem also exists, and the procedure to get it is the basis of the SMS design method. This procedure calculates the refractive (or reflective) surfaces point-by-point in a way similar to that used by Schulz in the design of aspheric lenses (Schulz, 1983, 1988). Each new point of one of the surfaces permits the calculation of another point of the other surface and so on. In this way both surfaces are calculated simultaneously.

Before applying this method we shall impose certain conditions on the transformation of the rays of ∂M_i into the rays of ∂M_o. These conditions derive from the statement of the problem. For instance, note that the rays reaching the extreme point of the lens N (or N') cannot be the same as the rays departing from the extreme point of the lens X (or X') unless the lens has zero thickness at its edges. Only the ray r_a (and its symmetric counterpart r_d) crosses N' and X' (r_d crosses N and X; see Figure 8.3). The trajectory of r_a reaches the point N' of the entry aperture with the most negative value of p (this ray comes from S). This ray must cross the most x-negative point of the exit aperture (point X'), and the value of p of this ray at the point X' must be the highest one when compared to the other rays of ∂M_o crossing X'. Then the x-p representation of the ray r_a at the exit aperture must be r'_a—that is, the ray linking the x-negative edge of the exit aperture and the x-positive edge of the receiver (R) (see Figure 8.4, the notation r and r' referring to the same ray before and after crossing the lens).

Because of the symmetry of the lens, the conditions are only stated for the rays crossing the x-positive side of the lens. These are the additional conditions to start the design.

1. The ray represented by one corner of ∂M_i (r_d) is transformed into a corner of ∂M_o (r'_d).
2. The ray represented by the other corner of ∂M_i (r_e) is transformed in a ray (r'_e) that crosses the lens exit aperture at a point Y different from X.
3. The other corner of ∂M_o (r'_c) represents a ray that come from a ray r_c that crosses the entry aperture at a point M different from N.
4. By symmetry, similar conditions hold for the rays r_a, r_b, and r_f.

The preceding conditions determine the portions MN (and $M'N'$) and XY (and $X'Y'$) of the two surfaces of the lens: The profile MN must be a portion of a Cartesian oval imaging the rays coming from S' (between r_c and r_d) at the point X; and the profile YX must also be a portion of a Cartesian oval that images N at R'.

The receiver points R and R' are assumed to have coordinates $x = 1$ and $x = -1$, respectively. The size and position of the source (relative to the receiver) are considered known data. The design procedure can be described as follows.

1. The étendue E of the input manifold M_i is chosen. Remember that this is a 2D design procedure, so E is the étendue of a two-dimensional bundle of rays. After this selection of M_i, the points N and X are chosen so that they set the same étendue of the manifolds M_i and M_o. This means that the point N must lie on the hyperbola $|NS'| - |NS| = E/2$, and the point X must lie on the hyperbola defined by $|XR'| - |XR| = E/2$, where $|XR|$ means the optical path length from X to R. The selection of N and X fully determines the Cartesian ovals MN and XY. The first Cartesian oval is the one crossing N and images S' at X. The second one is a Cartesian oval crossing X and imaging N at R'. Point M is the intersection of the ray r_c coming from R and crossing X (note that the normal to the refractive surface at X is known, since the Cartesian oval crossing X is known, and so it is possible to trace the trajectory of the ray r_c inside the lens and then calculate the point M). In a similar way the point Y can be calculated with the ray r_e coming from S and refracting at N.
2. Now consider an arbitrary point O of the lens surface between M and N (see Figure 8.6). The ray r_g impinging on O from S must be directed to R'. The tra-
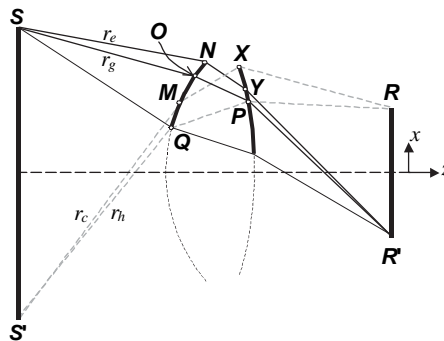
Figure 8.6   The remaining points of the lens are obtained with the point-by-point method, departing from the Cartesian ovals NM and XY.

jectory of r_g inside the lens can be easily calculated (with the refraction law), since the profile MN is known. Then, point P, where the ray r_g leaves the lens, can be calculated by specifying that the optical length l from S to R' be the same through r_e or through r_g—that is, $|SNYR'| = |SOPR'|$ (note that $|SNYR'|$ can be calculated because all the points S, N, Y, and R' are known). A new portion of the rightmost surface of the lens can be obtained by applying this procedure to all the points between N and M. The derivative of the surface at the point P can be calculated using its neighboring points or by the application of the refraction law to ray r_g to give the lens surface normal at point P.

3. Now consider ray r_h, which links P and R. The trajectory of r_h inside the lens can be calculated because the normal to the profile at P is known. This ray must impinge on the leftmost surface of the lens at a point Q in such a way that r_h comes from S'. If the lens is symmetric (with respect to the z-axis), the optical length along r_h from S' to R must be l, and so the point Q and the normal to the profile at Q can be calculated just as we did with the point P. A new portion of the leftmost surface of the lens can be calculated by repeating the preceding procedure with the rays linking R with the calculated portion of the rightmost surface of the lens (i.e., the portion XY and the portion calculated in step 2 of the procedure).

4. The remaining points of the lens are calculated by repeating steps 2 and 3 until the surfaces reach its center at the z-axis.

5. The x-negative side of the lens is obtained by symmetry.

Generally, the lens obtained with this procedure is untypical in that its surface is not normal to the z-axis at $x = 0$ (this is not a necessary condition for the design of the lens), so that there may be a discontinuity of the derivative of the profile there. To get lens profiles normal to the z-axis at $x = 0$, it is necessary to systematically repeat the design procedure with different initial points N and X. First, point N can be kept in its initial position, and point X is moved along the hyperbola $|XR'| - |XR| = E/2$ until the leftmost surface of the lens is normal to the z-axis at $x = 0$ (more than a single solution may exist). Second, the point X is kept constant and the point N is moved along the hyperbola $|NS'| - |NS| = E/2$ until the rightmost surface of the lens is normal to the z-axis at $x = 0$. By iteration of this procedure it is possible to find a lens having both surfaces normal to the z-axis at $x = 0$. Finally, when the x positive side of the lens has been designed with their surfaces normal to the z-axis at $x = 0$, the x negative side of the lens is obtained by symmetry, as we said before.

Generally, there is not a single solution. To choose the best, one must consider other features of the lenses, as their performance as 3D lenses, or the thickness at the center of the lens, and so on. The thinnest lens is found when N coincides with X, but mounting considerations may preclude its use.

This design procedure ensures that the bundles ∂M_i and ∂M_o (see Figure 8.4) contain the same rays, except for a small subset of rays crossing the central region of the lens. In order to study this small subset of rays, assume that the x positive side of the lens is designed so that both surfaces of the lens are normal to the z-axis at $x = 0$. The design procedure ensures that all the rays of ∂M_i impinging on the entry aperture from the points N to L (see Figure 8.7) are rays of ∂M_o after crossing the lens.

The same can be said of the rays that cross the lens through the portion XZ of the rightmost surface of the lens. The design method also ensures that the rays
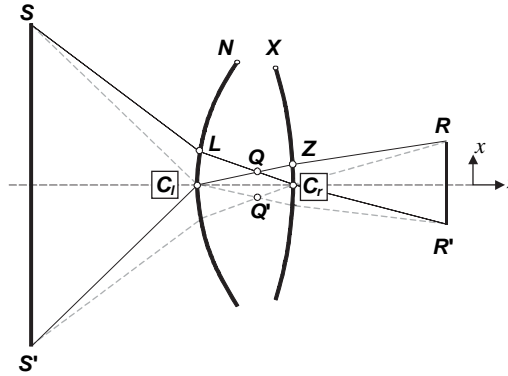
Figure 8.7   In the construction of the lens at the center, the method requires an additional degree of freedom to obtain rigorously an ideal nonimaging 2D concentrator.
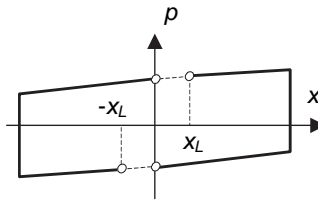
Figure 8.8   The method ensures that the part of ∂M_i represented with a solid line in this figure is transformed in rays of ∂M_o.

of ∂M_i coming from S' and impinging on the portion C_1L of leftmost surface of the lens become rays of ∂M_o (in the case shown in Figure 8.7 these rays are focused at R), and that the rays crossing C_rZ (rightmost surface) and focused to R' are rays of ∂M_i coming from S. Let us construct the x-negative side of the lens by symmetry with respect to the z-axis. The design method does not ensure that the rays from S impinging on the lens through C_1L (these rays belong to ∂M_i) will be imaged at R', except for of two of these rays: those crossing the points C_1 and L (see Figure 8.8). The same condition can be obtained at the exit aperture of the lens. There is no evidence that the rays focused to R from C_rZ are rays coming from S', except for the rays crossing the points C_r and Z. The portions C_1L and C_rZ are fully determined by the design procedure, so there are no more degrees of freedom for solving this problem, unless we accept, for example, the rare possibility of an additional small lens (with different refractive index) between Q and Q'.

In practice this problem does not occur when the number of times that steps 2 and 3 are done is high, which corresponds to a selection of initial points N and X close to each other. When this is so, the rays impinging on the lens through C_1L and coming from S come to be focused on R' (similarly for the rays focused on R and crossing C_rZ), so $\partial M_i = \partial M_o$ in 2D geometry. Nevertheless, we cannot establish rigorously that the method ensures that $\partial M_i = \partial M_o$.

In any case, from the energy-transfer point of view, it would not be critical even if $\partial M_i \neq \partial M_o$ at the central regions of the profiles, if these central regions do not generate a significant portion of area of the 3D concentrators generated by rotational symmetry around the z-axis.

The x-negative side of the lens can also be generated using the same procedure with which the x-positive side has been calculated—that is, following the procedure without stopping at $x = 0$. In this case the lens is, in general, asymmetric (except when the lens surfaces are normal to the z-axis at $x = 0$), so a 3D rotational symmetric lens (with rotational symmetry around the z-axis) cannot be constructed. Another possibility is to construct the x-negative side of the lens by symmetry even if the surfaces are not normal to the z-axis at $x = 0$ and so accepting that the lenses have a kink at the center.

As said before, these nonimaging lenses are also called RR concentrators (RRc), where RR means that the rays are twice refracted between the source and the receiver. We shall keep this notation to avoid confusion with other concentrators.

Hereafter the analysis is restricted to nonimaging lenses designed for a source placed at infinity. Therefore, the source is not characterized by the position of the points S and S' but by the angular spread of the source, $\pm\theta_a$.

8.4   THREE-DIMENSIONAL RAY TRACING OF ROTATIONAL SYMMETRIC RR CONCENTRATORS

In this section, the results of the ray tracing of several rotational symmetric RR concentrators are presented. The axis of symmetry is z.

As a part of the 3D ray tracing of the RRs, a 2D ray tracing can be done to verify that $M_i = M_o$ in 2D geometry—that is, to verify that (1) any meridion ray impinging on the entry aperture of the lens NN' with $-\sin(\theta_a) \leq p \leq \sin(\theta_a)$ reaches the receiver RR', and (2) any meridian ray linking the exit aperture of the lens XX' and the receiver RR' has crossed the entry aperture NN' with $-\sin(\theta_a) \leq p \leq \sin(\theta_a)$.

Figure 8.9 shows the results of the 3D ray-tracing analysis (neither absorption nor Fresnel losses have been considered). The curves in this figure are the
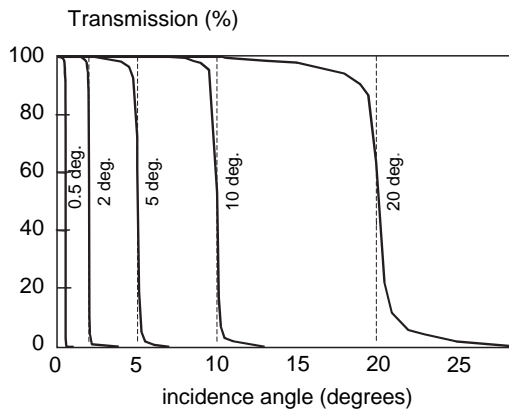
Figure 8.9   Transmission-angle curves for several three-dimensional RRc's designed for sources at the infinity. The number by each curve is the acceptance angle. (For other characteristics, see Table 8.1.)

Table 8.1.   Geometrical Characteristics, and 3D Ray-Tracing Results of Some Selected RR Concentrators. Lens refractive index 1.483; Receiver radius 1.

θ_a (Degrees)	0.5	2	5	10	20
Geometrical concentration, C_{g3D}	3,600	225	36	12.25	2.25
Total transmission, $T(\theta_a)$ (%)	97.5	98.0	96.9	96.6	95.9
Cut-off angular spread $\Delta\theta$ (degrees)	0.04	0.065	0.5	0.6	2
Thickness at the center	101.5	25.65	9.5	3.99	1.33
Length/entry aperture diameter, f	1.161	1.16	1.161	1.036	1.073
Exit aperture radius, R_o	20	5	2.5	2.5	1.2
Exit aperture to receiver distance $z_R - z_X$	32.53	8.09	3.98	3.16	1.81
Entry aperture to receiver distance $z_R - z_N$	72.35	17.95	7.84	4.28	2.27

transmission-angle curves $T(\theta, \theta_a)$ (other characteristics of these lenses are given in Table 8.1). The function $T(\theta, \theta_a)$ gives the percent of rays transmitted to the receiver out of those impinging upon the entry aperture at the incidence angle θ (with respect to the z-axis) for several RR concentrators that are characterized by the angular extension of the source used in the design (θ_a). A trace of 9000 rays was used for each incidence angle θ. The function $T(\theta, \theta_a)$ was more closely explored near the transition (around $\theta = \theta_a$) to ensure that $T(\theta_i, \theta_a) - T(\theta_{i+1}, \theta_a) < 0.1$ (θ_i and θ_{i+1} being two consecutive values of θ for which the function $T(\theta_i, \theta_a)$ is calculated).

If only meridian rays were considered, the transmission-angle curves would be $T(\theta, \theta_a) = 1$ if $\theta \le \theta_a$ and $T(\theta, \theta_a) = 0$ if $\theta > \theta_a$ because the method of design ensures this stepped transition. Nevertheless, since these concentrators are not ideal in 3D geometry (some skew rays with $\theta \le \theta_a$ are not sent to the receiver, and some other skew rays with $\theta > \theta_a$ are sent to it), the transition from $T(\theta, \theta_a) = 1$ to $T(\theta, \theta_a) = 0$ is not stepped. In other words, the set of rays impinging on the concentrator entry aperture with $\theta \le \theta_a$ (the set M_i) does not coincide with the set of rays linking the concentrator exit aperture and the receiver (M_o) in 3D geometry. Nevertheless, the étendue of these two sets E_{3D} is the same because of the method used for the construction of the concentrator. This étendue can be calculated either at the entry or at the exit apertures (Welford, 1989),

$$E_{3D} = \pi A_e \sin^2 \theta_a = \frac{\pi^2}{4}(|X'R| - |XR|)^2 \tag{8.2}$$

where A_e is the area of the concentrator's entry aperture ($A_e = \pi x_N^2$, x_N being the coordinate x of the point N).

Because the concentrator is not ideal in 3D, $M_i \ne M_o$. Let us call M_c to the set of collected rays—that is, the set of rays impinging the entry aperture within the cone defined by θ_a that finally reach the receiver. Obviously, M_c is a subset of M_i and M_o.

An important figure characterizing the transmission of the concentrators is the total transmission $T(\theta_a)$, which can be defined as the ratio of the étendue of M_c to E_{3D} (therefore $T(\theta_a) \le 1$). $T(\theta_a)$ is the total flux transmitted inside the design collecting angle (Welford, 1989). Its expression is

$$T(\theta_a) = \frac{A_e \pi}{E_{3D}} \int_0^{\theta_a} T(\theta, \theta_a) \sin 2\theta \, d\theta \tag{8.3}$$

$T(\theta_a)$ is a quality factor of the concentrator (an ideal 3D concentrator would have a rectangular cutoff at $\theta = \theta_a$, and so $T(\theta_a) = 1$). Another figure characterizing the transmission-angle curves is the angle difference $\Delta\theta = \theta_9 - \theta_1$, where θ_9 and θ_1 fulfill $T(\theta_9, \theta_a) = 0.9$ and $T(\theta_1, \theta_a) = 0.1$.

Table 8.1 shows the total transmission and $\Delta\theta$ obtained from the 3D ray tracing of the RRc whose transmission-angle curve is shown in Figure 8.9. The table gives also some other features of the RRc: the geometrical concentration C_{g3D} (ratio of the entry and exit aperture areas), the ratio f of the length of the concentrator (difference between the z-coordinate of the receiver and the center of the leftmost surface of the lens) to the lens diameter, the thickness at the center of the lens (thickness), the exit aperture radius R_o, and the z-coordinate of the points X and N relative to the receiver plane RR' ($z_R - z_X$ and $z_R - z_N$, respectively). The receiver radius is always 1, and the refractive index of the lens is 1.483.

Figure 8.10 shows the cross-section of an RR with $\theta_a = 10$ degrees (see Table 8.1 for other data). When the acceptance angle of design θ_a is small and the lens-to-receiver distance is greater than twice (approximately) the lens diameter, then the RRc is approximately equivalent to a conventional imaging lens.

Only in some particular cases is it possible to get RR close to the maximal concentration case (when M_o includes every ray reaching the receiver). Maximal concentration with an RR can be achieved if a rotational hyperbolic concentrator is attached to it. The attachment must be such that the reflector fits with the RR exit aperture and the RR receiver coincides with the circle generated by the hyperbola foci. The receiver of the new concentrator is the receiver of the FLC—that is, the circle generated by the apex of the hyperbola, which is smaller than the receiver of the RRc. This new concentrator has the same total transmission $T(\theta_a)$
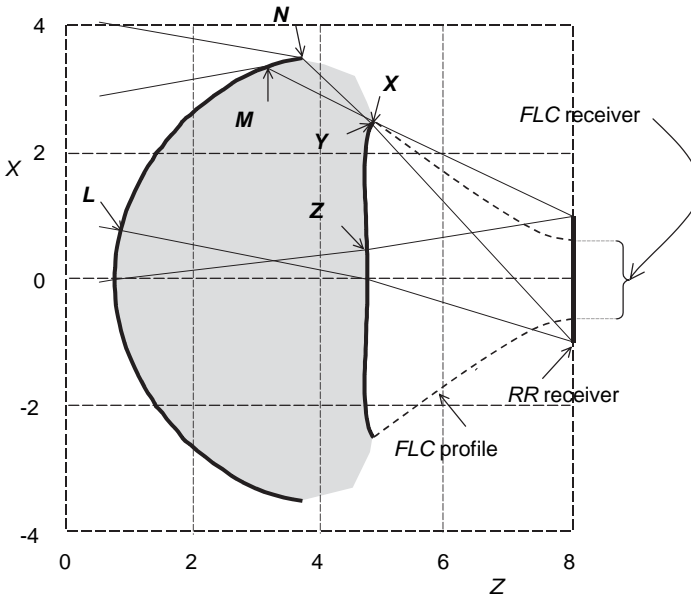
Figure 8.10   RRc-FLC combination with acceptance angle 10° and geometrical concentration 33.2. The portion of Cartesian ovals (*MN* and *XY*) and the points *L* and *Z* are also shown.

and angle transmission curves $T(\theta, \theta_a)$ as the RRc, since the FLC is ideal in 3D geometry, as we mentioned before. The geometrical concentration (ratio of entry aperture area to receiver area) reaches the upper bound ($C_g = 1/\sin^2\theta_a$), and so this new concentrator would be optimal if $T(\theta_a) = 100\%$. Notice in Table 8.1 that the values of $T(\theta_a)$ are quite close to 100%, but none is exactly equal to 100%. The total transmission increases when the acceptance angle decreases and the length from the receiver to the lens apex increases—that is, when the RR approaches a thin lens. Such combinations (thin lens-*FLC*) are detailed analyzed in Welford, O'Gallagher, and Winston (1987).

Imaging solutions using a single lens clearly don't give a satisfactory solution when maximal concentration is required. As we have seen, the *RR* also fails. Nevertheless, other devices designed with the same procedure (RX, XR, RXI, XX) attain the goal, or very close to it. The next sections are devoted to these devices.

8.5   THE XR CONCENTRATOR

There are other possibilities for designing 2D optimal concentrators with the method of the preceding section. Generally speaking, the method requires a minimum of two optically active surfaces to be designed. The two surfaces provide sufficient degrees of freedom to generate the design. These two surfaces need not be both refractive, as in the case of the *RR*; one can be refractive and another reflective, or both can be reflective. Additional restrictions can make some of these possibilities impractical. In this section we shall study the XR concentrator, one formed by a reflective (*X*) and a refractive (*R*) surface such that the rays coming from the source intercept first the reflective surface and then the refractive surface. The medium between the concentrator and the source is assumed to have a refractive index of 1, so the receiver is immersed in a medium of refractive index $n > 1$ (note that the rays intercept the refractive surface only once), and so the maximum achievable geometrical concentration C_g is increased by a factor n^2 with respect to the cases in which the receiver is immersed in a medium of refractive index 1.

The design procedure is qualitatively identical to that of the RR. The only difference is that now there is one reflective and one refractive surface instead of the two refractive surfaces of the RR. Figure 8.11 shows one of these concentrators designed for maximal concentration and for a source placed at infinity.

We shall describe the general procedure of design when the source is placed at a finite distance from the concentrator and when the set of rays M_o is not the one containing all possible rays reaching the receiver (see Figure 8.12). The location of the points N and X is done as before: taking into account that the conservation of étendue theorem has to be fulfilled. This means, in this case, that the point X must lie on the hyperbola $|XR'| - |XR| = E/2$ and the point N on the hyperbola $|NS'| - |NS| = E/2$ (the points R and R' are assumed to be on the straight line $z = 0$ with $x = \pm 1$). Note that the optical path lengths $|XR'|$ and $|XR|$ are the lengths between the corresponding points multiplied by the refractive index n.

Once the points N and X have been chosen, the design of the profiles can start with the portion XY of the lens, a Cartesian oval that images the points N and R'. Since the positions of the points N, X, and R' are known, such a Cartesian oval can be easily constructed. The point Y is obtained as the intersection of the Cartesian oval and the ray coming from S and reflecting at N.
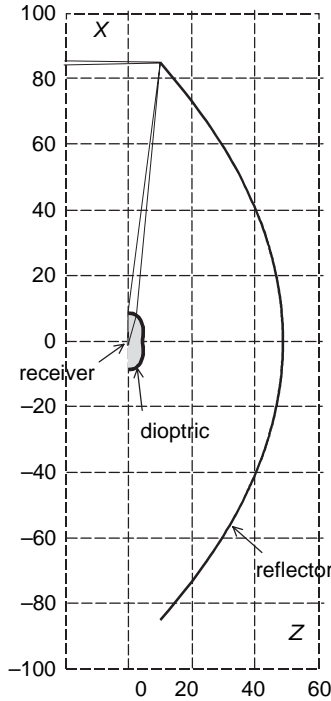
Figure 8.11   XRc for a source at the infinity subtending an angle of 1° with the z-axis.
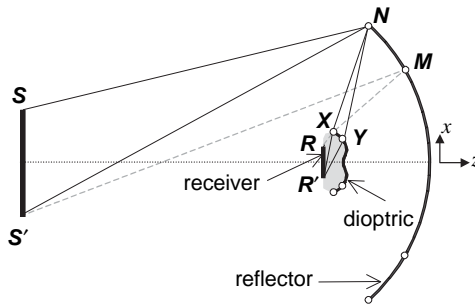
Figure 8.12   The construction of the XRc begins at the extreme points of the lens X and of the mirror N.

The portion NM of the mirror is obtained in a similar way: NM is a part of an ellipse imaging the points S' and X. The point M is the intersection of this ellipse and the ray departing from R and crossing X. Observe the qualitative similarity between the portions XY and NM of the XR and those portions of the RR (see Figures 8.3 and 8.12).

The design can now start from the portion XY: First, rays crossing XY are traced from R and a portion of mirror is calculated requiring that these rays be the reflection of those reaching the mirror from the point S'. Second, rays coming from S are reflected in the last calculated portion of the mirror, and another portion

of the lens surface is obtained by requiring that these rays be focused at R'. The procedure is repeated until the profiles at $x = 0$ are known. If it is desired to have surfaces normal to the z-axis at $x = 0$, then usually one must iterate the procedure with different starting points X and N until the surfaces are normal to the z-axis at $x = 0$.

The design procedure does not always produce a real device. Sometimes loops appear in the reflective or refractive surfaces caused by the caustics of the edge rays, which can be generated by an inappropriate selection of the points X and N. Whether the selection is proper or not is something that can be concluded only during the design. In some cases, these problems can be avoided by defining a virtual exit aperture that is not touching the refractive surface. Note that it has been implicitly assumed that the exit aperture of the concentrator is the segment NN'—that is, the exit aperture edges are on the refractive surface. This assignment is quite arbitrary and sometimes it can be useful to define an exit aperture that has no points in common with the refractive surface—for instance, an exit aperture that is fully inside the dielectric medium.

As in the case of the RR, there is a region around the center of the optical system where there are not enough degrees of freedom to ensure that all the edge rays are correctly directed. Again, it is found that, up to the accuracy of the 2D ray-tracing, these edge rays are correctly directed so $\partial M_i = \partial M_o$. Nevertheless, we cannot establish rigorously that this last condition is exactly fulfilled by a subset of the edge rays crossing the above-mentioned regions of the mirror and the lens.

8.6   THREE-DIMENSIONAL RAY TRACING OF SOME XR CONCENTRATORS

The 3D concentrators to be analyzed here are again constructed by rotational symmetry around the z-axis. The calculations are restricted to the case in which the source is at infinity, as it was done before. The angle formed by a set of parallel rays with respect to the z-axis is called the incidence angle θ. The calculations are also restricted to the designs made for an output bundle M_o comprising all possible rays hitting the receiver—that is, the point X is aligned with R and R' (see Figure 8.12). The points N and M of the mirror become the same point in this case. The 3D étendue of the bundle of rays crossing the entry aperture and reaching the receiver is $E_{3D} = \pi\, n^2 A_r$, where A_r is the receiver area.

Figure 8.13 shows the transmission angle curves of several XRc's, one of which is the concentrator shown in Figure 8.11 (the one corresponding to $\theta_a = 1°$). Other data on these concentrators appear in Table 8.2. For large values of θ_a (greater than 10–15 degrees), the design method may fail, or the resulting portion of dielectric is so big compared with the entry aperture that its shadow on the mirror makes the concentrator useless.

Table 8.2 includes two values of the total transmission $T(\theta_a)$. The first one takes into account the shadow made by the dielectric on the mirror—in other words, it is assumed that the light impinging on the entry aperture at the back side of the dielectric (or the back side of the receiver) is lost. The second one does not consider the shadow.

The dielectric thickness is measured at the center. $z_N - z_R$ is the distance from the plane of the entry aperture to the plane of the receiver. This distance can be
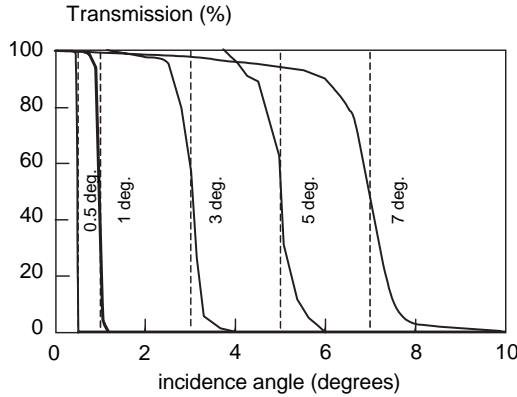
Figure 8.13   Transmission-angle curves for several 3D XRc designed for sources at the infinity with an angle θ_a. The curve numbering is the acceptance angle. (See Table 8.2.)

Table 8.2.   Geometrical Characteristics and 3D Ray-Tracing Results of Some Selected XR Concentrators. Refractive index 1.483; Receiver radius 1.

θ_a (Degrees)	0.5	1	3	5	7
Geometrical concentration, C_{g3D}	28,880	7,221	802.9	289.5	148
Total transmission, $T(\theta_a)$ (%)	94.43	94.62	90.22	90.03	80.74
Total transmission without shadow, $T(\theta_a)$	98.67	95.64	92.38	93.9	88.18
Cutoff angular spread, $\Delta\theta$ (degrees)	0.05	0.15	0.65	1.0	1.5
Concentrator depth/entry aperture diameter	0.257	0.287	0.313	0.356	0.348
Dielectric thickness at the center	13.56	4.56	2.57	2.57	2.33
Dielectric radius	35	8.5	4	3	3
Entry aperture to receiver distance $z_N - z_R$	−5.42	10.08	5.88	5.91	3.77

negative when θ_a is small, meaning that the dielectric and the receiver are fully inside the volume enclosed by the mirror surface and the plane of the entry aperture.

8.7   THE RX CONCENTRATOR

Conceptually identical to the XR case is the one introduced in this section. The RX is a concentrator in which the rays from the source are first refracted (R) and the reflected (X) (see Figure 8.14) (Miñano, Benítez, and González, 1995). Because of the reversibility of the rays, this case is not different from the XR. Nevertheless, the result is quite different, not only because from practical considerations (the RX can be made in a single solid piece, and its dielectric is bigger than the one of the XR), but also because the optical properties are different. In particular, an RX can sometimes behave also as a good imaging device in addition to its excellent properties as a nonimaging concentrator.
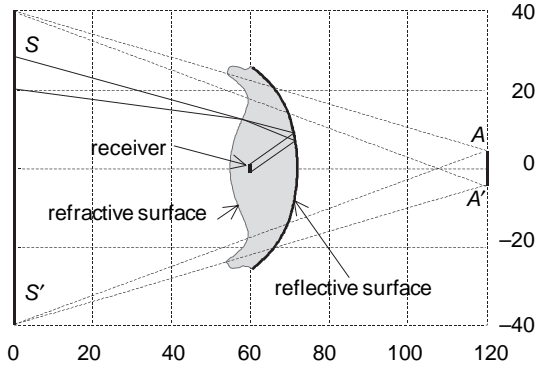
Figure 8.14   Rays emitted by the segment SS' toward the segment AA' are concentrated on the receiver by the RX concentrator in such a way that the back face of the receiver becomes isotropically illuminated by the input bundle.
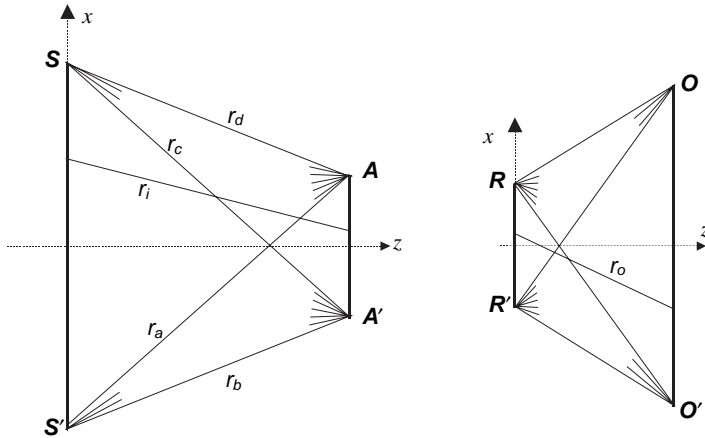
Figure 8.15   Input and output bundles of the RX concentrator in 2D geometry. The rays of the input bundle are those linking the segments SS' and AA', and the rays of the output bundle are those linking OO' and RR'.

The design procedure is conceptually identical to the previous cases. Figure 8.15 shows the input and output bundles of the 2D concentrator to be designed. Each one of the bundles (M_i and M_o) is formed by the rays linking two segments: SS' and AA' for the input bundle and OO', RR' for the output bundle. The segments are parallel to the x-axis and are symmetric with respect to the z-axis. Several usual cases are considered within this description of the input and output bundles: If OO' and RR' are coincident, then the output bundle is formed by all the rays impinging on RR' (the bundle M_o illuminates isotropically all the points of RR') which will henceforth be called the receiver. If OO' is far from RR' (assume it is toward the negative side of the z-axis), then the output bundle illuminates the receiver with restricted exit angle (strictly speaking this will occur only when
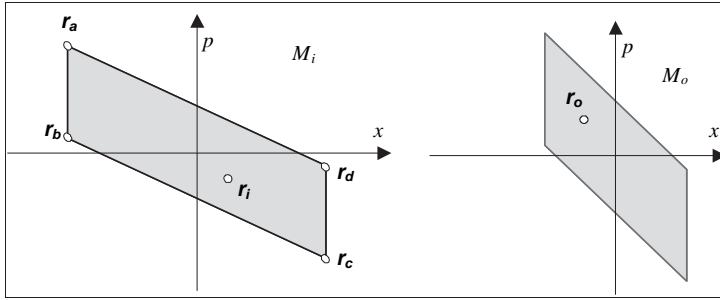
Figure 8.16 Representation of the input and output bundles in the plane x-p. The figure also shows some particular points (r_a, r_b, r_c, r_d) of the input bundle whose trajectories in the x-z plane can be seen in Figure 8.15.

OO' is at infinity). In a similar way if SS' is at infinity (for instance, toward the positive side of the z-axis), then the input bundle is equivalent to a source at infinity. The rays are assumed to proceed from SS' to AA' and from OO' to RR', so if a receiver is placed at RR', its active side must face rightwards.

Figure 8.16 shows the representation of the input and output bundles in the plane x-p. Each point of this plane represents a ray. The correspondence between points of this plane and rays may be done in many ways. For instance, in the representation of the input bundle in Figure 8.16, the abscissa (x) of a point of the plane is the coordinate x of the point of intersection of the ray, and the segment SS' and p is the refractive index times the cosine of the angle formed by the ray and the x-axis at that point of intersection. The refractive index is assumed to be 1 for the input bundle and n ($n > 1$) for the output bundle. For the output bundle, x and p are calculated in a similar way at the points of intersection of the rays and the segment RR'. The rays of the input and output bundles are those represented by the points inside the closed lines of Figure 8.16.

Since it is required that $M_i = M_o$, a necessary condition to start the design process is to ensure that the étendues of both bundles are the same—that is, the areas of the regions representing the bundles in the phase-space x-p must be the same. Analytically, $E(M_i) = E(M_o)$ can be expressed as

$$|SA'| - |SA| = n(|RO'| - |RO|) \tag{8.4}$$

where $|SA|$ denotes the length from the point S to the point A. Eq. (8.4) derives from the application of Hottel formula (Welford and Winston, 1989) to the calculations of the étendues of M_i and M_o.

The design procedure is the same as the one shown for the RR (see Section 8.3) except that now the rays undergo one refraction and one reflection instead of two refractions. The design ensures that the edge rays of the input bundle (∂M_i) are transformed into the edge rays of the output bundle (∂M_o). Figure 8.17 shows the trajectories of the input bundle edge rays in the x-z plane. Since there are no more than two edge rays crossing a point of this plane, the representation of these trajectories has been done in two separate drawings (left and right).
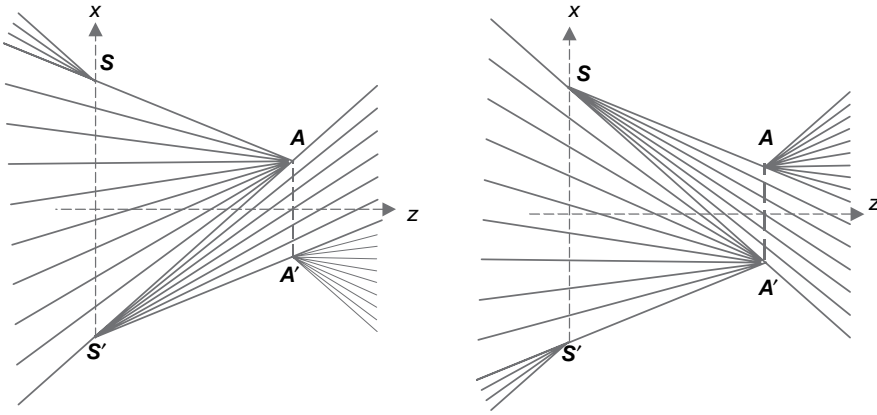
Figure 8.17   Trajectories in the *x-z* plane of the edge rays of ∂M_i.
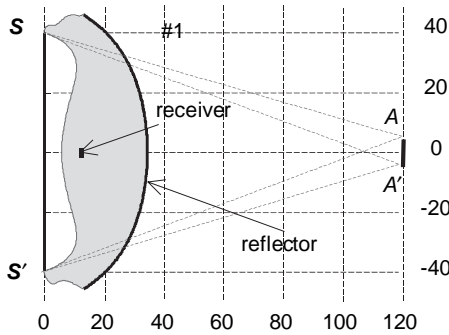
Figure 8.18   *RX* concentrator #1: The input bundle comprises the rays emitted by the disk *SS′* toward *AA′*. The output bundle comprises the rays impinging on the right face of receiver at any angle. Other data for this concentrator are shown in Table 8.3.

8.8   THREE-DIMENSIONAL RAY TRACING OF SOME RX CONCENTRATORS

The results of the 3D ray tracing of several RX concentrators are presented in this section. In the following examples, M_o contains every ray reaching the active side of the receiver. Rays of M_i, which, unlike RR and XR is not an infinite distance (the rays link the disks *SS′* and *AA′*), have been traced to identify which ones hit the receiver by its active side. The unit of length is the receiver radius ($|RR′| = 2$), and the refractive index of the dielectric medium is $n = 1.483$.

Figures 8.18 and 8.19 show the cross sections of several *RX* concentrators. The input bundle of these concentrators is the same: $|SS′| = 80$, the difference between the z coordinate of *SS′* and *AA′* is H = 120. $|AA′|$ can be derived from the preceding data with Eq. (8.4). This type of input bundle is similar to the one obtained as the output bundle of an $f/1.5$ lens when illuminated by an extended source at infinity.
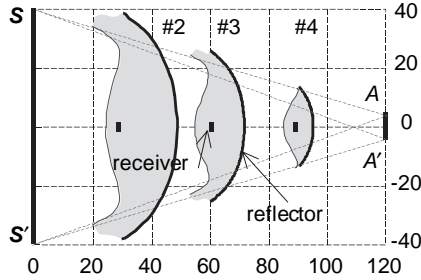
Figure 8.19   *RX* concentrator #2, #3, and #4. The input and output bundles are the same as in Figure 8.18. Other data of these concentrators are shown in Table 8.3.
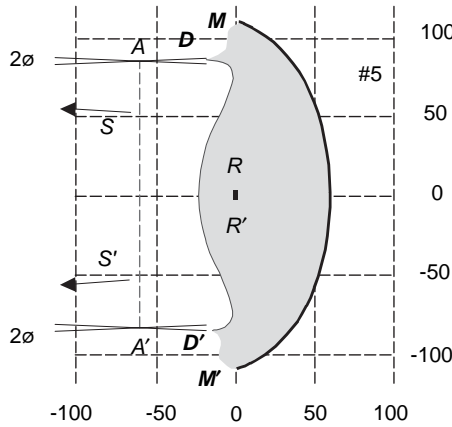
Figure 8.20   RX concentrator #5. The points S and S' of the input bundle of this concentrator are at infinity.

The points S and S' of the input bundle of the concentrator 5 (shown in Figure 8.20) are at infinity. This concentrator is not a typical case of one designed for a source at infinity because the points A and A' are not coincident with D and D'.

Table 8.3 shows some geometrical characteristics and the 3D ray-tracing results of the RX concentrators #1 to #4. The second row in this table shows the values T_t when the shadow of a surface placed at the receiver is taken into account. Observe that this surface intercepts some rays in their trajectories from the refractive surface to the reflective surface. The third row shows the values of T_t when the reflection losses at the mirror (reflection coefficient 0.91) and the Fresnel losses at the dielectric interface (refractive index $n = 1.483$) are also considered.

Figure 8.21 shows the transmission function of the RX #3 (Figure 8.19). This is a function of ρ_S/SS' and ρ_A/AA' (ρ_S is the distance from a point of the circle SS' to the symmetry axis, and ρ_A is the same for a point of the circle AA'). The transmission function gives the percent power reaching the receiver relative to the power emitted from the ring ρ_S, $\rho_S + d\rho_S$ (of the circle SS') toward the ring ρ_A, $\rho_A + d\rho_A$ (of the circle AA'). The radiance of all the rays is assumed to be identical for this calculation. If the 3D concentrator were ideal, then the transmission function would be 1 if $\rho_S/SS' \leq 1$ and $\rho_A/AA' \leq 1$, and it would be zero elsewhere. The

Table 8.3. Geometrical Characteristics and 3D Ray-Tracing Results of Selected RX Concentrators.

Concentrator Characteristics	Concentrator Number			
	1	2	3	4
Total transmission T_t (%)	99.4	98.2	98.3	95.7
T_t considering shadow losses (%)	99.4	98.2	97.8	94.3
T_t considering shadow reflection and Fresnel losses (%)	84.9	84.5	85.2	82.6
Concentrator aperture diameter	91.9	77.5	51.8	28.6
Dielectric thickness at the center	28.5	24.0	17.0	11.0
Distance from the SS' plane to the concentrator bottom	34.5	49.0	72.0	101.0

Concentrator parameters: Refractive index $n = 1.483$; receiver radius of 1; input bundle SS' with a diameter of 80; input bundle AA' with a diameter of 9.38; distance from SS' to AA' of 120; and output bundle yielding maximal concentration.
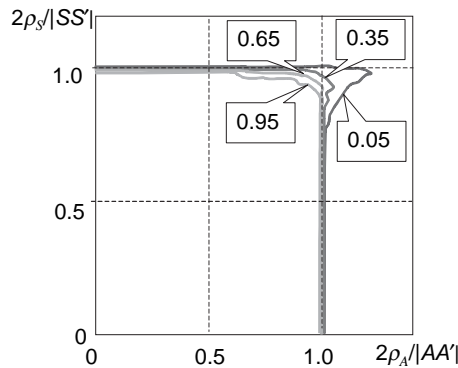
Figure 8.21   Transmission function of the RX #3 of Figure 8.19. The plotted function gives the power reaching the receiver over the power emitted by the points of the ring [ρ_S, $\rho_S + d\rho_S$] (of the circle SS') toward the ring [ρ_A, $\rho_A + d\rho_A$] (of the circle AA').

total transmission T_t appearing in Table 8.3 can also be defined as the percent power reaching the receiver relative to the power emitted by the circle SS' toward the circle AA'. Calculation of the transmission function (and T_t) by averaging the number of rays reaching the receiver must be carefully made. In this average each ray r_x should be weighted by the étendue of the pencil of rays represented by r_x. For instance, assume that the circles SS' and AA' of the input bundle are divided into small regions of area dA_S and dA_A, respectively, and that the rays traced are those linking the centers of the small regions of AA' with the center of the small regions of AA'. Then, each ray represents a pencil of rays whose étendue is

$$dE = \frac{\cos^2 \theta}{D^2} dA_S dA_A \qquad (8.5)$$

where D is the distance between the centers of both regions and θ is the angle formed by the ray and the z-axis.

As the XRs, the rotational symmetric RX performs well for maximal concentration when the average angular spread of the input bundle is small. The geo-

metrical aspect of the RX concentrator is also quite compact, although less than the XR concentrator. Moreover, the distance from SS' to the concentrator's bottom is smaller in an RX than in other classical nonimaging concentrators designed for a finite source as an input bundle. This implies that if a lens is placed at SS' as a first stage of concentration, then the aspect ratio of the whole concentrator is more compact in the case of the RX. The total transmission T_t is, in general, better for the RX than for an equivalent Compound Elliptical Concentrator (CEC) when the input bundle has a small angular spread, and worse for the RX than for the Flow Line Concentrator, which is one of the few known ideal 3D concentrators (defined by $T_t = 1$). In the case of a large angular spread of the input bundle, the CEC has better total transmission than the RX concentrator.

The RX uses much more dielectric material than an equivalent *XR* concentrator, but it can be made in a single piece, whereas the XR needs at least two pieces that must be assembled. Then, the RX seems more appropriate than the XR when the cost of the dielectric material is not critical or when the assembly may be a complex process. In both concentrators (RX and XR), the receiver faces in opposite direction as the concentrator's entry aperture. This is a problem in some applications, such as photovoltaics where the heat sink of the solar cell is in the back side of the cell and may introduce additional losses because of its shadowing.

8.9   THE XX CONCENTRATOR

Similarly, we can design a concentrator in which both surfaces are reflective. The design technique is the same, with no shadowing effect taken into account. In general, one of the shadows of the mirrors dramatically degrades the performance of the concentrators for maximal concentration on flat receivers (emitters). This problem can be solved for tubular receivers (see Chapter 14); in other words, the mirrors can be designed so the efficiency reduction due to shadowing is very small. For flat receivers there is also a solution in some cases. This solution is based on using layers of low refractive index material (n_1) immersed in a high refractive index layer (n_h) (see Figure 8.22). Such layers appeared also in the design of the
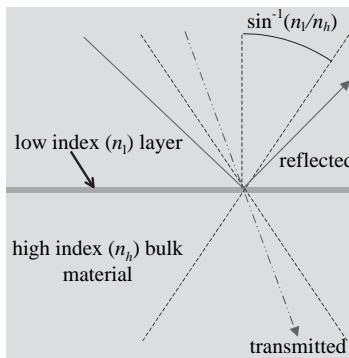
Figure 8.22   A low index layer immersed in a high index material is a mirror at grazing angles and does not modify the ray trajectories otherwise.
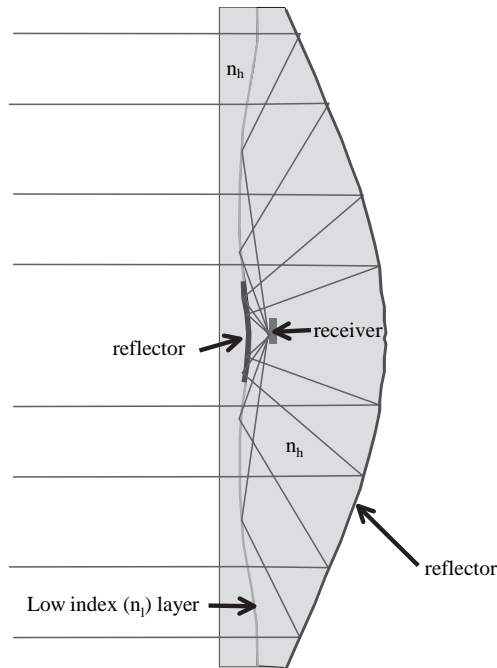
Figure 8.23   XX ($n_1 = 1.5$), refractive index of the low index layer $n = 1.33$. Half acceptance angle 1.5°, Geometrical Concentration 3283.

CTC (see Chapter 6). The basic idea is that if the layer is thin enough, it does not modify the trajectory of the rays if these hit the interface at an angle smaller than $\sin^{-1}(n_1/n_h)$, or otherwise the layer behaves like a reflector.

Fortunately some XX designs are such that most of the collected rays cross one of the mirror surfaces at small angles when they shouldn't be reflected according to the design technique, whereas they hit the mirror surface at great angles when a reflection is necessary. Thus, the low index layer is adequate for such designs. Figure 8.23 shows an example of an XX in which the low index layer technique applies. In the central part of the left mirror the incidence angle is not enough to produce total internal reflection. For this reason it should be mirrored, and this creates some losses by shadowing. These losses can be kept below 5% for most of the designs with maximal concentration and collimated (<±8 degrees) infinite source.

8.10   THE RXI CONCENTRATOR

As we saw before, both the XR and the RX have the active side of the receiver (emitter) facing in the opposite direction of the entry aperture. This creates a practical problem when, for example, we want to use an optoelectronic component. Heat sinks and electrical connections must be in the back side of the component. The RXI concentrator was initially designed to solve this problem. The RXI can also be made in a single dielectric piece but, unlike the RX, the receiver, which is
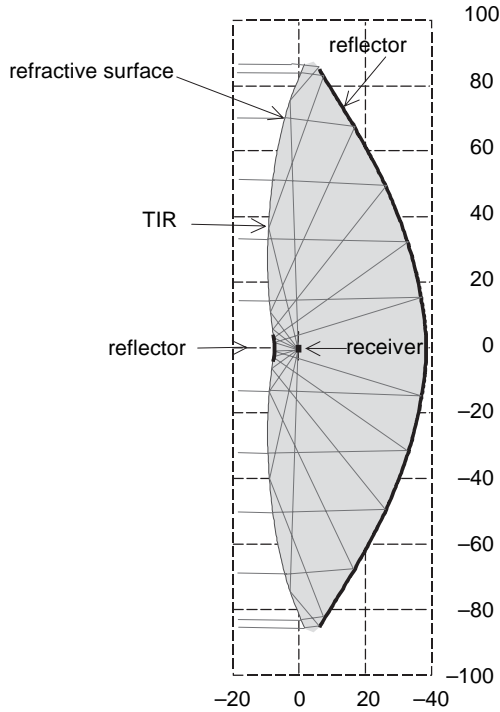
Figure 8.24   RXI concentrator with maximum concentration ($C_{g3D} = 7387$, $\theta_a = \pm 1°$ and $n = 1.5$).

also immersed in the dielectric, faces the entry aperture (Miñano, González, and Benítez, 1995).

There is an important difference between the RXI concentrator and other related concentrators: the RR, the XR, and the RX. All of these concentrators are formed by two optical surfaces. The rays hit one of the surfaces and then the other when they go from the source to the receiver. The RXI also has two optical surfaces, but unlike the other cases, the rays impinge on its entry aperture, twice. First they suffer a refraction then they are reflected on the second surface and finally they are reflected again (by total internal reflection) on the first surface and sent toward the receiver (see Figure 8.24). The difference with respect to the RX concentrator is that the rays interact twice with the first surface. This fact modifies substantially the design procedure.

The condition for total internal reflection cannot be fulfilled at every point of the entry aperture, and because of this, a small part of the entry aperture surface should be mirrored.

The main aim in the design of the RXI was to obtain (1) a highly compact concentrator for maximal concentration designs in which M_i has a small angular spread; (2) high values of the total transmission T_t (like in the XR and the RX); (3) a concentrator that can be made in a single solid piece (like the RX); and (4) the active side of the receiver facing the concentrator's aperture. This last condition is what differentiates the potential applications of the RX and the RXI.

The RXI can be designed in an iterative process of XX and RX designs. Here, the RXI is considered not as a two-surface optical device but as a three-surface

one. Two of these surfaces (named 1X and 2) are reflective, and the last one (1R)
is refractive. It is also assumed that the rays of the input bundle M_i are first
refracted on surface 1R and then reflected by surface 2 and finally reflected by
surface 1X in this order, no matter what would actually happen. For example, the
first change of trajectory of a M_i ray is calculated at the point of interception of
the ray with the surface 1R, even if this ray has previously intercepted another
surface. Henceforth, the surfaces 1R, 2, and 1X will be designated curves 1R, 2,
and 1X, respectively, because only meridian rays are considered in the design. If
the iterative process is successful, curves 1R and 1X converge to the same curve,
which will be designated curve 1.

It is assumed that the description of the bundles M_i and M_o, as well as the
refractive index n of the dielectric material, is known for the design. Of course, M_i
and M_o must fulfill $E(M_i) = E(M_o)$. In the examples shown in Figures 8.24, 8.27,
and 8.28, M_i is a source at infinity, and M_o yields maximal concentration (i.e., M_o
is formed by all the rays reaching the upper face of the receiver).

As an example of the design procedure, here are the steps to follow for design-
ing an RXI with maximal concentration (or close to it).

1.   The iterative process starts with an arbitrary 1X curve subject to the follow-
     ing restrictions: (a) every ray of the bundle ∂M_i intercepts the curve 1X once.
     The same must happen with the rays of ∂M_o when reversed (see Figure 8.25).
     It is assumed that the forward direction of the rays is from M_i and M_o. The
     rays of M_i and M_o in Figure 8.25 advance rightward toward increasing values
     of z. (b). The curve 1X is symmetric with respect to the z-axis.
2.   Choose a 1R curve coincident with the 1X curve.
3.   Trace the ray trajectory r_1 after being refracted at the point R of the curve 1R
     and the ray trajectory of r_2 before being reflected at the point I of the curve
     1X (see Figure 8.25). r_1 is the edge ray of ∂M_i whose point of interception with
     the curve 1R has the most negative coordinate x and whose direction cosine p
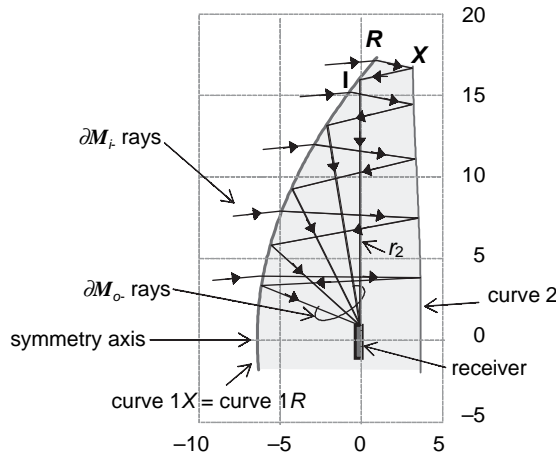
Figure 8.25   Curve 2 obtained in the first iteration of an RXI design. Curve 1X (coincident
with curve 1R in this design step) is the one proposed at the beginning of the design pro-
cedure. The RXI is designed for a source at infinity (subtending an angle of ±5°) as input
bundle and for maximal concentration.

(with respect to the *x*-axis) at this point is the most negative one when com-
pared with other edge rays reaching this point (this point is the point R in
Figure 8.25). When evaluating the sign of the direction cosine, it is important
to keep in mind the convention about the forward direction of the rays. r_2 is
the ray of ∂M_o whose point of interception with the curve $1X$ has the most neg-
ative coordinate x (point I in Figure 8.25). The trajectories of r_1 after refrac-
tion on $1R$ and that of r_2 before reflection on $1X$ cross at the point X (if not,
start the process again with a new $1R$ curve at step 1). The point X will be a
point of the curve 2.

4. Now make up the subset ∂M_{i-} by choosing, at every point of the curve $1R$, the
    ray of ∂M_i with the most negative value of p and form ∂M_{o-} in a similar way
    with the rays of ∂M_i when they intercept the $1X$ curve. The ray r_1 belongs to
    ∂M_{i-}, and r_2 belongs to ∂M_{o-}.

5. Calculate the reflector curve 2 that passes through X and couples ∂M_{i-} and
    ∂M_{o-} (Figure 8.25). The ray r_1 will become r_2 after reflection at X. Observe that
    this problem is qualitatively identical to a Cartesian oval problem. In order to
    solve it, it is necessary to have no more than one ray of ∂M_{i-} and no more than
    one ray of ∂M_{o-} at any point on the curve 2. If this condition is not fulfilled,
    the iterative process should be restarted again with another $1R$ curve (go to
    step 1). This condition is not fulfilled when, for instance, the ∂M_{i-} rays form a
    caustic after refraction on $1R$ or when the ∂M_{o-} rays form a caustic before
    reflection on $1X$. These cases can be easily recognized because a loop or a dis-
    continuity of the slope appears in the shape of the curve 2. When that occurs
    it is advisable to use a flatter curve $1R$ at step 1. Curve 2 must be calculated
    from the point X up to the symmetry axis.

6. Trace the rays of ∂M_{i+} (rays of ∂M_i not belonging to ∂M_{i-}) after refraction at
    curve $1R$ and then after reflection at curve 2. Calculate now the new $1X$ curve
    that passes through the point I and reflects the ∂M_{i+} rays into the ∂M_{o+} rays
    (these are the ∂M_o rays not belonging to ∂M_{o-}; see Figure 8.26). The calcula-
    tions needed to get the new $1X$ curve are again a Cartesian oval problem and
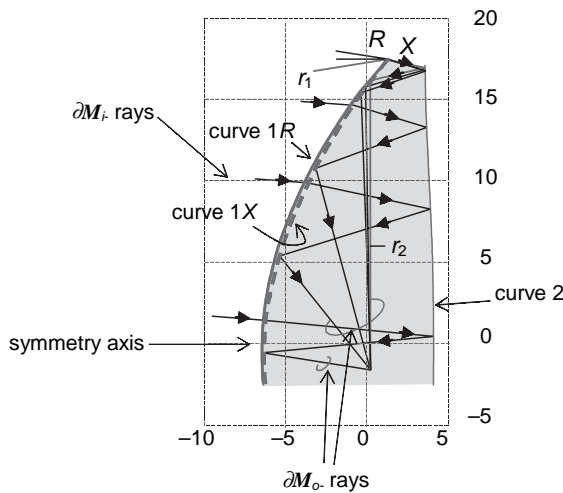
Figure 8.26   Curve $1X$ obtained at step 6 of the first iteration of the RXI design.

the same considerations done in step 5 apply to this step. The curve has to be calculated up to the symmetry axis.

7. If the old and new $1X$ curves are close enough, then the 2D design is finished; if not, let the new $1R$ curve be equal to the last calculated $1X$ curve, and continue the process in step 5.

8. Analyze the condition for total internal reflection (for the edge rays) on curve 1 (at this step $1X$ and $1R$ have converged to the same curve). Usually, this condition is not fulfilled in the central part of the curve, and thus this part should be mirrored (see Figure 8.24).

The preceding design procedure does not always converge, and the analysis of the conditions for convergence is a complex task. Obviously, convergence depends on the selection of the $1X$ curve in step 1. In general, it is advisable to choose a smooth curve in this step. When the starting curve $1X$ is very close to the receiver, then the process may converge in an RXI concentrator that has part or its entire receiver outside the dielectric. In this case the receiver is virtual.

Figures 8.24, 8.27 and 8.28 show three different RXI designs for maximal concentration and for a source at infinity (subtending angles $\pm1°$, $\pm3°$, and $\pm5°$, respectively). Figure 8.29 shows three RXI concentrators that have also been designed for maximal concentration but whose input bundle is a finite source formed by the rays issuing from every point of AA' toward every point of BB'.
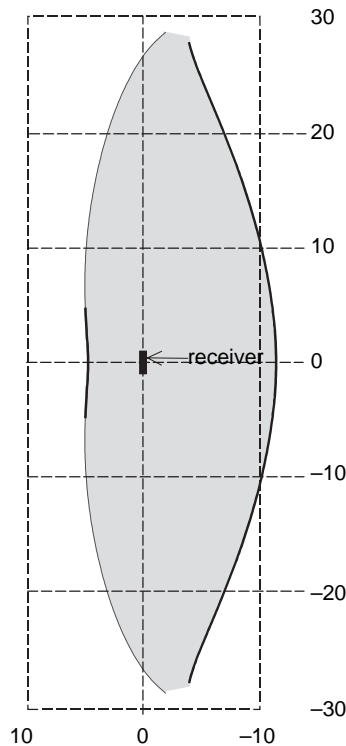
Figure 8.27   The same as in Figure 8.24 but with an angle $\theta_i = 3°$.
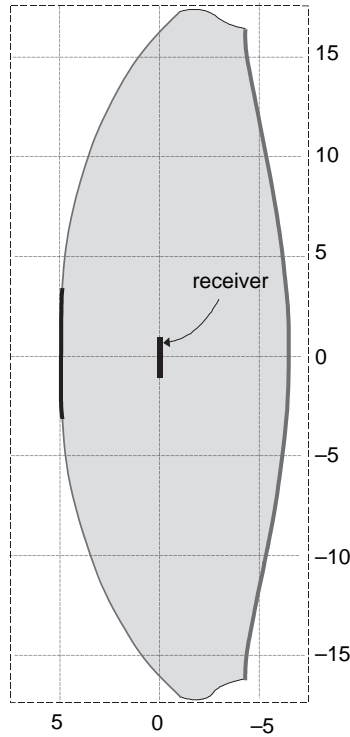
Figure 8.28   The same as in Figure 8.24 but with an angle $\theta_i = 5°$.
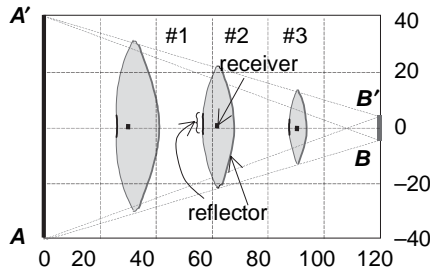
Figure 8.29   RXI concentrators for the input bundle formed by the rays issuing from any point of AA' toward any point of BB'. The output bundle contains all the rays that reach the receiver.

8.11   THREE-DIMENSIONAL RAY TRACING OF SOME RXI CONCENTRATORS

This section gives results of the 3D analysis of RXI obtained by rotational symmetry around the z-axis. Two types of RXI concentrators have been designed and analyzed. The input bundle of the first type of RXI concentrator is a source at infinity that comprises all the rays impinging on the concentrator aperture (surface

generated by curve 1) whose angle with the axis of symmetry is not greater than a given value θ_i. The input bundle for the other type of RXI concentrators comprises the rays issuing from the disk with diameter AA' toward the disk with diameter BB' (see Figure 8.29). The output bundle for both types of RXI concentrators comprises all the rays reaching the receiver. This is a flat disk centered at the symmetry axis and orthogonal to it.

Figure 8.30 shows the transmission-angle curves for several *RXI*s with sources at infinity and with different acceptance angles. Three of these RXIs are those whose cross section appears in Figures 8.24, 8.27, and 8.28. These curves do not take into account the shadowing of the front metallic area nor the one produced by rays intercepting the back side of the receiver. Both shadowing effects are taken into account in the calculations of the total transmission appearing in Tables 8.4 and 8.5—that is, for the calculations of the total transmissions in both tables, it is considered that a ray is lost if it intercepts the outer face of the mirrored region in the aperture surface or if it intercepts the back face of the receiver.

Table 8.4 also gives other results of the 3D analysis and some geometrical characteristics of the rotational RXI concentrators designed for a source at infinity. The geometrical concentration C_G is defined $C_G = A_e/A_r$, where A_e is the area of the projection of the concentrator entry aperture on a plane orthogonal to the
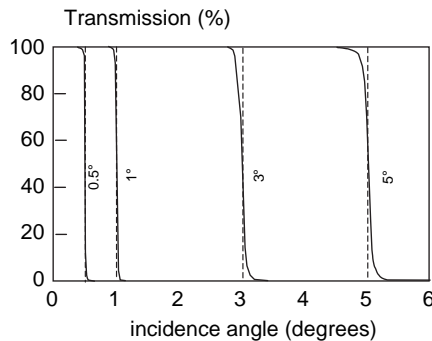
Figure 8.30   Transmission-angle curves for several 3D RXIc's for sources at the infinity with angle θ_a. The transmission includes the shadowing of the upper mirror.

Table 8.4. Geometric Characteristics and 3D Ray-Tracing Results of Selected RXI Concentrators for Maximum Concentration and Infinite Source. Refractive index $n = 1.5$, length unit is equal to receiver radius.

Parameters	Acceptance Angle θ_1 (degrees)			
	±0.5	±1	±3	±5
Geometric concentration C_G	29,546	7,387	821.5	296.2
Total transmission, $T_t(\theta_i)$ (%)	96.9	97.3	94.5	86.9
Total transmission without shadow $T_t(\theta_i)$ (%)	97.4	97.7	96.6	96.7
Thickness/entry aperture diameter	0.278	0.279	0.289	0.332
Dielectric thickness	95.6	47.9	16.6	11.4
Entry aperture diameter	343.8	171.9	57.3	34.4
Diameter of the front metallic reflector	23.2	11.2	8.2	10.8
Receiver to concentrator bottom distance	76.0	38.6	11.6	6.5

Table 8.5. Geometric Characteristics and 3D Ray-Tracing Results of Selected RXI Concentrators for Maximum Concentration and Finite Source. Refractive index $n = 1.5$; length unit is equal to receiver radius; input bundle AA' diameter, 80; input bundle BB' diameter, 9.5; distance from AA' to BB', 120.

Parameter	Concentrator Number		
	1	2	3
Total transmission, T_t (%)	96.0	93.2	86.7
Total transmission without shadow T_t (%)	97.8	96.8	95.1
Thickness/entry aperture diameter	0.252	0.256	0.226
Dielectric thickness	15.4	11.2	6.1
Entry aperture diameter	61.2	43.6	26.8
Diameter of the front metallic reflector	7.9	7.2	5.6
Receiver to concentrator bottom distance	11.3	8.2	4.0

symmetry axis, and A_r is the receiver area. Table 8.5 gives the results of the 3D analysis of the rotational RXI concentrators whose cross section appears in Figure 8.29.

8.12   COMPARISON OF THE SMS CONCENTRATORS WITH OTHER NONIMAGING CONCENTRATORS AND WITH IMAGE FORMING SYSTEMS

The fundamental advantages of nonimaging concentrators versus imaging ones are that those get concentrations closer (in general greater than 90%) to the thermodynamic limit (see Chapter 1) and that nonimaging concentrators can be very simple (one or two optical surfaces are usually enough). For instance, a single parabolic mirror of revolution does not achieve concentrations greater than 25% of the limit and not greater than 56% of the limit when the parabolic mirror is combined with a spherical refracting surface with $n = 1.5$. Increasing the number of components improves the concentration in an imaging device, but concentration levels as high as in the nonimaging concentrators are only achieved with systems or materials that are usually impractical, such as the Luneburg lens (see Appendix F). Imaging concentrators, formed, for instance, by a single lens, compete advantageously in the low concentration regime when the desired concentration is small compared with the thermodynamical limit.

The best-known nonimaging concentrator, the compound parabolic concentrator (CPC), is used as a reference for other nonimaging concentrators. The total transmission of the CPC of revolution is above 95% for all acceptance angles (for $\theta < 5°$ the total transmission is about 95%). Its most important disadvantage is its thickness for small acceptance angles, approximately $0.5(C_g)^{1/2}$. For instance, this ratio is 43 for a CPC of $n = 1.5$ and $\theta = \pm1°$ to be compared with the ratio of 0.28 of an XR or an RXI. Practical CPCs for this acceptance angle range are truncated, which only decreases slightly its total transmission, whereas the depth can be dramatically reduced (see Chapter 5). For larger acceptance angles, the CPC increases its total transmission and decreases its size strongly, whereas the XR and the RXI

decrease their total transmission and increase their size slightly, so the CPC becomes a better candidate for large acceptance angles (above ≈15°).

There are other concentrators of the CPC family with a smaller size. For instance, the dielectric-filled CPC with curved entry aperture, called DTIRC (Ning, Winston, and O'Gallagher, 1987), has a thickness to entry aperture ratio between 1 and 2 for acceptance angles above 10°.

When the acceptance angle is smaller, the combination of an imaging-forming device such as a parabolic mirror with a concentrator of the CPC family (usually a Compound Elliptical Concentrator, CEC) leads to more compact devices than a single CPC-type concentrator. The highest levels of solar concentration have been achieved with combinations of parabolic mirrors and nonimaging concentrators, similar to the aforementioned one (except that the nonimaging concentrator is not a CEC) (Gleckman, O'Gallagher, and Winston, 1989). Such combinations are thicker than the equivalent SMS concentrators. Their optical performance may be better or worse depending on the concentrator parameters. The theoretical limit of concentration can be attained with these two-stage systems if the imaging stage (parabolic mirror) is without aberrations—that is, when the mirror forms a sharp image of the source at the CEC entry aperture. This requires f-numbers much greater than 0.287. For instance, consider a combination of an $f/2$ parabolic mirror with a dielectric filled CEC ($n = 1.483$) to form a concentrator with design acceptance angle $\theta_a = \pm 1°$. The geometrical concentration of the first stage is $C_{g1} = \sin^2 \psi \cos^2 \phi / \sin^2 \theta_a = 181.76$ (where ϕ is the rim angle of the mirror as seen from the focal point—that is, $\phi = \tan^{-1}(1/2f)$). The second-stage concentration is approximately $C_{g2} = n^2/\sin^2 \phi = 37.39$, so $C_{g3D} = C_{g1}C_{g2} = 6796$, which is 94% of the maximum concentration ($C_{g3D} < n^2/\sin^2 \theta_a$). Considering this loss of concentration as a loss of total transmission and assuming that the total transmission of the CEC is similar to that of a CPC with acceptance angle $\theta_a \approx \phi$ (i.e., $T_{CEC} \approx 0.962$; see Figure 4.13), then the total transmission of the mirror-CEC combination is approximately 90%–which is 4.6% less than an *XR* (see Table 8.2). As the f number increases, the transmission angle curve of the mirror-CEC combination is sharper, but the concentration of the first stage decreases, and then the concentration (and the size) of the second stage increases. That reduces the total transmission of the CEC. When f is very high, the shadow of the second stage on the mirror can be the cause of the decrease of the total transmission.

The rays suffer a single metallic reflection in both systems (assuming that the CEC can work by total internal reflection). The Fresnel reflections at the CEC entry aperture are, approximately, those corresponding to a beam impinging on its entry aperture at angles (with the normal to the entry aperture) below φ ($\phi \approx 14°$). The angular distribution at the lens entry aperture of the XR is less homogeneous. At the center of the lens the beam is impinging on the lens at angles below 18.52°. Because of the quasi-spherical shape of the lens and because the center of the lens is the point of the surface closest to the receiver center, it can be expected that the rays will not form angles much greater than this with the normal to the lens surface, and thus concluding that there will not be significant differences in the Fresnel losses of both systems.

Optical losses depend on the type of materials used. Reflectivities between 85% and 90% for the mirror reflection are easily achieved with an evaporated aluminum coating. Fresnel losses can be reduced with an antireflection coating. Absorption and dispersion losses not only depend on the type of dielectric material to be used

but also on the average length of the rays, and thus it depends on the global size of the concentrator. Values of concentrator optical efficiencies between 70% and 85% can be expected, using very common materials (for instance, poly-methylmetacrylate as dielectric and aluminum coating as reflector) for average ray length no greater than 60 mm and acceptance angles greater than 1°. Another source of losses are the imperfections of the optical surface (including roughness). This type of loss depends on the technology used to make the concentrator and is more important for small acceptance angles. Since the surfaces of nonimaging concentrators are aspheric, polymer injection molding is usually the most appropriate technology. In that case, imperfections of the optical surfaces depend strongly on the mold accuracy, mold surface roughness, molding cycle time, and the maximum material thickness.

8.13   COMBINATION OF THE SMS AND THE FLOW-LINE METHOD

The Simultaneous Multiple Surfaces design method has been used in previous sections in the 2D design of several ideal concentrators, such as the RXI, RX, and XR. These concentrators transfer the rays emitted by a source to a receiver through a number of sequential incidences. For example, in the case of the RX all the rays of the source that impinge on the concentrator undergo first a refraction and then a reflection before reaching the receiver. In the case of the RXI, the rays undergo a refraction, a metallic reflection, and a total internal reflection on their way from source to receiver.

Let us take by way of example the RX concentrator that appears in Figure 8.31b. In the same way as the rays, the flow lines of the bundle transmitted by the concentrator are refracted on the refractive surface and reflected on the mirror. The functioning of the RX concentrator can thus be understood as the transformation, through two sequential incidences, of the flow lines of the source into those of the receiver.

The string method with which the CPC was designed (see Chapter 4) is used for calculating mirrors that coincide with the flow lines of the transmitted bundle (see Figure 8.31a). These mirrors act as a guide for the flow lines instead of reflecting them, as is the case with the SMS method. The behavior of these mirrors,
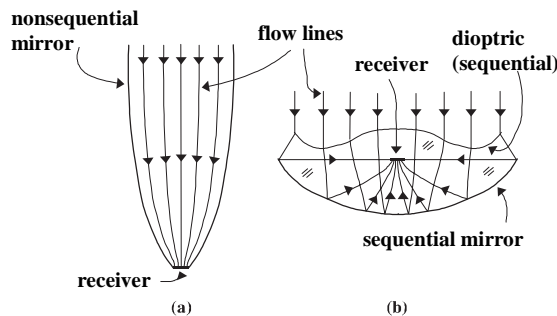
Figure 8.31   Flow lines of the bundle transmitted by (a) a CPC and (b) an RX concentrator.

moreover, is not sequential: The number of reflections that the rays from the source may undergo before reaching the receiver is not equal for all the rays. We shall call these mirrors nonsequential, in order to distinguish them from the sequential mirrors used up to now in the SMS method.

This section deals with the inclusion of nonsequential mirrors in the SMS design method. In principle the designs obtained are ideal in two dimensions (later, some of them are truncated for practical reasons, as occurs in the CPC case), and the nonsequential mirrors coincide with a flow line of the transmitted bundle, M_c. The nomenclature used for referring to the different designs is similar to that of the other concentrators obtained through the SMS method. Each concentrator is named with a succession of letters indicating the order and type of incidence of the optical surfaces that the bundle M encounters on its way from source to receiver. The following symbols are used: R = refraction, X = sequential reflection, X_F = nonsequential reflection, and I_F = nonsequential total internal reflection (the subindex F refers to the coincidence of the nonsequential surface with the flow line).

An example of a concentrator designed with the SMS method that contains nonsequential mirrors is presented next, the XRI_F. This design is made up of two surfaces to be calculated. Other examples are presented as applied designs in Chapter 14. Friedman, Gordon, and Ries (1995) and Friedman and Gordon (1996) published two designs that can be considered as degenerate cases of SMS designs with nonsequential mirrors. Using the nomenclature just described, the two-stage solar collector of Friedman et al. (1995) is an XX_F, whereas the monolithic concentrator of Friedman and Gordon (1996) is an RX_F.

8.14   AN EXAMPLE: THE XRI$_F$ CONCENTRATOR

The XRI_F concentrator is made up of a sequential primary mirror (X) and a secondary concentrator conceptually similar to that of Figure 5.31, composed of a refractive surface (R) and two nonsequential mirrors that work by total internal reflection (I_F). We will call these types of concentrators (like the secondary one) DTIRC (Dielectric Totally Internal Reflection Concentrator). Figure 8.32 shows an XRI_F design for an infinite source of acceptance $\pm\alpha$ and maximum concentration on the segment RR'. The size of the secondary is not to scale, in order to facilitate explanation. The points I and I', which are the rims of the primary mirror, define the entry aperture of the concentrator. Therefore, the bundle M_o, which coincides with M_R, is composed of all the rays that impinge on the segment RR' from below, and the bundle M_i by the rays that impinge on the segment II' with an angle of incidence less than α. We shall call rays $e(+)$ and $e(-)$ those rays of M_i with angle
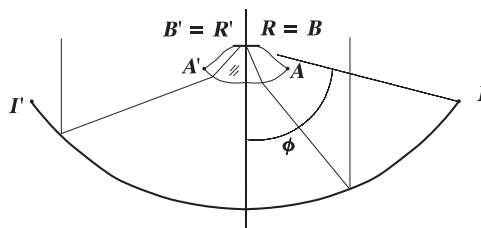
Figure 8.32   Description of an XRI$_F$ concentrator. The secondary is not to scale.
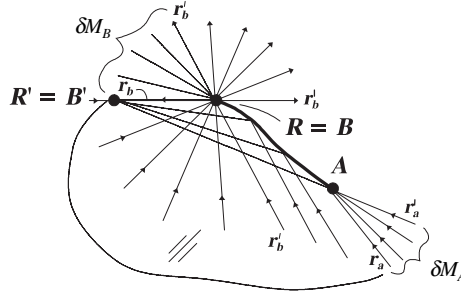
Figure 8.33   Edge rays associated to the nonsequential mirror I_F. The refraction on the refractive surface has been omitted for the sake of clarity.

of incidence $+\alpha$ and $-\alpha$, respectively (the clockwise angles are taken as positive). The secondary is considered as transparent for the rays of M_i before impinging on the primary (its shadow will be considered later on).

The input parameters of the design are (1) the length of the receiver RR', (2) the refractive index n of the secondary, (3) the acceptance angle α, (4) the rim angle of the primary ϕ, and (5) the profile of the two symmetric mirrors I_F. It is not necessary to design these mirrors because on the surfaces R and X there are sufficient degrees of freedom for practical purposes.[1] The edges of the mirror I_F are the points A (and A') and $B \equiv R$ (and $B' \equiv R'$). The profile of the surface I_F should be selected so that it produces total internal reflection and that all edge rays impinging between A and B undergo a single reflection on the mirror. This condition is used for the application of the edge ray theorem, as proven in Appendix B. Additionally, we will guarantee the following:

1.   The rays of M_c that pass through the point A are edge rays and form a connected bundle at A. The angle formed between the tangent to the mirror at A and the ray r_a is smaller than that it forms with r_a'. The rays of M_c at A situated between r_a and its symmetric with respect to the tangent are the edge rays δM_A, which do not extend beyond A (that is, they are edge rays only before reaching A).

2.   The rays of M_c that pass through the point B are edge rays and form a connected bundle at B. The angle formed between the tangent to the mirror at B and the ray r_b is smaller than that it forms with r_b'. The rays of M_c at B situated between r_b and its symmetric with respect to the tangent are the edge rays δM_B, which begin at B (that is, they are edge rays only after passing through B).

The positions of points I and I' are calculated with the angle ϕ and the length of segment II', which is equal to $n/\sin\alpha$ by conservation of the étendue.

Figure 8.33 shows the edge rays associated to one of the nonsequential mirrors (I_F), specifically those reflected toward R'. The definition and use of these edge rays are detailed in Appendix B. It should be noted that in this case the rays r_a and r_a'

---

[1]   Later on we shall see that, strictly speaking, there is a small portion of the edge rays that it is not possible to couple, as occurs in other SMS designs (Section 8.3). A subsequent ray tracing shows that the coupling of these rays is good, even though it has not been guaranteed by the design. The mirrors I_F may be used for guaranteeing this coupling, although it is unnecessary from a practical point of view.

bisect the bundle that impinges at A so that this bundle coincides with ∂M_A. Meanwhile, the edge rays ∂M_B couple directly with rays of ∂M_o because the points R (R') and B (B') coincide. Note that the ray r'_b is reflected at R toward R'.

The surfaces R and X are obtained by connecting the edge rays ∂M_i with ∂M_o and ∂M_A, a process illustrated in Figure 8.34 and that includes the following steps:

1. Trace r'_a, the ray $e(-)$ that impinges at point I, is reflected toward point A, and then refracted toward R' (see Figure 8.34a). Calculate the normal vectors to the mirror at I and to the refractive surface at A so that the laws of reflection and refraction are satisfied. Calculate the trajectory of the ray r'_b inside the secondary, knowing that it is reflected at point R toward R' (the normal at R is known).
2. Construct the first portion of the primary mirror, imposing the condition that the rays $e(-)$ are reflected toward point A (see Figure 8.34b). Obviously, the
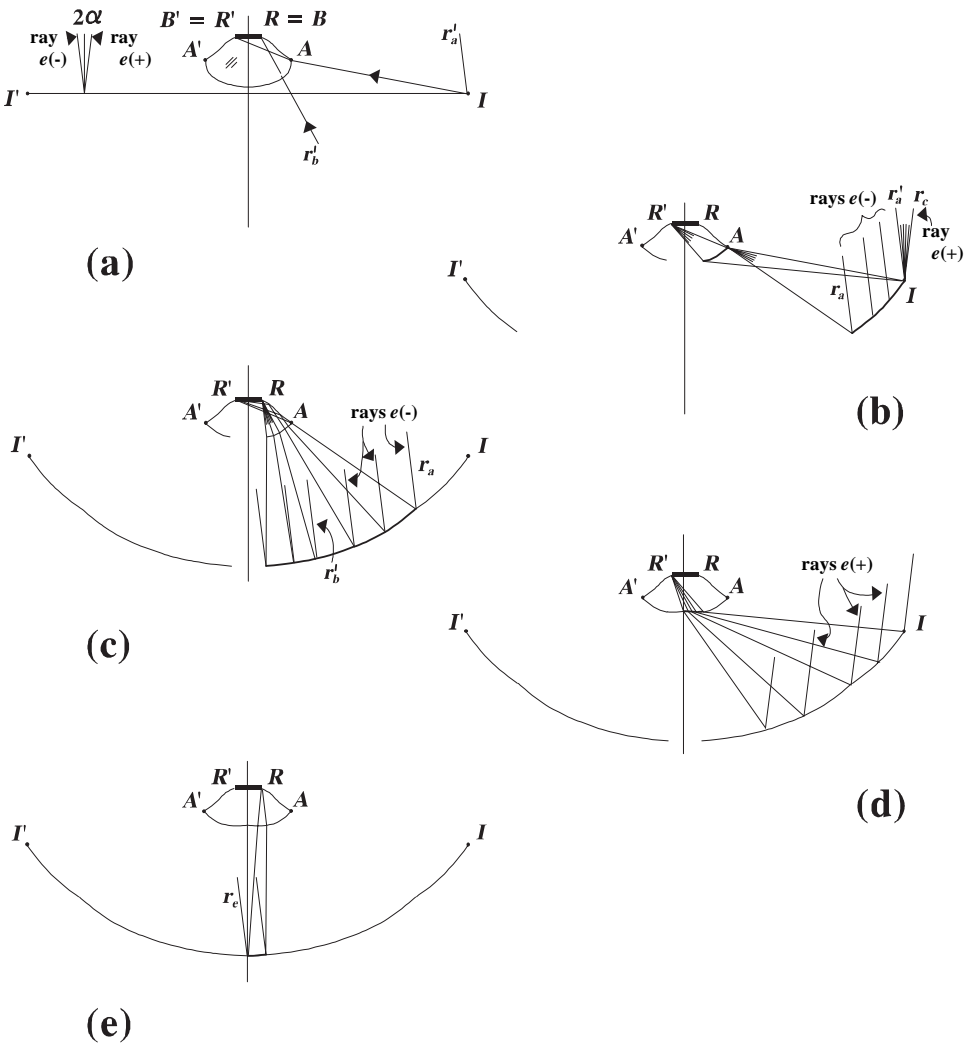
Figure 8.34   Design process of an XRI$_F$ concentrator.

solution is the parabola with focus at A and whose axis is parallel to the rays $e(-)$ passing through I. The final point of this portion is given by the ray $e(-)$ that, after undergoing a refraction (on the refractive surface) at A and a reflection (on the mirror I_F), reaches R'. Moreover, this ray is the ray r_a.

3. Calculate the first portion of the refractive surface with the condition that the rays reflected at I must be refracted toward R' (see Figure 8.34b). The solution is the Cartesian oval with foci I and R' that passes through A. The ray $e(+)$ that impinges at I marks the end of this portion.

4. Calculate the next stage of the primary mirror knowing that the rays $e(-)$, after being reflected on it, must be refracted on the refractive surface, reflected on the nonsequential mirror and impinge at R' (see Figure 8.34c). The solution can be calculated by applying the constancy of the optical path length that must exist, according to Fermat's principle, between the flat wavefront associated to the rays $e(-)$ and the spherical wavefront directed toward R'. The final point is that at which the impinging ray $e(-)$ coincides with the ray r_b' calculated in step 1.

5. Continue to calculate the primary mirror, imposing the condition that the rays $e(-)$, after being reflected on it, are refracted on the refractive surface toward point R (see Figure 8.34c). Again, the condition of constant optical path length determines the solution.

6. Calculate the following stage of the refractive surface, knowing that the rays $e(+)$, after being reflected on the mirror, are refracted toward point R' (see Figure 8.34d).

7. Repeat steps 5 and 6 until the mirror and the refractive surface intersect the symmetry axis (see Figures 8.34d and 8.34e).

The design is complete. In order to illustrate the procedure followed, Figure 8.35 shows the representation of the bundle M_i in the phase space of the entry aperture and of M_o in the receiver phase space. The design procedure does not
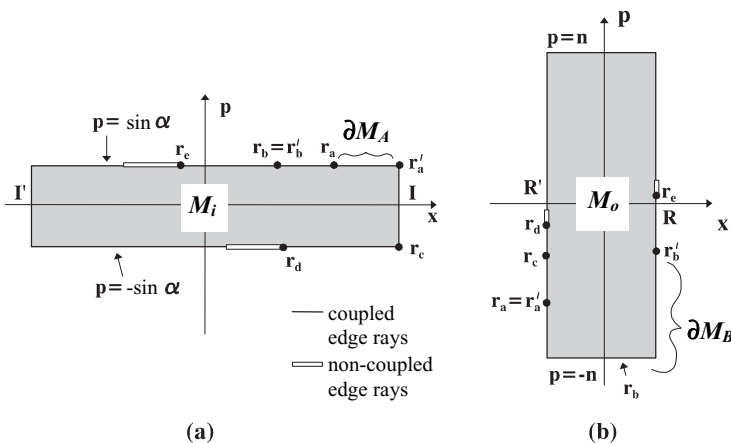
Figure 8.35   Representation of the entry bundle M_i in the phase space of the entry aperture (a) and the exit bundle M_o in the phase space of the receiver (b). Some of the important rays in the design process are shown, as is the portion of edge rays whose transformation is not guaranteed.

guarantee coupling between the two portions of edge rays shown in the figure (defined by the rays r_d, r_e, and their symmetrics with respect to the axis), although a subsequent ray tracing indicates satisfactory coupling. This situation is usual in two-surface SMS designs. No theoretical explanation for this fact is known.

Figure 8.36 shows an XRI_F concentrator designed with the following parameters: (a) the length unit is half the length of the receiver, (b) $n = 1.5$, (c) $\alpha = \pm 0.75°$, (d) $\phi = \pm 80°$, and (e) the surfaces I_F are planes inclined at $70°$ to the vertical. The design is optimal so that the collection efficiency is $\eta = 100\%$, and the concentration is $Cg = Cmax = n/\sin \alpha = 114.6\times$. The shadow produced by the secondary, calculated as the ratio between the apertures of the secondary and the primary, is 6.4%.

It is noteworthy that the XRI_F of Figure 8.36 is optimal in 2D and has a large angle ϕ ($\pm 80°$), which leads to an aspect ratio (i.e., ratio between height and aperture diameter) of 0.30 for the complete system. This is the main difference between the XRI_F and a conventional parabolic mirror-DTIRC combination. This combination behaves far from optimally if a large angle ϕ is selected, as shown in Figure 8.37.
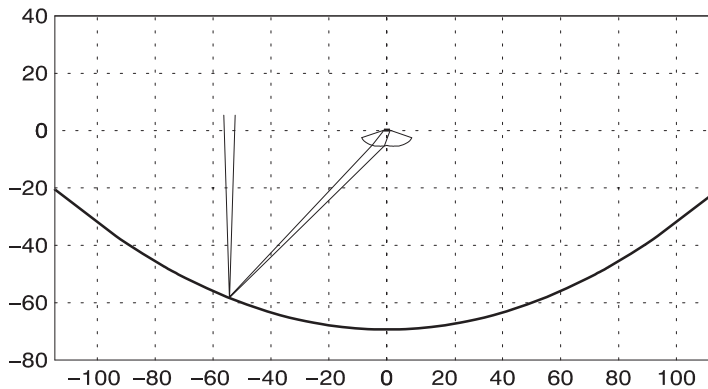
Figure 8.36 XRI_F concentrator designed with refractive index $n = 1.5$, geometrical concentration 114.6x′, acceptance angle $a = \pm 0.75°$, and primary rim angle $\phi = \pm 80°$.
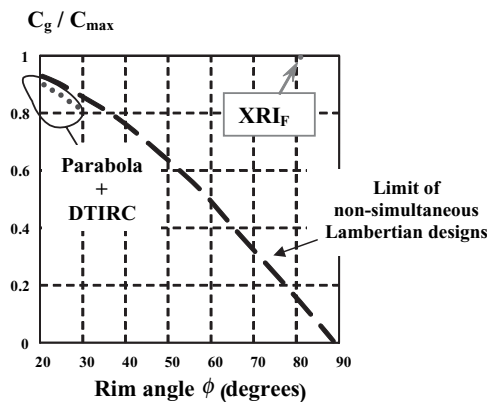
Figure 8.37   Comparison of the XRI_F concentrator with the conventional parabolic mirror-DTIRC secondary combination.

Note also that if the mirrors I_F were selected as horizontal planes, they would not reflect any edge rays, and the angular size of the bundles ∂m_A and ∂m_B would be null. This means that, at this extreme, the XRI$_F$ design coincides with an XR design of Section 8.5.

REFERENCES

Chaves, J., and Collares-Pereira, M. (1999). New ideas for new solar concentrators. In *Nonimaging Optics: Maximum Efficiency Light Transfer V* (Winston, R., ed.). Proceedings of SPIE, Vol. 3781, 174–182.

Friedman, R. P., and Gordon, J. M. (1996). Optical designs for ultra high-flux infrared and solar energy collection: Monolithic dielectric tailored edge-ray concentrators. *Appl. Opt*. **35**, 6684–6691.

Friedman, R. P., Gordon, J. M., and Ries, H. (1995). Compact high-flux two-stage solar collectors based on tailored edge-ray concentrators. In *Nonimaging Optics: Maximum Efficiency Light Transfer III* (Winston, R., ed.). Proceedings of SPIE, Vol. 2538, 30–41.

Gleckman, P., O'Gallagher, J., and Winston, R. (1989). Approaching the irradiance of the sun through nonimaging optics. *Optics News*, 33–36.

Luneburg, R. K. (1964). *Mathematical Theory of Optics*. University of California Press, Berkeley.

Miñano, J. C., and González, J. C. (1991). Design of nonimaging lenses and lens-mirror combinations. In *Nonimaging Optics: Maximum Efficiency Light Transfer* (Winston, R., and Holman, R. L., eds.). Proc. SPIE, Vol. 1528, 104–116.

Miñano, J. C., and González, J. C. (1992). New method of design of nonimaging concentrators. *Appl. Opt*. **31**, 3051–3060.

Miñano, J. C., Benítez, P., and González, J. C. (1995). RX: A nonimaging concentrator. *Applied Optics*, Vol. 34, **13**, 2226–2235.

Miñano, J. C., González, J. C., and Benítez, P. (1995). RXI: A high-gain, compact, nonimaging concentrator. *Applied Optics*, Vol. 34, **34**, 7850–7856.

Ning, X., Winston, R., and O'Gallagher, J. (1987). Dielectric totally internally reflecting concentrators. *Appl. Opt.*, Vol. 26, 300.

Schulz, G. (1983). Achromatic and sharp real imaging of a point by a single aspheric lens. *Appl. Opt.*, Vol. 22, 3242–3248.

Schulz, G. (1988). Aspheric surfaces. In *Progress in Optics* (Wolf, E., ed.). Vol. XXV. North Holland, Amsterdam, pp. 351–416.

Stavroudis, O. N. (1972). *The Optics of Rays, Wavefronts, and Caustics*. Academic Press, New York.

US Letters Patent 6,639,733. *High Efficiency Non-Imaging Optics*.

Welford, W. T., and Winston, R. (1989). *High Collection Nonimaging Optics*. Academic Press, New York.

Welford, W. T., O'Gallagher, J., and Winston, R. (1987). Axially symmetric nonimaging flux concentrators with the maximum theoretical concentration ratio. *J. Opt. Soc. Am. A.* **4**, 66–68.

Winston, R. (1980). Cavity enhancement by controlled directional scattering. *Appl. Opt*. **19**, 195–197.

Wolf, E. (1948). On the designing of aspheric surfaces. *Proc. Physical Soc.*, Vol. LXI, 494–503.

Wolf, E., and Preddy, W. S. (1947). On the determination of aspheric profiles. *Proc. Physical Soc.*, Vol. LIX, 704–711.

9

IMAGING APPLICATIONS OF NONIMAGING CONCENTRATORS

9.1   INTRODUCTION

The design of imaging optical systems is a classic field of research that has achieved a high level of development. There are on the market optical design programs that permit the numerical optimization of the design parameters, allowing us to obtain results that were unattainable with the analytical tools used before the development of computers.

In many imaging applications it would be interesting to use an optical system with an image-side numerical aperture (NA) as high as possible but satisfying a specified imaging quality. This is the case of most radiation sensors, since their sensitivity is proportional to the luminous power density (irradiance) and thus to NA^2. High values of the numerical aperture are especially desirable when working with long wavelengths because the diffraction effects limit the minimum achievable spot diameter, and this limit is proportional to λ/NA (λ is the wavelength of the light in vacuum).

The image-side numerical aperture for paraxial optical systems can be calculated as $NA = n' \sin\beta$, where n' is the index of refraction of the image space and β is the half-rim angle of illumination of the axial point of the image plane. When the angle β is not small, the system cannot be considered paraxial, and in general the preceding formula cannot apply. However, in the case of the optical systems appearing in this chapter, this formula is still valid (as will be shown later). Since $\sin\beta \leq 1$, the inequality $NA \leq n'$ defines an upper bound for NA, which is achieved only if $\beta = 90°$.

Designing optical systems for very high numerical apertures, even with modern design tools, is not an easy task. High-aperture camera objectives usually have six lenses and obtain NA values of between 0.4 and 0.5, with $n' = 1$ and $\beta = 25–30°$. The designs with the highest values of NA are microscope objectives, which need to be even more complex to avoid aberrations. In a simple oil immersion microscope design, 10 lenses are typically used, and $\beta \approx 60°$ is achieved, which implies NA around 1.35 with a refractive index $n' = 1.52$.

Nonimaging designs show that if image formation is not required, then it is possible to obtain simple optical systems that concentrate the incident light much

219

more than imaging systems. A well-known example of a nonimaging concentrator is the CPC (compound parabolic concentrator). This concentrator comes very close to the thermodynamic limit of concentration, which corresponds to the isotropic illumination (i.e., $\beta = 90°$) for every point of the receiver (as already seen in Chapter 4).

The limitations of language may lead to some confusion when using the term "nonimaging." It indicates that image formation is not required in the design process, although image formation is not necessarily excluded. Therefore, there is no contradiction if a nonimaging concentrator is also an imaging device. The Luneburg lens is an example of a perfect device in imaging and in nonimaging contexts (Luneburg, 1964).

In this chapter we show that the RX concentrator, introduced in Chapter 8, can also be considered a good imaging device. This characteristic, which was studied in Benítez and Miñano (1997), should be added to its simplicity, its good nonimaging performance, and its small thickness to aperture diameter ratio (it can be smaller than 0.5). It is just an example of an SMS design applied to imaging applications, but it is not necessarily the only one, nor is it the best one. Sections 9.2 and 9.3 present the analysis of the image formation capability of RX concentrators with rotational symmetry. All of them have been designed with $n' = 1.5$ and a geometrical concentration of 95% of the theoretical maximum. This 95% implies that $\beta = 77°$, and therefore the numerical aperture of these RXs is 1.46.

9.2   IMAGING PROPERTIES OF THE DESIGN METHOD

The RX is formed of two sequential optical surfaces, one of them refractive (R) and the other reflective (X). The concentrator is thus a single dielectric piece bounded by these two aspheric surfaces. The receiver is embedded in the dielectric with its active side facing the mirror. Figure 9.1 shows the cross section of an RX designed with an acceptance angle α of $\pm3°$.

The design method of the RX concentrator is two-dimensional—that is, only meridian rays are taken into account. Let us consider the design shown in Figure
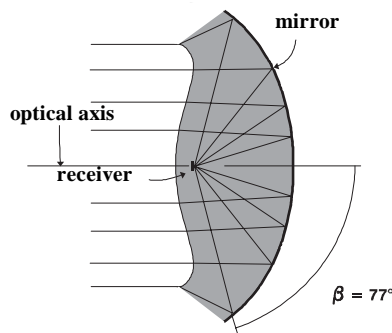
Figure 9.1   Cross section of an RX concentrator designed with $n' = 1.5$, acceptance $\pm3°$, and 95% of maximum concentration. Its numerical aperture is $NA = 1.46$.
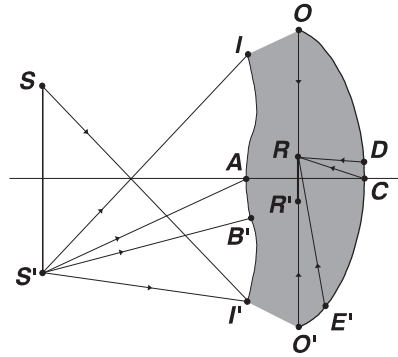
Figure 9.2   Description of the RX design process.

9.2, in which the source is the segment SS' and the receiver is RR'. The concentrator is symmetric with respect to the optical axis. The profiles of the aspherics are calculated with the following conditions:

(a)   The rays emitted from S' and incident on the refractive surface between I' and B' are coupled with the rays reflected between E' and C toward R.

(b)   The rays emitted from S' incident between A and I are coupled to the rays reflected between D and O toward R.

(c)   The rays emitted between S' and S passing through I' are coupled to the rays reflected between O' and E' toward R.

Nevertheless, the design process does not have enough degrees of freedom to guarantee the following:

(a′)   The rays emitted from S' incident between the points B' and A are coupled to the rays reflected on the surface X between the points C and D toward R.

A 2D ray-tracing was made to establish whether (a′) is accomplished, even though it is not guaranteed by the design. In all the RXs considered in this chapter, the maximum deviation of the points of incidence of the rays from the point R is smaller than 6.10^{-6} times the aperture diameter. This means that if an aperture diameter as large as 50 mm is considered, a maximum deviation of 0.3μm is obtained. This value is below the diffraction limit in the visible range even for a numerical aperture as high as that of the RX.

As a conclusion, it can be considered that the condition (a′) is also fulfilled from a practical point of view. The reasons for such a fortunate result are still under study. Observe that, because of the symmetry of the RX, the symmetric rays of those mentioned from (a) to (a′) are also coupled.

In nonimaging optics terms, the fulfillment of conditions (a), (b), (c), and (a′) means that the edge rays of the source and those of the receiver have been coupled (Benítez and Miñano, 1997) and that the radiation emitted by SS' between I and I' has been concentrated on RR' up to the thermodynamic limit. Observe that the conditions (a), (b), and (a′) imply that the meridian rays emitted from S' (and S) are sharply focused onto the point R (and R'). Welford and Winston (1979) provides a proof of the impossibility, in general, of forming a perfect image of a nonaxial object point if skew rays are also considered, although the possibility of

forming a perfect image for only meridian rays is admitted (this conclusion corrects a previous analysis made in Welford and Winston, 1978). The RX is not a formal example for this last statement due to the uncertainty related to condition (a′). Appendix E presents a rigorous example of a system that perfectly forms an image of a nonaxial point for meridian rays.

Condition (c) implies that not all meridian rays impinging on the image points R and R' come from the object points S' and S, respectively. The rays passing through the points I' from the interior of the segment SS' are also focused onto R, and the symmetric rays (passing through I) onto R'. This does not affect the image formation of S' and S' but implies that the points between S and S' are not perfectly imaged onto points of the segment RR'. This assignment of rays corresponds to the solution of the problem of nonimaging design with maximum concentration, and is not adequate from the imaging point of view.

As these rays are located at the entry aperture rim II', the RX concentrators can be truncated to improve the global imaging performance, although this implies a reduction of the concentration. A truncation has been made for all the RX concentrators considered in this chapter. This truncation is such that the new entry aperture diameter is $0.95^{1/2}$ times the initial one. The resulting geometrical concentration is thus 95% of the upper bound.

Note that, although the truncated outer ring of the RX does not form an image, the imaging quality of the RX for maximum concentration (before the truncation) will be only slightly poorer than that of the truncated RX. This is because the 5% of extra luminous power is not enough to degrade the imaging performance.

9.2.1   Conditions on the Wave Aberration Function W

The geometrical performance of an optical system with rotational symmetry can be described in terms of the wave aberration function $W(r, \rho, \psi)$ (Born and Wolf, 1975). The variables r, ρ, and ψ define a ray emitted from the source as shown in Figure 9.3. For the imaging analysis, the source and receiver planes are the object and image planes, respectively. The variable ρ is situated at the entry aperture
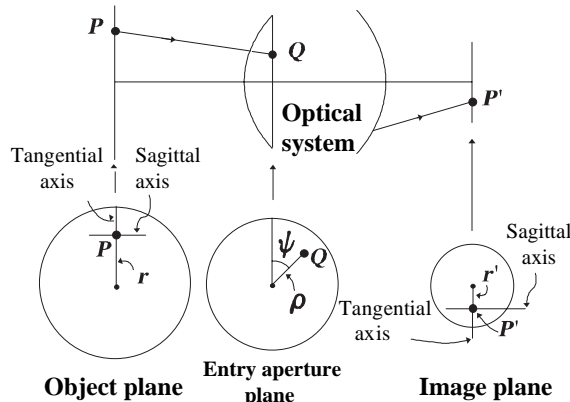

Figure 9.3   Definition of the system of coordinates in the object and image spaces. The entry aperture of the concentrator is the entry pupil of the optical system.

because this is the object-side pupil in all the optical systems considered in this chapter. W is defined as the distance along the ray from the image-side pupil sphere to the actual wave front, multiplied by the index of refraction n' of the image space. From the axial symmetry of the system, it is deduced (Born and Wolf, 1975) that W is only a function of the variables r^2, ρ^2, and $r\rho\cos\psi$.

The function W is usually expressed as a power series expansion. Let us suppose that W is expanded in power series of the single variable r. It can thus be written as

$$W(r, \rho, \psi) = W_o(\rho, \psi) + W_1(\rho, \psi)r + W_2(\rho, \psi)r^2 + \cdots \qquad (9.1)$$

A perfect imaging system—that is, one with no aberrations—would have $W(r, \rho, \psi) \equiv 0$. This is equivalent to $W_K(\rho, \psi) \equiv 0$ for all the subindices K. An optical system is said to attain axial stigmatism if $W_0(\rho, \psi) \equiv 0$. This means that $W(r = 0, \rho, \psi) \equiv 0$, so that the axial object point is imaged perfectly (also, stigmatically). It is also said that the system is free of spherical aberration of all orders. Cartesian ovals are examples of these systems.

An optical system is said to be aplanatic if $W_0(\rho, \psi) \equiv 0$ and $W_1(\rho, \psi) \equiv 0$. This is equivalent to having axial stigmatism with fulfillment of the Abbe sine condition (Born and Wolf, 1975). The aplanatic systems are free of spherical aberration and circular coma of all orders. In conventional designs of high numerical aperture systems, aplanatism is considered a design goal.

In our case, the RX concentrator forms a stigmatic image with meridian rays of the off-axis point S (and S') so that it verifies the following conditions:

$$W(r = r_s, \rho, \psi = 0) \equiv 0 \quad W(r = r_s, \rho, \psi = \pi) \equiv 0 \qquad (9.2)$$

where r_S is the distance of S (and S') to the optical axis.

Because the value of r_S is a design parameter, it can be chosen arbitrarily small. Let us prove that when r_S tends to zero, then the RX becomes aplanatic. We will follow a similar procedure to those given by Schulz (1982; 1985) in the design of aspheric lenses. Another proof, based on the concepts of flow lines of the transmitted bundle, has been given (Benítez and Miñano, 1997).

Observe that when r_S varies, the RX profiles change, and therefore so does the wave aberration function W. The power series Eq. (9.1) can be rewritten with that dependence:

$$W(r, \rho, \psi; r_s) = W_o(\rho, \psi; r_s) + W_1(\rho, \psi; r_s)r + W_2(\rho, \psi; r_s)r^2 + \cdots \qquad (9.3)$$

As W depends only on (r^2, ρ^2, $r\rho\cos\psi$; r_S), the coefficients W_0 and W_1 can be expressed as

$$W_o(\rho, \psi; r_s) = a(\rho; r_s) \quad W_1(\rho, \psi; r_s) = b(\rho; r_s)\cos\psi \qquad (9.4)$$

and grouping the remaining terms in Eq. (9.3), it leads to

$$W(r, \rho, \psi; r_s) = a(\rho; r_s) + b(\rho; r_s)\cos\psi r + c(r, \rho, \psi; r_s)r^2 \qquad (9.5)$$

Using Eq. (9.5), the conditions Eq. (9.2) of the RX are

$$\begin{aligned} a(\rho; r_s) + b(\rho; r_s)r_s + c(r_s, \rho, 0; r_s)r_s^2 = 0 \\ a(\rho; r_s) - b(\rho; r_s)r_s + c(r_s, \rho, \pi; r_s)r_s^2 = 0 \end{aligned} \qquad (9.6)$$

From equation Eq. (9.6) the functions $a(\rho; r_S)$ and $b(\rho; r_S)$ can be found:

$$a(\rho; r_s) = r_s^2 \; \frac{c(r_s, \rho, 0; r_s) + c(r_s, \rho, \pi; r_s)}{2}$$

$$b(\rho; r_s) = r_s \; \frac{c(r_s, \rho, 0; r_s) - c(r_s, \rho, \pi; r_s)}{2}$$

$$(9.7)$$

Note that the function $c(r, \rho, \psi; r_S)$ is continuous in $r = r_S$ for all r_S if we admit that W can be expanded in power series of the variable r. Therefore, it is deduced from Eq. (9.7) that $a(\rho; r_S) \to 0$ and $b(\rho; r_S) \to 0$ when $r_S \to 0$, which implies the vanishing of W_0 and W_1 by Eq. (9.4).

   This proof implies that the design method of the RX converges to the conventional two-asphere design method to attain aplanatism, first published by Schwarzschild (Born and Wolf, 1975) in 1905 for a two-mirror configuration. Later Wassermann and Wolf gave a method for designing the two-refractive aspheric for achieving aplanatism (Wassermann and Wolf, 1949).

9.2.2   Distortion and Focal Length

An optical system forming a perfect image of infinity on the receiver plane is assumed to have a constant ratio $r'/\tan\theta$, where r' is the distance from the image point to the optical axis and θ is the angle of incidence on the aperture with respect to that axis. This ratio is called object-side focal length of the system (Born and Wolf, 1975) and is denoted by f. Thus, the dependence of r' on θ is of the form $r'(\theta) = f\tan\theta$. An optical system is said to present distortion if, even forming a stigmatic image of all points, it has a functional relationship between r' and θ that is other than ideal. From the conservation of the étendue it is easy to deduce that a system that forms a stigmatic image for all of the object points and that attains maximum concentration must verify that $r'(\theta) = C\sin\theta$, for a certain constant C. As $\sin\theta \le \tan\theta$, this distortion is called barrel distortion. For such a system the object-side focal length is defined as the constant $r'/\sin\theta$, so that $f = C$. Figure 9.4 shows the variation of $r'/\sin\theta$ and $r'/\tan\theta$ with respect to θ for the RX concentrator of the figure. In this case r' refers to the midpoint of the geometrical spot for each θ, calculated by means of a ray-tracing in 3D. The function $r'/\sin\theta$ is more constant than $r'/\tan\theta$, so it can be written as

$$f \approx \frac{r'}{\sin\theta} \tag{9.8}$$
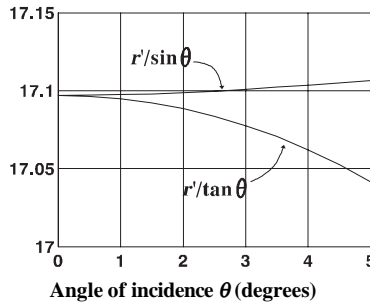
Figure 9.4   Definition of the focal distance in RX concentrators.

9.3   RESULTS

Because of the rotational symmetry, we can restrict the analysis to object points located in a meridian plane without loss of generality. The coordinates are defined as shown in Figure 9.3. The axis containing the object points is called the tangential axis, and the orthogonal axis is the sagittal axis. These coordinates are defined in analogue form in the image plane.

Figure 9.5 shows the geometrical spot diameters as a function of the angle of incidence for four RX concentrators designed for different acceptances. The calculations have been made with a 3D ray-tracing procedure. These RXs have a common entry aperture diameter of 50 mm, which implies an object-side focal length f of 17.1 mm, in accordance with Eq. (9.8). All of these RXs have the same distance between the vertices of the two aspherics (points A and C in Figure 9.2) and the image plane. These distances are $d_A = 21.4$ mm and $d_C = 17.6$ mm. The parameters f, d_A, and d_C, along with the acceptance angle α of the source, determine the RX design univocally (see Chapter 8). The values of α for each RX are $0°$, $1.5°$, $3°$, and $4.5°$. The RX with $\alpha = 0$ is the conventional aplanatic design (see Section 9.2), and is labeled "aplanatic RX" in Figure 9.5. The design method needs a finite α, so this design has actually been calculated with $\alpha = 0.001°$. Figure 9.5a shows the tangential spot diameter ($2\sigma_T$), defined as twice the standard deviation of the tangential coordinate x_T of the points of incidence of the rays on the image plane. The sagittal spot diameter ($2\sigma_S$), represented in Figure 9.5b, is defined in a similar way, but the sagittal coordinate x_S applies. The tangential spot diameter for the incidence angle of design $\theta = \alpha$ (different for each RX) is not null because the skew rays are not perfectly focused. Note that the distance D from the point of incidence of a ray to the midpoint ($<x_T>$, $<x_S>$) verifies $D^2 = (x_T - <x_T>)^2 + (x_S - <x_S>)^2$. Thus, the spot area can be estimated as $\pi(\sigma_T^2 + \sigma_S^2)$.

9.3.1   Calculation of the Modulation Transfer Function (MTF)

The most accepted criterion for evaluating the imaging performance of an optical system is the Modulation Transfer Function (MTF). The MTF can be defined as

Figure 9.5   Tangential and sagittal spot diameters as a function of the angle of incidence for various RX concentrators.

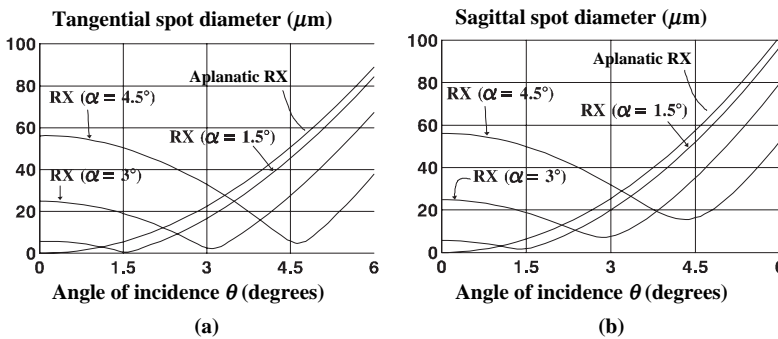the ratio of the contrast of the image to the contrast of the object, and it can be calculated as the normalized modulus of the Fourier Transform of the Point Spread Function (PSF) (Williams and Becklund, 1989). Although in systems with large numerical apertures the effects of polarization of the light may not be negligible (Stamnes, 1986), we shall consider that the Scalar Wave Theory is valid for the analysis. MTF is defined for each point of the object plane, and, as a consequence, the MTF does not inform about the distortion (already seen in Section 9.2.2). In rotational symmetric systems, two points of the object plane at the same distance from the optical axis have the same MTF.

MTF is a function of two variables, the spatial frequencies (tangential frequency f_T and sagittal frequency f_S, in line pairs per millimeter), which are the Fourier variables corresponding to the tangential and sagittal spatial coordinates in the image plane. For an MTF it is usual to represent only the two sections $MTF(f_T, 0)$ and $MTF(0, f_S)$, which are also the Fourier Transforms of the tangential and sagittal line spread functions, respectively (Williams and Becklund, 1989). These two sections of the MTF are referred to as tangential MTF and sagittal MTF. In the case of the axial point, the MTF has rotational symmetry, and thus it is only necessary to represent one section.

A perfect imaging system (i.e., one with no aberrations) without diffraction effects (which is physically impossible) would have an MTF equal to one for all spatial frequencies. The MTF of a perfect imaging system with diffraction effects (usually called perfect MTF) decreases with the increasing spatial frequencies until becoming null beyond the ellipse $(f_T/NA_T)^2 + (f_S/NA_S)^2 = (2/\lambda)^2$, where NA_T and NA_S are the tangential and sagittal image-side numerical apertures of the optical system for the point of analysis, and λ is the wavelength of the light in vacuum (Williams and Becklund, 1989).

These image-side numerical apertures are defined for nonparaxial optical systems and for each point of the image plane. At the axial point the two numerical apertures are equal and can be calculated with the classic formula $n' \sin \beta$, where β is the half-rim angle of illumination of the axial point. According to their definitions (Williams and Becklund, 1989), NA_T and NA_S have been computed for the RX concentrators in this chapter for the corresponding object points inside the source. The maximum deviation of NA_T and NA_S from the value $n' \sin \beta$ is below 2% in all cases, so the $NA = n' \sin \beta$ can be assumed as the numerical aperture of the complete system.

This small variation of NA_T and NA_S is due to the similarity of the cones of rays illuminating the points of the RX image plane. In the case of NA_T, this implies that the meridian rays impinging on the entry aperture with incidence angles within $\pm \alpha$ will reach the image plane with incidence angles approximately within $\pm \beta$ (when this is exact the system is called telecentric). Thus, the conservation of the étendue in two-dimensions implies that

$$r_I \sin \alpha \approx n' r_R \sin \beta \qquad (9.9)$$

where r_I and r_R are the entry aperture radius and receiver radius, respectively. As the maximum geometrical concentration is $(n'/\sin \alpha)^2$, the preceding expression states that $\sin^2 \beta$ is the ratio of $(r_I/r_R)^2$ to this maximum. For example, a fraction of 95% of the maximum concentration implies $\beta \approx \sin^{-1}(0.95^{1/2}) = 77.1°$. Also, the focal length introduced in Eq. (9.8) can be calculated as

$$f \approx \frac{r'}{\sin\theta} \approx \frac{r_I}{NA} \qquad (9.10)$$

When a single value of NA can be assumed for the whole system (which is the case of the RX concentrators in this chapter), the perfect MTF is the same for all object points and, moreover, has rotational symmetry and is null outside the circle $f_T^2 + f_S^2 = (2NA/\lambda)^2$. The value $2NA/\lambda$ is called the spatial cutoff frequency. It means that if a sinusoidal intensity pattern with a frequency greater than the cutoff frequency is located in the object plane, the image will have no contrast. The perfect MTF for low numerical apertures (Williams and Becklund, 1989) depends only on NA/λ, but when the NA is high and no paraxial approximation can be assumed (Barakat and Lev, 1963), the perfect MTF also depends explicitly on the rim angle β.

    As the cutoff frequency for a perfect MTF is proportional to NA, the optimum imaging system must have no aberrations and maximum $NA = n'$ ($\beta = 90°$). An example of such a system is the Luneburg lens (Luneburg, 1964) (which has $n' = 1$ and a nonflat receiver).

    The method used here to calculate the MTF of a concentrator is the one that appears in Smith (1966). This approximate method has been considered acceptable especially when the optical system is far from being diffraction-limited. The results of the method are usually pessimistic. The method consists of the following steps:

(a) Calculate the Geometrical Point Spread Function (GPSF) of the concentrator with a 3D ray tracing.
(b) Compute the Geometrical MTF (GMTF) as the Fourier Transform of the GPSF normalized to its value at $f_T = f_S = 0$.
(c) Multiply this GMTF by the perfect MTF with the same values of NA_T, NA_S, and β.

Steps (b) and (c) can also be calculated as (b′) convolve the GPSF of the concentrator with the perfect Point Spread Function (PSF) with the same NA_T, NA_S, and β (which is the Airy pattern when NA is low), and (c′) apply the Fourier Transform and normalize to its value at $f_T = f_S = 0$.

    Observe that the GPSF changes proportionally with the scale of the system. If the scale is reduced sufficiently, the GPSF approaches the Dirac δ function when compared with the corresponding perfect PSF (which is not scaled). In consequence, the MTF of the system approaches the perfect MTF. This situation means that the optical system is diffraction limited.

    Let us consider once again the RX concentrators discussed in Section 9.3, with a common entry aperture diameter of 50mm and a wavelength of $\lambda = 950$nm.

    The focal length for all of these RXs is 17.1mm. Figure 9.6 shows the MTF for normal incidence and for an incidence angle $\theta = 3°$ of several devices. These devices are the aforementioned RX concentrators and an $f/4.5$ planoconvex spherical lens with the same object-side focal length $f = 17.1$mm and optimum defocusing. The MTF of this lens is shown for comparison purposes. The entry aperture diameter of this lens is $17.1/4.5 = 3.8$mm, and its numerical aperture is $NA = 1/(2f/\#) = 0.11$. Table 9.1 summarizes the data of these systems.

    The image formation quality of the RX designed for $\alpha = 3°$ and that of the lens are similar for normal incidence. Note that if the scale of both devices is
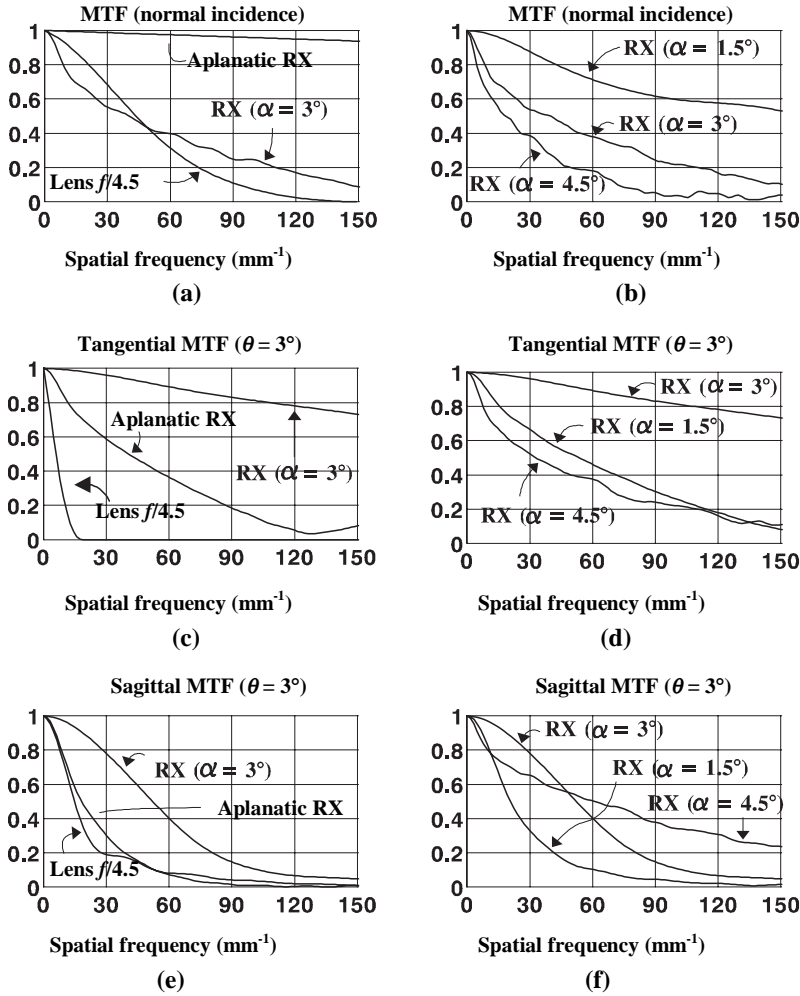
Figure 9.6   MTF for normal incidence and for incidence angle of 3° of a spherical planoconvex lens with $f/4.5$, the aplanatic RX, and the RX designed with acceptance angles of 1.5°, 3°, and 4.5°.

Table 9.1. Comparative Data Between the RX and the Plano-Convex Spherical Lens.

Concentrator	NA	$f/\#$	f (mm)	Entry aperture (mm)
RX	1.46	0.34	17.1	50
Lens	0.11	4.5	17.1	3.8

reduced, then the performance of the RX for normal incidence will be superior to that of the lens because the RX is further than the lens from being diffraction limited. The MTFs of the lens degrade when the incidence angle increases from zero. As will be explained following, that is not the case for the RX.

To compare the imaging performance of different systems and for different object points a single number derived from the MTF is more useful than the full curves. One possible criterion is the equivalent bandwidth, which is defined for the tangential MTF and the sagittal MTF as

$$f_{c,T} = \int_0^\infty \mathrm{MTF}^2(f_T, 0)df_T \quad f_{c,s} = \int_0^\infty \mathrm{MTF}^2(0, f_s)df_s \qquad (9.11)$$

Observe that the axial symmetry of the optical systems implies $f_{c,T} = f_{c,S}$ for the axial object point. As an example, both the RX designed with $\alpha = 3°$ and the $f/4.5$ planoconvex spherical lens (optimally defocused) have $f_{c,T} = f_{c,S} = 32.1\,\mathrm{mm}^{-1}$ for normal incidence.

When a single NA value can be applied to the image points to be studied, then $f_{c,T} = f_{c,S}$ for all the object points in a perfect MTF because the perfect MTF has rotational symmetry in the variables (f_T, f_S). Direct calculations of the equivalent bandwidth of the perfect MTF, which depends only on the values of NA and β, demonstrate that dependence on β is small when NA is fixed. Neglecting this dependence, it is easy to calculate the f_c of the perfect MTF as $0.55 \times NA/\lambda$. For instance, $A = 1.46$ and $\beta = 77°$ (which are the values of the RXs) gives $f_c = 845\,\mathrm{mm}^{-1}$.

Figure 9.7 shows $f_{c,T}$ and $f_{c,S}$ as a function of the angle of incidence θ for the RX concentrators designed with $\alpha = 1.5°$, $3°$, and $4.5°$, the aplanatic RX and their diffraction limit of $845\,\mathrm{mm}^{-1}$. All the RXs attain a maximum of $f_{c,T}$ for $\theta = \alpha$. For $\theta = 0$, the smaller α, the greater $f_{c,T}$. The same behaviour is observed in $f_{c,S}$, although the maximum is obtained at $\theta \approx 2\alpha/3$ and is less abrupt. In the case of the aplanatic RX, the value of $f_{c,T}$ (and $f_{c,S}$) equals the diffraction limit for $\theta = 0$.

Let us call $f_c(\theta)$ the smallest (i.e., the poorest) of $f_{c,T}$ and $f_{c,S}$ for each angle of incidence. It is a global indicator of the imaging performance at the incidence angle θ. In the case of the former RXs, $f_{c,T}(\theta) > f_{c,S}(\theta)$. Thus, $f_c(\theta) = f_{c,S}(\theta)$. This f_c is shown in Figure 9.7b. Let us consider the one-parametric family of RX concentrators designed with the input parameters $f = 17.1\,\mathrm{mm}$, $d_A = 21.4\,\mathrm{mm}$, and $d_B = 17.6\,\mathrm{mm}$ but with variable α. The RX concentrators designed for $\alpha = 0°$, $1.5°$, $3°$, and $4.5°$ belong to this family. For this family of concentrators the function $f_c(\theta, \alpha)$ can be calculated, and its performance can be summarized as
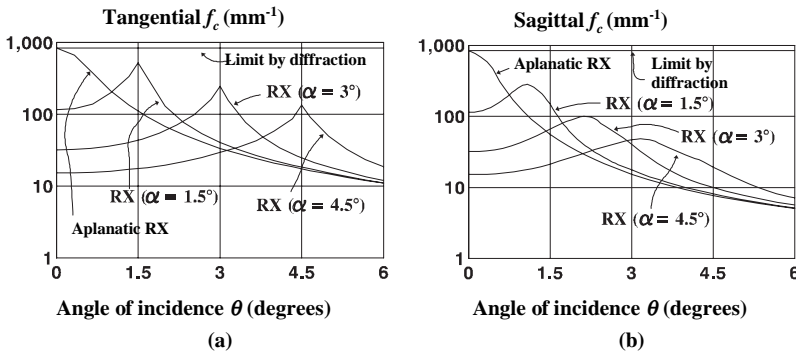

Figure 9.7   (a) Tangential and (b) sagittal equivalent wavelengths as a function of angle of incidence for the selected RX concentrators.
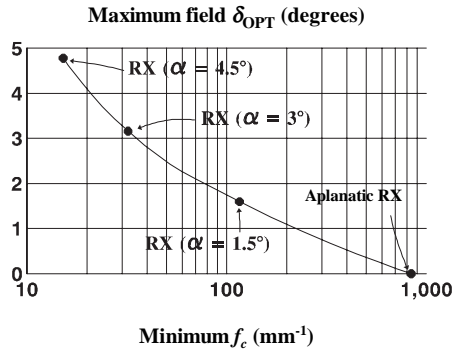
Maximum field δ_{OPT} (degrees)

Minimum f_c (mm^{-1})

Figure 9.8   Angular field of view (semiangle) as a function of the minimum specified equivalent wavelength f_{MIN} for the optimum RX concentrator designed with parameter $d_A = 21.4$ mm and $d_C = 17.6$ mm.

1.   For $\theta = 0$, f_c decreases when α increases.
2.   For a given α, f_c reaches a maximum at $\theta \approx 2\alpha/3$.

The imaging quality required for a certain application can be specified using $f_c(\theta)$, imposing the following condition:

$$f_c(\theta) \geq f_{MIN} \qquad (9.12)$$

where f_{MIN} is the minimum resolution allowed. On inspection of Figure 9.7, it is easy to see that the RX of the one-parametric family verifying the Eq. (9.12) for $0 \leq \theta \leq \delta$, with maximum field of view δ, is precisely that verifying $f_c(\theta = 0) = f_{MIN}$. Let us call δ_{OPT} the maximum field of view corresponding to a given value of f_{MIN}. Figure 9.8 shows δ_{OPT} as a function of f_{MIN}. Each point of this curve corresponds to a different RX concentrator. Observe that this curve is wavelength and scale dependent and that it is associated only with the one-parametric family of RXs considered (only α has been varied, while the other design parameters have been kept constant). Note that for any value of α, the concentrator designed with this α is optimum for $f_{MIN} = f_c(\theta = 0, \alpha)$, achieving the field of view δ_{OPT} given by the curve represented in Figure 9.8. The design for $\alpha = 0$ is optimum when $\delta_{OPT} = 0$. As null fields of view are of no practical interest, this means that the aplanatic RX is not the optimum for any practical case.

Remember that the RX design method implies that the meridian rays of two symmetric off-axis object points (S and S' in Figure 9.2) are focused stigmatically on their image points (R' and R in Figure 9.2, respectively). In the preceding example, this strategy leads to better results than the conventional aplanatism. Schulz described an algorithm to obtain second-order aplanatism that included a similar construction but not with the same strategy: The off-axis points were located as close as possible to the optical axis (Schulz, 1982).

9.3.2   Global Merit Function

The RX designed for $\alpha = 3$ degrees and the $f/4.5$ lens, both with a focal length of 17.1 mm, have a similar image quality for normal incidence but their luminosities are very different. Ignoring optical losses, the ratio of the average irradiances on the receivers of the RX concentrator and of the $f/4.5$ lens is $(50\text{mm}/3.8\text{mm})^2 = 173$.
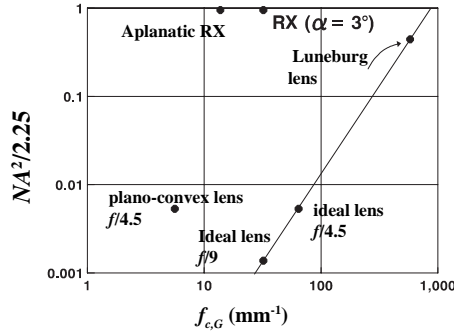
Figure 9.9   The performance of any concentrator for imaging detection is represented by a point in the $f_{c,G} - NA^2$ plane ($f_{c,G}$ is the global equivalent bandwidth of the concentrator in the field of view of $\delta = \pm 3.2$ degrees and indicates the imaging quality. The square of the image-side numerical aperture NA^2 indicates the concentrator luminosity).

In order to compare the global performance for imaging detection of an RX designed for an infinite source with other optical systems, we shall assume that the light is again monochromatic with wavelength $\lambda = 950$ nm, the object-side focal length of the systems to compare is $f = 17.1$ mm, and the field of view is $\delta = \pm 3.2°$. Thus, the detector diameter is fixed at $17.1 \times 2 \times \sin 3.2° = 1.79$ mm. A concentrator may be characterized for imaging detection by two numbers: (1) the global equivalent bandwidth $f_{c,G}$, defined as the minimum value of $f_c(\theta)$ when θ varies in the range $(0, \delta)$, and (2) the square of the numerical aperture NA. The parameter NA^2 quantifies the luminosity of the concentrator (if no optical losses are considered), while $f_{c,G}$ quantifies its imaging quality.

Figure 9.9 shows the $f_{c,G} - NA^2$ plane using logarithmic scales for both axes. The performance for imaging detection of any concentrator is represented by a point ($f_{c,G}$, NA^2) of this plane, which will be called the performance point. The continuous line represents the locus of the performance points of the perfect (or diffraction-limited) imaging devices. These points fulfill $f_{c,G}^2 = (0.55 \times NA/\lambda)^2$. A high $f_{c,G}$ means good imaging quality, and a high NA^2 means high luminosity. The figure also shows the points corresponding to (1) the RX designed for $\alpha = 3°$, (2) the aplanatic RX, (3) the $f/4.5$ planoconvex lens, (4) the $f/4.5$ ideal lens, (5) the $f/9$ ideal lens, and (6) the Luneburg lens.

For this wavelength, this focal length, and this field of view, the $f/4.5$ planoconvex lens has poorer imaging performance and much poorer luminosity than the RX designed for $\alpha = 3°$. This RX has more than double the luminosity of the Luneburg lens (which has $n' = 1$) and an imaging quality similar to the $f/9$ ideal lens (diffraction limited). Finally, the aplanatic RX is as luminous as the RX with $\alpha = 3°$ but with poorer imaging performance.

9.4   NONIMAGING APPLICATIONS

The image formation capability of RX concentrators gives it an interesting property as a nonimaging concentrator: The same concentrator can be used for different acceptance angles (within a certain range) simply by changing the receiver

diameter. Curve A in Figure 9.10 is the angle transmission curve $T(\theta)$ of the RX of Figure 9.1 (designed for $\alpha = 3°$). This curve is very stepped around $\theta = \alpha$, which means that the concentrator's performance is close to ideal.

Curves B to G in the figure are the transmission curves for the same RX using different receiver diameters. Observe that these curves are also very stepped, implying that the nonimaging performance of the concentrator is also good. The calculations of $T(\theta)$ take into account the receiver shadowing but not optical losses.

Figure 9.11 shows the collection efficiency as a function of the resulting semi-acceptance angle, which is calculated for each receiver as the value of θ for which $T(\theta) = 1/2$. In all the cases, the geometrical concentration is 95% of the maximum possible for each acceptance angle.

The RX concentrators achieve concentrations close to the thermodynamic limit, even with receiver diameters that are quite different from that of the design. This feature is not present in the classic nonimaging designs such as the CPC, whose performance suffers if the receiver is changed.
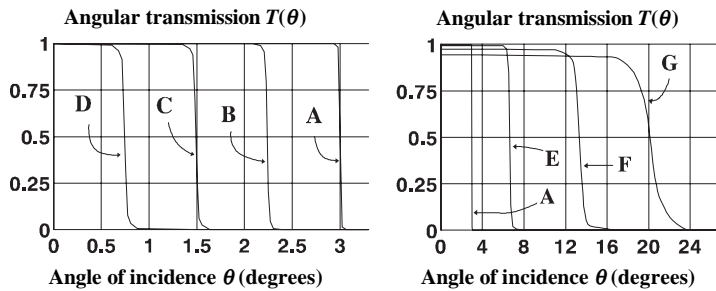
Figure 9.10   Curve A is the angle transmission curve of the concentrator of insert Figure 9.1. Each one of the other transmission curves corresponds to the same concentrator but with a different receiver diameter d. If the entry aperture diameter is 50 mm, then $d(A) = 1.79$ mm, $d(B) = 1.33$ mm, $d(C) = 890 \, \mu m$, $d(D) = 445 \, \mu m$, $d(E) = 3.95$ mm, $d(F) = 7.9$ mm, and $d(G) = 11.86$ mm.
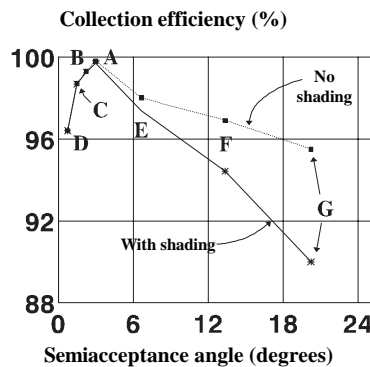
Figure 9.11   Collection efficiency for the RX concentrator of Figure 9.1 for different receiver diameters, as a function of the resulting semiacceptance angle. The upper curve considers a transparent receiver, while the lower one takes into account the shadow losses it introduces. In both cases optical losses have been ignored.

Moreover, if the receiver and the source of an RX design are tailored in any shape (the same for both), the RX still couples very well the rays of the source onto the receiver. This means, for example, that any stepped angular transmission response without rotational symmetry can be achieved with a rotational symmetric RX if the receiver is tailored with the proper shape.

9.5   SMS METHOD AND IMAGING OPTICS

In this chapter, the RX concentrators have been analyzed as imaging devices and have been found to have good image formation capability. For example, for a field of view of $\delta = \pm 3.2°$, an RX with 50 mm aperture diameter and $n' = 1.5$ has an image quality similar to that of an $f/9$ ideal thin lens of 1.9 mm aperture diameter ($\lambda = 950$ nm). This image formation capability is added to its excellent performance as a nonimaging concentrator, which means that its NA is close to the maximum possible ($NA = 1.46$ for the preceding example). As a nonimaging concentrator, the RX is a simple device that achieves concentration levels close to the thermodynamic limit. The combination of the RX's imaging and high concentration properties with its simplicity and compactness means that it is almost unique and makes it an excellent optical device for low-cost, high-sensitivity Focal Plane Array applications.

The strategy used to design the RX (sharp imaging of meridian rays of two off-axis points) suggests that aplanatism (traditionally used in the design of systems with large NA) is not the best solution when a minimum imaging quality is required within a non-null field of view. Moreover, aplanatism has been shown to be a particular case in the RX design procedure when the two off-axis points tend to an axial point.

Observe that if the design method is extended to three aspherics, the axial object point could also be imaged stigmatically. In general, if $2N$ aspherics are designed, the sharp imaging of the meridian rays of $2N$ symmetric off-axis points can be achieved. With $2N + 1$ aspherics the axial point could also be imaged. It seems also possible to design to provide stigmatic imaging of skew rays. Theoretically, designing two aspherics would allow the focusing of a one-parameter bundle of skew rays emitted from an off-axis object point symmetricly with respect to the meridian plane. An example of such a bundle is that formed by the skew rays $\psi = \pm 90°$ in insert Figure 9.3. Analogously, $2N$ aspherics would focus N symmetric skew ray bundles. Combining meridian and skew rays along with using different object points in the design may be more effective and an interesting strategy for imaging optical system design.

REFERENCES

Barakat, R., and Lev, D. (1963). Transfer functions and total illuminance of high numerical aperture systems obeying the sine condition. *J. Opt. Soc. Am.* **53**, 324–332.

Benítez, P., and Miñano, J. C. (1997). Ultrahigh-numerical-aperture imaging concentrator. *J. Opt. Soc. Am. A.* **14**, 1988–1997.

Born, M., and Wolf, E. (1975). *Principles of Optics.* Pergamon, Oxford.

Luneburg, R. K. (1964). *Mathematical Theory of Optics*. University of California, Berkeley.

Schulz, G. (1982). Higher order aplanatism. *Optics Communications* **41**, 315–319.

Schulz, G. (1985). Aberration-free imaging of large fields with thin pencils. *Optica Acta* **32**, 1361–1371.

Smith, W. J. (1966). *Modern Optical Engineering*. McGraw-Hill, New York.

Stamnes, J. J. (1986). *Waves in Focal Regions*. Adam Hilger, Boston.

Wassermann, G. D., and Wolf, E. (1949). On the theory of aplanatic aspheric systems. Proceeds of the Physical Society, B, Vol. LXII, 2–8.

Welford, W. T., and Winston, R. (1978). On the problem of ideal flux concentrators. *J. Opt. Soc. Am.* **68**, 531–534.

Welford, W. T., and Winston, R. (1979). On the problem of ideal flux concentrators: Addendum. *J. Opt. Soc. Am.* **69**, 367.

Williams, C. S., and Becklund, O. A. (1989). *Introduction to the Optical Transfer Function*. Wiley, New York.

10

CONSEQUENCES OF SYMMETRY

Narkis Shatz and John C. Bortz

Science Applications International Corporation, San Diego, CA

10.1   INTRODUCTION

The flux-transfer efficiency of passive nonimaging optical systems—such as lenses, reflectors, and combinations thereof—is limited by the principle of étendue conservation. As a practical matter, many nonimaging optical systems possess a symmetric construction, translational and rotational symmetries being the most common. In this chapter, we find that for such symmetric optical systems a further, more stringent limitation on flux-transfer efficiency is imposed. This performance limitation, which may be severe, can only be overcome by breaking the symmetry of the optical system.

In the geometrical optics approximation, the behavior of a nonimaging optical system can be formulated and studied as a mapping $g : S^{2n} \rightarrow S^{2n}$ from input phase space to output phase space, where S is an even-dimensional piecewise differentiable manifold and n (= 2) is the number of generalized coordinates. The starting point for this formulation is the generalization of Fermat's variational principle, which states that a ray of light propagates through an optical system in such a manner that the time required for it to travel from one point to another is stationary. Applying the Euler-Lagrange necessary condition to Fermat's principle, followed by the Legendre transformation, we obtain a canonical Hamiltonian system that defines a vector field on a symplectic manifold. A vector field on a manifold determines a phase flow—that is, a one-parameter group of diffeomorphisms (transformations that are differentiable and also possess a differentiable inverse). The phase flow of a Hamiltonian vector field on a symplectic manifold preserves the symplectic structure of phase space and consequently is canonical.

The performance limitations imposed on nonimaging optical systems by rotational and translational symmetry are a consequence of Noether's theorem, which relates symmetry to conservation laws (Arnold, 1989). Noether's theorem states that to every one-parameter group of diffeomorphisms of the configuration manifold of a Lagrangian system that preserves the Lagrangian function, there corresponds a first integral of the equations of motion. In Newtonian mechanics, the imposition of rotational and translational holonomic constraints (hence

symmetries) results in the conservation of angular and linear momentum, respectively. In geometrical optics, the imposition of these constraints results in the conservation of quantities known as the rotational and translational skew invariants, which are analogous to, respectively, angular and linear momentum. In this chapter we derive formulas for computing the performance limits of rotationally and translationally symmetric nonimaging optical devices from distributions of the rotational and translational skew invariants of the optical source and the target to which flux is to be transferred.

10.2   ROTATIONAL SYMMETRY

Due to the inherent constraints of image formation, imaging optical systems typically are rotationally symmetric. Many nonimaging optical systems are also rotationally symmetric. In some cases this design choice is suggested by the inherent rotational symmetry of the source and target. However, even when both the source and target are nonaxisymmetric, the optics are often rotationally symmetric due to the ease of designing and manufacturing such components.

We have already seen that the conservation of étendue places an upper limit on the performance of nonimaging optical systems. In this section we explore a further, more stringent performance limitation that is imposed on the important class of nonimaging optical concentrators having rotational symmetry. This limitation can be derived from the fact that the rotational skew invariant of each ray propagating through such a system is conserved. For purposes of brevity, the rotational skew invariant will be referred to as the skew invariant, or simply as the skewness, for the remainder of this section. The performance limitations of translationally symmetric optical systems will be discussed in Section 10.3.

A ray of light emitted by a light source will have a certain value of the skew invariant, or skewness, defined relative to a specified symmetry axis. An optical system having one or more optical surfaces that are symmetric about this axis will not alter the skewness of the ray, no matter how many times the ray is reflected or refracted by the optical system. Since propagation through a uniform medium also maintains skewness, the ray's skewness will be preserved even when it fails to intersect some or all of the optical surfaces, due to the presence of holes and/or apertures in any of these surfaces. This is true even for holes or apertures that are not themselves rotationally symmetric, as long as all the optical surfaces are rotationally symmetric about the specified axis. Rotationally symmetric gradient-index lenses will also preserve the skewness of the ray.

An extended source will emit rays having a range of skewness values. We define the skewness distribution of a source as the differential étendue per unit skewness occupied by all regions of the source that lie within a differential skewness interval centered on the value s. In other words, the skewness distribution is the derivative of étendue with respect to skewness. It should be noted that the skewness distribution is a function of the skewness. The functional form of the skewness distribution obtained for a given light source will depend on the orientation of the symmetry axis relative to the source. The skewness distribution will be zero for skewness values greater than the source's maximum skewness value

or less than its minimum skewness value. Since the skewness of each ray emitted by a source is conserved by an axisymmetric optical system, the source's skewness distribution must also be conserved.

We can also compute the skewness distribution of a desired output light distribution to be produced from the light distribution of the source by means of the nonimaging optical system. We refer to such a desired output light distribution as a target. Since a target is simply a desired distribution of light, it can be treated as just another source. Thus, the formulas derived below for computing the skewness distribution of a source apply equally well for use in computing the skewness distribution of a target. It is worth noting that skewness is conserved by an axisymmetric optical system regardless of whether the light source or target are themselves axisymmetric.

10.2.1   Definition of the Skew Invariant

To define the skew invariant of a light ray, we consider an arbitrary vector $\vec{r}$ linking the optical axis with the light ray. The skew invariant, or skewness, of the ray is defined as

$$s \equiv \vec{r} \cdot \left(\vec{k} \times \hat{a} \right), \tag{10.1}$$

where $\hat{a}$ is a unit vector oriented along the optical axis, and $\vec{k}$ is a vector of magnitude equal to the refractive index, oriented along the ray's propagation direction. The preceding formula for the skewness can easily be simplified to the form

$$s = r_{min} k_t, \tag{10.2}$$

where r_{min} is the magnitude of the shortest vector $\vec{r}_{min}$ connecting the optical axis with the ray, and k_t is the component of $\vec{k}$ in the tangential direction perpendicular to both the optical axis and $\vec{r}_{min}$. It is apparent from Eq. (10.2) that the skewness is always zero for meridional rays, since the vector $\vec{k}$ for such rays always has a tangential component of zero.

10.2.2   Derivation of the Skewness Distribution of an Axisymmetric Surface Emitter

We now derive a formula for the skewness distribution of a source that emits light from an axisymmetric surface, under the assumption that the symmetry axis of the optical system is coincident with that of the source. As depicted in Figure 10.1, we consider a differential source patch of surface area dA. The x,y,z-axes in Figure 10.1 comprise a right-handed Cartesian coordinate system, where the y,z-plane corresponds to the meridional plane, and the x-axis represents the tangential direction. The differential-area patch lies in the x,z-plane with its unit-surface-normal vector $\hat{b}$ pointing in the y-direction. Although the z-axis is coplanar with the symmetry axis, it is not necessarily parallel to the symmetry axis. We assume the differential-area patch is located a distance r from the symmetry axis. Based on the definition of the skew invariant, the skewness of a ray
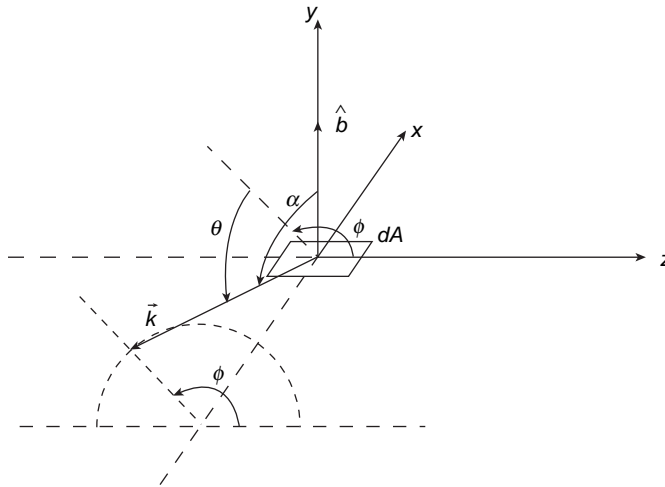
Figure 10.1   Geometry of ray emission from differential-area patch on surface of axisymmetric source.

emitted from this patch at tangential angle θ measured relative to the meridional plane is

$$s = nr \sin(\theta), \tag{10.3}$$

where n is the index of refraction of the material in which the ray is propagating. As shown in Figure 10.1, to completely specify the emission direction of a ray we must specify not only the value of the tangential angle θ but also of the azimuthal angle ϕ.

The differential solid angle can be expressed in the form

$$d\Omega = \cos(\theta)d\theta d\phi. \tag{10.4}$$

The differential étendue can be expressed as

$$d\varepsilon = n^2 d\Omega dA \cos(\alpha), \tag{10.5}$$

where α is the angle between the surface normal of the patch and the ray. It is not difficult to demonstrate that

$$\cos(\alpha) = \cos(\theta)\sin(\phi). \tag{10.6}$$

Substitution of Eqs. (10.4) and (10.6) into Eq. (10.5) produces the following expression for the differential étendue:

$$d\varepsilon = n^2 \cos^2(\theta)\sin(\phi)d\phi dA d\theta. \tag{10.7}$$

Taking the derivative with respect to θ of Eq. (10.3), we find that

$$n \cos(\theta)d\theta = \frac{ds}{r}. \tag{10.8}$$

Again using Eq. (10.3), we find that

$$\cos(\theta) = \sqrt{1-\sin^2(\theta)} = \sqrt{1-\frac{s^2}{n^2r^2}}. \tag{10.9}$$

Substitution of Eqs. (10.8) and (10.9) into Eq. (10.7) gives the result

$$d\varepsilon = \frac{n}{r}\sqrt{1-\frac{s^2}{n^2r^2}}\,\sin(\phi)d\phi dAds. \tag{10.10}$$

Integrating over the angle ϕ and the source surface area, we obtain the skewness distribution as a function of skewness:

$$\frac{d\varepsilon(s)}{ds} = \int_S \int_{\phi_{min}}^{\phi_{max}} \frac{n}{r}\sqrt{1-\frac{s^2}{n^2r^2}}\,\sin(\phi)d\phi dA, \tag{10.11}$$

where S is the region of the source's surface area over which the integrand is defined, and ϕ_{min} and ϕ_{max} are the minimum and maximum values of the azimuthal angle ϕ. We first determine the values of ϕ_{min} and ϕ_{max}. For the case in which the source surface emits into a full 2π steradians of solid angle, the values of ϕ_{min} and ϕ_{max} are simply 0 and 180°, respectively. However, to make our derivation somewhat more general, we consider the case in which the source radiance is zero for all values of the local emission angle α greater than the cutoff angle α_{max}, where $0 < \alpha_{max} \leq 90°$. Combining Eqs. (10.3) and (10.6), we obtain

$$\cos(\alpha) = \sqrt{1-\frac{s^2}{n^2r^2}}\,\sin(\phi). \tag{10.12}$$

Solving for ϕ after substitution of α_{max} for α in the above formula produces the following expressions for the lower and upper limits on ϕ:

$$\phi_{min} = \arcsin\left[\frac{\cos(\alpha_{max})}{\sqrt{1-\dfrac{s^2}{n^2r^2}}}\right] \tag{10.13}$$

and

$$\phi_{max} = 180° - \arcsin\left[\frac{\cos(\alpha_{max})}{\sqrt{1-\dfrac{s^2}{n^2r^2}}}\right]. \tag{10.14}$$

We now determine the region S to be used for the area integration. Solving Eq. (10.12) for r, we obtain

$$r = \frac{|s|}{n\sqrt{1-\dfrac{\cos^2(\alpha)}{\sin^2(\phi)}}}. \tag{10.15}$$

Examining the preceding equation, it is apparent that the minimum allowed value of r for a given value of s will occur when $\alpha = \alpha_{max}$ and $\phi = 90°$. Making these substitutions leads to the following expression for the minimum allowable r-value:

$$r_{min} = \frac{|s|}{n\sin(\alpha_{max})}. \tag{10.16}$$

Thus, for a given value of s, the region S over which the area integration is to be performed corresponds to all source regions satisfying the inequality

$$nr\sin(\alpha_{max}) > |s|. \tag{10.17}$$

This inequality tells us that only source regions having greater than some minimum r-value can contribute rays having a particular skewness value. To take an extreme example, a ray emitted from a surface region at a radius value of zero can only have a skewness value of zero. We now perform the integral over ϕ in Eq. (10.11):

$$\text{Int}_\phi(s,r) \equiv \int_{\phi_{min}}^{\phi_{max}} \sin(\phi)\,\mathrm{d}\phi = 2\cos(\phi_{min}) = 2\sqrt{1 - \frac{\cos^2(\alpha_{max})}{1 - \dfrac{s^2}{n^2 r^2}}}. \tag{10.18}$$

Substitution of Eqs. (10.17) and (10.18) into Eq. (10.11) then gives

$$\frac{\mathrm{d}\varepsilon(s)}{\mathrm{d}s} = \int_{nr\sin(\alpha_{max})>|s|} \frac{2n\sin(\alpha_{max})}{r} \sqrt{\left[1 - \frac{s^2}{n^2 r^2 \sin^2(\alpha_{max})}\right]}\,\mathrm{d}A. \tag{10.19}$$

This formula is a general expression for the skewness distribution of an axisymmetric surface-emitting source. To compute the distribution for a source having a particular shape, the integral over the surface area must be evaluated for the geometry of interest. For the case of a disk-shaped emitter of radius R, Eq. (10.19) can be shown to reduce to the form

$$\frac{\mathrm{d}\varepsilon(s)}{\mathrm{d}s} = \begin{cases} 4\pi nR\sin(\alpha_{max})\left[\sqrt{1-\tilde{s}^2} - \tilde{s}\arccos(\tilde{s})\right], & \text{for } \tilde{s}\leq 1, \\ 0, \text{ otherwise} \end{cases} \tag{10.20}$$

where

$$\tilde{s} \equiv \frac{|s|}{nR\sin(\alpha_{max})}. \tag{10.21}$$

For a spherical emitter of radius R, we have

$$\frac{\mathrm{d}\varepsilon(s)}{\mathrm{d}s} = \begin{cases} 4\pi^2 nR\sin(\alpha_{max})(1-\tilde{s}), & \text{for } \tilde{s} \leq 1 \\ 0, \text{ otherwise} \end{cases}. \tag{10.22}$$

For a cylinder of radius R and axial length H, we have

$$\frac{\mathrm{d}\varepsilon(s)}{\mathrm{d}s} = \begin{cases} 4\pi nH\sin(\alpha_{max})\sqrt{1-\tilde{s}^2}, & \text{for } \tilde{s} \leq 1 \\ 0, \text{ otherwise} \end{cases}. \tag{10.23}$$

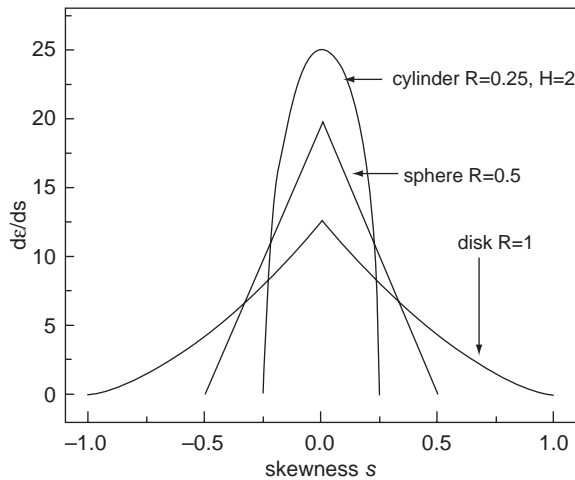Figure 10.2 depicts plots of the skewness distributions corresponding to the disk, sphere, and cylinder.

Figure 10.2   Skewness distributions for disk-shaped, cylindrical, and spherical sources (Ries, Shatz, Bortz, and Spirkl, 1997).

Since each source has the same étendue as the other two, the area under each curve is identical.

10.2.3 Homogeneous Versus Inhomogeneous Sources and Targets

An important consideration in evaluating the performance limits of a nonimaging optical system is the homogeneity of the source and target with which it is to be used. A homogeneous source is one for which the radiance is constant throughout the phase space of the source. An inhomogeneous source, on the other hand, is a source for which radiance varies over its phase space.

Targets may also be classified according to whether they are homogeneous or inhomogeneous. An inhomogeneous target is one for which certain regions of its phase space are considered more important to fill with flux than are other regions. The relative desirability of filling different regions of the phase space of an inhomogeneous target with flux can be expressed by means of a weight function that varies with position over the phase space. Two candidate optical systems that transfer the same amount of flux to the phase space of an inhomogeneous target will *not* in general provide the same level of performance, since one of the two systems will likely transfer a greater amount of flux into regions of the target's phase space having higher-weight values. An example of a design problem in which it would be advantageous to utilize an inhomogeneous target model is the problem of coupling light into an optical fiber having transmission losses that are a function of the position and angle of rays incident on the entrance aperture of the fiber.

A homogeneous target is one for which all regions of its phase space have been assigned the same importance weight. For such targets, performance is maximized simply by transferring as much flux as possible from the source to the target without regard to where in the target's phase space the highest radiance values occur.

10.2.4   Étendue, Efficiency, and Concentration Limits for Homogeneous Sources and Targets

We now derive the upper limits on étendue, efficiency, and concentration imposed by the axisymmetric nature of the optical system for the case in which both the source and target are homogeneous. Efficiency in this case is defined as the total étendue transferred to the target divided by the total étendue of the source. Similarly, concentration is defined as the total étendue transferred to the target divided by the total étendue of the target. Under these definitions, the values of both efficiency and concentration will always be less than or equal to unity. It should be noted that this definition of concentration differs from its other common definition as the ratio of the source's surface area to that of the target.

We consider a homogeneous source and target having skewness distributions $d\varepsilon_1(s)/ds$ and $d\varepsilon_2(s)/ds$, respectively. Since the axisymmetric optical system cannot alter the skewness of any ray, the principle of étendue conservation must apply not only for the integrated source and target étendue but also within each skewness interval. Thus, for all skewness intervals for which $d\varepsilon_1/ds < d\varepsilon_2/ds$, all of the source étendue may be transferred to the target, but some regions of the phase space of the target within those skewness intervals will not be filled with radiation. Within the differential skewness interval between s and $s + ds$, this represents a dilution of the radiation transferred to the target's phase space by a factor of $d\varepsilon_2/d\varepsilon_1$. On the other hand, for those skewness intervals within which $d\varepsilon_1/ds > d\varepsilon_2/ds$, not all of the source étendue may be transferred because it would not fit into the target's phase space available within those skewness intervals. The fraction $1 - d\varepsilon_2/d\varepsilon_1$ of the source étendue is lost within a given differential skewness interval. Figure 10.3 illustrates the mechanisms of dilution and loss due to mismatch of the source and target skewness distributions.

The maximum étendue per unit skewness that can be transferred from the source to the target for any given skewness value s is min $(d\varepsilon_1/ds, d\varepsilon_2/ds)$. The maximum total étendue that can be transferred is obtained by integrating this quantity over all skewness values:

$$\varepsilon_{max} = \int_{-\infty}^{\infty} \min\left[\frac{d\varepsilon_1(s)}{ds}, \frac{d\varepsilon_2(s)}{ds}\right]ds. \tag{10.24}$$
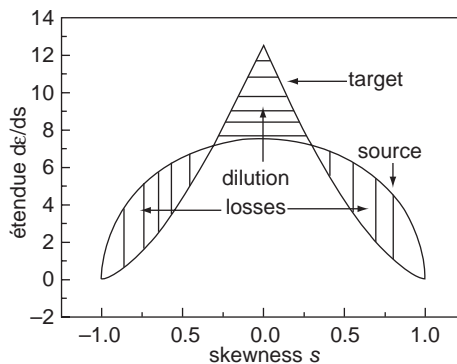

Figure 10.3   Mismatch of the skewness distributions of a source and target, leading to dilution and losses (Ries, Shatz, Bortz, and Spirkl, 1997).

The quantity ε_{max} is the fundamental upper limit imposed by skewness on the performance of axisymmetric nonimaging devices with homogeneous sources and targets.

It is also convenient to define two normalized versions of the transferred étendue, which we refer to as the efficiency and concentration. We define the efficiency η as the transferred étendue divided by the total source étendue. Similarly, we define the concentration C as the transferred étendue divided by the total target étendue. The upper limit on the efficiency is then the ratio of the maximum étendue that can be transferred divided by the étendue of the source:

$$\eta_{max} = \frac{\varepsilon_{max}}{\varepsilon_{src}}, \tag{10.25}$$

where the source étendue is given by the formula

$$\varepsilon_{src} = \int_{-\infty}^{\infty} \frac{d\varepsilon_1(s)}{ds}\,ds. \tag{10.26}$$

The upper limit on the concentration is the ratio of the maximum étendue that can be transferred divided by the étendue of the target:

$$C_{max} = \frac{\varepsilon_{max}}{\varepsilon_{trg}}, \tag{10.27}$$

where the target étendue is given by the formula

$$\varepsilon_{trg} = \int_{-\infty}^{\infty} \frac{d\varepsilon_2(s)}{ds}\,ds. \tag{10.28}$$

Without imposing the constraint that the optical system be axisymmetric, the upper limits on efficiency and concentration are both unity for the étendue-matched case of a source having the same étendue as the target. However, for axisymmetric optics, the upper limits on both efficiency and concentration will be less than unity for the étendue-matched case, except when the source and target happen to have the same skewness distribution. As is apparent from their definitions, efficiency will always equal concentration for the étendue-matched case, when both the source and target are homogeneous. It should emphasized that the performance limits given in Eqs. (10.24), (10.25), and (10.27) are theoretical upper limits, which it will not necessarily be possible to achieve in practice.

By varying the size of the target relative to the source, or vice versa, one can compute the upper limit of achievable efficiency as a function of concentration, $\eta_{max}(C)$, or, equivalently, the upper limit of achievable concentration as a function of efficiency, $C_{max}(\eta)$. The performance of a specific concentrator for a given source and target is represented by a single point on the C, η-plane. Due to the way η and C are defined, this performance point always lies on the line

$$\varepsilon_{src}\eta = \varepsilon_{trg}C. \tag{10.29}$$

When the concentrator is axisymmetric, the distance along this line from the origin to the performance point will always be less than or equal to the distance to the curve $\eta_{max}(C)$ along the same line. Thus, the $\eta_{max}(C)$-curve provides a convenient way to visualize the flux-transfer performance envelope for axisymmetric optics when the relative sizes of a given source and target are varied.

10.2.4.1   Example: Flux Transfer Between Two Homogeneous Disks

In two dimensions, the compound parabolic concentrator (CPC) is known to be an ideal concentrator owing to its ability to transfer 100% of the flux from a rectangular Lambertian source to an equal-étendue rectangular aperture within its acceptance half angle. In three dimensions, however, the CPC is known to be nonideal, because it transfers less than 100% of the flux from a Lambertian disk emitter to an equal-étendue disk target within its acceptance half angle. Furthermore, the 3D CPC is also known to be nonoptimal considering that numerically optimized axisymmetric reflective concentrators have been developed that reduce the performance gap relative to ideality by approximately 15% (Shatz and Bortz, 1995).

It is reasonable to ask whether performance limitations arising from mismatched skewness distributions are responsible for the non-ideality of the three-dimensional CPC. The answer to this question becomes immediately apparent upon examining the formula for the skewness distribution of a disk emitter or target, given in Eqs. (10.20) and (10.21). We see that the skewness distribution of a Lambertian disk emitter is identical to that of an equal-étendue disk target having any acceptance half angle α_{max}. Therefore, the nonideality of the CPC is not due to a skewness mismatch between the source and target.

10.2.4.2   Example: Flux Transfer Between an Open-Ended Cylinder and a Disk

Using the previously derived formulas for the skewness distributions of an open-ended cylinder and a disk, the $\eta_{max}(C)$-curves in Figure 10.4 were obtained for flux transfer from a cylindrical source to a disk target. Cylinders having three different aspect ratios (H/R) were considered: 5, 10, and 20. The curves were generated by computing the upper limit on efficiency and concentration for different values of the ratio of the cylinder's radius to that of the disk.
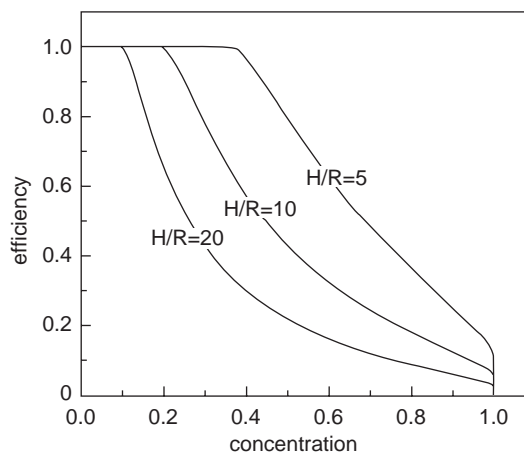

Figure 10.4   Upper limit of efficiency versus concentration for an open-ended cylindrical source and a disk target (Ries, Shatz, Bortz, and Spirkl, 1997). Cylinders having three different values of the aspect ratio H/R were considered: 5, 10, and 20.
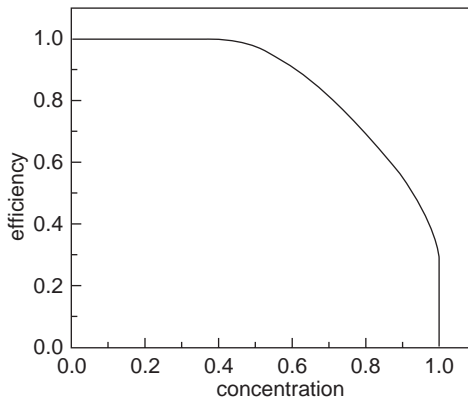
Figure 10.5   Upper limit of efficiency versus concentration for a spherical source and a disk target (Ries, Shatz, Bortz, and Spirkl, 1997).

10.2.4.3   Example: Flux Transfer Between a Sphere and a Disk

A plot of the upper limit of efficiency versus concentration for the case of a spherical source and a disk target is shown in Figure 10.5. The curve was generated by computing the upper limit on efficiency and concentration for different values of the ratio of the sphere's radius to that of the disk. The maximum achievable efficiency and concentration for the equal-étendue case is 75.3%. In the following chapter, an axisymmetric reflective concentrator will be presented that provides efficiency and concentration at the performance limit for the case of flux transfer from a spherical source to a disk target.

10.2.5   Performance Limits for Inhomogeneous Sources and Targets

We now consider the general case in which the target and source are inhomogeneous (Bortz, Shatz, and Ries, 1997). The source is assumed to emit a total flux of $P_{src,tot}$ with radiance distribution $L_{src}(\mathbf{x})$, where the vector $\mathbf{x}$ represents a point in the source's phase space S. For the target, we define the weight function $W_{trg}(\mathbf{x}')$, where the vector $\mathbf{x}'$ represents a point in the target's phase space S'. Our goal in designing a nonimaging system for use with this source and target is to maximize the weighted flux P^{wgt} transferred from the source to the target:

$$P^{wgt} \equiv \int_{\mathbf{x}' \in S'} d\varepsilon(\mathbf{x}') L_{trg}(\mathbf{x}') W_{trg}(\mathbf{x}'), \qquad (10.30)$$

where $L_{trg}(\mathbf{x}')$ is the radiance as a function of position transferred to the target's phase space and $d\varepsilon(\mathbf{x}')$ is the differential element of the target's étendue. We now derive a formula for the maximum achievable value of P^{wgt}, given the simultaneous constraints of étendue and skewness conservation.

We assume that rays emitted by the source have skewness $s_{src}(\mathbf{x})$ as a function of position $\mathbf{x}$ in the source's phase space. The cumulative sorted source étendue per unit skewness at a skewness value of s is given by

$$\frac{\partial \varepsilon_{src}}{\partial s}(L, s) = \int_{\mathbf{x} \in S; L_{src}(\mathbf{x}) > L} d\varepsilon(\mathbf{x}) \delta[s_{src}(\mathbf{x}) - s], \tag{10.31}$$

where L is the source-radiance threshold, $d\varepsilon(\mathbf{x})$ is the differential element of the source's phase-space volume, and the Dirac δ-function ensures that only radiation of skewness s is allowed to contribute to the integral. For purposes of brevity, we now introduce a dot to represent partial differentiation with respect to the skewness:

$$\dot{\varepsilon}_{src}(L, s) \equiv \frac{\partial \varepsilon_{src}}{\partial s}(L, s). \tag{10.32}$$

The function $\dot{\varepsilon}_{src}(L, s)$ represents the amount of source étendue per unit skewness associated with regions of the source's phase space for which the radiance is greater than the radiance threshold value L and for which the skewness equals s. Since increasing the radiance threshold reduces the étendue contributing to the integral in Eq. (10.31), it is apparent that $\dot{\varepsilon}_{src}(L, s)$ is a monotonically decreasing function of L for $L_{min,s}(s) \leq L \leq L_{max,s}(s)$, where $L_{min,s}(s)$ and $L_{max,s}(s)$ are defined as the minimum and maximum values of the source radiance for a given skewness value s. This monotonically decreasing function ranges from a maximum value of $\dot{\varepsilon}_{src}(L_{min,s}(s), s)$ at $L = L_{min,s}(s)$ to a minimum value of zero at $L = L_{max,s}(s)$.

Because $\dot{\varepsilon}_{src}(L, s)$ decreases monotonically with L, it can be inverted to obtain the monotonically decreasing inverse function $L_s(\dot{\varepsilon}_s, s)$, which ranges from $L_{max,s}(s)$ at $\dot{\varepsilon}_{src} = 0$ to $L_{min,s}(s)$ at $\dot{\varepsilon}_{src} = \dot{\varepsilon}_{src}(L_{min,s}(s), s)$. The inverse function $L_s(\dot{\varepsilon}_s, s)$ represents the sorted source radiance as a function of $\dot{\varepsilon}_{src}$ for skewness value s. The sorting process forces the largest radiance value to occur at $\dot{\varepsilon}_{src} = 0$ with decreasing values of radiance as $\dot{\varepsilon}_{src}$ is increased.

For the target, the cumulative sorted étendue per unit skewness takes the form

$$\dot{\varepsilon}_{trg}(W, s) \equiv \frac{\partial \varepsilon_{trg}}{\partial s}(W, s) = \int_{\mathbf{x}' \in S'; W_{trg}(\mathbf{x}') > W} d\varepsilon(\mathbf{x}') \delta[s_{trg}(\mathbf{x}') - s], \tag{10.33}$$

where W is the weight-function threshold and $s_{trg}(\mathbf{x}')$ is the skewness value corresponding to the position $\mathbf{x}'$ in the phase space of the target. The function $\dot{\varepsilon}_{trg}(W, s)$ is a monotonically decreasing function of W for $W_{min,s}(s) \leq W \leq W_{max,s}(s)$, where $W_{min,s}(s)$ and $W_{max,s}(s)$ are defined as the minimum and maximum values of the target weight function for a given skewness value s. This monotonically decreasing function ranges from a maximum value of $\dot{\varepsilon}_{trg}(W_{min,s}(s), s)$ at $W = W_{min,s}(s)$ to a minimum value of zero at $W = W_{max,s}(s)$. It is therefore possible to invert the function $\dot{\varepsilon}_{trg}(W, s)$ to obtain the monotonically decreasing inverse function $W_s(\dot{\varepsilon}_{trg}, s)$, which ranges from $W_{max,s}(s)$ at $\dot{\varepsilon}_{trg} = 0$ to $W_{min,s}(s)$ at $\dot{\varepsilon}_{trg} = \dot{\varepsilon}_{trg}(W_{min,s}(s), s)$. The inverse function $W_s(\dot{\varepsilon}_{trg}, s)$ represents the sorted target weight as a function of the étendue per unit skewness for any given value of the skewness s. The sorting process has forced the largest weight value to occur at $\dot{\varepsilon}_{trg} = 0$ with decreasing weight values as $\dot{\varepsilon}_{trg}$ is increased.

We can now write down an expression for the maximum possible value of P^{wgt} for a rotationally symmetric optical system. This maximum occurs when the source's phase space is transferred to the target's phase space in such a way that the largest source-radiance values preferentially fill the regions of the target's phase space having the largest weight values. Thus, referring to Eq. (10.30), we find that the maximum value of P^{wgt} can be expressed in the form

$$P_{max,s}^{wgt} = \int_{s_{lower}}^{s_{upper}} ds \int_{0}^{\dot{\varepsilon}_{upper}(s)} d\dot{\varepsilon} L_s(\dot{\varepsilon}, s) W_s(\dot{\varepsilon}, s), \tag{10.34}$$

where

$$\dot{\varepsilon}_{upper}(s) \equiv \min[\dot{\varepsilon}_{src}(L_{min,s}(s), s), \dot{\varepsilon}_{trg}(W_{min,s}(s), s)], \tag{10.35}$$

$$s_{lower} \equiv \max[s_{src,min}, s_{trg,min}], \tag{10.36}$$

$$s_{upper} \equiv \min[s_{src,max}, s_{trg,max}], \tag{10.37}$$

and $s_{src,min}$, $s_{src,max}$, $s_{trg,min}$, and $s_{trg,max}$ are the minimum and maximum skewness values for the phase spaces of the source and the target.

Now that we have derived the upper limit on the transferred flux, we are in a position to write down formulas for the upper limits on efficiency and concentration. We define the efficiency η for inhomogeneous sources and targets as the actual weighted flux transferred to the target divided by the weighted flux that would be achieved if all of the source radiation were to be transferred to the region of the target's phase space for which the weight function has its maximum value:

$$\eta \equiv \frac{P^{wgt}}{W_{max} P_{src,tot}}, \tag{10.38}$$

where W_{max} is the maximum target weight value and $P_{src,tot}$ is the total flux emitted by the source. Similarly, the concentration C is defined as the actual weighted flux transferred to the target divided by the weighted target flux level that would be achieved if all of the target's phase space were to be filled with radiation having radiance equal to the maximum radiance level of the source:

$$C \equiv \frac{P^{wgt}}{L_{max} \varepsilon_{trg,tot}^{wgt}}, \tag{10.39}$$

where L_{max} is the maximum radiance value of the source and $\varepsilon_{trg,tot}^{wgt}$ is the total *weighted* target étendue

$$\varepsilon_{trg,tot}^{wgt} = \int_{s_{trg,min}}^{s_{trg,max}} ds \int_{0}^{\varepsilon_{upper}(s)} d\dot{\varepsilon} W_s(\dot{\varepsilon}, s). \tag{10.40}$$

The upper limits on efficiency and concentration are obtained by substitution of $P_{max,s}^{wgt}$ for P^{wgt} in Eqs. (10.38) and (10.39):

$$\eta_{max,s} \equiv \frac{P_{max,s}^{wgt}}{W_{max} P_{src,tot}} \tag{10.41}$$

and

$$C_{max,s} \equiv \frac{P_{max,s}^{wgt}}{L_{max} \varepsilon_{trg,tot}^{wgt}}. \tag{10.42}$$

10.3 TRANSLATIONAL SYMMETRY

Just as the rotational skew invariant is conserved for each ray propagated through an axisymmetric optical system, an analogous quantity—known as the

translational skew invariant—is conserved for each ray propagated through a translationally symmetric optical system. When the translational skewness distribution of a source is different from that of the target, translationally symmetric nonimaging systems are subject to performance limitations analogous to those discussed in the last section for axisymmetric optical systems. (Bortz, Shatz, and Winston, 2001)

10.3.1   The Translational Skew Invariant

We define a translationally symmetric nonimaging device as a nonimaging optical system for which all refractive and reflective optical surfaces have surface normal vectors that are everywhere perpendicular to a single Cartesian coordinate axis, referred to as the symmetry axis. We consider an optical ray incident on a translationally symmetric optical surface, where the symmetry axis is assumed to be the z-axis of a Cartesian x,y,z-coordinate system. The incident ray is assumed to propagate through a medium of refractive index n_0. We define the incident optical direction vector as

$$\vec{S}_0 \equiv n_0 \vec{Q}_0, \tag{10.43}$$

where $\vec{Q}_0$ is a unit vector pointing in the propagation direction of the incident ray. It is well known that the component of the optical direction vector along the symmetry axis is conserved for all rays propagating through a translationally symmetric optical system. This follows from the vector formulation of the laws of reflection and refraction, in which the optical direction vector of a ray reflected or refracted by the optical surface is

$$\vec{S}_1 = \vec{S}_0 + \Gamma \vec{M}_1, \tag{10.44}$$

where $\vec{M}_1$ is the unit vector normal to the surface at the point of intersection of the incident ray with the surface. The formula for the quantity Γ is

$$\Gamma = -2n_0 \cos(I) \tag{10.45}$$

for reflection and

$$\Gamma = -n_0 \cos(I) + n_1 \sqrt{\left(\frac{n_0}{n_1}\right)^2 \cos^2(I) - \left(\frac{n_0}{n_1}\right)^2 + 1} \tag{10.46}$$

for refraction of the ray into a material of refractive index n_1.

In the preceding two formulas for Γ, the quantity I is the angle of incidence of the ray relative to the surface-normal vector. The unit vector $\vec{M}_1$ in the above formulation is, by definition, perpendicular to the z-axis, meaning that its z-component equals zero. It therefore follows from Eq. (10.44) that the incident and reflected (or refracted) optical direction vectors—$\vec{S}_0$ and $\vec{S}_1$—must have the same z-component. Since the z-axis is the symmetry axis, we conclude that the component of the optical direction vector along the symmetry axis is invariant for any ray propagated through a translationally symmetric optical system. We refer to this invariant component of the optical direction vector as the *translational skew invariant*, or simply the *translational skewness*. The fact that a translationally symmetric nonimaging system cannot alter the translational skew invariant places a fundamental limitation on the flux-transfer efficiency achievable by such a system.

The translational skewness is a unitless quantity with absolute value less than or equal to the refractive index. Its sign can be positive or negative depending on the ray direction relative to the z-axis. A ray that is perpendicular to the symmetry axis always has zero translational skewness. It should be noted that the only requirement for the translational skewness to be an invariant quantity is that the optical system be translationally symmetric. In particular, there is no requirement that either the radiation source or the target to which flux is to be transferred be symmetric.

The translational skewness is analogous to the component of linear momentum of a particle along the symmetry axis of a translationally symmetric force field. If we imagine a small unit-mass particle traveling along the ray path with speed equal to the index of refraction, then the linear-momentum component of the particle along the translational symmetry axis is equal to the translational skewness of the ray. Refraction or reflection at a translationally symmetric surface is analogous to imposition of a force having direction perpendicular to the symmetry axis. Such a force alters the momentum vector of the particle while preserving its momentum component perpendicular to the symmetry axis.

If the translational skewness of each individual ray entering an optical system is conserved, then the complete distribution of translational skewness for all emitted rays must also be conserved. This places a much stronger condition on achievable performance than the conservation of phase-space volume (étendue), which is merely a scalar quantity. For purposes of brevity, we use the term *skewness* or *skew invariant* to refer to the translational skew invariant in the remainder of this section.

10.3.2   Upper Limits on Efficiency and Concentration

We now consider a translationally symmetric optical system having the z-axis as its symmetry axis. The skewness of any given ray is defined as the z-component, S_z, of the ray's optical direction vector. Suppose that the optical system transfers flux from some extended optical source to an extended target. We use the notation $d\varepsilon_{src}(S_z)$ to refer to the differential source étendue as a function of skewness contributed by all source flux having skewness values between S_z and $S_z + dS_z$. Similarly, $d\varepsilon_{trg}(S_z)$ refers to the available differential target étendue in the same skewness interval. Since the skewness cannot be altered by the symmetric optical system, it is impossible to transfer étendue from one skewness interval on the source to a different skewness interval on the target. Therefore, the principle of étendue conservation applies simultaneously to each differential skewness interval. This places the following two constraints on the differential skewness $d\varepsilon_{tran}(S_z)$ transferred from the source to the target within any given differential skewness interval:

$$d\varepsilon_{tran}(S_z) \le d\varepsilon_{src}(S_z) \qquad\qquad (10.47)$$

and

$$d\varepsilon_{tran}(S_z) \le d\varepsilon_{trg}(S_z). \qquad\qquad (10.48)$$

Eq. (10.47) is a statement of the fact that the optical system cannot transfer more étendue to the target than is available in the source for any given differential skewness interval. Eq. (10.48) is a statement of the fact that the optical system

cannot transfer to the target more source étendue than can be accommodated by the available target phase-space volume for any given differential skewness interval. It is convenient to combine the preceding two constraints into the single constraint:

$$\frac{\mathrm{d}\varepsilon_{tran}(S_z)}{\mathrm{d}S_z} \le \min\left[\frac{\mathrm{d}\varepsilon_{src}(S_z)}{\mathrm{d}S_z}, \frac{\mathrm{d}\varepsilon_{trg}(S_z)}{\mathrm{d}S_z}\right],$$ (10.49)

where we have also divided the right and left sides by $\mathrm{d}S_z$ in order to express the inequality as a constraint on the derivative of étendue with respect to skewness. We refer to this étendue derivative as a function of skewness as the skewness distribution. The quantities $\mathrm{d}\varepsilon_{src}(S_z)/\mathrm{d}S_z$, $\mathrm{d}\varepsilon_{trg}(S_z)/\mathrm{d}S_z$ and $\mathrm{d}\varepsilon_{tran}(S_z)/\mathrm{d}S_z$ are the skewness distributions of the source, target, and transferred source-to-target flux, respectively. It should be emphasized that the inequality of Eq. (10.49) only applies for translationally symmetric optical systems.

In skewness intervals for which $\mathrm{d}\varepsilon_{src}(S_z)/\mathrm{d}S_z < \mathrm{d}\varepsilon_{trg}(S_z)/\mathrm{d}S_z$, there is insufficient source étendue to completely fill the available target étendue so étendue dilution occurs. In skewness intervals for which $\mathrm{d}\varepsilon_{trg}(S_z)/\mathrm{d}S_z < \mathrm{d}\varepsilon_{src}(S_z)/\mathrm{d}S_z$, there is insufficient target étendue to accommodate the source étendue, so that losses inevitably occur. Only when $\mathrm{d}\varepsilon_{trg}(S_z)/\mathrm{d}S_z = \mathrm{d}\varepsilon_{src}(S_z)/\mathrm{d}S_z$ over the entire range of allowable skewness values does the possibility exist of avoiding both losses and dilution using a translationally symmetric optical system.

The total étendue is obtained by integrating the skewness distribution over all skewness values. It follows from Eq. (10.49) that the upper limit on the total étendue that can be transferred from the source to the target is

$$\varepsilon_{max} = \int_{-\infty}^{\infty} \min\left[\frac{\mathrm{d}\varepsilon_{src}(S_z)}{\mathrm{d}S_z}, \frac{\mathrm{d}\varepsilon_{trg}(S_z)}{\mathrm{d}S_z}\right]\mathrm{d}S_z.$$ (10.50)

Following the terminology of Ries et al. (1997), we define the efficiency η as the ratio of the transferred étendue to the total source étendue. Similarly, the concentration C is defined as the ratio of the transferred étendue to the total target étendue. The maximum efficiency achievable by a translationally symmetric optical system is therefore

$$\eta_{max} = \frac{\varepsilon_{max}}{\varepsilon_{src}},$$ (10.51)

where

$$\varepsilon_{src} = \int_{-\infty}^{\infty} \frac{\mathrm{d}\varepsilon_{src}(S_z)}{\mathrm{d}S_z}\mathrm{d}S_z$$ (10.52)

is the total source étendue. The maximum achievable concentration is given by the formula

$$C_{max} = \frac{\varepsilon_{max}}{\varepsilon_{trg}},$$ (10.53)

where

$$\varepsilon_{trg} = \int_{-\infty}^{\infty} \frac{\mathrm{d}\varepsilon_{trg}(S_z)}{\mathrm{d}S_z}\mathrm{d}S_z$$ (10.54)

is the total target étendue.

10.3.3   Inhomogeneous Sources and Targets

The formalism of the previous subsection provides a means of computing the upper limits on the transfer of étendue from a source to a target. For homogeneous sources and targets, the efficiency and concentration limits also apply to the transfer of flux. For inhomogeneous sources and targets, however, the formulas for the flux-transfer limits are not the same as the formulas for the étendue-transfer limits. The flux-transfer limits, in the case of a translationally symmetric optical system with an inhomogeneous source and target, are obtained by substitution of the translational skewness for the rotational skewness in the flux-transfer-limit formulas for rotationally symmetric optics used with inhomogeneous sources and targets. These formulas are provided in the previous section of this chapter.

10.3.4   Calculation of the Translational Skewness Distribution

We now derive an expression useful for calculating the skewness distributions for certain types of sources and targets. As depicted in Figure 10.6, we consider a differential patch of area d^2A on the surface of an extended source or target. The unit vector normal to the patch's surface is designated as $\hat{b}$. As long as the z-axis remains parallel to the symmetry axis, we are free to reorient the Cartesian x,y,z-coordinate system without loss of generality. It is convenient to place the origin at the center of the patch and to orient the x- and y-axes such that the unit surface-normal vector $\hat{b}$ lies in the y,z-plane. We define $\vec{S}_{xy}$ to be the projection of the optical direction vector $\vec{S}$ onto the x,y-plane. The skewness S_z can be expressed as

$$S_z = n\sin(\theta), \tag{10.55}$$

where n is the refractive index and θ is the angle between the vectors $\vec{S}$ and $\vec{S}_{xy}$. The differential étendue is given by

$$\mathrm{d}^4\varepsilon = n^2\mathrm{d}^2\Omega\mathrm{d}^2A\cos(\alpha), \tag{10.56}$$
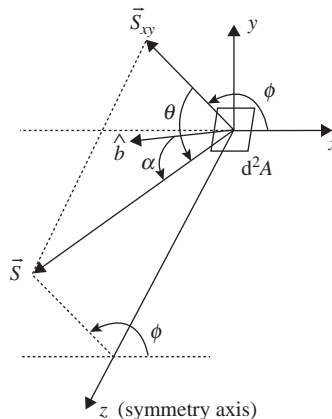

Figure 10.6   Geometry for calculation of the translational skewness distribution of a source or target.

where α is the angle between the vectors $\hat{b}$ and $\vec{S}$. The quantity $d^2\Omega$ in Eq. (10.56) is the differential solid angle. This can be written in the form

$$d^2\Omega = \cos(\theta)d\theta d\phi, \qquad\qquad (10.57)$$

where ϕ is the angle between the vector $\vec{S}_{xy}$ and the x-axis. The cosine of α can be computed from the dot product of $\vec{S}$ and $\hat{b}$:

$$\cos(\alpha) = \frac{|\vec{S}\cdot\hat{b}|}{n}. \qquad\qquad (10.58)$$

Using the fact that $\hat{b}$ lies in the y,z-plane, we can evaluate the dot product in Eq. (10.58) to obtain the formula

$$\cos(\alpha) = \cos(\theta)\sin(\phi)\cos(\beta) + \sin(\theta)\sin(\beta), \qquad\qquad (10.59)$$

where β is the angle between $\hat{b}$ and the y-axis. The sign convention for β is that it always has the same sign as the z-component of $\hat{b}$. Substitution of Eqs. (10.57) and (10.59) into Eq. (10.56) produces the result

$$d^4\varepsilon = n^2[\cos^2(\theta)\sin(\phi)\cos(\beta) + \cos(\theta)\sin(\theta)\sin(\beta)]d\theta d\phi d^2A. \qquad\qquad (10.60)$$

Taking the derivative of Eq. (10.55), we find that

$$dS_z = n\cos(\theta)d\theta. \qquad\qquad (10.61)$$

Since $\cos(\theta) = \sqrt{1-\sin^2(\theta)}$, Eq. (10.55) allows us to express the cosine of θ in terms of the skewness:

$$\cos(\theta) = \sqrt{1 - \frac{S_z^2}{n^2}}. \qquad\qquad (10.62)$$

The θ-dependence in Eq. (10.60) can be converted into S_z-dependence through substitution of Eqs. (10.55), (10.61), and (10.62):

$$\frac{d^4\varepsilon}{dS_z} = \left[\sqrt{n^2 - S_z^2}\,\sin(\phi)\cos(\beta) + S_z\sin(\beta)\right]d\phi d^2A. \qquad\qquad (10.63)$$

Integrating over the angle ϕ and the total surface area, we obtain the desired expression for the skewness distribution as a function of the skewness:

$$\frac{d\varepsilon(S_z)}{dS_z} = \iint\left[\sqrt{n^2 - S_z^2}\,\sin(\phi)\cos(\beta) + S_z\sin(\beta)\right]d\phi d^2A, \qquad\qquad (10.64)$$

where the integrals are taken over all ϕ-values and regions of surface area within the phase-space volume. The limits on the ϕ-integral of Eq. (10.64) can be a function of both S_z and position on the surface. In addition, the angle β can be a function of position on the surface.

For the important special case of a translationally symmetric source or target having the same symmetry axis as the optical system, the skewness distribution can be expressed in a particularly simple form. Noting that α is the angle between the direction vector and the surface normal, we find that Eq. (10.56) can be rewritten as

$$d^4\varepsilon = n^2 d^2\Omega_p d^2A, \qquad\qquad (10.65)$$

where Ω_p is the projected solid angle:

$$\Omega_p = \Omega \cos(\alpha). \tag{10.66}$$

For a translationally symmetric source or target, the surface-normal vector $\vec{b}$ coincides with the y-axis so that the projected solid angle becomes a projection of the solid angle onto the x,z-plane. This allows us to express the differential projected solid angle in terms of the x and z-components of the optical direction vector:

$$\mathrm{d}^2\Omega_p = \frac{\mathrm{d}S_x \mathrm{d}S_z}{n^2}. \tag{10.67}$$

The factor of n^2 in the denominator of Eq. (10.67) is required to properly normalize the projected solid angle since the optical direction vector is of length n. From Eqs. (10.65) and (10.67) we obtain the desired formula for the skewness distribution of a translationally symmetric source or target having the same symmetry axis as the optical system:

$$\frac{\mathrm{d}\varepsilon(S_z)}{\mathrm{d}S_z} = \iint \mathrm{d}S_x \mathrm{d}^2 A. \tag{10.68}$$

For the general case of non-translationally-symmetric sources and targets, it is not difficult to demonstrate that Eq. (10.64) can be rewritten in the form

$$\frac{\mathrm{d}\varepsilon(S_z)}{\mathrm{d}S_z} = \int \left[\cos(\beta) \int \mathrm{d}S_x + S_z \sin(\beta) \int \mathrm{d}\phi \right] \mathrm{d}^2 A, \tag{10.69}$$

which reduces to Eq. (10.68) upon setting β equal to zero.

10.3.5 Examples of Translational Skewness Distributions

We now apply the formulas derived in the last section to determine the skewness distributions for four specific sources. In each case we assume the source is Lambertian over a particular angular region and zero outside that region. In addition, we assume the angular distribution of emitted radiation is independent of position over the surface of the source and that the source is translationally symmetric, with symmetry axis coincident with that of the optical system. To prevent emitted source flux from being reabsorbed by the source, we require that the emitting surface be convex. Each skewness distribution derived below also applies in the case of an analogous translationally symmetric convex target, as long as the allowed angular region of absorption for the target is identical to the angular region of emission of the source for which the skewness distribution was derived.

10.3.5.1 Lambertian Source

As our first example, we derive the skewness distribution for the case of a source that is Lambertian for all possible emission directions. Due to the translational symmetry of the source, the surface-normal orientation angle β equals zero over the entire surface area. The limits of the ϕ-integral in Eq. (10.64) are 0 to π radians for all allowed values of S_z and all positions on the surface. Eq. (10.64) therefore reduces to

$$\frac{\mathrm{d}\varepsilon(S_z)}{\mathrm{d}S_z} = \begin{cases} 2A_{tot}\sqrt{n^2 - S_z^2} & \text{when } |S_z| \le n, \\ 0 & \text{otherwise} \end{cases}, \tag{10.70}$$
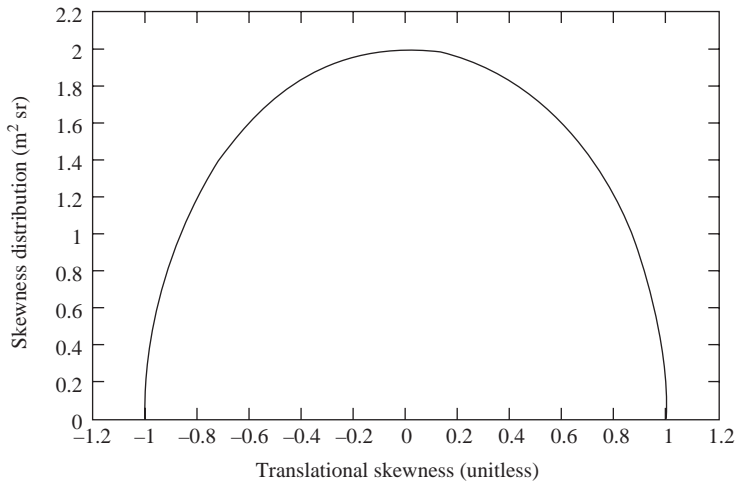
Figure 10.7   Translational skewness distribution of a Lambertian source with 1m^2 surface area in a medium of unit refractive index.

where A_{tot} is the total surface area of the source. A plot of this skewness distribution is provided in Figure 10.7. It is easily verified by integrating the preceding skewness distribution over all allowed skewness values that the total étendue for this source is

$$\varepsilon_{tot} = \pi n^2 A_{tot}. \tag{10.71}$$

The skewness distribution of Eq. (10.70) can also be easily derived using Eq. (10.68) rather than Eq. (10.64). We first note that the projection onto the x,z-plane of the allowed optical direction vectors for this source is a circular region of radius n centered on the origin. Due to the circular geometry, the integration range for the S_x-integral in Eq. (10.68) will be from $-\sqrt{n^2 - S_z^2}$ to $\sqrt{n^2 + S_z^2}$, for S_z-values such that $|S_z| \leq n$. Thus, the S_x-integral equals $2\sqrt{n^2 - S_z^2}$ for $|S_z| \leq n$. For $n < |S_z|$ the S_x-integral equals zero, since these S_z-values are outside of the allowed circular region. Multiplying the S_x-integral by the total area-integral contribution of A_{tot}, we obtain Eq. (10.70).

10.3.5.2   Fixed Longitudinal Cutoffs Parallel to Symmetry Axis

As our next example, we consider a source that is Lambertian within a fixed range of ϕ-values centered on the y,z-plane. As is apparent from Figure 10.6, the angle ϕ is analogous to longitude. We therefore refer to this type of source as one having fixed longitudinal angular cutoffs. In addition, since the set of optical direction vectors $\vec{S}$ having the same ϕ-value all lie in a plane passing through the z-axis, we say that these longitudinal angular cutoffs are parallel to the symmetry axis.

The ϕ-integration in Eq. (10.64) must now be performed over the range

$$\frac{\pi}{2} - \phi_0 \leq \phi \leq \frac{\pi}{2} + \phi_0, \tag{10.72}$$

where ϕ_0 is the angular half width of the emitted flux distribution. Setting β equal to zero in Eq. (10.64) and using the above range of ϕ-values, we immediately obtain

$$\frac{d\varepsilon(S_z)}{dS_z} = \begin{cases} 2A_{tot}\sin(\phi_0)\sqrt{n^2 - S_z^2} & \text{when } |S_z| \le n \\ 0 & \text{otherwise} \end{cases}. \tag{10.73}$$

The total étendue in this case is

$$\varepsilon_{tot} = \pi n^2 A_{tot} \sin(\phi_0). \tag{10.74}$$

Comparing Eqs. (10.70) and (10.73), we find that the skewness distributions for the Lambertian case and the fixed-longitudinal-angular-cutoff case are identical as long as the following two conditions are met:

- The indices of refraction for the two cases are equal.
- The surface area for the fixed-longitudinal-cutoff case divided by the surface area for the Lambertian case equals $1/\sin(\phi_0)$.

   The fact that the two skewness distributions match when these two conditions are satisfied is consistent with the well-known fact that both the translationally symmetric compound parabolic concentrator (CPC) and the translationally symmetric involute CPC are ideal concentrators capable of transferring all of the source phase space to the phase-space volume of an equal-étendue target when the source has fixed longitudinal angular cutoffs. When the indices of refraction for the source and target are different, there is an inherent skewness mismatch between the Lambertian and fixed-longitudinal-angular-cutoff cases, meaning that a translationally symmetric concentrator cannot be ideal in this case.

10.3.5.3   Fixed Latitudinal Cutoffs Parallel to Symmetry Axis

As our third example of a skewness distribution, we consider another case of a source that is Lambertian within a fixed angular band. However, we now consider emission over a particular band of latitudinal angles rather than the longitudinal band analyzed in the previous subsection. We define Θ as the latitudinal angle between the optical direction vector and the y,z-plane, where the y,z-plane is analogous to the equatorial plane. The sign of Θ is assumed to equal the sign of the x-component of the optical direction vector. We assume that the angular band is centered on the y,z-plane with half width Θ_0. Because the set of optical direction vectors $\vec{S}$ having the same Θ-value all lie in a plane parallel to the z-axis, we say that the latitudinal angular cutoffs are parallel to the symmetry axis.

   Due to the assumed latitudinal angular cutoffs, the source emits radiation only for $\vec{S}$-values within the angular interval

$$|\Theta| \le \Theta_0. \tag{10.75}$$

The shaded region in Figure 10.8 depicts the projection onto the x,z-plane of the allowed optical direction vectors within this angular band. The two straight edges of this projected angular band intersect the circle of radius n at skewness values of $\pm S_{z,bnd}$, where

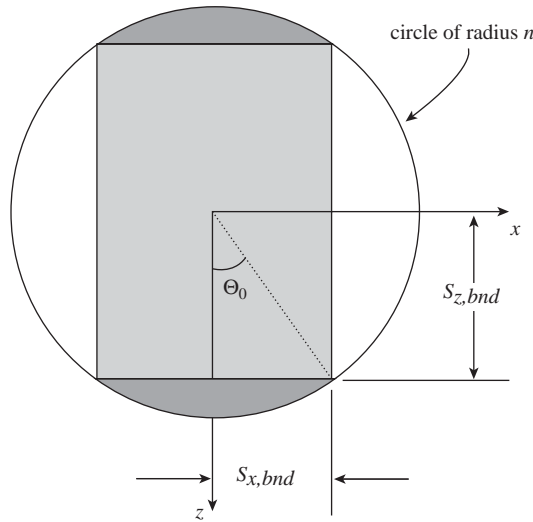$$S_{z,bnd} = n\cos(\Theta_0). \tag{10.76}$$

Figure 10.8   Projection onto the x,z-plane of the allowed optical direction vectors of a source having fixed latitudinal angular cutoffs parallel to the symmetry axis and centered on the y,z-plane.

The two darker-shaded regions in Figure 10.8 represent the subset of the projected angular band for which the skewness values have magnitude greater than or equal to $S_{z,bnd}$. The upper and lower limits of the allowed values of S_x in these two regions occur on the perimeter of the circle. From the equation of this circle, we find that these upper and lower limits as a function of S_z are $\pm\sqrt{n^2 - S_z^2}$. Incorporating these limits into the S_x-integral of Eq. (10.68), we find that

$$\frac{d\varepsilon(S_z)}{dS_z} = 2A_{tot}\sqrt{n^2 - S_z^2} \quad \text{when } S_{z,bnd} \leq |S_z| \leq n. \qquad (10.77)$$

The rectangular, lighter-shaded region of the projected angular band of Figure 10.8 represents the portion of the band having skewness values of magnitude less than $S_{z,bnd}$. The upper and lower S_x-limits within this rectangular region are simply $\pm S_{x,bnd}$, where $S_{x,bnd}$ is the half width of the rectangle in the x-direction

$$S_{x,bnd} = \sqrt{n^2 - S_{z,bnd}^2}. \qquad (10.78)$$

Using these limits, Eq. (10.68) gives the formula

$$\frac{d\varepsilon(S_z)}{dS_z} = 2A_{tot}\sqrt{n^2 - S_{z,bnd}^2} \quad \text{when } |S_z| < S_{z,bnd}. \qquad (10.79)$$

Combining Eqs. (10.77) and (10.79), along with the fact that the skewness distribution must equal zero for skewness values greater than n, we find that the skewness distribution over the entire range of skewness values is of the form

$$\frac{d\varepsilon(S_z)}{dS_z} = \begin{cases} 2A_{tot}\sqrt{n^2 - S_{z,bnd}^2} & \text{when } |S_z| < S_{z,bnd} \\ 2A_{tot}\sqrt{n^2 - S_z^2} & \text{when } S_{z,bnd} \leq |S_z| \leq n. \\ 0 & \text{otherwise} \end{cases} \qquad (10.80)$$
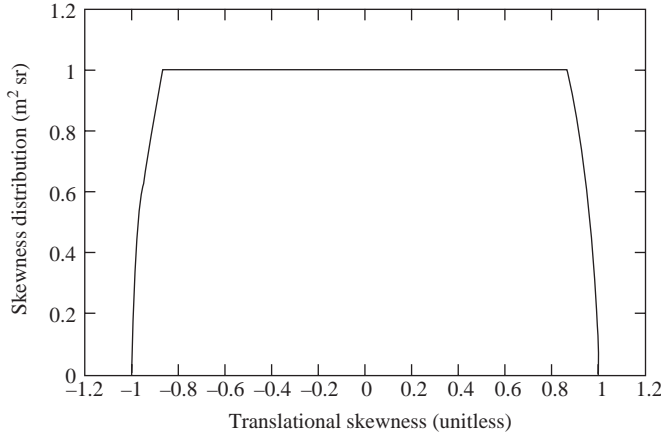
Figure 10.9 Translational skewness distribution in a unit-refractive-index medium for a 1-m^2 source having fixed latitudinal angular cutoffs parallel to the symmetry axis and centered on the y,z-plane. An angular half width of $\Theta_0 = 30°$ is assumed.

A plot of this skewness distribution is provided in Figure 10.9. Integrating Eq. (10.80) over all S_z-values, we find that the total étendue is

$$\varepsilon_{tot} = n^2 A_{tot}[2\Theta_0 + \sin(2\Theta_0)]. \tag{10.81}$$

For skewness values of magnitude greater than $S_{z,bnd}$, the skewness distribution of Eq. (10.80) has the same functional form as the Lambertian case discussed in Section 10.3.5.1. However, for skewness values of lower magnitude, the skewness distribution has a constant value, and therefore fails to match the $\sqrt{n^2 - S_z^2}$-dependence of a Lambertian source or target. This is consistent with the fact that translationally symmetric concentrators are inherently incapable of transferring all of the source's phase space to the phase-space volume of an equal-étendue Lambertian target, when the source has fixed latitudinal angular cutoffs.

10.3.5.4 Fixed Longitudinal Cutoffs Parallel to Symmetry Axis with Orthogonal Fixed Latitudinal Cutoffs

As a final example, we consider a slight modification of the case discussed in Section 10.3.5.2. We retain the longitudinal cutoffs of that section, which are parallel to the symmetry axis and centered on the y,z-plane. In addition to these angular cutoffs, we impose the restriction that emission is only allowed for optical direction vectors having θ-values satisfying

$$|\theta| \leq \theta_0, \tag{10.82}$$

where θ_0 is the latitudinal angular half width. The latitude is now defined as the angle θ between the optical direction vector and the x,y-plane, as depicted in Figure 10.6. Using Eq. (10.55), the inequality of Eq. (10.82) becomes

$$|S_z| \leq n \sin(\theta_0). \tag{10.83}$$

For skewness values satisfying this inequality, the skewness distribution is identical to that of Eq. (10.73). However, for skewness values outside the interval of
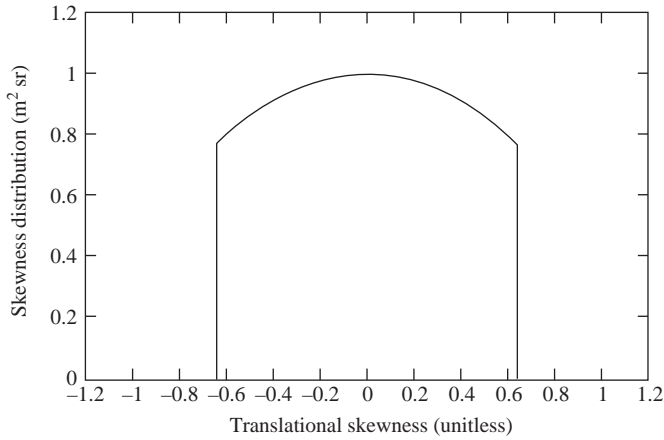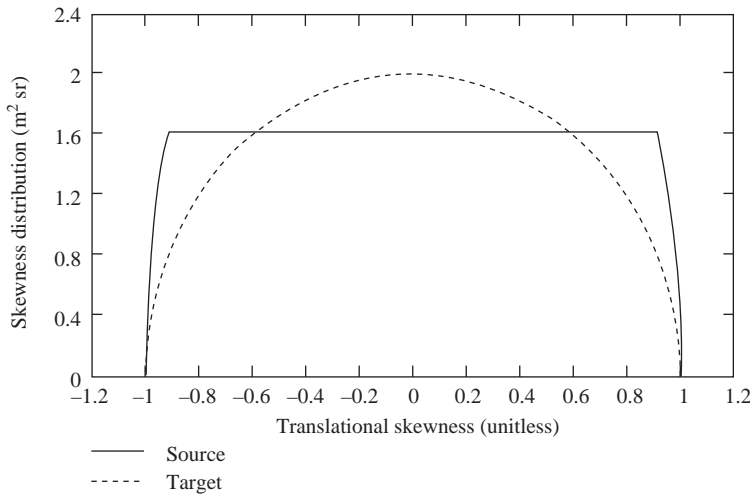
Figure 10.10   Translational skewness distribution in a unit-refractive-index medium for a unit-area source having fixed longitudinal angular cutoffs parallel to the symmetry axis and centered on the y,z-plane, combined with orthogonal fixed latitudinal angular cutoffs centered on the x,y-plane. The longitudinal and latitudinal angular half widths are $\phi_0 = 30°$ and $\theta_0 = 40°$, respectively.

Eq. (10.83), the skewness distribution is identically zero. The distribution for all skewness values is therefore

$$\frac{\mathrm{d}\varepsilon(S_z)}{\mathrm{d}S_z} = \begin{cases} 2A_{tot}\sin(\phi_0)\sqrt{n^2 - S_z^2} & \text{when } |S_z| \le n\sin(\theta_0) \\ 0 & \text{otherwise} \end{cases}. \qquad (10.84)$$

A plot of this distribution is provided in Figure 10.10. Integrating Eq. (10.84) over the full range of skewness values, we find that the total étendue for this source is

$$\varepsilon_{tot} = n^2 A_{tot}\sin(\phi_0)[2\theta_0 + \sin(2\theta_0)]. \qquad (10.85)$$

10.3.6   Examples of Computed Efficiency and Concentration Limits

Once the skewness distributions of a source and target have been computed, Eqs. (10.51) and (10.53) can be used to compute the upper limits on efficiency and concentration achievable for that combination of source and target, when the optical system is translationally symmetric. We now compute the performance limits for flux transfer to a translationally symmetric Lambertian target from translationally symmetric sources having fixed angular cutoffs.

10.3.6.1   Flux Transfer to Lambertian Target from Source Having Fixed Latitudinal Cutoffs Parallel to Symmetry Axis

For this example, we consider a source of the type analyzed in Section 10.3.5.3. We set the latitudinal half-angle, Θ_0, equal to 23.45° with refractive index $n = 1$ for both the source and the target. With these choices, the case is representative of an east-west-oriented non-tracking solar concentrator located at the equator and

operating every day of the year from sunrise to sunset. The formulas for the target and source étendue are given in Eqs. (10.71) and (10.81), respectively. We assume the total surface area of the target is

$$A_{trg} = 1\,\text{m}^2. \tag{10.86}$$

Eq. (10.71) then allows us to calculate the target étendue:

$$\varepsilon_{trg} = \pi\,\text{m}^2\text{sr}. \tag{10.87}$$

When the source étendue is set equal to the target étendue, Eqs. (10.81) and (10.87) give the following value for the total surface area of the source:

$$A_{src} = 2.029\,\text{m}^2. \tag{10.88}$$

For this equal-étendue case, Figure 10.11 depicts the skewness distributions of Eqs. (10.70) and (10.73). The upper limit on étendue that can be transferred from the source to the target by a translationally symmetric concentrator is computed using the integral of Eq. (10.50). For the equal-étendue case, this upper limit on transferred étendue turns out to be

$$\varepsilon_{max} = 2.832\ \text{m}^2\text{sr}. \tag{10.89}$$

With reference to Figure 10.11, this étendue limit is equal to that portion of the étendue region contained under the source's skewness distribution that is intersected by the étendue region contained under the target's skewness distribution. As indicated by Eqs. (10.51) and (10.53), the upper limits on efficiency and concentration are computed by dividing ε_{max} by the total source and target étendue, respectively. For the equal-étendue case, the efficiency and concentration limits have the same value:

$$\eta_{max} = C_{max} = 90.14\%. \tag{10.90}$$

Figure 10.11 Translational skewness distributions in a unit-refractive-index medium for a Lambertian target and a source having fixed latitudinal cutoffs parallel to the symmetry axis. The angular half width of the source is $\Theta_0 = 23.45°$. The source étendue equals that of the target.

Using Eqs. (10.51) and (10.53), we can also compute the efficiency and concentration limits for source-to-target étendue ratios other than unity. The resulting efficiency limit is plotted as a function of the concentration limit in Figure 10.12. The diamond-shaped marker on this plot indicates the efficiency and concentration limits for the equal-étendue case. It is also useful to plot the efficiency and concentration limits as a function of the source-to-target étendue ratio itself, as shown in Figure 10.13. Note that the equal-étendue case corresponds to the crossing point of the two curves in this plot.
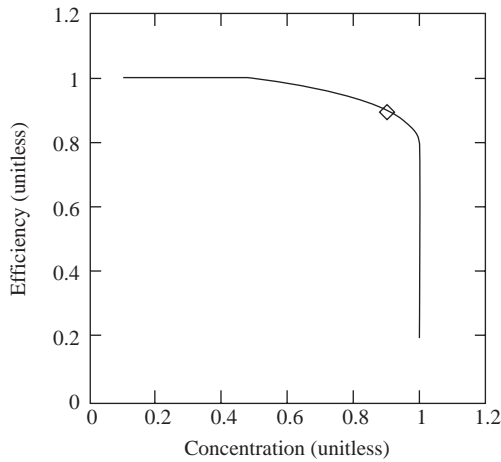
Figure 10.12   Plot of the efficiency limit as a function of the concentration limit for translationally symmetric nonimaging devices that transfer flux to a Lambertian target from a source having fixed latitudinal cutoffs parallel to the symmetry axis. The angular half width of the source is $\Theta_0 = 23.45°$. The diamond-shaped marker indicates the performance limit for the equal-étendue case.
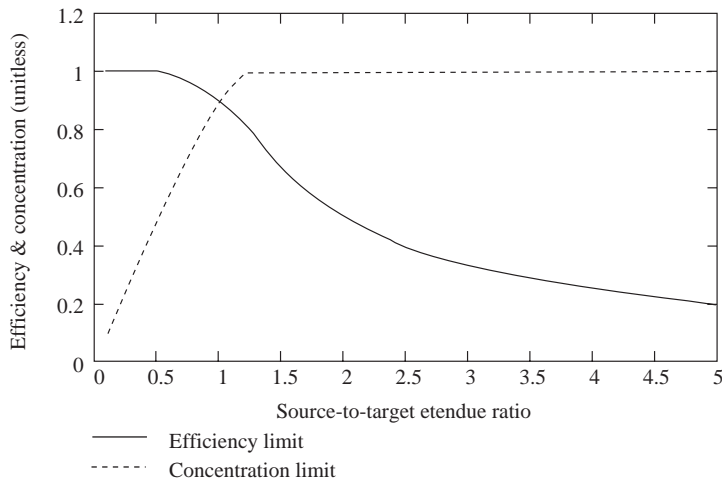
Figure 10.13   Plot of the efficiency and concentration limits as a function of the source-to-target étendue ratio for translationally symmetric nonimaging devices that transfer flux to a Lambertian target from a source having fixed latitudinal cutoffs parallel to the symmetry axis. The angular half width of the source is $\Theta_0 = 23.45°$. The equal-étendue case corresponds to the crossing point of the two curves.

10.3.6.2 Flux Transfer to Lambertian Target from Source Having Fixed Longitudinal Cutoffs Parallel to Symmetry Axis with Orthogonal Fixed Latitudinal Cutoffs

As a second and final example, we now consider a source of the type analyzed in Section 10.3.5.4. We set the latitudinal half angle equal to the latitudinal half width of incident solar radiation for non-tracking solar concentrators:

$$\theta_0 = 23.45°. \tag{10.91}$$

The refractive index is assumed to be unity for both the source and the target. With these choices, this case is representative of a north-south-oriented non-tracking solar concentrator. For such a concentrator, the appropriate value of the longitudinal half angle ϕ_0 depends on the daily hours of operation. We assume that

$$\phi_0 = 90°, \tag{10.92}$$

which corresponds to daily operation from sunrise to sunset at an equatorial location. As is apparent from Eq. (10.84), the choice of ϕ_0 affects only the vertical scaling of the skewness distribution. It therefore has no effect on the efficiency and concentration limits as a function of étendue. The formulas for the target and source étendue are given in Eqs. (10.71) and (10.85). As in Section 10.3.6.1, we assume the total surface area of the target is

$$A_{trg} = 1\,\mathrm{m}^2, \tag{10.93}$$

so that the target étendue is

$$\varepsilon_{trg} = \pi\,\mathrm{m}^2 sr. \tag{10.94}$$

When the source and target étendue are equal, Eqs. (10.85) and (10.94) give the following value for the total surface area of the source:

$$A_{src} = 2.029\ \mathrm{m}^2. \tag{10.95}$$

For this equal-étendue case, the skewness distributions of Eqs. (10.70) and (10.84) are as depicted in Figure 10.14. The upper limit on étendue that can be transferred from the source to the target by a translationally symmetric concentrator is computed using the integral of Eq. (10.50). For the equal-étendue case, this upper limit on transferred étendue turns out to be

$$\varepsilon_{max} = 1.549\ \mathrm{m}^2 sr. \tag{10.96}$$

The upper limits on efficiency and concentration are computed by dividing ε_{max} by the total source and target étendue, respectively. For the equal-étendue case, the efficiency and concentration limits have the same value:

$$\eta_{max} = C_{max} = 49.30\%. \tag{10.97}$$

As in Section 10.3.6.1, we can also compute efficiency and concentration limits for source-to-target étendue ratios other than unity. The resulting efficiency limit as a function of the concentration limit is shown in Figure 10.15. As before, the diamond-shaped marker on this plot indicates the efficiency and concentration limits for the equal-étendue case. Figure 10.16 provides plots of the efficiency and concentration limits as a function of the source-to-target étendue ratio. Concen-
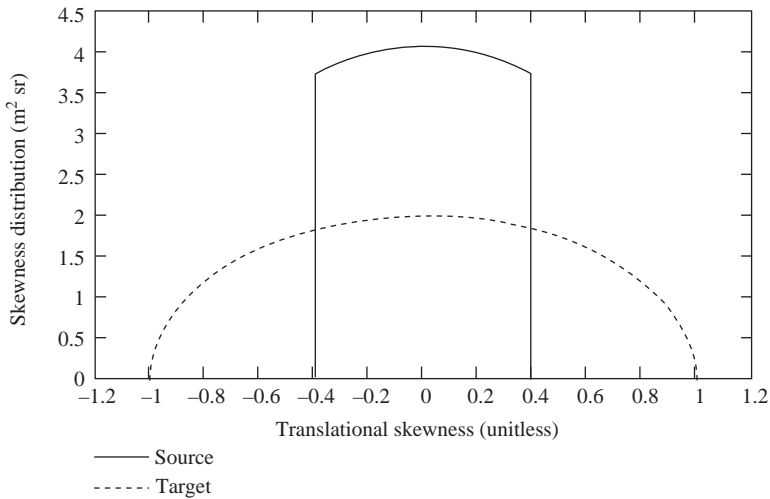
Figure 10.14   Translational skewness distributions in a unit-refractive-index medium for a Lambertian target and a source having fixed longitudinal cutoffs parallel to symmetry axis with orthogonal fixed latitudinal cutoffs. The latitudinal angular half width of the source is $\theta_0 = 23.45°$. The source étendue equals that of the target.
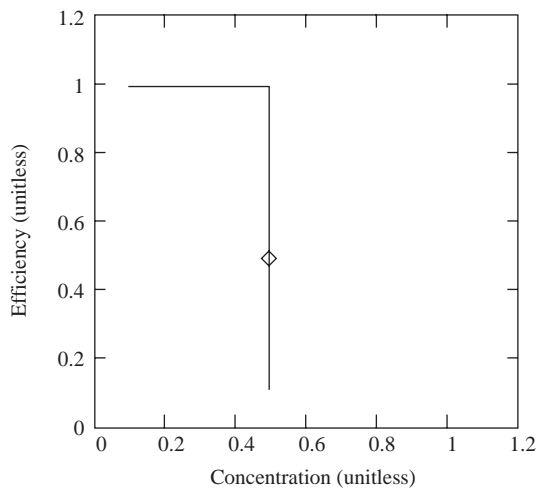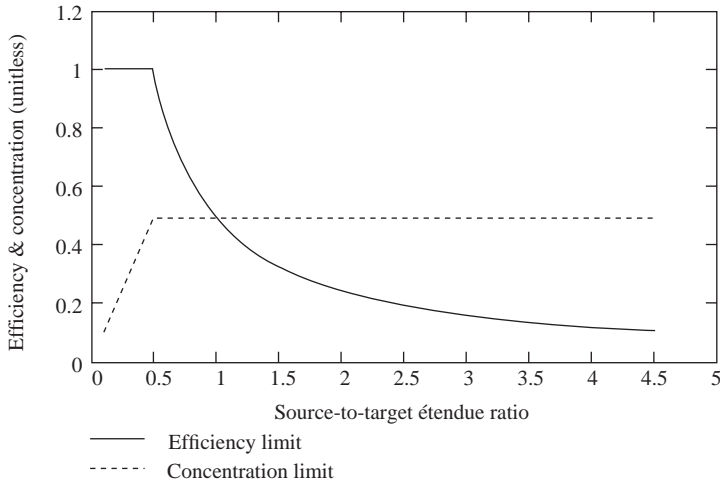

Figure 10.15   Plot of the efficiency limit as a function of the concentration limit for translationally symmetric nonimaging devices that transfer flux to a Lambertian target from a source having fixed longitudinal cutoffs parallel to symmetry axis with orthogonal fixed latitudinal cutoffs. The latitudinal angular half width of the source is $\theta_0 = 23.45°$. The diamond-shaped marker indicates the performance limit for the equal-étendue case.

tration limit in this figure never exceeds the 49.30% value of Eq. (10.97). This is because the half width of the skewness distribution of the source is always $\sin(\theta_0)$ = 0.3979, independent of the value of the source-to-target étendue ratio. Thus, no matter what value of the source-to-target étendue ratio is used, no étendue can be transferred from the source to the target by a translationally symmetric nonimaging device for skewness values satisfying $0.3979 < |S_z| \leq 1$.

Figure 10.16   Plot of the efficiency and concentration limits as a function of the source-to-target étendue ratio for translationally symmetric nonimaging devices that transfer flux to a Lambertian target from a source having fixed longitudinal cutoffs parallel to symmetry axis with orthogonal fixed latitudinal cutoffs. The latitudinal angular half width of the source is $\theta_0 = 23.45°$. The equal-étendue case corresponds to the crossing point of the two curves.

REFERENCES

Arnold, V. I. (1989). *Mathematical Methods of Classical Mechanics.* Springer Verlag, New York.

Bortz, J., Shatz, N., and Ries, H. (1997). Consequences of étendue and skewness conservation for nonimaging devices with inhomogeneous sources and targets. Proceedings of SPIE, Vol. 3139, 59–75.

Bortz, J., Shatz, N., and Winston, R. (2001). Performance limitations of translationally symmetric nonimaging devices. Proceedings of SPIE, Vol. 4446, 201–220.

Ries, H., Shatz, N., Bortz, J., and Spirkl, W. (1997). Performance limitations of rotationally symmetric nonimaging devices. *J. Opt. Soc. Am. A.,* Vol. 14, **10**, 2855–2862.

Shatz, N., and Bortz, J. (1995). An inverse engineering perspective on nonimaging optical design. Proceedings of SPIE, Vol. 2538, 136–156.

11

GLOBAL OPTIMIZATION OF HIGH-PERFORMANCE CONCENTRATORS

Narkis Shatz and John C. Bortz

Science Applications International Corporation, San Diego, CA

11.1 INTRODUCTION

Many nonimaging optical design problems encountered in practice have no known closed-form solutions. When this is the case, it is possible to obtain high-performance designs by means of global optimization. When applied to the design problem, this process is referred to as inverse engineering. This computationally intensive numerical design approach sequentially modifies the reflective and/or refractive surfaces of the optical system within a given parameterization scheme and constraint set until performance objective global optimality, evaluated upon a system radiometric model, is achieved. Global optimization can be used to determine reflector and lens configurations that achieve maximal flux transfer to a given target or that produce a desired radiometric distribution, such as an irradiance or intensity pattern.

This chapter provides an overview of the application of inverse engineering to the problem of nonimaging optical design. We begin with a brief summary of the behavior of nonimaging optical systems in terms of properties of mappings. An understanding of this behavior is central to the design problem, since it affects the limits on system performance, the choice of parametrization schemes, and the choice of the class of global-optimization algorithms to be used. We review the various factors affecting the performance of nonimaging optical systems, and present generalized formulations of étendue limits for use with inhomogeneous sources and/or targets. Following this preparatory material, we review the formalism of inverse engineering and include references to global-optimization algorithms that are applicable in the domain of nonimaging optical design. Finally, we provide five case studies of globally optimized designs, including three designs that use symmetry breaking to overcome the limitations imposed by skewness mismatches between the source and target.

11.2   MATHEMATICAL PROPERTIES OF MAPPINGS IN NONIMAGING OPTICS

In the geometrical optics approximation, the behavior of a nonimaging optical system can be formulated and studied as a mapping g: $S^{2n} \to S^{2n}$ from input phase space to output phase space, where S is an even-dimensional piecewise differentiable manifold and n is the number of generalized coordinates. The starting point for this formulation is the generalization of Fermat's variational principle, which states that a ray of light propagates through an optical system in such a manner that the time required for it to travel from one point to another is stationary.

Let g be a differentiable mapping. The mapping g is called canonical (Arnold, 1989) if g preserves the differential 2-form $w^2 = dp_i \wedge dq_i\; i = 1 \ldots n$, where q is the generalized coordinate and p is the generalized momentum. Applying the Euler-Lagrange necessary condition to Fermat's principle and then the Legendre transformation, we obtain a canonical Hamiltonian system, which defines a vector field on a symplectic manifold (a closed nondegenerate differential 2-form). A vector field on a manifold determines a phase flow—that is, a one-parameter group of diffeomorphisms (transformations that are differentiable and also possess a differentiable inverse). The phase flow of a Hamiltonian vector field on a symplectic manifold preserves the symplectic structure of phase space and consequently is canonical.

The properties of these mappings can be summarized as follows:

1. The mappings from input phase space to output phase space are piecewise diffeomorphic. Consequently they are one-to-one and onto.
2. The transformation of phase space induced by the phase flow is canonical—that is, it preserves the differential 2-form.
3. The mappings preserve the integral invariants, known as the Poincaré-Cartan invariants. Geometrically, these invariants are the sums of the oriented volumes of the projections onto the coordinate planes.
4. The mappings preserve the phase-space volume element (étendue). The volume of gD is equal to the volume of D, for any region D.

In order to uniquely determine a mapping, we require, in addition to Fermat's principle, knowledge of material properties and the stipulation of boundary conditions, which constitute the geometry of the reflective and/or refractive surfaces. Determining these boundary conditions is the subject of the nonimaging optical design problem.

Because the mappings in nonimaging optics are diffeomorphic (i.e., smooth), the solution topology for nonimaging optical design problems is Lipschitz continuous. This is an advantageous property that can be exploited as regards the selection of a global optimization search algorithm. Certain choices of parameterization schemes, particularly in designs involving multiple optical components, may lead to a finite number of discontinuities in the topology. Since the optical source will, in practice, be represented by a finite-sized ray set, the solution topology will be mildly stochastic.

11.3   FACTORS AFFECTING PERFORMANCE

We now summarize the key factors affecting the flux-transfer performance of nonimaging optical designs—such as collection, projection or coupling optics—that utilize three-dimensional sources and/or targets. Performance limits are associated with étendue matching, skewness matching, source and target inhomogeneities, design constraints, design goals, and nonidealities. The ultimate performance of the nonimaging system will generally be driven by some combination of these performance-limiting factors.

11.3.1   Étendue Matching

Étendue matching is a dominant consideration in the design of nonimaging optical systems. We define the target-to-source étendue ratio (TSER) as the total target étendue divided by the total source étendue. When the target étendue is smaller than that of the source, the upper limit of the fraction of source étendue that can be transferred to the target's phase space is equal to the TSER. When the target étendue is greater than that of the source, the upper limit on the fractional étendue that can be transferred is unity, but the phase space of the target will be diluted. As a practical matter, the étendue of the source should be computed based on an integration of experimental measurements or a valid source model, whereas the étendue of the target can usually be computed using an analytical integration.

11.3.2   Skewness Matching

A skewness mismatch between the source and target may cause severe performance limitations. A comparison of skewness mismatch between candidate sources and the target can be useful during the design selection process—for example, a common dilemma is whether to select an on-axis or a transverse source orientation. Losses due to skewness mismatch may be recovered to a large extent by employing an optimized nonrotationally symmetric design, which actively attempts to match the skewness of the source to that of the target.

11.3.3   Source and Target Inhomogeneities

Source and target inhomogeneities will affect the performance limits. In order to assess these limits, weight functions need to be introduced based on the source's specific spatial and angular radiance distributions and the target's preferential characteristics. A detailed method to accomplish this was introduced in the previous chapter for nonimaging optical systems exhibiting rotational or translational symmetry. An analogous method for the case of nonsymmetric systems is the subject of Section 11.4.

11.3.4   Design Constraints

Real world designs are often subject to constraints. If constraints are active at the optimal design point, then the performance of the system will be adversely affected. Typical design constraints may include minimum source-to-reflector clearance, reflector (or lens) diameter, and length constraints.

11.3.5   Competing Design Goals

We identify two main classes of design goals: maximum flux transfer and beam shaping. Weighted combinations of these goals may also be constructed. Flux-transfer performance limits may be adversely affected by the inclusion of beam shaping considerations.

11.3.6   Nonidealities

In the real world, nonidealities exist that can limit performance. These include reflectivity and variation of reflectivity with angle of incidence, as well as degrees of nonspecularity in the reflecting surface. Low reflectivities tend to discourage the use of multiple reflections in the design.

After a nominal design is achieved, off-nominal operating conditions should be examined. These may include design robustness to variations in TSER and skewness mismatch, as well as analysis of tolerance to optical-surface-figure errors and component misalignments.

11.4   THE EFFECT OF SOURCE AND TARGET INHOMOGENEITIES ON THE PERFORMANCE LIMITS OF NONSYMMETRIC NONIMAGING OPTICAL SYSTEMS

For many optical sources of interest, the radiance varies over the phase space of the source. The radiance as a function of position along the length of an incandescent filament, for example, is typically greater near the center of the filament than near either end. Optical sources also commonly exhibit angular radiance distributions that differ significantly from a Lambertian distribution. Sources having nonuniform spatial and/or angular distributions are said to be inhomogeneous.

The radiance of a given source is often an extremely complicated function of position within the source's phase space. Fortunately, with the development in recent years of experimental techniques that can achieve full phase-space source characterization, high-fidelity virtual-source representations are now commercially available (Rykowski and Wooley, 1997). Using a two-axis computer-controlled goniometer on which a CCD camera is mounted, thousands of images of the source are captured from view angles over 4π steradians. The captured images contain detailed information about the spatial and angular distributions of light radiating from the source, enabling a detailed model of the source's radiance distribution to be developed. From this model a radiometric ray set virtually representing the three-dimensional physical source can be constructed. With additional processing to provide variance reduction, a stochastic ray-set model of the

source can be readily prepared for insertion into a global optimization design loop. A large variety of sources have been accurately characterized using this technique, including filament lamps, xenon arc lamps, light emitting diodes, compact fluorescent lamps, metal halide lamps, and so on.

A target may also be inhomogeneous. For example, a target could have an absorptance function that decreases with increasing angle of incidence with respect to the target's local surface normal. To achieve high-performance flux transfer to such a target, it would be preferable to transfer as much flux as possible to the regions of phase space corresponding to smaller, rather than larger, angles of incidence. In general, we refer to a target as inhomogeneous when a nonconstant weight function has been defined over the phase space of the target in order to specify the relative usefulness of flux transferred to different regions of the target's phase space.

When the source and/or target are inhomogeneous, the principle of étendue conservation still applies, but the computation of the upper limit on performance is complicated by the fact that more importance is attached to some regions of the source and/or target phase spaces than to others. In this section we derive the upper limits on flux-transfer performance for nonimaging optical systems designed for use with inhomogeneous sources and targets. We consider the general case of nonsymmetric optical systems, for which performance limitations due to the skew invariant do not apply. The special cases of performance limits for rotationally and translationally symmetric optics were considered in the previous chapter.

11.4.1   Flux Transfer from an Inhomogeneous Source to an Inhomogeneous Target

We consider an inhomogeneous source having a total emitted flux of $P_{src,tot}$. The source radiance is represented by the function $L_{src}(\mathbf{x})$, where the vector $\mathbf{x}$ represents a point in the phase space S of the source. For the target we define the weight function $W_{trg}(\mathbf{x}')$, where $\mathbf{x}'$ represents a point in the phase space S' of the target. Our goal in designing a nonimaging system is to maximize the weighted flux

$$P^{wgt} \equiv \int_{\mathbf{x}' \in S'} d\varepsilon(\mathbf{x}') L_{trg}(\mathbf{x}') W_{trg}(\mathbf{x}') \tag{11.1}$$

transferred from the source to the target, where $L_{trg}(\mathbf{x}')$ is the radiance as a function of position in the target's phase space, and $d\varepsilon(\mathbf{x}')$ is the differential element of the target's phase-space volume. We now derive a formula for the maximum achievable value of P^{wgt}.

To begin, we derive some useful formulas related to the source. We define the cumulative sorted source étendue as

$$\varepsilon_{src}(L) \equiv \int_{\mathbf{x} \in S; L_{src}(\mathbf{x}) > L} d\varepsilon(\mathbf{x}), \tag{11.2}$$

where L is the source-radiance threshold, and $d\varepsilon(\mathbf{x})$ is the differential element of the source's phase-space volume. The function $\varepsilon_{src}(L)$ represents the amount of source phase-space volume associated with regions of the source's phase space for which the radiance is greater than the radiance threshold value L. Since increasing the threshold reduces the phase-space volume contributing to the integral in Eq. (11.2), it is apparent that $\varepsilon_{src}(L)$ must be a monotonically decreasing function

of L for $L_{min} \leq L \leq L_{max}$, where L_{min} and L_{max} are the minimum and maximum values of the source radiance. Specifically, $\varepsilon_{src}(L)$ must decrease monotonically from a maximum value of $\varepsilon_{src}(L_{min})$ at $L = L_{min}$ to a minimum value of 0 at $L = L_{max}$. It is also apparent that

$$\varepsilon_{src}(L_{min}) = \varepsilon_{src,tot}, \tag{11.3}$$

where $\varepsilon_{src,tot}$ is the total phase-space volume associated with the source.

Because $\varepsilon_{src}(L)$ decreases monotonically with L, it can be inverted to obtain the monotonically decreasing inverse function $L(\varepsilon_{src})$, which ranges from a value of L_{max} at $\varepsilon_{src} = 0$ to L_{min} at $\varepsilon_{src} = \varepsilon_{src,tot}$. The inverse function $L(\varepsilon_{src})$ represents the sorted source radiance as a function of the source étendue. The sorting process has forced the largest radiance value to occur at $\varepsilon_{src} = 0$, with decreasing values of radiance as ε_{src} is increased. The integral of $L(\varepsilon_{src})$ over the range, $0 \leq \varepsilon_{src} \leq \varepsilon_{src,tot}$ equals the total source flux $P_{src,tot}$.

We now consider quantities related to the target. In the phase space of the target, the weight function $W_{trg}(\mathbf{x}')$ plays a role analogous to that played by the radiance function $L_{src}(\mathbf{x})$ in the phase space of the source. We define the cumulative sorted target étendue as

$$\varepsilon_{trg}(W) = \int_{\mathbf{x}' \in S'; W_{trg}(\mathbf{x}') > W} d\varepsilon(\mathbf{x}'), \tag{11.4}$$

where W is the weight-function threshold. The function $\varepsilon_{trg}(W)$ is a monotonically decreasing function of W for $W_{min} \leq W \leq W_{max}$, where W_{min} and W_{max} are the minimum and maximum values of the target weight function. This monotonically decreasing function ranges from a maximum value of $\varepsilon_{trg}(W_{min})$ at $W = W_{min}$ to a minimum value of 0 at $W = W_{max}$. It is therefore possible to invert the function $\varepsilon_{trg}(W)$ to obtain the monotonically decreasing inverse function $W(\varepsilon_{trg})$, which ranges from W_{max} at $\varepsilon_{trg} = 0$ to W_{min} at $\varepsilon_{trg} = \varepsilon_{trg,tot}$, where $\varepsilon_{trg,tot} = \varepsilon_{trg}(W_{min})$ is the total target étendue. The inverse function $W(\varepsilon_{trg})$ represents the sorted target weight as a function of the target étendue. The sorting process forces the largest weight value to occur at $\varepsilon_{trg} = 0$, with decreasing weight values as ε_{trg} is increased.

We can now write down an expression for the maximum possible value of the weighted flux P^{wgt} transferred from the source to the target for a nonsymmetric optical system. This maximum occurs when the source radiation is transferred to the target in such a way that the largest source-radiance values preferentially fill the regions of the target's phase space having the largest weight values. Thus, referring to Eq. (11.1), we find that the maximum value of P^{wgt} can be expressed in the form

$$P^{wgt}_{max} = \int_0^{\varepsilon_{upper}} d\varepsilon\, L(\varepsilon) W(\varepsilon), \tag{11.5}$$

where

$$\varepsilon_{upper} = \min(\varepsilon_{src,tot}, \varepsilon_{trg,tot}) \tag{11.6}$$

is the lesser of the total target and source étendue values.

Now that we have derived the upper limit on the transferred flux, we are in a position to write down formulas for the upper limits on efficiency and concentration. For inhomogeneous sources and targets, the efficiency η is defined as the actual weighted flux transferred to the target divided by the weighted flux level that would be achieved if all of the source radiation could be transferred to the

region of the target's phase space for which the weight function has its maximum value. In other words, the efficiency η is defined as P^{wgt} divided by the value of P^{wgt}_{max} that would be obtained by replacing $W(\varepsilon)$ and ε_{upper} by W_{max} and $\varepsilon_{src,tot}$, respectively, in Eq. (11.5):

$$\eta \equiv \frac{P^{wgt}}{W_{max}P_{src,tot}}, \tag{11.7}$$

where

$$P_{src,tot} = \int\limits_{0}^{\varepsilon_{src,tot}} d\varepsilon\, L(\varepsilon) \tag{11.8}$$

is the total flux emitted by the source. Analogously, the concentration C is defined as the actual weighted flux transferred to the target divided by the weighted target flux level that would be achieved if all of the target's phase space could be filled with radiation having radiance equal to the maximum radiance level of the source. Thus, the concentration C is defined as P^{wgt} divided by the value of P^{wgt}_{max} that would be obtained by replacing $L(\varepsilon)$ and ε_{upper} by L_{max} and $\varepsilon_{trg,tot}$, respectively, in Eq. (11.5):

$$C \equiv \frac{P^{wgt}}{L_{max}\varepsilon^{wgt}_{trg,tot}}, \tag{11.9}$$

where

$$\varepsilon^{wgt}_{trg,tot} = \int\limits_{0}^{\varepsilon_{trg,tot}} d\varepsilon W(\varepsilon) \tag{11.10}$$

is the total weighted target étendue. It should be noted that given the way η and C have been defined, the values of both must always fall within the range from 0 to 1. The upper limits on the efficiency and concentration can now be obtained simply by substitution of P^{wgt}_{max} for P^{wgt} in Eqs. (11.7) and (11.9), respectively:

$$\eta_{max} = \frac{P^{wgt}_{max}}{W_{max}P_{src,tot}} \tag{11.11}$$

and

$$C_{max} = \frac{P^{wgt}_{max}}{L_{max}\varepsilon^{wgt}_{trg,tot}}. \tag{11.12}$$

11.4.2   Flux Transfer from an Inhomogeneous Source to a Homogeneous Target

This case is identical to the general case discussed in the previous section, except that the assumption of target homogeneity allows us to set the weight function equal to unity throughout the target's phase space:

$$W_{trg}(\mathbf{x}') = 1. \tag{11.13}$$

Since the weight function equals unity for all points $\mathbf{x}'$ in the phase space of the target, the sorted target weight function $W(\varepsilon_{trg})$ must also equal unity over the range of allowed values of ε_{trg}:

$$W(\varepsilon_{trg}) = 1. \tag{11.14}$$

Thus, Eqs. (11.1) and (11.5) can be rewritten in the form

$$P^{wgt} = \int_{\mathbf{x}' \in S'} d\varepsilon(\mathbf{x}') L_{trg}(\mathbf{x}') \tag{11.15}$$

and

$$P_{max}^{wgt} = \int_0^{\varepsilon_{upper}} d\varepsilon L(\varepsilon). \tag{11.16}$$

The efficiency and concentration reduce to

$$\eta = \frac{P^{wgt}}{P_{src,tot}} \tag{11.17}$$

and

$$C = \frac{P^{wgt}}{L_{max}\varepsilon_{trg,tot}}, \tag{11.18}$$

where we have used the facts that, for a homogeneous target, $W_{max} = 1$ and $\varepsilon_{trg,tot}^{wgt} = \varepsilon_{trg,tot}$. Similarly, the formulas for the maximum efficiency and concentration become

$$\eta_{max} = \frac{P_{max}^{wgt}}{P_{src,tot}} \tag{11.19}$$

and

$$C_{max} = \frac{P_{max}^{wgt}}{L_{max}\varepsilon_{trg,tot}}. \tag{11.20}$$

11.4.3 Flux Transfer from a Homogeneous Source to an Inhomogeneous Target

This case is identical to the general case discussed in Section 11.4.1, except that the assumption of source homogeneity allows us to set the radiance function equal to a constant throughout the source's phase space:

$$L_{src}(\mathbf{x}) = L_0, \tag{11.21}$$

where L_0 is the constant source radiance. Since the source radiance is constant for all points $\mathbf{x}$ in the source phase space, the sorted source radiance function $L(\varepsilon_{src})$ must also be constant over the range of allowed values of ε_{src}:

$$L(\varepsilon_{src}) = L_0. \tag{11.22}$$

The expression for the weighted flux transferred to the target [Eq. (11.1)] cannot be simplified, since source homogeneity does not guarantee that the target radiance $L_{trg}(\mathbf{x}')$ will be homogeneous. However, Eq. (11.22) allows us to simplify Eq. (11.5) to obtain the following formula for the upper limit on the weighted flux transferred to the target:

$$P_{max}^{wgt} = L_0 \varepsilon_{max}^{wgt}, \tag{11.23}$$

where

$$\varepsilon_{max}^{wgt} \equiv \int_0^{\varepsilon_{upper}} d\varepsilon W(\varepsilon) \tag{11.24}$$

is the maximum possible weighted target étendue that can be filled by flux transferred from the source.

The efficiency and concentration can then be expressed in the form

$$\eta = \frac{P^{wgt}}{W_{max}L_0\varepsilon_{src,tot}}$$ (11.25)

and

$$C = \frac{P^{wgt}}{L_0\varepsilon_{trg,tot}^{wgt}},$$ (11.26)

where we have used the fact that $P_{src,tot} = L_0\varepsilon_{src,tot}$ for a homogeneous source. Similarly, the formulas for the maximum efficiency and concentration become

$$\eta_{max} = \frac{\varepsilon_{max}^{wgt}}{W_{max}\varepsilon_{src,tot}}$$ (11.27)

and

$$C_{max} = \frac{\varepsilon_{max}^{wgt}}{\varepsilon_{trg,tot}^{wgt}},$$ (11.28)

where Eq. (11.23) has been used.

11.4.4   Flux Transfer from a Homogeneous Source to a Homogeneous Target

This case is the same as the case discussed in the last section, except that the assumption of target homogeneity allows us to set the weight function equal to unity. This allows us to simplify the expression for the weighted on-target flux by setting $W_{trg}(\mathbf{x}') = 1$ in Eq. (11.1) to give

$$P^{wgt} = \int_{\mathbf{x}' \in S'} d\varepsilon(\mathbf{x}')L_{trg}(\mathbf{x}'),$$ (11.29)

which is equivalent to Eq. (11.15). Despite the homogeneity of the source, it does not follow that the radiance produced in the target's phase space is necessarily homogeneous. Therefore, it is not possible to further simplify the expression for P^{wgt} by setting $L_{trg}(\mathbf{x}')$ equal to a constant in Eq. (11.29). Setting $W(\varepsilon)$ equal to unity, Eq. (11.23) reduces to the following expression for the upper limit on the weighted flux transferred to the target:

$$P_{max}^{wgt} = L_0\varepsilon_{upper},$$ (11.30)

where Eq. (11.24) has been used. The formulas for the efficiency and concentration reduce to

$$\eta = \frac{P^{wgt}}{L_0\varepsilon_{src,tot}}$$ (11.31)

and

$$C = \frac{P^{wgt}}{L_0\varepsilon_{trg,tot}}.$$ (11.32)

The maximum efficiency and concentration are obtained by substitution of the right-hand side of Eq. (11.30) for P^{wgt} in Eqs. (11.31) and (11.32):

$$\eta_{max} = \frac{\varepsilon_{upper}}{\varepsilon_{src,tot}} \qquad (11.33)$$

and

$$C_{max} = \frac{\varepsilon_{upper}}{\varepsilon_{trg,tot}}. \qquad (11.34)$$

11.5   THE INVERSE-ENGINEERING FORMALISM

Suppose we wish to design an engineering system to achieve certain objectives. Let us assume that this system supports a set of I piecewise continuous objective functions P_i that accept a set of J inputs d_j and that map a set of N independent design parameters $x_n \in A \in R^N$ into a set of I objective-function values $F_i \in B \in R^I$, where R^N is the N-dimensional Euclidean space, A is the feasible space formed by the action of a set of Q constraints h_q on R^N, and B is the co-domain of P_i. We also assume that we have available a set of I computational models (algorithms) M_i, such that M_i are equal to P_i under ideal conditions. Notice that the M_i are functions and therefore generate multidimensional topologies $f_i \in b \in R^I$. Consequently we may write

$$f_i = M_i(d_j, x_n) \quad R^N \rightarrow R^I. \qquad (11.35)$$

We can now write the inverse transformations M_i^{-1} of the functions $M_i \subset b \times A$:

$$\{x_n\} = M_i^{-1}(d_j, f_i) \quad R^I \rightarrow R^N. \qquad (11.36)$$

In the general case the inverse transformations M_i^{-1} are not functions because the M_i may be multimodal.

The basic axiom of inverse engineering can now be stated as follows: An engineering design problem on a physical system, for which there exist ideal computational models $M_i = P_i$, can be reduced to a set of objectives and a set of constraints. From this point on we shall only consider a single objective function ($I = 1$, a single topology). One possible method for computing $\{x_n\}$ is to seek extremal points on the solution topology, such as

$$\{x_n\} = \arg\min(f) \quad R^1 \rightarrow R^N. \qquad (11.37)$$

In our experience, the solution topology for a broad range of nonimaging optical design problems has been found to be multimodal and Lipschitz continuous. Consequently, the use of local optimization techniques to accomplish nonimaging optical system design has limited application. Over the past two decades, a branch of optimization has emerged, which has come to be known as global optimization. Global optimization algorithms (Otten and van Ginneken, 1989; Ratschek and Rokne, 1988; Torn and Zilinskas, 1987) can find, under certain regularity conditions (such as Lipschitz continuity), the locations of global optima in multimodal topologies. Such techniques are therefore generally useful for engineering design, and their use in this context has engendered the term inverse engineering.

The global optimization problem can be formally stated as follows: Minimize $f(d_j, x_n)$, $x_n \in R^N$ subject to the constraint set $h_q(d_j, x_n) = True$, assuming that f is piecewise continuous and exists almost everywhere and that the h_q are piecewise

continuous. The problem is to find the global minimum f^*, which may not be unique. In other words,

$$f^* = f\left(d_j, x_n^*\right) \leq f(d_j, x_n) \quad \text{for all} \quad x_n \in R_\varepsilon, \tag{11.38}$$

where R_ε, for any positive ε, is given by

$$R_\varepsilon\left(x_n^*\right) = \left\{ x_n \middle| h_q(d_j, x_n) = True, \left\| x_n - x_n^* \right\| \leq \varepsilon, x_n \in R^N \right\}. \tag{11.39}$$

In the context of nonimaging optical design, the quantities P_i, M_i, d_j, h_q, and x_n take on the following meanings:

- The P_i represent one or more radiometric or photometric quantities to be optimized, such as transmitted flux incident on a target or a measure of the difference between a delivered irradiance distribution and a required irradiance distribution.
- The M_i represent the computational models available for calculating the corresponding functions P_i. The functions M_i differ from the P_i in the sense that the M_i can only approximate the physics (e.g., the geometrical optics approximation) and typically are computed using numerical techniques (e.g., a quadrature employing a finite number of rays), which introduce a variance into the calculation of each M_i. This causes the solution topologies to become stochastic.
- The d_j are problem-specific inputs describing known characteristics of the optical system, such as surface reflectance, the spatial and angular radiance distributions of the radiation source, and so forth.
- The h_q are system design constraints, such as length and diameter limitations on the optics, or more complex requirements such as a restriction of the solution to concave or nonreentrant optical forms, and so on.
- The x_n are the independent parameters describing the design of the optical system, such as reflector shape parameters, source location relative to the optical system, choice of material for a refractive component, and so on.

The global optimization approach can search only within a finite dimensional space. This means that we need to define a parametrization scheme in order to approximately represent the general solution form and then find optimal values for a finite-sized set of independent parameters. In order to reduce the time complexity of the search process we desire a parsimonious representation. At the same time, we require an ability to accurately approximate the form of the true solution. In order to achieve both of these goals, we seek a natural coordinate system for the problem, which helps regularize the solution topology, and a set of basis functions in that coordinate system.

If no information is available about the general form of the solution, then the best that can be done is to choose a fixed Cartesian or polar coordinate system. If an approximate solution is known—for example, through application of the edge-ray principle—then that solution could be used to define a coordinate system. The basis functions of the parameterization are then expressed in that coordinate system. Good results have been obtained using natural splines as basis functions, although many other parametric approximations are possible (in particular, the now popular nonuniform rational B-Splines). Note that the concept of a coordinate
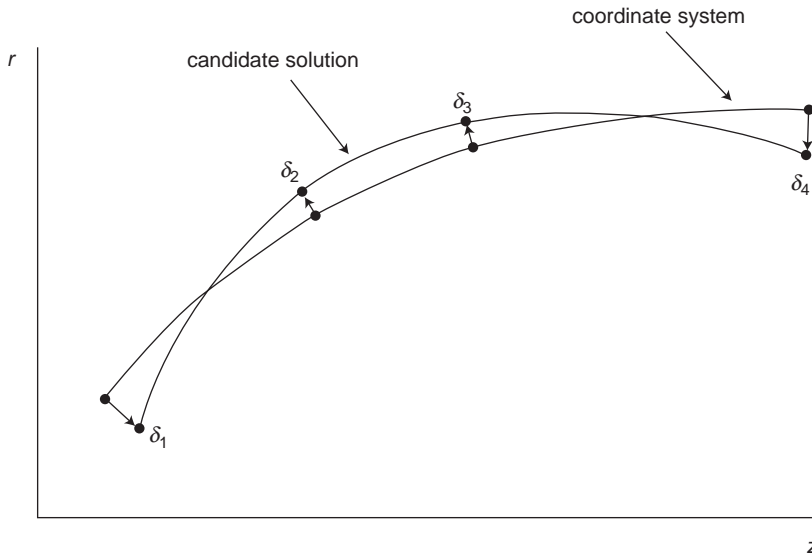
Figure 11.1.   Definition of a coordinate system and a parameterization scheme for representation of an axisymmetric reflective surface.

system here is distinct from the concept of a "starting point," which is required for local optimization. In global optimization a starting point has little or no meaning; a closed hypercube is generally used to confine the search space.

Figure 11.1 demonstrates, by way of example, a coordinate system specified by means of a fixed arc and by deviations normal to it. A candidate reflector solution can now be represented in that coordinate system by means of the deviations, which are provided by a natural cubic spline defined using four knots. The coordinate system does not necessarily have to be fixed during the search process for a given design problem. For example, p of the design degrees of freedom may be allocated to create a dynamic embedding—that is, a p-parameter group of coordinate systems.

Finally, for the purposes of the ray-trace calculations, the solution form may be transformed into a reference Cartesian or polar coordinate system and then represented in that system. For this final representation we often use a natural parametric spline system. This system uses a separate natural cubic spline to represent the dependence of each coordinate on the polygonal arc length s (e.g., in polar coordinates these would be $r(s)$ and $\phi(s)$), computed by connecting the knots with straight-line segments. Since s varies monotonically with the true arc length along the shape profile of the optical surface, this system can provide a compact, continuous, twice differentiable representation of complex nonconcave and reentrant forms as necessary.

11.6   EXAMPLES OF GLOBALLY OPTIMIZED CONCENTRATOR DESIGNS

In this final section, we present five examples of globally optimized concentrator designs. We first consider axisymmetric designs for use in transferring flux to a disk target from a disk source as well as a spherical source. Three optimized

designs are then presented that use symmetry breaking to provide flux-transfer efficiency beyond the skewness limit. The first two of these nonsymmetrical designs transfer flux to a disk target from spherical and cylindrical sources, respectively. The third is a non-tracking solar concentrator design for transferring flux from a rectangular aperture to a cylindrical target.

11.6.1 Axisymmetric Concentrators for a Disk Source and Disk Target

The compound parabolic concentrator (CPC), discussed elsewhere in this book, is a basic edge-ray design. In its three-dimensional form, the CPC is used to transfer flux between two equal-étendue disks, one having an emission (or acceptance) half angle of 90° and the other having an acceptance (or emission) half angle of less than 90°. The two-dimensional version of the CPC has been proven to be ideal and, consequently, optimal. In three dimensions it is known to be nonideal, since it transfers less than 100% of the flux from source to target. The question of whether the three-dimensional form is also nonoptimal within the design space comprising continuously differentiable axisymmetric perfectly specular reflective forms had been an open problem for many years following the development of the CPC. This question was eventually resolved by using global optimization to obtain reflective concentrators having performance superior to that of the 3D CPC (Shatz and Bortz, 1995). We now describe these globally optimized concentrators.

Our problem statement is as follows: Determine the profile of a rotationally symmetric 3D optimized spline concentrator (OSC) that maximizes the flux transferred from a 10-cm-diameter disk source having a specified emission half angle of less than 90°, to an equal-étendue disk target having a 90° acceptance half angle.

The shape of the 3D OSC is represented by using a CPC to define a coordinate system. A unitless axial scaling parameter C_{scal} is used to scale the length of the concentrator, thus extending the coordinate system to form a 1-parameter group. In this coordinate system we employ a radial natural cubic-spline basis function $\Delta r(x, \mathbf{r}_{dev})$, where x is the axial coordinate and $\mathbf{r}_{dev}$ is a vector of parameters controlling the radial deviation. Using this formulation, the radial coordinate of the OSC as a function of x is given by

$$r_{OSC}(x) = r_{CPC}(x/C_{scal}) + \Delta r(x/C_{scal}, \mathbf{r}_{dev}), \qquad (11.40)$$

where $r_{CPC}(x)$ is the radial coordinate of the CPC profile as a function of x. The components of the vector $\mathbf{r}_{dev}$ are the radial coordinate values of the N_{dev} cubic-spline knots that define the radial cubic-spline deviation function. The corresponding x-coordinates of the knots are equally spaced over the range of x-values between the minimum and maximum x-coordinates of the CPC. To ensure that the entrance and exit-aperture diameters of the OSC remain equal to the corresponding aperture diameters of the CPC, the radial coordinates of the initial and final knots on the deviation spline are set equal to zero and are not allowed to vary. Therefore, the optimization parameters comprise the $(N_{dev} - 2)$ internal radial deviations plus the axial scaling parameter C_{scal}, providing a total of $N_{param} = N_{dev} - 1$ independent parameters. It is apparent from Eq. (11.40) that when $C_{scal} = 1$ and the radial deviation vector is equal to the null vector, the shape function $r_{OSC}(x)$ reduces to $r_{CPC}(x)$. Thus, the CPC shape is an allowed solution in the N_{param}-dimensional space of possible solutions. A total of $N_{param} = 8$ optimization parame-
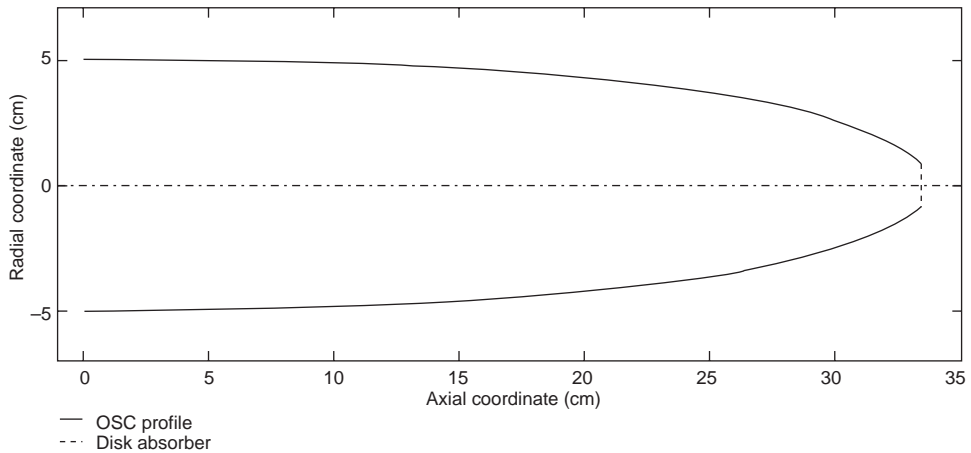
Figure 11.2   The 3D OSC with a 10° acceptance half angle.
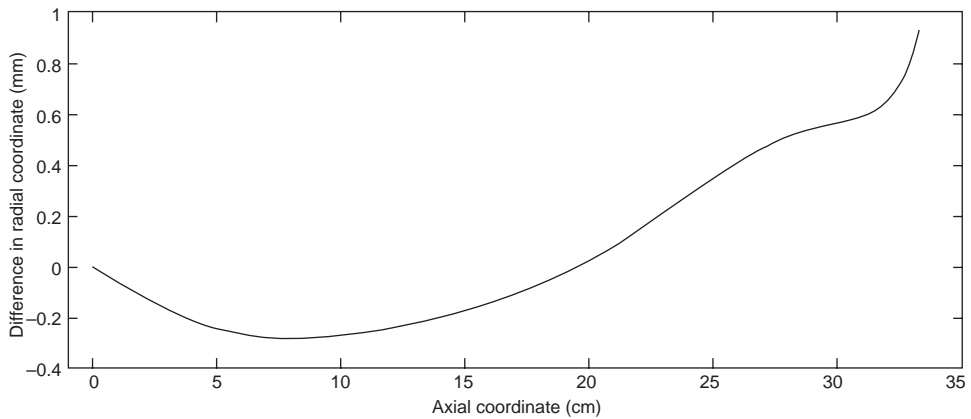
Figure 11.3   Radial shape difference between the 3D OSC and the 3D CPC as a function of the axial coordinate for a 10° acceptance angle. The difference is plotted out to the maximum length of the CPC. The 3D OSC is 1.13 mm longer than the 3D CPC.

ters were used, consisting of seven radial deviation values plus the scaling parameter C_{scal}. Each radial deviation value was allowed to vary within the range of ±1.0 cm, subject to the additional constraint that $r_{OSC}(x)$ was not allowed to exceed the exit-aperture diameter. The axial scaling parameter C_{scal} was allowed to vary between 0.9 and 1.1.

For the case of a 10° acceptance half angle, the global optimization procedure produced the 3D OSC design depicted in Figure 11.2. The shape difference between the 3D OSC and the 3D CPC is shown in Figure 11.3. Total transmission within the acceptance angle is shown in Figure 11.4. For the 10° case, the 3D OSC transmits 96.36% of the energy into the 10° beam, as compared to 95.68% for the 3D CPC. The accuracy of these computations is estimated at ±0.002%. This improvement accounts for 15.9% of the residual energy rejected by the 3D CPC. A comparable improvement is achieved for the 20° case. From Figure 11.5 it is apparent
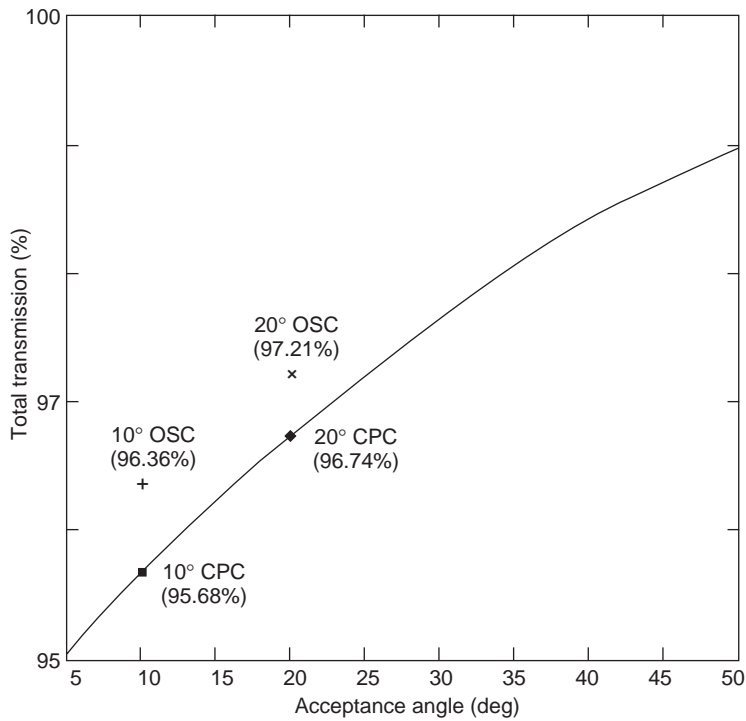
Figure 11.4   Total transmission within the acceptance angle as a function of the acceptance angle for the 3D OSC and the 3D CPC.
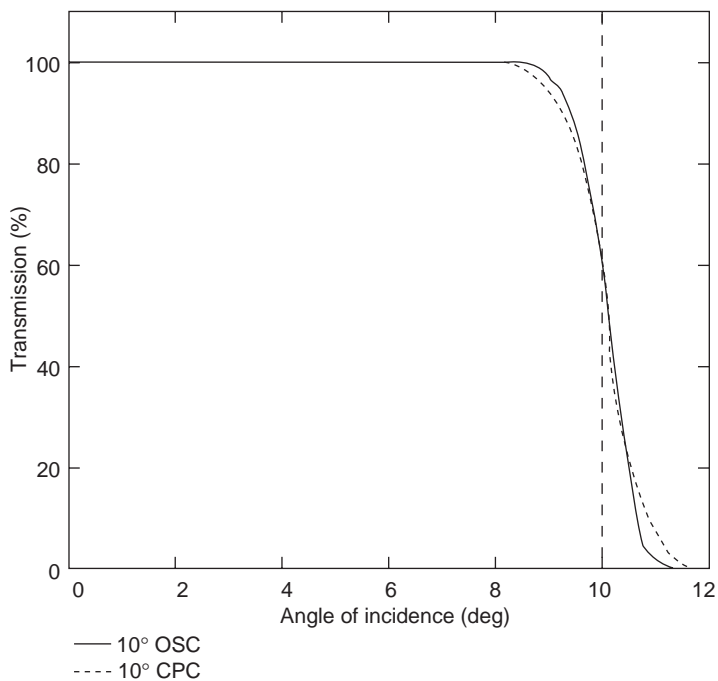
Figure 11.5   Plot of transmission versus angle of incidence for the 3D OSC and the 3D CPC with a 10° acceptance angle.
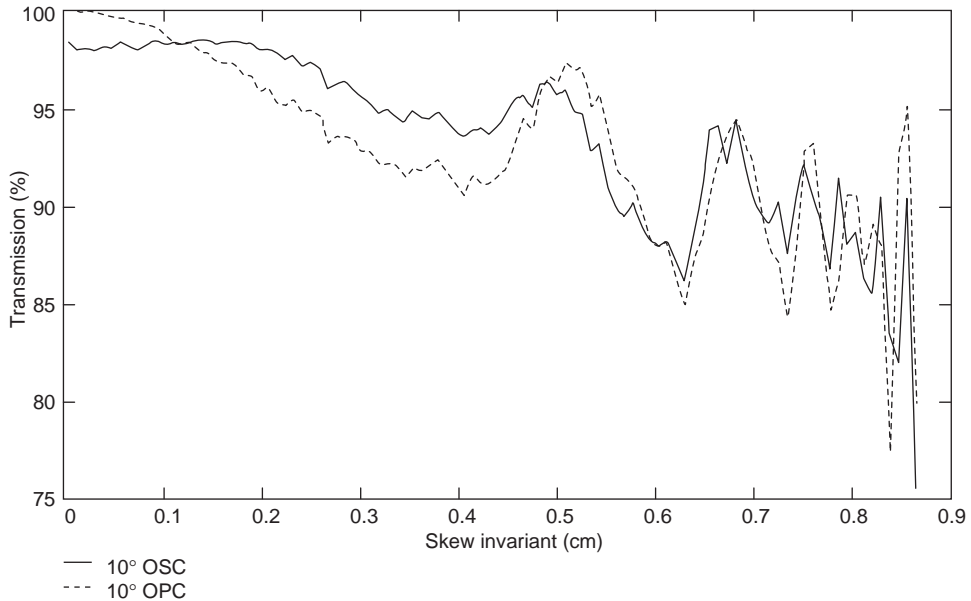
Figure 11.6   Transmission versus the skew invariant for the 3D OSC and the 3D CPC with a 10° acceptance angle.

that the 3D OSC has a sharper transmission cutoff than the 3D CPC and that, relative to the 3D CPC, it takes some of the input phase-space volume from beyond 10° and folds it into the 10° regime. Plots of transmission efficiency versus the skew invariant are shown in Figure 11.6 for the two concentrators. Since the CPC is an ideal 2D concentrator, we might expect the performance of the 3D OSC to be inferior to that of the 3D CPC for low values of the skew invariant—that is, for rays that are approximately meridional. Figure 11.6 shows that this is indeed the case: the transmission curve for the 3D OSC lies below that of the 3D CPC in the skew-invariant range 0.00–0.12 cm. In the range 0.12–0.49 cm, however, the 3D OSC transmission efficiency exceeds that of the 3D CPC. It is the superior performance of the 3D OSC for this range of values of the skew invariant that accounts for its overall improved performance relative to the 3D CPC.

It is particularly interesting to note that a local optimization, using the 10° 3D CPC as a starting point, does not arrive at the solution of Figure 11.6. Another solution is found, which, although it improves upon the 3D CPC and constitutes a local maximum, achieves only 70% of the performance gain reported above. From this we conclude that, although they appear to have similar shapes, the solution of Figure 11.2 and the 3D CPC do not occupy the same mode in the solution topology. This is consistent with our experience that the solution topology for nonimaging design problems is generally multimodal.

11.6.2   Optimized Axisymmetric Concentrator for Use with a Spherical Source

In the previous chapter the upper limits on performance for axisymmetric concentrators were presented. We now describe a numerically optimized axisymmetric concentrator that achieves performance virtually identical to the theoretical
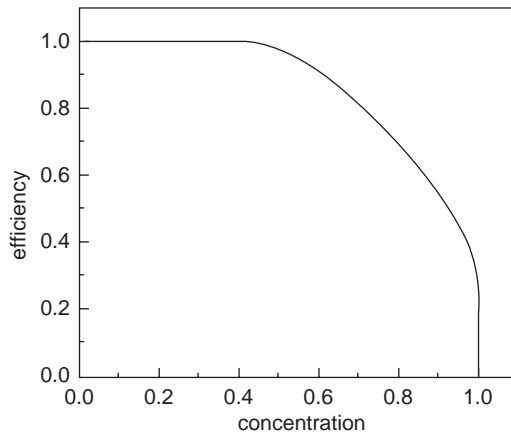
Figure 11.7   Upper limit on efficiency as a function of concentration for an axisymmetric optical system transferring flux from a spherical source to a disk target, or vice versa.

upper limit on flux-transfer efficiency from a Lambertian spherical source to a disk-shaped target having a fixed acceptance half angle.

A plot of the upper limit on achievable efficiency as a function of concentration for a spherical source and a disk target is shown in Figure 11.7. The equal-étendue case occurs at the point on this curve where efficiency and concentration are equal, in which case the efficiency and concentration are both equal to 75.3%. It is apparent from the efficiency-versus-concentration curve that an efficiency limit of 100% can only be achieved for concentrations less than or equal to about 40%. It can be shown that the exact value of this concentration limit is $4/\pi^2 \cdot 100\%$ $\approx 40.5\%$. It can also be shown that a concentration limit of 100% can only be achieved for efficiency values less than or equal to exactly 25%.

As an example, we consider the problem of determining the shape of a numerically optimized reflective concentrator that provides maximal flux-transfer efficiency from a 10-mm-diameter spherical Lambertian emitter to a disk-shaped target that accepts flux only within a 30° half angle relative to its surface normal. Additionally, we require that the source and target have equal étendue, which occurs when the diameter of the target equals 40mm. The reflective concentrator is assumed to have a loss-free, purely specular reflective coating. Its shape is assumed to be continuously differentiable. Finally, we assume the reflector is allowed to come in contact with both the source and the target. We refer to the optimized reflective concentrator as a three-dimensional optimized spline reflector (3D OSR).

We now describe a parameterization scheme that defines the shape of the concentrator as a function of a set of shape parameters. This parameterization scheme provided a solution that achieved performance very close to the theoretical upper limit of 75.3% efficiency and concentration. Other parameterization schemes could also be suggested.

The shape of the 3D OSR is represented by using a truncated involute CPC (TICPC) to define a coordinate system. To enforce the requirement that the exit-aperture radius of the TICPC must remain constant as its design parameters are varied, we introduce the scaled source radius $r_s(\tau, \theta_i, R_{exit})$, as a function of the truncation parameter τ, the acceptance angle θ_i, and the required exit-aperture

radius R_{exit}. The function $r_s(\tau, \theta_i, R_{exit})$ is defined as the design source radius that keeps the exit-aperture radius of the TICPC equal to R_{exit} for any given pair of values of τ and θ_i. The function $r_s(\tau, \theta_i, R_{exit})$ is used only in specifying the size of the TICPC and is not to be confused with the actual radius of the spherical source, which is a constant designated as r. For purposes of notational brevity, we now drop the explicit τ, θ_i, and R_{exit} dependence of $r_s(\tau, \theta_i, R_{exit})$, referring to it simply as r_s. The radial and axial coordinate pairs on the profile of the TICPC can be conveniently expressed in parametric form as the functions

$$r_{CPC}(\theta) = r_s \cdot \sin(\theta) - t(\theta) \cdot \cos(\theta) \tag{11.41}$$

and

$$x_{CPC}(\theta) = -r_s \cdot \cos(\theta) - t(\theta) \cdot \sin(\theta), \tag{11.42}$$

respectively, where θ is the angular parameter ranging from a minimum value of

$$\theta_{min} = 0 \tag{11.43}$$

to a maximum value of

$$\theta_{max} = \left(\frac{3\pi}{2} - \theta_i\right) \cdot (1 - \tau), \tag{11.44}$$

and $t(\theta)$ is the edge-ray distance as a function of the angular parameter

$$t(\theta) = \begin{cases} r_s \cdot \theta, & \theta \leq \theta_i + \dfrac{\pi}{2} \\[2mm] r_s \cdot \dfrac{\theta + \theta_i + \dfrac{\pi}{2} - \cos(\theta - \theta_i)}{1 + \sin(\theta - \theta_i)}, & \theta_i + \dfrac{\pi}{2} < \theta \end{cases}. \tag{11.45}$$

The value of the truncation parameter τ must lie in the range from 0 to 1, where $\tau = 0$ corresponds to the untruncated case. Using this parametric shape representation, the arc length $s_{CPC}(\theta)$ of the TICPC as a function of θ can easily be derived. The arc-length function $s_{CPC}(\theta)$ equals zero for $\theta = \theta_{min}$ and increases monotonically for values of θ greater than θ_{min}. We also define the adjusted arc-length,

$$s'_{CPC}(\theta) \equiv [s_{CPC}(\theta)]^\rho, \tag{11.46}$$

where ρ is a positive real number, referred to as the deviation-knot spacing control parameter. Since the adjusted arc length increases monotonically with θ, the inverse function $\theta(s'_{CPC})$ can also be defined. In practice, we use a natural cubic-spline representation of this inverse function. We now consider the N_{dev} equally spaced points

$$s'_n = \delta s \cdot (n - 1), \quad n = 1, \dots, N_{dev} \tag{11.47}$$

along the adjusted arc length, where

$$\delta s' = \frac{s'_{CPC}(\theta_{max})}{N_{dev} - 1} \tag{11.48}$$

is the sampling interval. The radial and axial coordinates of the N_{dev} points on the TICPC profile corresponding to these adjusted arc-length samples are $r_n = r_{CPC}(\theta(s'_n))$ and $x_n = x_{CPC}(\theta(s'_n))$. Similarly, the values of the angular parameter and edge-ray distance for these N_{dev} points are $\theta_n = \theta(s'_n)$ and $t_n = t(\theta(s'_n))$. We define

N_{dev} deviation axes, each of which passes through a sampling point and is perpendicular to the local tangent to the TICPC profile at that point. Associated with each of these N_{dev} deviation axes, we define a deviation knot, which must lie on its corresponding deviation axis. The position of deviation knot n along its associated deviation axis is given by the deviation-knot position parameter w_n, which represents the distance of the deviation knot from the sampled coordinate (r_n, x_n) on the TICPC profile. The allowed range of variation for each deviation-knot position parameter w_n is $\pm W_n/2$, with W_n obtained from the formula

$$W_n = a + b \cdot \left(\frac{s'_n}{s'_{N_{dev}}} \right)^{\frac{1}{\rho}}, \quad n = 1, \ldots, N_{dev}, \tag{11.49}$$

where a and b are real deviation-knot range-control constants. The OSR profile intersects each of the deviation knots. The radial and axial coordinates of the deviation knots are designated as R_n and X_n, respectively. The perturbed edge-ray distance corresponding to each of the N_{dev} deviation knots is of the form

$$T_n = \sqrt{X_n^{\,2} + R_n^{\,2} - r_s^{\,2}}. \tag{11.50}$$

Similarly, the perturbed angular parameter values Θ_n are given by the formulas

$$\sin(\Theta_n) = \frac{R_n \cdot r_s - X_n \cdot T_n}{r_s^2 + T_n^2} \tag{11.51}$$

and

$$\cos(\Theta_n) = \frac{X_n \cdot r_s + R_n \cdot T_n}{r_s^2 + T_n^2}. \tag{11.52}$$

We define the knot deviations as the difference between the perturbed and unperturbed values of the edge-ray distance and the angular parameter

$$\delta T_n \equiv T_n - t_n \tag{11.53}$$

and

$$\delta \Theta_n \equiv \Theta_n - \theta_n. \tag{11.54}$$

We can now generate the continuous edge-ray-distance deviation function $\delta t(\theta)$ as a function of θ by means of cubic-spline interpolation between the N_{dev} Cartesian coordinate pairs $(\theta_n, \delta T_n)$. Similarly, we obtain the continuous angular-parameter deviation function $\delta\theta(\theta)$ as a function of θ by means of cubic-spline interpolation between the N_{dev} Cartesian coordinate pairs $(\theta_n, \delta\Theta_n)$. Using these continuous deviation functions [$\delta t(\theta)$ and $\delta\theta(\theta)$], we can derive a formula for the radial and axial coordinates on the OSR profile, based on a straightforward generalization of the formulas given in Eqs. (11.41) and (11.42) for $r_{CPC}(\theta)$ and $\mathbf{x}_{CPC}(\theta)$:

$$r_{OSR}(\theta) = r_s \cdot \sin(\theta + \delta\theta(\theta)) - [t(\theta) + \delta t(\theta)] \cdot \cos(\theta + \delta\theta(\theta)) \tag{11.55}$$

and

$$x_{OSR}(\theta) = -r_s \cdot \cos(\theta + \delta\theta(\theta)) - [t(\theta) + \delta t(\theta)] \cdot \sin(\theta + \delta\theta(\theta)). \tag{11.56}$$

During the optimization process, the coordinates of the first deviation knot were kept equal to the coordinate values of the first sampling point on the unperturbed TICPC. The purpose of this was to force the initial knot to lie on the axis

Table 11.1.   Lower and Upper Ranges of 3D OSR Optimization Parameters.

Parameter Description	Symbol	Units	Lower Range	Upper Range
Truncation	τ	Unitless	0.00	0.05
Acceptance angle	θ_i	Degrees	20.0	30.0
Spacing-control parameter	ρ	Unitless	0.4	1.0
Axial source position shift	δx_{src}	mm	−2.0	2.0
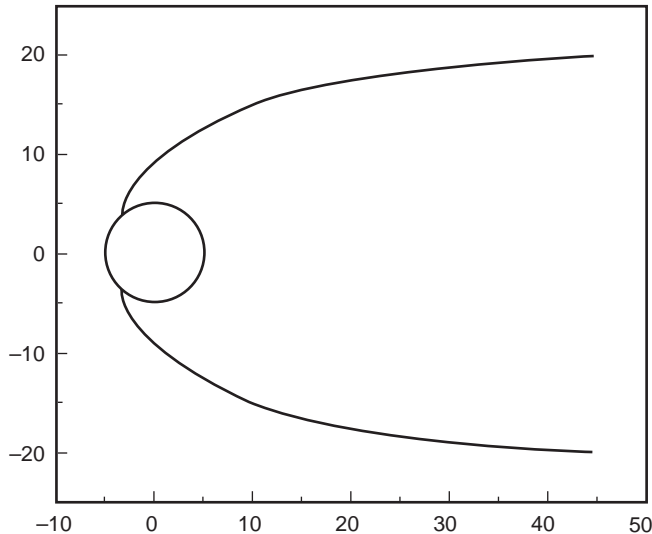
Figure 11.8   3D OSR designed to transfer flux from a 10-mm-diameter sphere to an equal-étendue disk having an acceptance half angle of 30°.

of symmetry of the OSR. Similarly, the coordinates of the last deviation knot were kept equal to the coordinate values of the last sampling point on the unperturbed TICPC. This was done to force the exit-aperture radius of the OSR to have the required value. Since the first and last deviation knots were kept unperturbed, only the (N_{dev}-2) intermediate deviation knots between the first and last values were allowed to vary as part of the optimization procedure. In addition to the (N_{dev}-2) deviation knots, the values of τ, θ_i, ρ, and δx_{src} were varied, where δx_{src} is the position shift of the spherical source along the axis of symmetry of the OSR. The source position shift is defined such that δx_{src} equals zero when the source surface just touches the on-axis point of the OSR on the left side of the source, with the exit pupil assumed to be on the right. For the OSR optimization discussed in this paper, the number of knots was N_{dev} = 15 and the values of the deviation-knot range-control parameters were a = 4mm and b = 0mm, respectively. The radius of the actual source was r = 5mm. A required exit-aperture-radius value of R_{exit} = 20mm was used. Table 11.1 gives the lower and upper ranges used in varying the parameters τ, θ_i, ρ, and δx_{src}.

The resulting shape profile of the numerically optimized 3D OSR design obtained using the above-described parameterization scheme for a 30° half-angle beam is shown in Figure 11.8. An unusual feature of this design is that the source protrudes from the back of the reflector. The performance of this solution was computed by tracing 200,000 rays, which provides an accuracy of better than 0.1% in concentra-

tion and efficiency. The solution provides both a concentration and efficiency of 75.1%, which is virtually identical to the theoretical performance limit of 75.3%.

For purposes of comparison, we also consider the performance of two edge-ray solutions. The two-dimensional (2D) involute CPC is an edge-ray solution capable of collecting light incident on a rectangular aperture within a given input acceptance angle and transferring it to a cylindrical absorber with ideal concentration and efficiency. The 3D involute CPC is defined as the shape swept out when the shape profile of the 2D involute CPC is rotated about its symmetry axis. Operated in reverse, the 3D involute CPC constitutes the edge-ray solution for the projection of energy from a spherical source into an emergent conical beam. We consider two cases: (1) a 3D involute CPC designed for an acceptance half-angle of 30° and then truncated to provide the required exit-aperture diameter of 40mm, and (2) an untruncated 3D involute CPC having a design acceptance half angle of 51.76°. The acceptance half angle of 51.76° was chosen in the latter case because this happens to be the value of the half angle for which the exit-aperture diameter of the untruncated concentrator equals 40.0mm.

A comparison of the efficiency versus concentration curve for the 3D OSR solution with the corresponding curves for the two edge-ray solutions is shown in Figure 11.9. The upper-limit curve in Figure 11.9 is the same curve shown previously in Figure 11.7. The curves shown for the 3D OSR and for the two edge-ray solutions were each obtained by computing both the efficiency and the concentra-
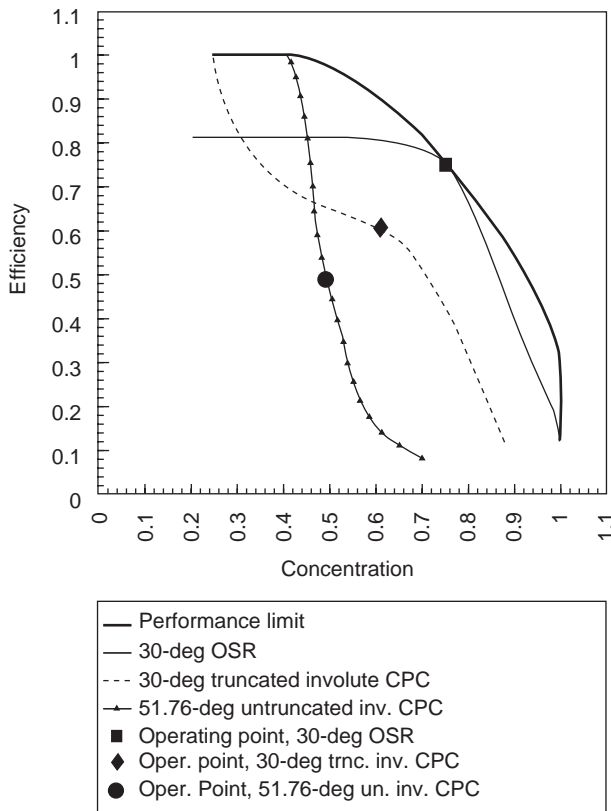
Figure 11.9   Efficiency versus concentration for 3D OSR and two edge-ray designs.
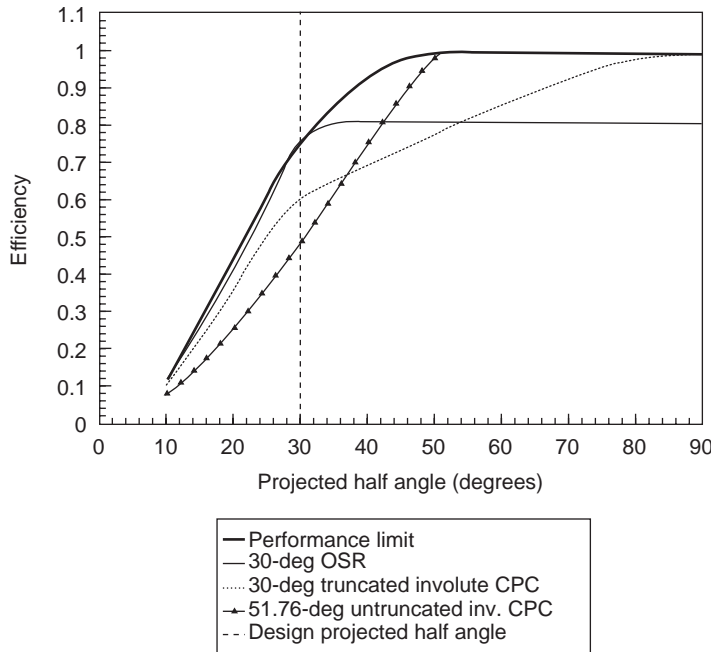
Figure 11.10   Efficiency versus projected half angle for 3D OSR and two edge-ray designs.

tion as a function of the target's phase-space volume and then plotting the result-ing efficiency values as a function of the corresponding concentration values. The target's phase-space volume was varied by changing the acceptance half angle of the target. All other system characteristics were held constant, including the radii of both the source and the target, as well as the shape of each concentrator. It should be noted that the performance of the 3D OSR is superior to both in-volute CPCs throughout most of the efficiency-versus-concentration regime. The optimized nature of the 3D OSR is particularly evident by the tangency of its performance curve to the limiting curve at the design point. A comparison of the efficiency versus the acceptance half angle for the 3D OSR and the two involute CPCs is shown in Figure 11.10. It can be seen from this figure that the 3D OSR tracks the limiting curve closely for operating half angles less than 30°, whereas the deficiencies of the edge-ray solutions are clearly evident.

It is worth noting that the involute CPC, which is an ideal concentrator in two dimensions, performs poorly when converted into a three-dimensional concentra-tor. The reason for its poor three-dimensional performance is that when the two-dimensional source and target are étendue-matched, the corresponding three-dimensional source and target are not. This is not the case for the conven-tional (noninvolute) CPC, which performs quite well, although not optimally, as a three-dimensional concentrator.

Figure 11.11 shows the skewness distribution of all rays transferred from the sphere to the disk by the 3D OSR. Also shown, for comparison, are the theoreti-cal skewness distributions of the sphere and the disk. (The slightly jagged appear-ance of the 3D OSR's output skewness distribution is due to the finite number of rays employed in the computation.) As shown in Figure 11.12, when we exclude
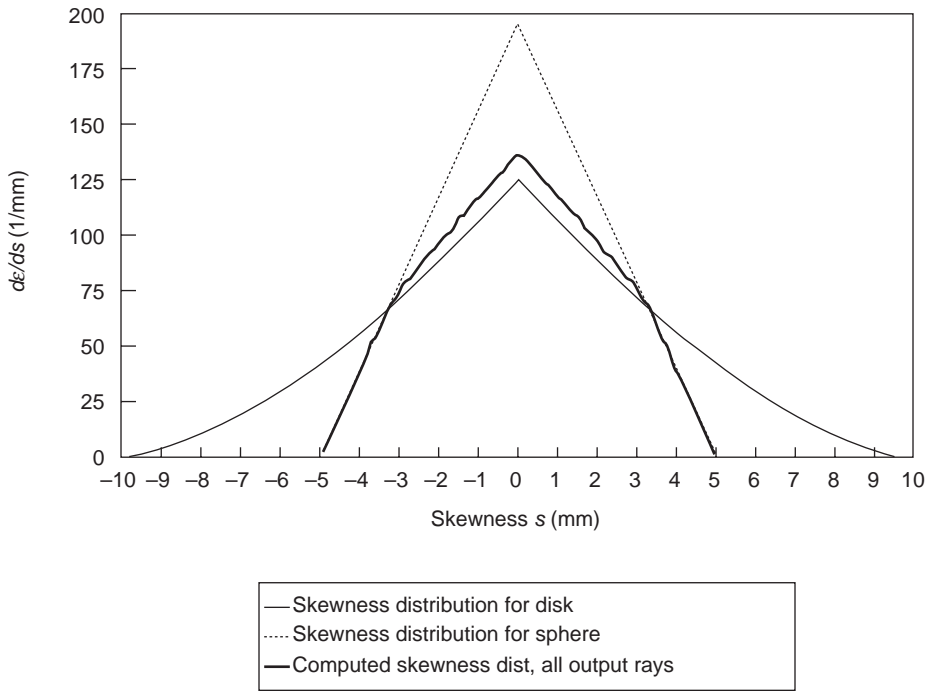
Figure 11.11.   Skewness distributions for spherical source, disk target, and all rays transferred to the disk by the 3D OSR.
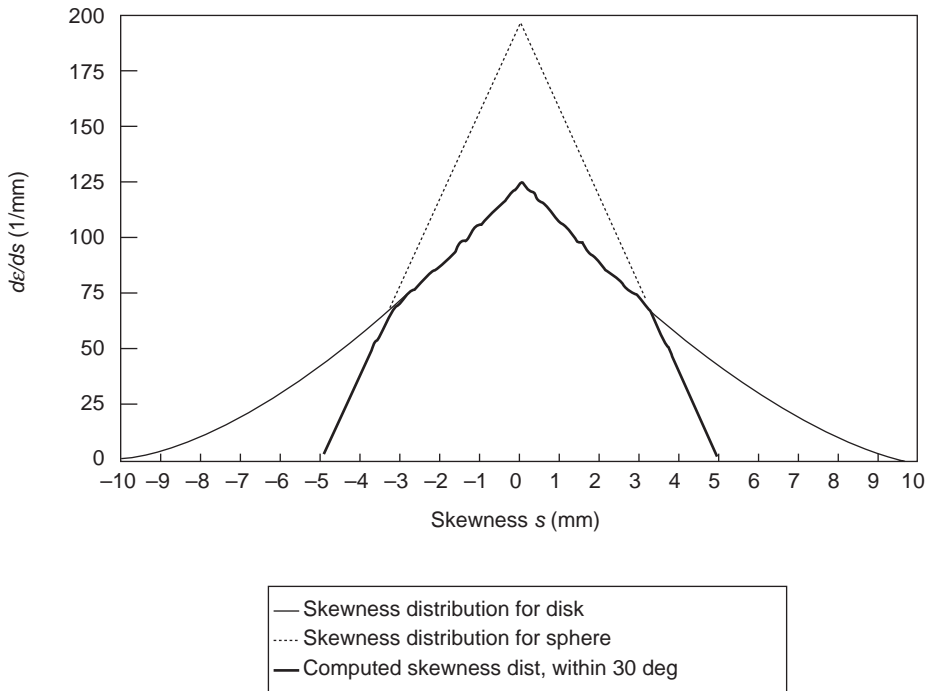
Figure 11.12   Skewness distributions for spherical source, disk target, and subset of rays transferred to disk within 30° half angle by 3D OSR.

from the flux output of the 3D OSR all rays beyond the projected half angle of 30°, we find that the skewness distribution closely approximates the boundary of the region of overlap between the skewness distributions of the sphere and the disk. This is consistent with the fact that the efficiency and concentration of the optimized concentrator are virtually at the upper limit on performance for these quantities.

11.6.3   Optimized Nonaxisymmetric Concentrator for a Spherical Source and a Disk Target

We now address the problem of designing a nonaxisymmetric nonimaging reflector that maximizes the flux transferred from a homogeneous spherical source to a disk target of equal étendue (Shatz, Bortz, Ries, and Winston, 1997). The design described in the previous subsection was achieved by applying numerical optimization to a solution space, defined by means of a parameterization scheme, resulting in an axisymmetric optimized spline reflector. That optimized solution achieved an efficiency of 75.1%, which is practically at the analytically computed theoretical limit of 75.3%. This performance represents a substantial improvement over two edge-ray solutions, which provide efficiencies of 49.2% and 61.0%, respectively. We now consider the generalized problem of designing a nonimaging reflector that is allowed to assume a non-axisymmetric form. Specifically, we consider a class of forms designated as three-dimensional star concentrators. We refer to the optimized design as a three-dimensional optimized star concentrator (3D OSC). To begin, we develop a star-concentrator parameterization scheme, intended for use in conjunction with numerical optimization procedures. The same parameterization scheme is also utilized in designing a 3D OSC for transferring flux from a cylindrical source to a disk target, as discussed in the following subsection. We initially consider a rotationally symmetric reflector. To generate this reflector shape we define a two-dimensional profile, which is then rotated about an axis of symmetry. It is convenient to specify the two-dimensional profile parametrically, in the form:

$$r = r(\theta) \tag{11.57}$$

and

$$x = x(\theta), \tag{11.58}$$

where r is the distance from the symmetry axis, x is the coordinate along the symmetry axis, and θ is a parameter that varies over some convenient finite interval. The functions $r(\theta)$ and $x(\theta)$ are assumed to be continuous functions of the parameter θ.

Now suppose we want to break the symmetry of the reflector shape by introducing a dependence on the angular variable ϕ, where ϕ specifies the amount of rotation about the x-axis. We define the ϕ-coordinate such that it conforms to the right-hand rule: When the thumb of the right hand points along the x-axis, the fingers point in the positive ϕ-direction. The ϕ-dependence of the desired nonrotationally symmetric shape is described in terms of constant-angle spirals on conical surfaces. This will enable a parsimonious parameterization. We now consider characteristics of such spirals.

To define a constant-angle spiral, we consider a cone of half-angle β having the x-axis as its axis of symmetry. Suppose that a continuous curve C has been defined on the surface of this cone. We designate as T the line that is tangent to

the curve C at some point P on C. We also define the circle c that passes through the point P and is centered on the x-axis. The line tangent to c at point P is designated as T'. The angle between T and T' is designated as α. A constant-angle spiral is defined as any continuous curve C for which the angle α is a constant for all points on the curve.

We define s to be the distance of a point P on a constant-angle spiral from the vertex of the cone. We consider the differential change ds in the distance s produced by a differential change $d\phi$ in the angle ϕ. The constant-angle criterion leads to following relationship between ds and $d\phi$:

$$\frac{ds}{rd\phi} = \tan(\alpha). \tag{11.59}$$

The radial coordinate r can be expressed in the form

$$r = r_0 + (s - s_0)\sin(\beta), \tag{11.60}$$

where r_0 and s_0 are the r and s-coordinates, respectively, of the starting point on the spiral. Similarly, the axial coordinate x can be expressed as

$$x = x_0 + (s - s_0)\cos(\beta), \tag{11.61}$$

where x_0 is the x-coordinate of the starting point on the spiral. Substitution of Eq. (11.60) into Eq. (11.59) gives

$$\frac{ds}{r_0 + (s - s_0)\sin(\beta)} = \tan(\alpha)d\phi. \tag{11.62}$$

Integrating both sides of Eq. (11.62), we obtain an expression for s as a function of ϕ:

$$s(\phi) = s_0 + \frac{r_0}{\sin(\beta)}\left[e^{\tan(\alpha)\sin(\beta)(\phi-\phi_0)} - 1\right], \tag{11.63}$$

where ϕ_0 is the value of the angular coordinate ϕ at the starting point of the spiral. Substitution of Eq. (11.63) into Eq. (11.60) gives the r-coordinate as a function of ϕ:

$$r(\phi) = r_0 e^{\tan(\alpha)\sin(\beta)(\phi-\phi_0)}. \tag{11.64}$$

Substitution of Eq. (11.63) into Eq. (11.61) gives the x-coordinate as a function of ϕ:

$$x(\phi) = x_0 + \frac{r_0}{\tan(\beta)}\left[e^{\tan(\alpha)\sin(\beta)(\phi-\phi_0)} - 1\right]. \tag{11.65}$$

For the special case of $\beta = 0°$, Eqs. (11.64) and (11.65) become

$$r(\phi) = r_0 \tag{11.66}$$

and

$$x(\phi) = x_0 + r_0 \tan(\alpha)(\phi - \phi_0). \tag{11.67}$$

For the special case of $\beta = 90°$, Eqs. (11.64) and (11.65) become

$$r(\phi) = r_0 e^{\tan(\alpha)(\phi-\phi_0)} \tag{11.68}$$

and

$$x(\phi) = x_0. \tag{11.69}$$

For the special case of $\beta = 180°$, Eqs. (11.64) and (11.65) become

$$r(\phi) = r_0 \tag{11.70}$$

and

$$x(\phi) = x_0 - r_0 \tan(\alpha)(\phi - \phi_0). \tag{11.71}$$

We now combine segments of constant-angle spirals to construct continuous curves that are periodic in the variable ϕ. If the number of periods per 360° is the integer N, then the angular period of the curve is

$$\Delta\phi = \frac{360\deg}{N}. \tag{11.72}$$

Each of the N segments of the curve is assumed to consist of two subsegments having α-values which are equal in magnitude but opposite in sign. The angular width of each subsegment is

$$\delta\phi = \frac{\Delta\phi}{2}. \tag{11.73}$$

We define the two indices m and n, which refer to the mth subsegment of the nth segment, where $m = 0$ or 1 and $n = 0, 1, \ldots, N-1$. The central ϕ-value of the mth subsegment of the nth segment is

$$\phi_{m,n}^{cent} = \left(n + \frac{1}{2}m + \frac{1}{4}\right)\Delta\phi. \tag{11.74}$$

It is also convenient to compute the value of m and n corresponding to any particular value of ϕ between 0° and 360°:

$$n(\phi) = Int\left(\frac{\phi}{\Delta\phi}\right) \tag{11.75}$$

and

$$m(\phi) = Int\left(2\frac{\phi}{\Delta\phi}\right) - 2Int\left(\frac{\phi}{\Delta\phi}\right), \tag{11.76}$$

where the function $Int(x)$ is defined as the greatest integer that is less than or equal to x. Substitution of Eqs. (11.75) and (11.76) into Eq. (11.74) allows us to compute the central ϕ-value of the subsegment in which any given ϕ-value lies

$$\phi^{cent}(\phi) = \left[\frac{1}{2}Int\left(2\frac{\phi}{\Delta\phi}\right) + \frac{1}{4}\right]\Delta\phi. \tag{11.77}$$

When ϕ_0 in Eqs. (11.64) and (11.65) is replaced by $\phi^{cent}(\phi)$, we find that the periodic segmented spiral curve is specified by the formulas

$$r(\phi) = r_0 e^{\tan\{a[2m(\phi)-1]\}\sin(\beta)[\phi-\phi^{cent}(\phi)]} \tag{11.78}$$

and

$$x(\phi) = x_0 + \frac{r_0}{\tan(\beta)}\left\{e^{\tan\{a[2m(\phi)-1]\}\sin(\beta)[\phi-\phi^{cent}(\phi)]} - 1\right\}, \tag{11.79}$$

where the constant angle α has been multiplied by $[2m(\phi) - 1]$ to produce the desired alternation of the sign of the spiral angle for adjacent subsegments. Similarly, for the special case of $\beta = 0$°, we obtain

$$r(\phi) = r_0 \tag{11.80}$$

and

$$x(\phi) = x_0 + r_0 \tan\{\alpha[2m(\phi) - 1]\}[\phi - \phi^{cent}(\phi)]. \tag{11.81}$$

For $\beta = 90°$, we find that

$$r(\phi) = r_0 e^{\tan\{\alpha[2m(\phi)-1]\}[\phi-\phi^{cent}(\phi)]} \tag{11.82}$$

and

$$x(\phi) = x_0. \tag{11.83}$$

For $\beta = 180°$, we have

$$r(\phi) = r_0 \tag{11.84}$$

and

$$x(\phi) = x_0 - r_0 \tan\{\alpha[2m(\phi) - 1]\}[\phi - \phi^{cent}(\phi)]. \tag{11.85}$$

By converting r_0, x_0, α, and β into the continuous functions $r_0(\theta)$, $x_0(\theta)$, $\alpha(\theta)$, and $\beta(\theta)$ of the parameter θ, Eqs. (11.78) through (11.85) then provide us with the desired specification of the entire reflector surface. The specific forms of the functions $r_0(\theta)$ and $x_0(\theta)$ are equivalent to the radial and axial coordinates for the parameterization scheme described in the previous subsection for rotationally symmetric reflectors. The cone-angle function $\beta(\theta)$ is defined as the slope angle with respect to the x-axis of the local surface normal of the parametric curve defined by $r_0(\theta)$ and $x_0(\theta)$. We employ a finite-dimensional cubic-spline parameterization to describe α as a function of θ.

We now present a design example in which the parameters used to control the shape deviation, truncation, acceptance angle, and axial source position shift were the same 17 parameters used in the previous subsection. In addition, 10 degrees of freedom (knots) were allocated to the description of α as a function of θ, for a total of 27 design degrees of freedom in the optimization. The lower and upper parameter ranges for the original 17 variables were the same as used in the previous subsection. The lower and upper parameter ranges for the knots used to define α as a function of θ were 0° and 35°, respectively. When the shape deviations of the star lobes as a function of ϕ are sufficiently small, the ray tracing can be simplified by modeling the ϕ-dependent shape perturbations purely as slope perturbations imposed on a rotationally symmetric surface. This approximation was utilized in the ray tracing to obtain the results presented below.

Global optimization was used to determine the form of 3D OSC that maximizes the flux transferred from a 10-mm-diameter spherical Lambertian source into an emergent conical beam subtending a 30° half angle. The constraint that the source and the target must have equal étendue was enforced, which leads to an aperture-to-source area ratio of $1/\sin^2(30°) = 4$. Thus, a target disk diameter of 40 mm was used. The reflector surface was assumed to be loss-free and specular.

The resulting shape of the 3D OSC design is shown in Figures 11.13 and 11.14. In contrast to the axisymmetric solution discussed in the previous subsection, the source does not protrude behind the back of the reflector, and there exists a small clearance gap between the source and the reflector. The performance of this solution was computed using 200,000 rays, which provides accuracy better than 0.1% in efficiency and concentration. The design produces a value of 84.9% for both flux-
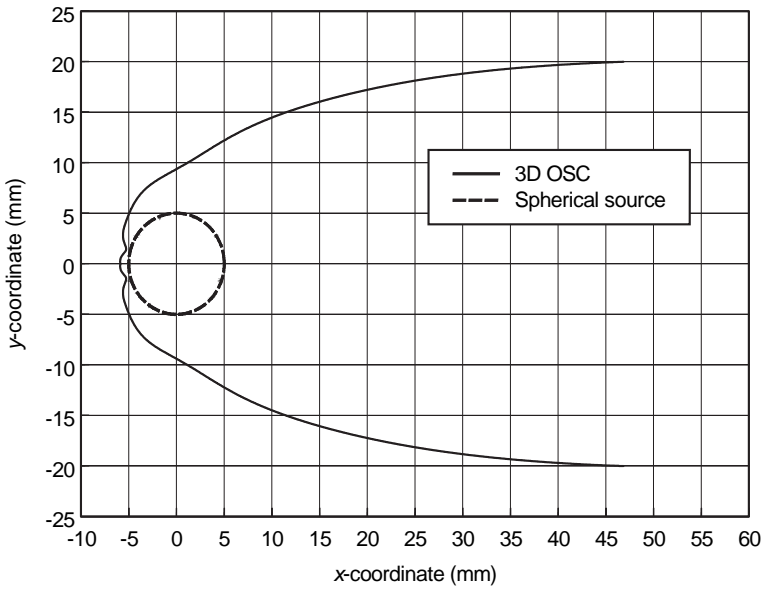
Figure 11.13   Shape profile of star concentrator optimized to transfer flux from a sphere to a 30° disk.
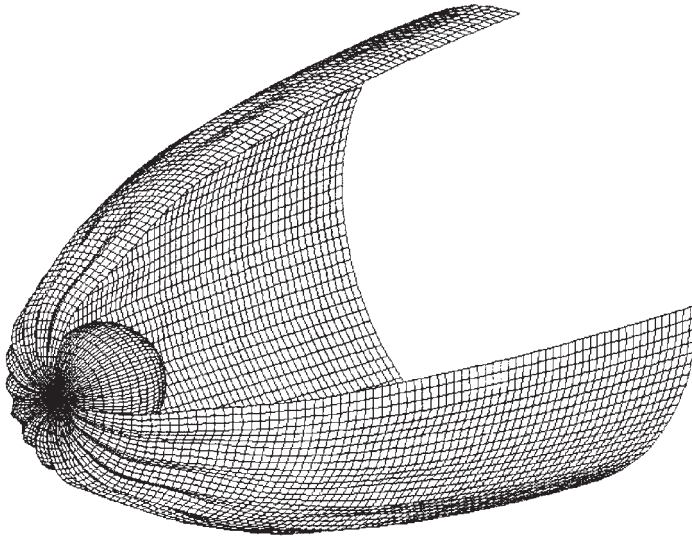
Figure 11.14   3D cut-away view of star concentrator.

transfer efficiency and concentration. The variation of the star-lobe angle along the arc-length of the reflector is depicted in Figure 11.15. The skewness distributions for the sphere, the disk, and the rays output within the required 30° half angle are depicted in Figure 11.16. Note that the output skewness distribution extends over the full skewness range of the disk target's skewness distribution. The symmetry-breaking star cross section has allowed the original skewness distribution of the source to be transformed into an output skewness distribution
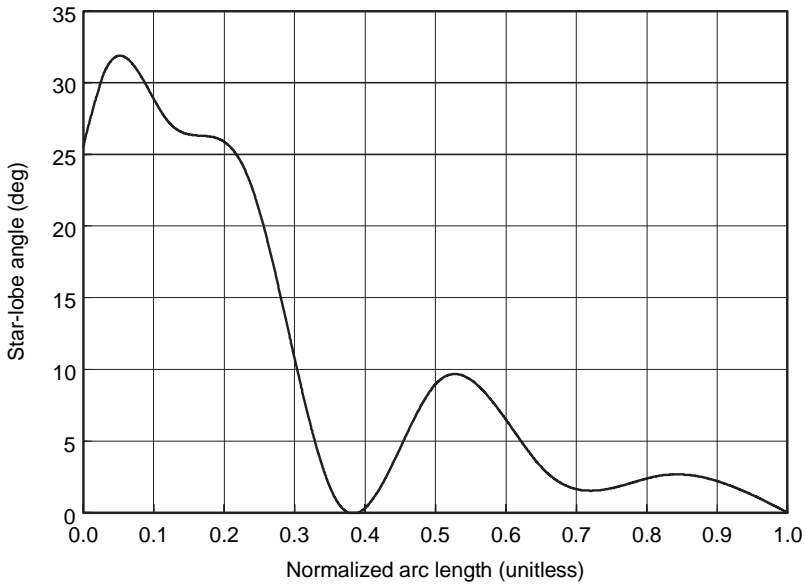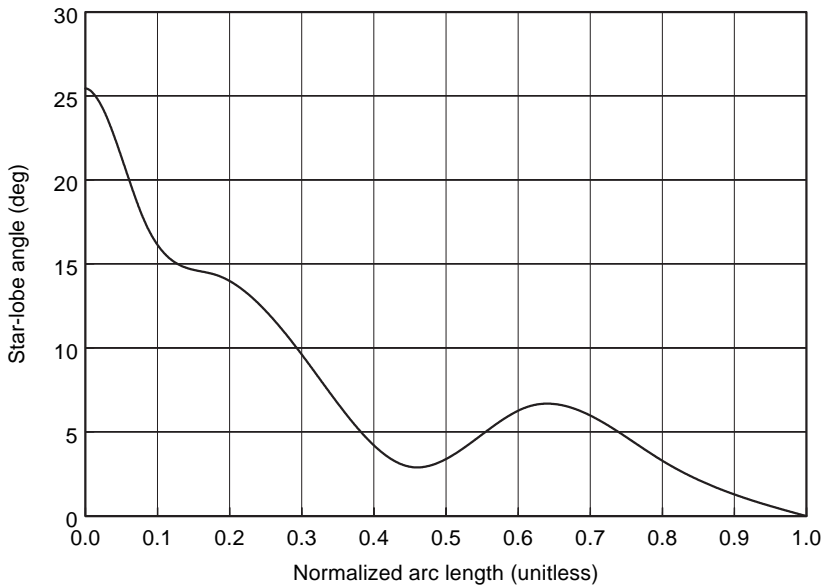
Figure 11.15   Star-lobe angle versus normalized arc length along reflector profile for 3D OSC optimized to transfer flux from a sphere to a 30° disk.
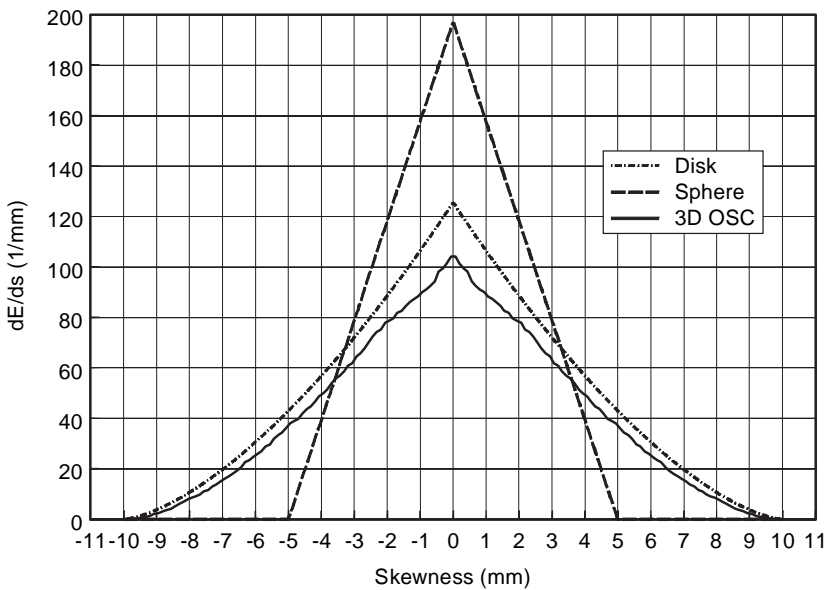
Figure 11.16   Skewness distributions for spherical source, disk target, and flux transferred by 3D OSC to a disk within a 30° half angle.

providing a superior match to that of the target. Figure 11.17 depicts the efficiency versus concentration curve for the 3D OSC in comparison with the curve for the truncated 30° involute CPC and the theoretical limiting curve for rotationally symmetrical systems. As previously discussed, the efficiency versus concentration curves for the two concentrators were generated by varying the
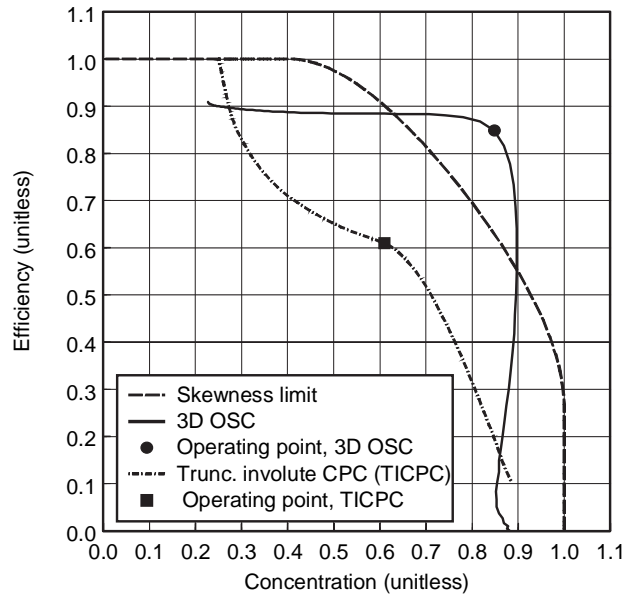
Figure 11.17   Efficiency versus concentration for 3D OSC optimized to transfer flux from a sphere to a disk within a 30° half angle.

acceptance half angle of the disk target, while keeping all other source, target, and concentrator characteristics fixed. The large range of concentration values over which the efficiency of the optimized design exceeds the limit for axisymmetric designs is an indication of the robustness of the design relative to variations in either efficiency or concentration from the nominal design point of equal étendue. In Figure 11.18 we compare the far-field intensity profile produced by the 3D OSC with that produced by the truncated 30° involute CPC. It is of interest to note that the 3D OSC provides significantly improved far-field intensity uniformity relative to the involute CPC design.

11.6.4   Nonaxisymmetric Concentrator for a Cylindrical Source and a Disk Target

We now consider the problem of designing a 3D OSC that maximizes the flux transfer from a homogeneous cylindrical source to a coaxial disk target of equal étendue (Shatz, Bortz, Ries, and Winston, 1997). We choose a height-to-radius ratio $H/R = 10$, which is typical of an incandescent filament. In this case the skewness mismatch between the source and target is strongly pronounced. The theoretical performance limit on efficiency for rotationally symmetric optics is 46.8% for the equal-étendue case, which means that this problem is more strongly affected by skewness mismatch than the case involving the spherical source.

We seek a star concentrator solution and state the design problem as one of determining the form of a 3D OSC that maximizes the flux transferred from a cylindrical Lambertian source into an emergent conical beam subtending a 30° half angle. As in the previous subsection, we use a target disk diameter of 40mm.
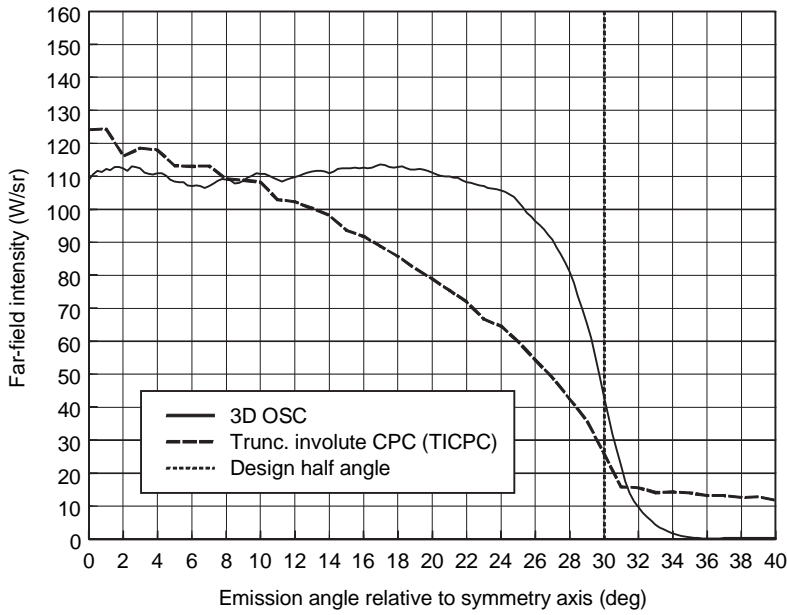
Figure 11.18   Far-field intensity versus emission angle for 3D OSC optimized to transfer flux from a sphere to a 30° disk.
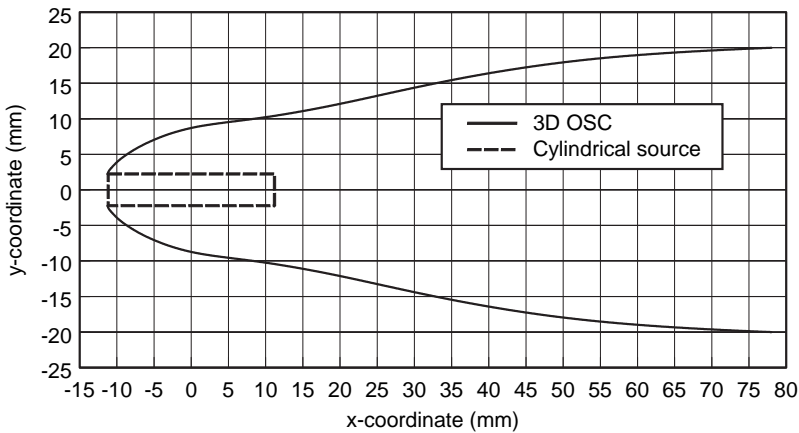
Figure 11.19   Shape profile of star concentrator, optimized to transfer flux from a cylinder to a 30° disk.

The equal-étendue requirement, along with the chosen height-to-radius ratio, leads to a source height and radius of 22.36 mm and 2.236 mm, respectively. The optimization was performed using the same parametrization scheme, number of optimization parameters, and parameter ranges as in the last subsection.

The resulting optimized shape of the 3D OSC design is shown in Figures 11.19 and 11.20. Note that the end of the source coincides precisely with the beginning of the reflector. The performance of this design was computed using 200,000 rays,
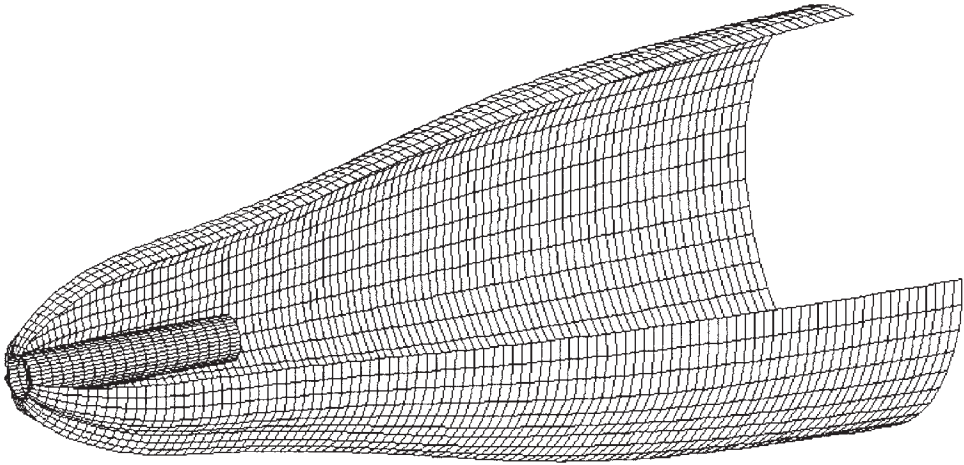
Figure 11.20   3D cutaway view of star concentrator.

Figure 11.21   Star-lobe angle versus normalized arc length along reflector profile, for 3D OSC optimized to transfer flux from a cylinder to a 30° disk.

which provided accuracy better than 0.1% in dilution and efficiency. The solution provides a 75.1% value of both efficiency and concentration. The variation of the star-lobe angle along the arc-length of the reflector is depicted in Figure 11.21. As was the case for the 3D OSC with the spherical source, we observe a generally decreasing behavior of the angle as arc length is increased. The skewness distributions for the cylinder, disk target, and flux transferred to the target within the 30° acceptance half angle are depicted in Figure 11.22. As with the 3D OSC for the spherical source, the output skewness distribution extends the full range of the disk's skewness distribution. Figure 11.23 depicts the efficiency versus
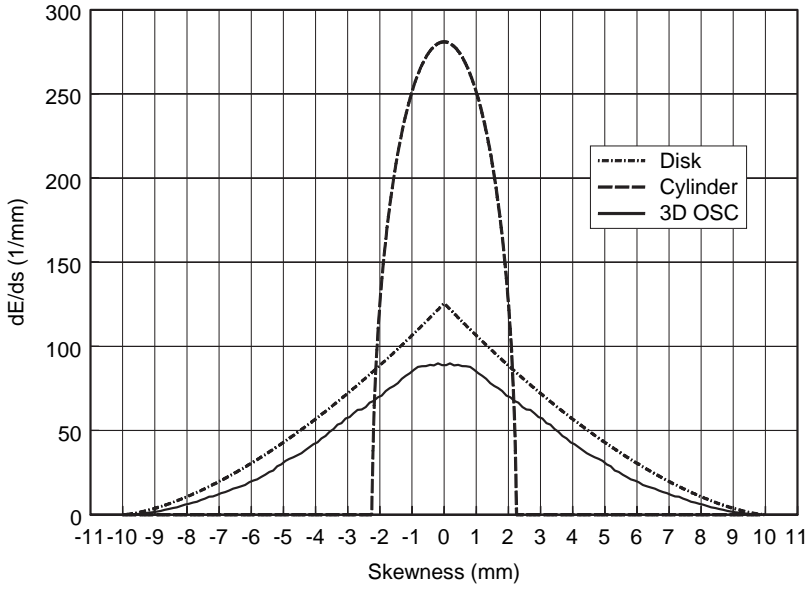
Figure 11.22   Skewness distributions for cylindrical source, disk target, and flux transferred by 3D OSC to disk within its 30° acceptance half angle.
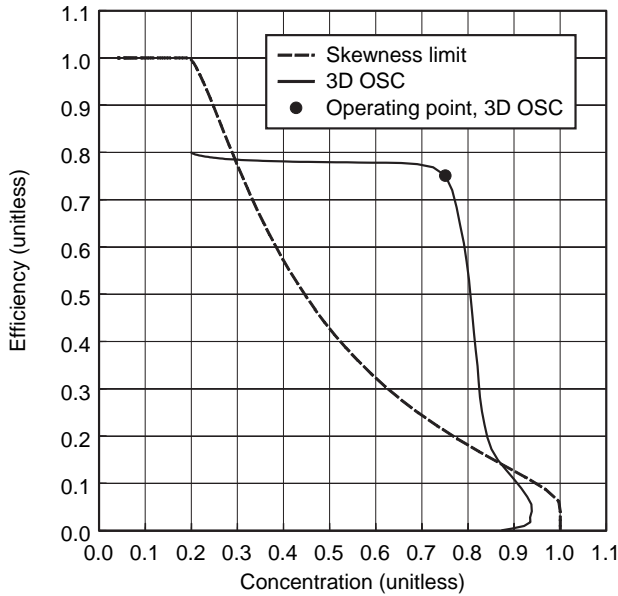
Figure 11.23   Efficiency versus concentration for 3D OSC optimized to transfer flux from a cylinder to a 30° disk.
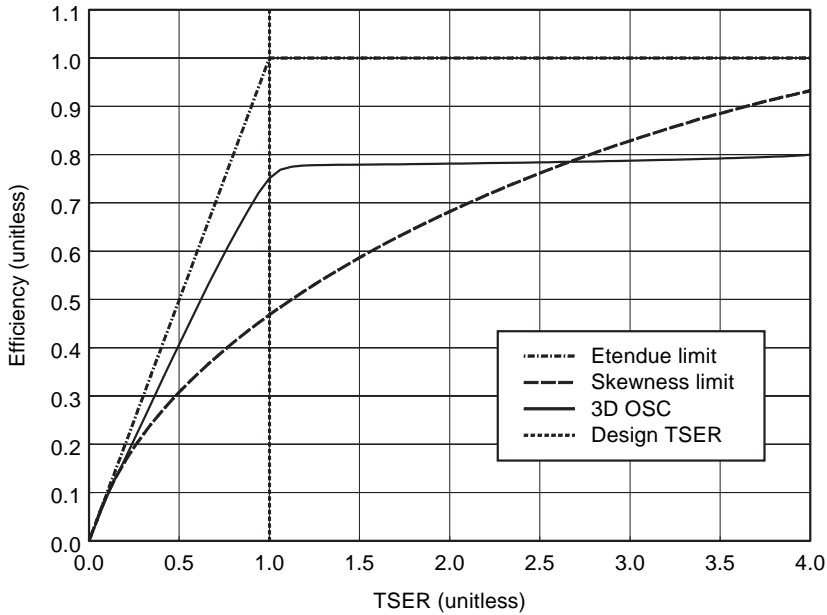
Figure 11.24   Efficiency versus target-to-source étendue ratio (TSER) for 3D OSC optimized to transfer flux from a cylinder to a 30° disk.

concentration curve for the 3D OSC design in comparison with the theoretical limiting curve for rotationally symmetrical systems. The efficiency versus concentration curve was generated by computing the efficiency and concentration for different values of the target's acceptance half angle, while holding all other system characteristics constant. It is of interest to note that the 3D OSC provides efficiency superior to the performance limit for axisymmetric systems over a large range of concentration values. Figure 11.24 depicts the efficiency versus TSER for the 3D OSC in relation to both the étendue limit, which is the upper limit for nonaxisymmetric optics, as well as the skewness limit, which is the upper limit for axisymmetric optics.

11.6.5 Nontracking Solar Concentrator with Broken Translational Symmetry

We now consider an example of a globally optimized design that uses microstructure ridges to overcome the performance limits imposed by translational symmetry (Bortz, Shatz, and Winston, 1997). The problem to be solved is that of designing a north-south-oriented nontracking solar concentrator in a material of unit refractive index, as discussed in Section 10.3.6.2. The latitudinal half angle of the solar radiation is 23.45°. The target is assumed to be a 20-mm-diameter cylindrical tube. We consider the equal-étendue case, with longitudinal half angle

$$\phi_0 = 50°. \tag{11.86}$$

Since the sun's angular position changes by 15° every hour, this half angle corresponds to a total daily operation interval of 6hr, 40min. We recall that the trans-

lational skew invariant places an upper limit on efficiency and concentration of 49.30%. It can be shown that the involute CPC operates at this upper limit for the case considered here, and therefore represents an optimal translationally symmetric solution. The equal-étendue assumption—and Eq. (11.86)—gives a source-to-target area ratio of

$$\rho_{stc,trg} = 2.648. \tag{11.87}$$

With this source-to-target area ratio, the involute CPC has a collection half angle of

$$\chi_{CPC} = \arcsin\left(\frac{1}{\rho_{src,trg}}\right) = 22.19°. \tag{11.88}$$

To improve performance beyond the 49.30% limit, we use a type of perturbed involute CPC. To break the symmetry, microstructure ridges are added to the surface such that the ridge peaks and valleys are aligned with planes oriented perpendicular to the translational symmetry axis. Rönnelid, Perers, and Karlsson (1994) considered the effect of constant tilt-angle microstructure ridges, superimposed on a conventional CPC. In the present case, however, the ridge tilt angle is a function of position along the reflector's shape profile. The ridge slopes alternate in sign as a function of position along the symmetry axis. The ridges are assumed to be sufficiently small that they can be considered to alter only the surface-normal vector as a function of position over the surface of the concentrator, without altering the macroscopic shape. In addition to the symmetry-breaking microstructure, the macroscopic shape profile itself is perturbed relative to the involute CPC. The parameterization scheme for the macroscopic shape perturbations is the same as that described in Section 11.6.2. The microstructure tilt angle of the ridges as a function of position along the shape profile is modeled as a cubic spline. The absolute value of the tilt angle is constrained to be less than or equal to 30°. Imposition of this angular limit improves manufacturability. In addition, it simplifies the ray tracing and reduces reflection losses by preventing more than two reflections from occurring for each ray incident on the region between two ridges.

Global optimization was used to maximize the flux-transfer efficiency of the nontranslationally symmetric solar collector. The computed efficiency and concentration of the optimized collector were found to be

$$\eta_{optim} = C_{optim} = 72.7\%, \tag{11.89}$$

which represents a 47.4% performance improvement relative to the baseline translationally symmetric concentrator. The performance of this optimized design is indicated by the square marker on the efficiency versus concentration plot shown in Fig. 11.25. The shape profile of the optimized collector, with and without traced rays, is shown in Fig. 11.26. For comparison, the profile of the baseline unperturbed involute CPC is depicted in Fig. 11.26 as a dashed line. A ray trace through the baseline concentrator is shown in Fig. 11.27. The rays visible to the right of the aperture in Figs. 11.26 and 11.27 represent rays that have been rejected by the concentrator. As expected from the higher efficiency of the optimized concentrator, fewer rejected rays are visible in Fig. 11.26 than in Fig. 11.27. A three-dimensional depiction of the optimized concentrator is shown in Fig. 11.28. To illustrate the microstructure geometry, the relative size of the symmetry-breaking ridges has been magnified by a large factor in this figure. The tilt angle of the
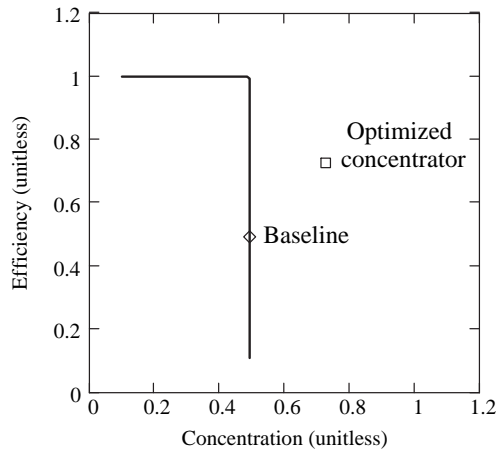
Figure 11.25   Plot of the efficiency limit as a function of the concentration limit for translationally symmetric nonimaging devices that transfer flux to a Lambertian target from a source having fixed longitudinal cutoffs parallel to symmetry axis with orthogonal fixed latitudinal cutoffs. The latitudinal angular half width of the source is $\theta_0 = 23.45°$. The performance achieved by the optimized nontranslationally symmetric concentrator and baseline concentrator for the case of $\phi_0 = 50°$ is indicated by the square and diamond markers, respectively.
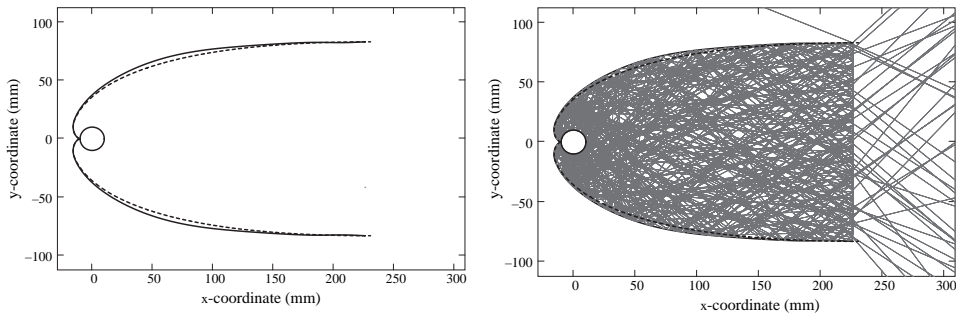
Figure 11.26   Shape profile of the optimized solar collector, with and without traced rays. The dashed line indicates the shape profile of the baseline translationally symmetric concentrator, which is an involute CPC.

ridges as a function of the normalized position along the optimized concentrator's shape profile is plotted in Fig. 11.29.

Plots of the transferred skewness distributions for the baseline and optimized concentrators are provided in Figs. 11.30 and 11.31. As expected, the skewness distribution for the translationally symmetric baseline design precisely matches the region of overlap of the skewness distributions of the source and target. The nontranslationally symmetric optimized design, however, has produced a broadening of the skewness distribution of the flux transferred to the target, thereby providing a better match to the target's distribution.

The ridge microstructure used to break the symmetry of the optimized concentrator can be thought of as a form of diffuser. As such, it has the effect of intro-
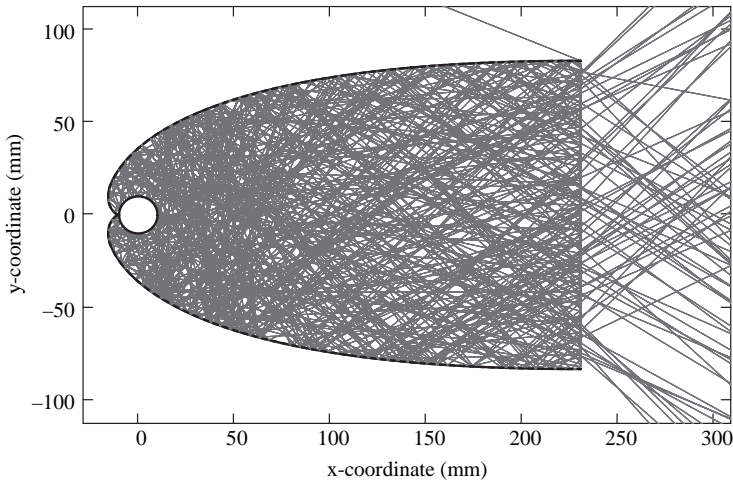
Figure 11.27   Shape profile of the baseline translationally symmetric concentrator (involute CPC), with traced rays.
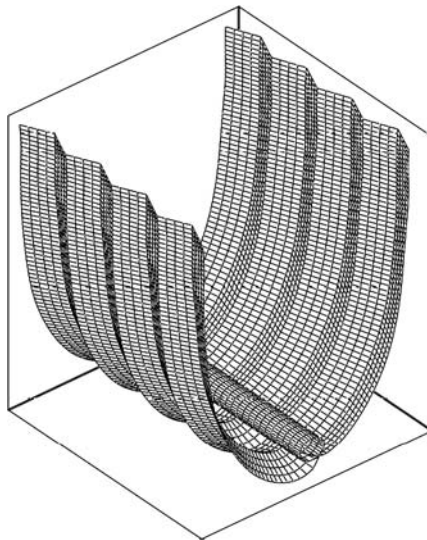
Figure 11.28   Three-dimensional depiction of the optimized non-translationally-symmetric concentrator. The symmetry-breaking ridge microstructure has been magnified by a large factor.

ducing a large number of small holes in the phase-space volume transferred to the target. Since these holes are not filled with radiation from the source, their presence reduces the achievable flux-transfer efficiency. The introduction of holes into the phase-space volume is analogous to the production of froth by injecting air bubbles into a liquid. In the same way that the presence of froth reduces the amount of liquid that can be poured into a container of a given volume, the presence of phase-space froth reduces the amount of étendue that can be transferred from a source to a target. For this reason, 100% flux-transfer efficiency from a
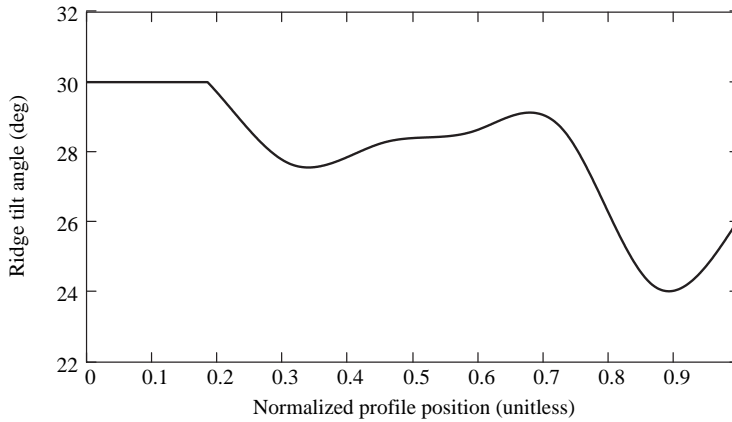
Figure 11.29   Tilt angle of the microstructure ridges as a function of the normalized position along the optimized concentrator's shape profile.
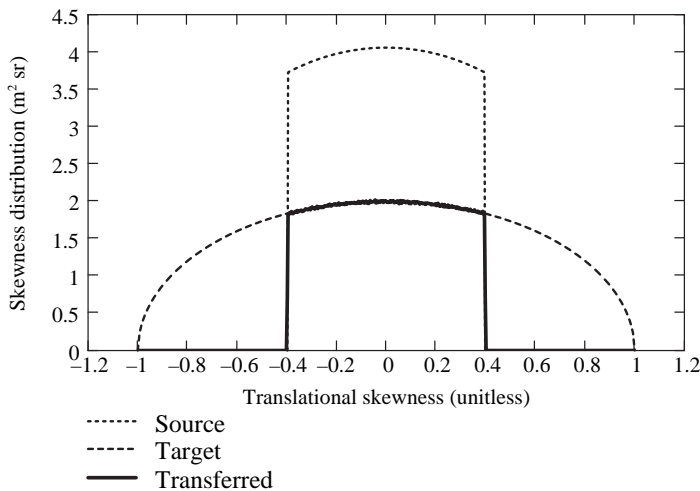
Figure 11.30   Source, target, and transferred skewness distributions for the baseline translationally symmetric concentrator.

source to a target of equal étendue can probably never be achieved by means of concentrators using microstructure ridges of the type considered herein. A related observation is that, since it is a diffuser, the microstructure can only produce a spreading of the skewness distribution. When the target's skewness distribution is narrower than that of the source, this spreading of the skewness distribution will only exacerbate the skewness mismatch between the source and the target, thereby reducing the achievable efficiency and concentration.

A second nontranslationally symmetric north-south-oriented nontracking solar collector was designed for the equal-étendue case with a longitudinal half-angle of
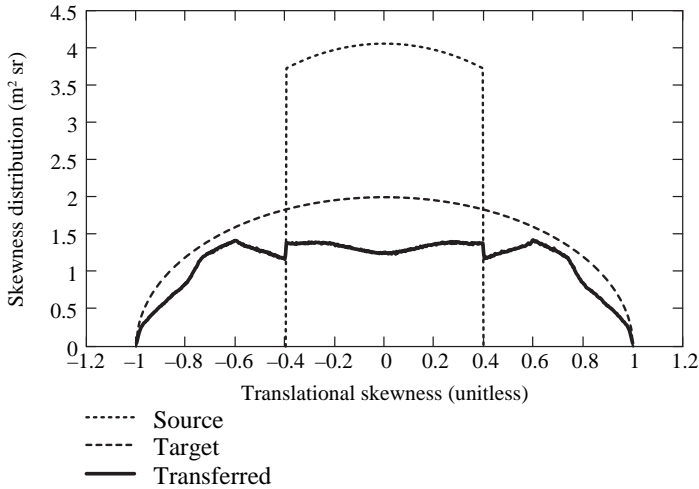
Figure 11.31   Source, target, and transferred skewness distributions for the optimized non-translationally symmetric concentrator.

$$\phi_0 = 90°. \tag{11.90}$$

This half angle is representative of dawn-to-dusk operation of the concentrator. The computed flux-transfer efficiency and concentration for this second optimized design were found to be

$$\eta_{optim} = C_{optim} = 68.4\%, \tag{11.91}$$

which represents a 38.7% performance improvement relative to the optimal translationally symmetric design.

REFERENCES

Arnold, V. I. (1989). *Mathematical Methods of Classical Mechanics*. Springer Verlag, New York.

Bortz, j., Shatz, N., and Winston, R. (2001). Performance limitations of translationally symmetric nonimaging devices. Proceedings of SPIE, Vol. 4446, 201–220.

Otten, R. H. J. M., and van Ginneken, L. P. P. P. (1989). *The Annealing Algorithm*. Kluwer Academic Publishers, Dordrecht.

Ratschek, H., and Rokne, J. (1998). *New Computer Methods for Global Optimization*. Ellis Horwood and John Wiley, Hoboken.

Rönnelid, M., Perers, B., and Karlsson, B. (1994). Optical properties of nonimaging concentrators with corrugated reflectors. Proceedings of SPIE, Vol. 2255, 595–602.

Rykowski, R., and Wooley, B. (1997). Source modeling for illumination design. Proceedings of SPIE, Vol. 3130, 204–208.

Shatz, N., and Bortz, J. (1995). An inverse engineering perspective on nonimaging optical design. Proceedings of SPIE, Vol. 2538, 136–156.

Shatz, N., Bortz, J., Ries, H., and Winston, R. (1997). Nonrotationally symmetric nonimaging systems that overcome the flux-transfer performance limit imposed by skewness conservation. Proceedings of SPIE, Vol. 3139, 76–85.

Torn, A., and Zilinskas, A. (1987). *Global Optimization, Lecture Notes in Computer Science*. Springer Verlag, New York.

A PARADIGM FOR A WAVE DESCRIPTION OF OPTICAL MEASUREMENTS

12.1   INTRODUCTION

Radiance, which is the density of radiative power in phase space has been the subject of a rich literature over the past 40 and more years (Littlejohn and Winston, 1993; Walther, 1968; 1973; Wolf, 1978). It is a tribute to the work of Adrian Walther, Emil Wolf, and many others that the development of a wave theory of radiance, known as "generalized radiance" continues to the present time. This chapter is complementary to this line of development in that we attempt to bridge the gap between theory and practical radiometry. Radiometric measurements are important in many branches of science and technology. For example, in illumination engineering, the visibilty of displays is quantified by radiometers. In astrophysics, radiometry in the far infrared has played a critical role in understanding the large space-time structure of the universe (Mather, Toral, and Memmati, 1986). Of course, in the short wave length limit where diffraction effects can be neglected, geometrical optics suffices and one can dispense with the technical difficulties that the wave property of light introduces. But it is precisely in the regime where diffraction effects cannot be neglected that the properties of the measuring instrument have to be taken into account. Moreover, in the absence of a consistent formalism which does take the diffraction property of the instrument into account, it may be difficult to assess the significance of such effects. "Back of the envelope" estimates of diffraction effects may not be reliable and the practical radiometer is left with little guidance as to the magnitude of such effects. For example, an excellent text on radiometry famously states that diffraction effects are "beyond the scope" of the book. In recent papers (Sun et al., 2002; Winston, Sun, and Littlejohn, 2002) we showed how the measurement of radiance can be understood in terms of the statistical properties of the electromagnetic field and the properties of the instrument. However the utility of this approach was limited by the availability of accessable instrument functions that represent the measuring apparatus. In the process we exhibited a remarkable analogy between the result of measuring radiance and the van Cittert-Zernike (VCZ) Theorem. In this

chapter we first give an overview of a wave description of the measurement of radiance. referring details to previous publications. This approach is validated by experimental results with highly sophisticated radiometers. The excellent agreement with the analytical model suggests that while our demonstration was confined to the measurement of radiance, it is likely that similar considerations apply to a wide class of optical measuerements, where diffraction effects are significant.

12.2   THE VAN CITTERT-ZERNIKE THEOREM

The well-known van Cittert-Zernike theorem states that for an incoherent, quasi-monochromatic source of radiation, the equal-time degree of coherence (2-point correlation function) $\Gamma(\mathbf{r}, \mathbf{r}')$ is proportional to the complex amplitude in a certain diffraction pattern: the amplitude at $\mathbf{r}$ formed by a spherical wave converging to $\mathbf{r}'$ and diffracted by an aperture the same size, shape, and location as the source (Born and Wolf, 1999; Mandel and Wolf, 1995). The source could, for example, be a thermal black body followed by a filter that selects a small wavelength range. A familiar geometry is a circular source. Then, apart from a normalizing factor. $\Gamma(\mathbf{r}, \mathbf{r}')$ in a transverse plane becomes the well-known Airy diffraction amplitude

$$\Gamma(\mathbf{r},\mathbf{r}') \ = \ (\text{const})F(ks\theta_s) \tag{12.1}$$

where $F(x) = 2J_1(x)/x$, $k = 2\pi/\lambda$, θ_s is the angle subtended by the source at $\mathbf{r}$ or $\mathbf{r}'$, and where $s = |\mathbf{r} - \mathbf{r}'|$. Recall that $\Gamma(\mathbf{r}, \mathbf{r}')$ has its first zero at $s_1 = 0.61\lambda/\theta_s$. For a numerical example, we consider terrestrial sunlight. Then θ_s is 4.7 mrad, so that for $\lambda = 0.5\,\mu\text{m}$, s_1 is approximately $65\,\mu\text{m}$. This is the scale of the transverse correlation of sunlight.

12.3   MEASURING RADIANCE

In a previous paper (Littlejohn and Winston, 1995), we examined the relationship between the generalized radiance and the measuring process. We showed how this process can be quantified by introducing the *instrument function*, which is a property of the measuring apparatus (Sun et al., 2002; Winston, Sun, and Littlejohn, 2002). We showed that the result of the measurement is represented by the quantity

$$Q = \text{Tr}\bigl(\hat{M}\hat{\Gamma}\bigr) \tag{12.2}$$

where $\hat{M}$ is a nonnegative-definite Hermitian operator that characterizes the measuring apparatus, and $\hat{\Gamma}$ is the 2-point correlation function of the incident light, viewed as an operator. The instrument function itself is a coordinate representation of the measurement operator $\hat{M}$—for example, its matrix element or its Weyl transform. The Weyl transform maps an operator to a Wigner function (for a discussion of the Wigner-Weyl formalism in optics, see Littlejohn and Winston, 1995). It is appropriate to associate Q with the signal. We then derived an analytical form for the instrument function for a simple radiometer in one space dimension.

One difficulty in using Eq. (12.2) is that it may not be easy to compute the instrument function. Although the one-dimensional calculation in Littlejohn and

Winston (1995) was not too hard, we do not expect it to be easy to compute the instrument function for many realistic radiometers, which are two-dimensional in cross section and that may have complicated geometry. Therefore, we have considered other means for determining the instrument function. In a previous publication (Winston and Littlejohn, 1997) we considered the possibility that the instrument function could be measured. In this chapter we present an alternative approach. That is, we point out a physical interpretation of the instrument function that is similar to the van Cittert-Zernike theorem. We do this initially by working through the example of a simple "pin hole" radiometer, and then we comment about generalizations.

Radiance is the power per unit volume in phase space. Therefore, an instrument for measuring radiance (called a radiometer) has to select a window function in phase space. For measurements close to the diffraction limit, the exact shape of the window function is not critical. For this reason we examine a simple radiometer, illustrated in Figure 12.1. The dotted line in the figure is the axis of the radiometer. Light enters from the left and passes through the circular pin hole of radius a. It then passes through a drift space of length L, before passing through another circular aperture of radius b. We assume $L \gg a$, b, so the rays are paraxial. The detector is assumed to measure the total power passing through the aperture b and can be thought of as composed of tiny, densely packed, independent absorbing particles (which is a fairly good approximation to what commonly used detectors like photon detectors, thermal detectors, or photographic film do).

As explained in Littlejohn and Winston (1995) the effect of the radiometer on the radiation field is described by the operator

$$\hat{P} = \hat{A}(b)\hat{D}(L)\hat{A}(a) \tag{12.3}$$

which maps the wave field at the entrance aperture a into the wave field at the exit aperture b. Here $\hat{A}(a)$ is the aperture or "cookie cutter" operator representing the pinhole, $\hat{D}(L)$ is the Huygens-Fresnel operator representing the drift space, and $\hat{A}(b)$ is the aperture operator for the aperture b. In the approximation $L \gg \lambda$ the drift operator has the kernel (or matrix element)
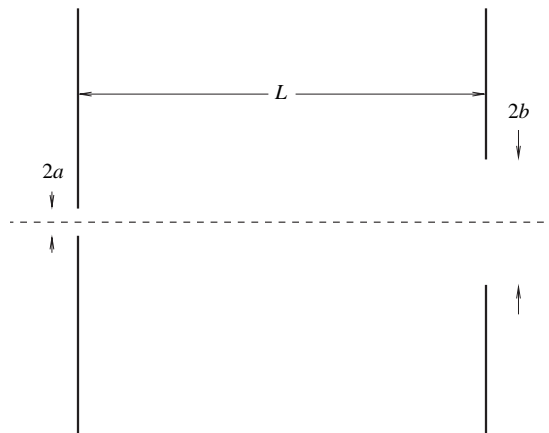
Figure 12.1   A pinhole radiometer. The dotted line is the axis. Light enters from the left, passing through circular pinhole a, drift space of length L, and finally circular aperture b.

$$\left\langle r_\perp \middle| \hat{D}(L) \middle| r_\perp' \right\rangle = -\frac{ikL}{2\pi} \frac{e^{ikR}}{R^2} \tag{12.4}$$

where $\mathbf{r}_\perp = (x, y)$, $\mathbf{r}_\perp' = (x', y')$, $\mathbf{r} = (x, y, z)$, $\mathbf{r}' = (x', y', z')$, $L = z - z'$ and $R = |\mathbf{r} - \mathbf{r}'|$. Here z is the coordinate along the optical axis and it is assumed that $z > z'$. If in addition we assume rays are paraxial ($r_\perp$, $r_\perp' \ll L$), then the kernel can be written

$$\left\langle r_\perp \middle| \hat{D}(L) \middle| \mathbf{r}_\perp' \right\rangle = -\frac{ik}{2\pi L} e^{ikL} \exp\left(\frac{ik}{2L} |\mathbf{r}_\perp - \mathbf{r}_\perp'|^2 \right) \tag{12.5}$$

Following the methods of Littlejohn and Winston (1995), the instrument is represented by the operator $\hat{M} = \hat{P}^\dagger \hat{P}$ whose matrix elements are

$$\left\langle \mathbf{r}_\perp \middle| \hat{M} \middle| \mathbf{r}_\perp' \right\rangle = \left(\frac{kL}{2\pi} \right)^2 \int_{|\mathbf{r}_\perp''| \leq b} d^2 \mathbf{r}_\perp'' \frac{\exp[ik(R_1 - R_2)]}{R_1^2 R_2^2} \tag{12.6}$$

where $R_1 = |\mathbf{r}' - \mathbf{r}''|$, $R_2 = |\mathbf{r} - \mathbf{r}''|$, and the integration is over the detector area. In Eq. (12.6) the transverse variables $\mathbf{r}_\perp$ and $\mathbf{r}_\perp'$ are understood to lie in the entrance plane (the pinhole), so that $r_\perp$, $r_\perp' \leq a$. If this condition is not met, the matrix element is understood to be zero.

The expression (12.6) is identical (up to constants) to the mutual intensity evaluated at aperture a of a uniform, delta-correlated source at aperture b (the location of the detector). Thus, to form the van Cittert-Zernike interpretation of the instrument function, we replace the detector (at aperture b, in this example) by a delta-correlated source, and measure the radiation field emanating from the entrace aperture of the instrument (the pinhole in this example). The instrument function (at a given plane) is then proportional to the amplitude at $\mathbf{r}'$ formed by a spherical wave converging to $\mathbf{r}$ and diffracted by an aperture the same size, shape, and location as the detector. The detector emulates a delta-correlated source. In a sense, this model involves running the radiometer backwards (exchanging the detector for a source).

This interpretation applies also to other radiometers, for example, those with lenses (see Figure 12.2). The essential property is that the operator $\hat{P}^\dagger$ should serve
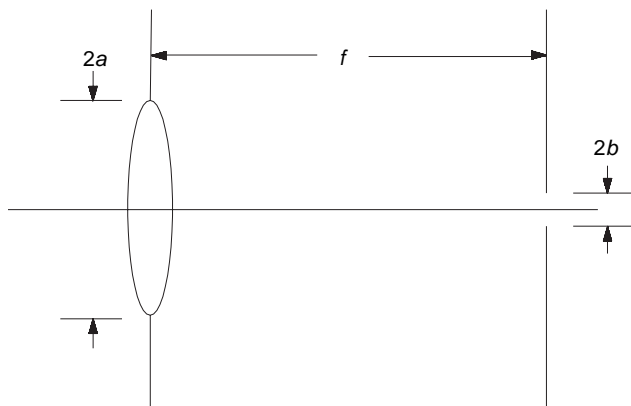
Figure 12.2   A practical radiometer. The pinhole is replaced by a lens. The drift space length is essentially the focal length for distant objects $L \approx f$.

as a propagator for wave fields traveling to the left (in the negative z direction), just as $\hat{P}$ serves as a propagator for waves travelling to the right. The situation is rather much like time reversal in quantum mechanics. Not all time evolutions in quantum mechanics are time reversal invariant (only those for which the Hamiltonian commutes with time reversal). In the case of optical fields, it is a kind of "z-reversal" that we need. Lenses, drift spaces, and apertures are "z-reversal invariant," as long as evanescent waves can be ignored. We remark that the same conditions apply to the usual van Cittert-Zernike theorem.

12.4   NEAR-FIELD AND FAR-FIELD LIMITS

It is useful to examine the pinhole radiometer and the formalism we have presented in two limiting cases. First, however, we explain some notation regarding the correlation operator and function (see Littlejohn and Winston, 1995, for more details). The mutual coherence is defined by $\Gamma(\mathbf{r}, \mathbf{r}') = \overline{\psi(\mathbf{r})\psi^*(\mathbf{r}')}$, where the overbar means a statistical or ensemble average, and where we use a scalar model for the wave field ψ, which in electromagnetic applications can be loosely identified with one of the components of the electric field. When $z = z'$ we can associate the mutual coherence with an operator $\hat{\Gamma}(z)$ by $\Gamma(\mathbf{r}, \mathbf{r}') = \Gamma(\mathbf{r}_\perp, z; \mathbf{r}'_\perp, z) = \langle \mathbf{r}_\perp | \hat{\Gamma}(z) | \mathbf{r}'_\perp \rangle$, so that $\hat{\Gamma}(z) = \overline{|\psi(z)\psi(z)|}$. Thus,

$$\operatorname{Tr}\hat{\Gamma}(z) = \int d^2\mathbf{r}_\perp \overline{|\psi(\mathbf{r}_\perp, z)|^2} \tag{12.7}$$

Thus, if we identify $|\psi|^2$ with 4π times the energy density, then for paraxial rays $(c/4\pi)\operatorname{Tr}\hat{\Gamma}(z)$ is the power crossing the plane z.

Now let us consider the case that a uniform, thermal source is very close to the radiometer, which is useful for normalizing the signal. Then at the entrance aperture the mutual intensity is proportional to a delta function, $\Gamma(\mathbf{r}, \mathbf{r}') = I_0\lambda^2\delta(\mathbf{r}_\perp - \mathbf{r}'_\perp)$, where I_0 is a constant with dimensions of energy/vol. A thermal source is not really delta correlated, of course; the spatial correlation is really a sine function with a width of the order of a wavelength. It is for this reason that we insert the factor of λ^2 into the formula for the mutual coherence, so that if we need to set $\mathbf{r}_\perp = \mathbf{r}'_\perp$ for the purposes of taking the trace, we can interpret $\lambda^2\delta(0)$ as being of order unity. In any case, when we compute the signal according to Eq. (12.2), we obtain

$$\operatorname{Tr}(\hat{M}\hat{\Gamma}) = I_0\lambda^2\operatorname{Tr}\hat{M} = I_0\lambda^2 N_i \tag{12.8}$$

where we set

$$\operatorname{Tr}\hat{M} = N_i \tag{12.9}$$

for the number of phase space cells in the acceptance region of the instrument. The trace of $\hat{M}$ is easy to compute. We first make the paraxial approximation in Eq. (12.6), which gives

$$\left\langle \mathbf{r}_\perp | \hat{M} | \mathbf{r}'_\perp \right\rangle = \left(\frac{k}{2\pi L}\right)^2 \int d^2\mathbf{r}''_\perp \exp\left[\frac{ik}{L}\mathbf{r}''_\perp \cdot (\mathbf{r}_\perp - \mathbf{r}'_\perp)\right] \tag{12.10}$$

where the $\mathbf{r}''_\perp$ integration is taken over a circle of radius b. Now setting $\mathbf{r}_\perp = \mathbf{r}'_\perp$ and integrating $\mathbf{r}_\perp$ over a circle of radius a, we obtain

$$N_i = \mathrm{Tr}\hat{M} = \left(\frac{kab}{2L}\right)^2 = \left(\frac{\pi a \theta_0}{\lambda}\right)^2 \tag{12.11}$$

where $\theta_0 = b/L$.

The number N_i has a simple interpretation. A phase space cell in the 4-dimensional $\mathbf{k}_\perp - \mathbf{r}_\perp$ phase space has volume $(2\pi)^2$. As viewed from the standpoint of the exit aperture b of the radiometer, the rays passing through each point of the aperture b occupy a solid angle of $\pi(a/L)^2$, or a region of $\mathbf{k}_\perp$-space of area $\pi(ka/L)^2$. The region of $\mathbf{r}_\perp$-space is just the exit aperture, of area πb^2. Multiplying these areas and dividing by $(2\pi)^2$ gives precisely N_i.

Next we consider the case of a very distant thermal source, which effectively produces a coherent plane wave at the entrance aperture, say, $\sqrt{I_0}\sqrt{\pi\theta_s^2}\,e^{ikz}$, so that $\langle\mathbf{r}_\perp|\hat{\Gamma}|\mathbf{r}'_\perp\rangle = I_0(\pi\theta_s^2)$. The dimensionless factor $\pi\theta_s^2$ will be explained below. Using this and Eq. (12.10), we obtain

$$Q = \mathrm{Tr}\left(\hat{M}\hat{\Gamma}\right) = I_0\pi\theta_s^2\left(\frac{k}{2\pi L}\right)^2\int d^2\mathbf{r}_\perp d^2\mathbf{r}'_\perp d^2\mathbf{r}''_\perp \exp\left(\frac{ik}{L}\mathbf{r}_\perp\cdot\mathbf{r}''_\perp\right)\exp\left(-\frac{ik}{L}\mathbf{r}'_\perp\cdot\mathbf{r}''_\perp\right). \tag{12.12}$$

It is easiest to do the $\mathbf{r}_\perp$ and $\mathbf{r}'_\perp$ integrals first (both of which go out to radius a). These are identical, and are given by

$$\int d^2\mathbf{s}\exp\left(\pm\frac{ik}{L}\mathbf{s}\cdot\mathbf{r}''_\perp\right) = \pi a^2 F(kar''_\perp/L) \tag{12.13}$$

where $\mathbf{s} = \mathbf{r}_\perp$ or $\mathbf{r}'_\perp$. This leaves only the $\mathbf{r}''_\perp$ integration (taken out to radius b)

$$Q = \mathrm{Tr}\left(\hat{M}\hat{\Gamma}\right) = I_0\pi\theta_s^2\left(\frac{\pi a^2}{\lambda L}\right)^2\int d^2\mathbf{r}''_\perp\left[\frac{2J_1(kar''_\perp/L)}{(kar''_\perp/L)}\right] \tag{12.14}$$

which agrees with the expected fraction of the Airy diffraction pattern contained by the detector of radius b.

12.5   A WAVE DESCRIPTION OF MEASUREMENT

We begin with the statistical properties of the incident wave field. The two-point correlation function at the source plane is

$$\langle\mathbf{r}_\perp|\hat{\Gamma}(0)|\mathbf{r}'_\perp\rangle = I_0\delta(\mathbf{r}_\perp - \mathbf{r}'_\perp) \tag{12.15}$$

with $\mathbf{r}_\perp$ and $\mathbf{r}'_\perp$ inside the source, σ, and equal to 0 otherwise. The matrix elements of the Fresnel free space propagator of a distance L are given by

$$\langle\mathbf{r}_\perp|\hat{D}(L)|\mathbf{r}'_\perp\rangle = \frac{-ik}{2\pi L}e^{ikL}\exp\left(\frac{ik}{2L}|\mathbf{r}_\perp - \mathbf{r}'_\perp|^2\right) \tag{12.16}$$

where $k = 2\pi/\lambda$. The two-point correlation at a plane a distance L from the source in the Fresnel diffraction regime is

$$\left\langle \mathbf{r}_\perp \big| \hat{\Gamma}(L) \big| \mathbf{r}_\perp' \right\rangle = I_0 \int d^2 \mathbf{s}_\perp d^2 \mathbf{s}_\perp' \left\langle \mathbf{r}_\perp \big| \hat{D}(L) \big| \mathbf{s}_\perp \right\rangle \left\langle \mathbf{s}_\perp \big| \hat{\Gamma}(0) \big| \mathbf{s}_\perp' \right\rangle \left\langle \mathbf{s}_\perp' \big| \hat{D}^\dagger(L) \big| \mathbf{r}_\perp' \right\rangle$$

$$= I_0 \int_\sigma d^2 \mathbf{s}_\perp \left\langle \mathbf{r}_\perp \big| \hat{D}(L) \big| \mathbf{s}_\perp \right\rangle \left\langle \mathbf{s}_\perp \big| \hat{D}^\dagger(L) \big| \mathbf{r}_\perp' \right\rangle$$

$$= I_0 \exp\!\left(\frac{ik}{2L} \left(|\mathbf{r}_\perp|^2 - |\mathbf{r}_\perp'|^2 \right) \right) \left(\frac{k}{2\pi L} \right)^2 \int_\sigma d^2 \mathbf{s}_\perp \exp\!\left(\frac{ik}{L} \mathbf{s}_\perp \cdot (\mathbf{r}_\perp - \mathbf{r}_\perp') \right)$$

$$= I_0 \exp\!\left(\frac{ik}{2L} \left(|\mathbf{r}_\perp|^2 - |\mathbf{r}_\perp'|^2 \right) \right) F_r(\mathbf{r}_\perp - \mathbf{r}_\perp'). \tag{12.17}$$

In Eq. (12.17), $F_r(\mathbf{r}_\perp - \mathbf{r}_\perp')$ is the Fourier transform of the source area; it is the VCZ result for the far field two-point correlation function.

We will be testing two source geometries, circular and square. For the square geometry F_r is

$$F_i(\mathbf{r}_\perp - \mathbf{r}_\perp') = \left(\frac{1}{\pi} \right)^2 \frac{1}{(x - x')(y - y')} \sin\!\left(\frac{kd}{2L}(x - x') \right) \sin\!\left(\frac{kd}{2L}(y - y') \right) \tag{12.18}$$

with d the linear dimension of the source. For the circular geometry F_r is

$$F_i(\mathbf{r}_\perp - \mathbf{r}_\perp') = \left(\frac{kd}{2L} \right) \frac{J_1\!\left(\dfrac{kd}{2L} |\mathbf{r}_\perp - \mathbf{r}_\perp'| \right)}{2\pi |\mathbf{r}_\perp - \mathbf{r}_\perp'|} \tag{12.19}$$

with d the diameter of the source.

12.6 FOCUSING AND THE INSTRUMENT OPERATOR

The matrix elements for the lens operator are given by

$$\left\langle \mathbf{r}_\perp \big| \hat{L}(f) \big| \mathbf{r}_\perp' \right\rangle = \exp\!\left(\frac{-ik}{2f} |\mathbf{r}_\perp|^2 \right) \delta(\mathbf{r}_\perp - \mathbf{r}_\perp') \tag{12.20}$$

The instrument operator is

$$\hat{M} = \hat{P}^\dagger \hat{P} \tag{12.21}$$

where $\hat{P}$ is the propagator from the aperture, $\hat{A}(a)$, to the detector $\hat{A}(b)$.

$$\hat{P} = \hat{A}(b)\hat{D}(l)\hat{L}(f)\hat{A}(a) \tag{12.22}$$

The matrix elements for $\hat{A}(a)$ are given by

$$\left\langle \mathbf{r}_\perp \big| \hat{A}(a) \big| \mathbf{r}_\perp' \right\rangle = \delta(\mathbf{r}_\perp - \mathbf{r}_\perp') \tag{12.23}$$

with $\mathbf{r}_\perp$ and $\mathbf{r}_\perp'$ inside the aperture, a, and equal to 0 otherwise. The matrix elements for $\hat{A}(b)$ are given by

$$\left\langle \mathbf{r}_\perp \big| \hat{A}(b) \big| \mathbf{r}_\perp' \right\rangle = \delta(\mathbf{r}_\perp - \mathbf{r}_\perp') \tag{12.24}$$

with $\mathbf{r}_\perp$ and $\mathbf{r}_\perp'$ inside the detector, b, and equal to 0 otherwise.

From Eqs. (12.16) and (12.20) we have

$$\left\langle \mathbf{r}_\perp \middle| \hat{D}(l)\hat{L}(f) \middle| \mathbf{r}'_\perp \right\rangle = \frac{-ik}{2\pi d} e^{ikl} \exp\!\left(\frac{ik}{2l} |\mathbf{r}_\perp - \mathbf{r}'_\perp|^2 \right) \exp\!\left(\frac{-ik}{2f} |\mathbf{r}'_\perp|^2 \right) \qquad (12.25)$$

This gives

$$\left\langle \mathbf{r}_\perp \middle| \hat{P} \middle| \mathbf{r}'_\perp \right\rangle = \frac{-ik}{2\pi d} e^{ikl} \exp\!\left(\frac{ik}{2l} |\mathbf{r}_\perp - \mathbf{r}'_\perp|^2 \right) \exp\!\left(\frac{-ik}{2f} |\mathbf{r}'_\perp|^2 \right) \qquad (12.26)$$

with $\mathbf{r}_\perp$ inside the detector, $\mathbf{r}'_\perp$ inside the aperture, and equal to 0 otherwise.

The matrix elements of $\hat{M}$ are given by

$$\left\langle \mathbf{r}_\perp \middle| \hat{M} \middle| \mathbf{r}'_\perp \right\rangle = \int_b d^2\mathbf{s}_\perp \left\langle \mathbf{s}_\perp \middle| \hat{P} \middle| \mathbf{r}_\perp \right\rangle^{*} \left\langle \mathbf{s}_\perp \middle| \hat{P} \middle| \mathbf{r}'_\perp \right\rangle \qquad (12.27)$$

The integration is over the detector area.

$$\left\langle \mathbf{r}_\perp \middle| \hat{M} \middle| \mathbf{r}'_\perp \right\rangle = \exp\!\left(\frac{ik}{2f} \left(|\mathbf{r}_\perp|^2 - |\mathbf{r}'_\perp|^2 \right) \right) \exp\!\left(\frac{-ik}{2l} \left(|\mathbf{r}_\perp|^2 - |\mathbf{r}'_\perp|^2 \right) \right)$$

$$\left(\frac{k}{2\pi d} \right)^2 \int_b d^2\mathbf{s}_\perp \exp\!\left(\frac{ik}{l} \mathbf{r}_\perp \cdot (\mathbf{r}_\perp - \mathbf{r}'_\perp) \right) \qquad (12.28)$$

$$= \exp\!\left(\frac{ik}{2f} \left(|\mathbf{r}_\perp|^2 - |\mathbf{r}'_\perp|^2 \right) \right) \exp\!\left(\frac{-ik}{2l} \left(|\mathbf{r}_\perp|^2 - |\mathbf{r}'_\perp|^2 \right) \right) F_i(\mathbf{r}_\perp - \mathbf{r}'_\perp) \quad (12.29)$$

with $\mathbf{r}_\perp$ and $\mathbf{r}'_\perp$ inside the aperture, a, and equal to 0 otherwise. Here F_i is

$$F_i(\mathbf{r}_\perp - \mathbf{r}'_\perp) = \left(\frac{k}{2\pi d} \right)^2 \int_b d^2\mathbf{s}_\perp \exp\!\left(\frac{ik}{l} \mathbf{s}_\perp \cdot (\mathbf{r}_\perp - \mathbf{r}'_\perp) \right) \qquad (12.30)$$

If the camera is focused on the source

$$\frac{1}{l} + \frac{1}{L} = \frac{1}{f} \qquad (12.31)$$

where L is the distance between the camera entrance aperture and the source. In this case the non-zero matrix elements are given by

$$\langle \mathbf{r}_\perp | M' | \mathbf{r}'_\perp \rangle = \exp\!\left(\frac{ik}{2L} \left(|\mathbf{r}_\perp|^2 - |\mathbf{r}'_\perp|^2 \right) \right) F_i(\mathbf{r}_\perp - \mathbf{r}'_\perp) \qquad (12.32)$$

The VCZ result for the nonzero matrix elements of $\hat{M}$ are given by Eq. (12.30) with the camera focused at infinity, $l = f$. This results in a slight difference between the VCZ result and the general result in time phase space volume of $\hat{M}$ defined by

$$N_i = \mathrm{Tr}\!\left(\hat{M} \right) \qquad (12.33)$$

with N_i being time number of phase space cells. The infrared camera we used in the experiment has a circular entrance aperature and a square detector. We thus have

$$F_i(\mathbf{r}_\perp - \mathbf{r}'_\perp) = \left(\frac{1}{\pi} \right)^2 \frac{1}{(x - x')(y - y')} \sin\!\left(\frac{kb}{l} (x - x') \right) \sin\!\left(\frac{kb}{l} (y - y') \right) \qquad (12.34)$$

with $\mathbf{r}_\perp$ and $\mathbf{r}'_\perp$ inside the circular entrance aperture of radius a and equal to zero otherwise. The linear dimension of the square detector is $2b$. From Eq. (12.32),

$$N_i = \pi a^2 \left(\frac{2\theta_i}{\lambda} \right)^2 \tag{12.35}$$

where $2\theta_i = 2b/l$ is the full acceptance angle. For all practical purposes, the difference in the values of N_i for the focused case and with $l = f$ is negligible. Henceforth, we will mean $l = f$ when referring to N_i.

12.7   MEASUREMENT BY FOCUSING THE CAMERA ON THE SOURCE

From Eqs. (17) and (32), the detected signal, Q, is

$$Q = \mathrm{Tr}\left(\hat{M} \hat{\Gamma}(L) \right) = I_0 \int d^2\mathbf{r}_\perp d^2\mathbf{r}'_\perp F_i(\mathbf{r}'_\perp - \mathbf{r}_\perp) F_r(\mathbf{r}_\perp - \mathbf{r}'_\perp) \tag{12.36}$$

The normalized signal, Q_n, is defined as

$$Q_n = \frac{Q}{I_0 N_i} \tag{12.37}$$

12.8   EXPERIMENTAL TEST OF FOCUSING

The matrix elements of the instrument operator modeling the infrared camera used in the experiment are given by Eq. (12.32). The experiment was conducted by focusing the camera on the source. The results of the experiments are compared with Eqs. (12.36) and (12.37).

   The following is the protocol for processing the data. The value of the signal when the detector of the camera is flood illuminated by the blackbody radiation subtracted by the value of the signal when the detector is flood illuminated by the background is used as the normalization. The measured normalized signal is obtained by first subtracting the detected signal by the background and then divided by the normalization. The normalized signal is compared with theory.

   The camera has an interference filter and a HgCdTd detector giving a wavelength window with a peak at $\lambda = 8.8\,\mu$m and $\Delta\lambda \simeq \pm 0.75\,\mu$m. The square HgCdTd detector is of dimensions $75\,\mu$m $\times 75\,\mu$m. The focal length of the camera is $f = 18.99$mm. In Eq. (12.32) F_i is then given by Eq. (12.30), with $b = (75/2)\mu$m.

   Two circular aperture plates of radii $a = 0.136$in and $a = 0.272$in were placed in front of the camera aperture. We have $N_i = 1.888$ for the $a = 0.136$ in aperture and $N_i = 7.551$ for the $a = 0.272$ in aperture. The face of each plate facing the lens was painted with high emissive paint to provide the background. Time source size and shape were controlled by placing aluminum masks with either square or circular apertures over a blackbody source set at $T \sim 500°$C.

   Define N_r as,

$$N_r = \int_a d^2\mathbf{r}_\perp \left\langle \mathbf{r}_\perp \left| \hat{\Gamma}(L) \right| \mathbf{r}_\perp \right\rangle, \tag{12.38}$$

where the integration is over the camera aperture. N_r can be interpreted as the number of phase space cells of the radiation field intercepted by the camera aperture. Therefore, N_r can be much less than one without violating the uncertainty principle. For a square source
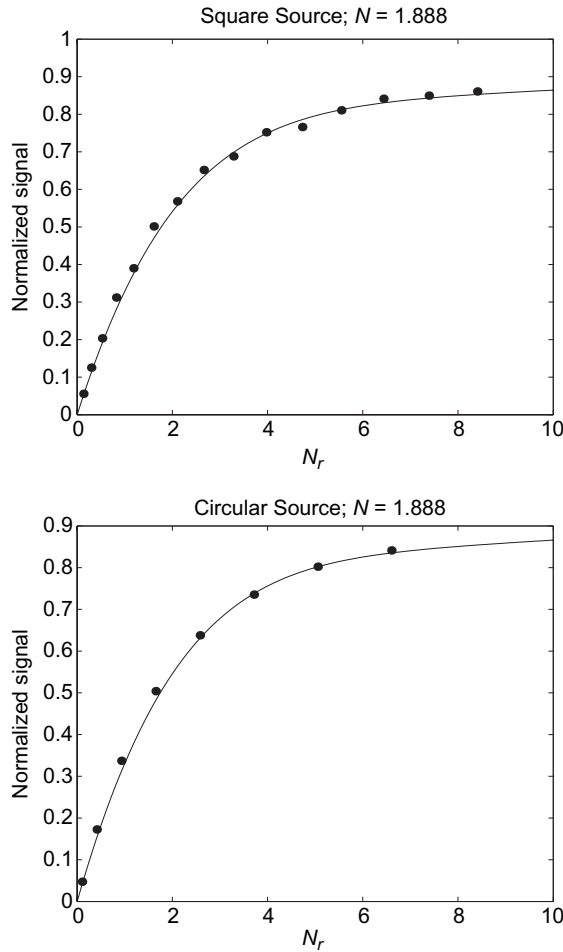
Figure 12.3   Comparison of experiment with theory for the $N = 1.888$ peak measurements. Filled circles are data points.

$$N_r = \pi\left(\frac{2a\theta_s}{\lambda}\right)^2 \tag{12.39}$$

and for a circular source

$$N_r = \left(\frac{\pi a\theta_s}{\lambda}\right)^2 \tag{12.40}$$

Here θ_s is the half angle subtended by the source at the camera's entrance aperture. For the square source $\theta_s = d/2L$ with d being the linear dimension for the square source. For the cirular source $\theta_s = d/2L$ with d being the diameter of the circular source.

   The results of the measured normalized signal are plotted with theory versus N_r in Figures 12.3 and 12.4. The agreement with experiment is highly satisfactory.
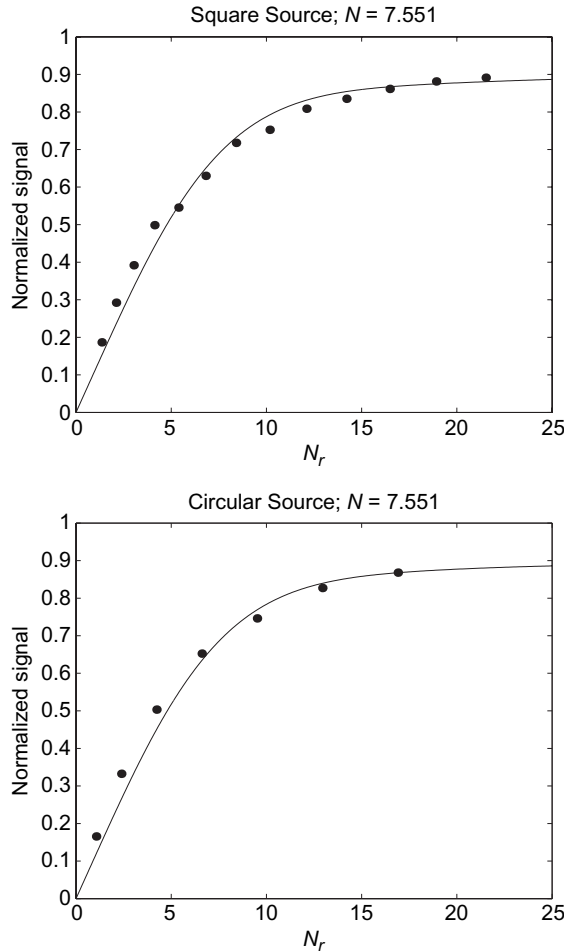
Figure 12.4   Comparison of experiment with theory for the $N = 7.551$ peak measurements. Filled circles are data points.

12.9   CONCLUSION

We have demonstrated a consistent approach to incorporating the diffraction properties of the instrument in optical measurements. We used a remarkable analogy between the result of measuring radiance and the van Cittert-Zernike Theorem that exploits the symmetry between an incoherent source whose radiance is being measured and the detector whose signal represents the measurement. It turns out that the measured radiance is represented (up to an overall constant) by the double integral over time instrument aperture of the mutual intensity of the field and the mutual intensity of a delta correlated source the same size, shape, and location as the detector. While we have expressed our results in time context of radiometry, one would go through a similar analysis in analyzing the detection of any partially coherent wave. The signal is represented by the double integral of two mutual coherence functions. One of these is for the incident wave, the other arising from

the detector considered as a source. It is likely that entirely similar considerations may apply to other signal detection processes where diffraction effects are important.

REFERENCES

Born, M., and Wolf, E. (1999). *Principles of Optics*, 7th Ed. Cambridge University Press.

Littlejohn, R. G., and Winston, R. (1993). Corrections to classical radiometry. *J. Opt. Soc. Am. A.* **10**, 2024–2037.

Littlejohn, R. G., and Winston, R. (1995). Generalized radiance and measurement. *J. Opt. Soc. Am. A.* **12**, 2736–2743.

Mandel, L., and Wolf, E. (1995). *Optical Coherence and Quantum Optics*. Cambridge University Press.

Mather, J. C., Toral, M., and Memmati, H. (1986). Heat trap with flare as multimode antenna. *Applied Optica* **25**, 2826–2830.

Sun, Y., Winston, J., O'Gallagher, J., and Snail, K. A. (2002). Statistical optics and radiance measurement in the diffraction limit. *Optics Communications* **206**, 243–251.

Walther, A. (1973). Radiometry and coherence. *J. Opt. Soc. Am. A.* **63**, 1622–1623.

Winston, R., and Litteljohn, R. G. (1997). Measuring the instrument function of radiometers. *J. Opt. Soc. Am. A.* **14**, 3099–3101.

Winston, R., Sun, Y., and Littlejohn, R. G. (2002). Measuring radiance and the van Cittert-Zernike theorem. *Optics Communications* **207**, 41–48.

Wolf, E. (1978). Coherence and radiometry. *J. Opt. Soc. Am. A.* **68**, 6–17.

13

APPLICATIONS TO SOLAR ENERGY CONCENTRATION

13.1 REQUIREMENTS FOR SOLAR CONCENTRATORS

It may be appropriate to remind ourselves why it is advantageous to concentrate the solar flux. After all, it is well known that flat-plate collectors are adequate for heating a working fluid at temperatures up to about 80°C, and this is at present the most economical way to use solar power for domestic hot water purposes. It is probably also best for domestic space heating in moderate climates, with the reservation that larger glass windows applying the greenhouse effect are also good when appropriate to the use of the building. However, even modest levels of concentration (say, 1.5–2) can dramatically increase the temperature at which heat from a solar collector can be extracted efficiently. The economics of photovoltaic conversion can improve significantly with intermediate levels of concentration (say, 10–100). At high flux levels (of, say, 1,000) efficient generation of electricity is technically feasible. And at the very highest flux levels achievable (say, 40,000 or more) certain processes (e.g., efficient laser pumping) become accessible to solar energy utilization, which may be of considerable interest. Therefore, while one may argue over the appropriate flux level for a particular application, there is general agreement that some degree of concentration would be desirable for most applications. Why then are most practical solar installations nonconcentrating flat plates? The answer must be that the penalties in cost and complexity associated with optical concentration outweigh its advantages. It is the task of nonimaging optics to mitigate these penalties—that is, to concentrate the solar flux at the least possible "cost" in tracking tolerance, precision of components, and so on.

In effect the optics of solar flux concentration has usually been "nonimaging," since the image quality is rarely an issue. But what nonimaging optics has come to signify is the systematic study and development of design techniques that optimize angular acceptance and throughput efficiency for a given flux concentration—that is, "high collection optics."

317

Concentrators are needed wherever we want to use solar power for applications requiring heat at temperatures above about 80°C.[1] There are numberless applications requiring temperatures up to about 300°C for both domestic and industrial applications—for example, space cooling, cooking, desalination, electricity generation (via steam or some working fluid for turbines). Among the factors that influence the design of concentrators for solar energy are the following:

1. Cost and ease of manufacture on an appropriate scale
2. Extent of guidance required for following the sun
3. Durability and maintenance
4. Required working temperature of absorber
5. Preferred geometry of absorber in relation to the mode of utilization of the energy
6. Susceptibility to contamination and durability under ultraviolet irradiation

Most of these are, of course, interlinked, and their relative weightings are changing in response to economic factors rather rapidly at the time of writing. We cannot, therefore, lay down specific guidance, and we restrict ourselves to descriptions of different systems in which we call attention to specific points of advantage or disadvantage.

13.2   SOLAR THERMAL VERSUS PHOTOVOLTAIC CONCENTRATOR SPECIFICATIONS

Besides the common requirements for solar thermal and photovoltaic concentrators presented in the previous section, there are also important differences between these two application fields. These specificities may make it so that a good concentrator design for one field is not appropriate for the other field. The most remarkable differences are the receiver characteristics, the requirements of uniformity of irradiance on the receiver, and the effect of dispersion on the optical efficiency corresponding to different receivers.

13.2.1   Receiver Characteristics

The thermal and the photovoltaic are completely different approaches to exploit the solar energy, and these differences are clear in the receiver nature and its operating conditions. For instance, in photovoltaics, the cell temperature must be kept as low as possible because the cell conversion efficiency decreases with temperature. As a rule of thumb, the photovoltaic system should be designed to provide a temperature at the cell $p - n$ junction smaller than 65°C at standard operating conditions (ambient temperature = 25°C; wind speed = 1 m/s; irradiance = 850 W/m^2). On the contrary, medium- to high-temperature solar thermal systems operate with the absorber at a temperature over 300°C. In this situation, the optics cannot be in touch with the receiver to avoid conduction losses and high thermal stresses. In case it is necessary to place the optics close to the receiver (as in the

---

[1]   For verify flat-plate collectors, the upper limit of useful operating temperature is somewhat higher. However, their performance is improved significantly by concentration.

case of the CPC), at least a small gap must be provided. Moreover, these thermal aspects are responsible for the difference between thermal and photovoltaics with respect to the use of optically dense transparent media ($n > 1$) surrounding the receiver. This is difficult to apply for thermal concentrators, and if excluded, the concentration limit is reduced by a factor n^2 for solar thermal systems (which is $n^2 = 1.5^2 = 2.25$ for typical materials).

Another aspect to consider is the electric nature of the photovoltaic cell. The necessity of extracting the photogenerated current and of interconnecting the different cells must be considered at the optical design stage. Also, the use of metallic mirrors in contact with the cell should be avoided to prevent short circuits. Obviously, these problems do not matter for solar thermal systems.

Also, the spectral sensitivity of the receivers is different in both fields. For example, the quantum efficiency of the silicon solar cells is very low over 1,200 nm. This means that the further infrared radiation of the sun, which usually is nearly 25% of the sun power, is not useful, contrary to the solar thermal case. These spectral differences may make it so that a good optical material for photovoltaics may be unsuitable for solar thermal applications due to its poor infrared response.

Finally, the last aspect to be considered here is the receiver geometry. This geometry affects the design importantly, as we have seen in Chapter 5. In the case of medium to high concentration in photovotaics, the active surface of the solar cell is always flat, typically with a round or squared contour of the active area. Sometimes the cells tessellate to build a strip receiver. The inactive face of the cell is used to evacuate the heat by fixing a heat-sink to this face. Only low-concentration systems (typically $C < 5$) can use bifacial solar cells—that is, cells that can collect the sunlight on their two faces. The bifacial geometry must be obviously considered in the concentrator design.

On the contrary, in solar thermal concentration there is a wider variety of receiver geometries. Flat (monofacial and bifacial) and tubular are the most common ones. In the case of a tubular receiver, the conventional cross section is circular. However, the contour can also been considered a surface to design, which constitutes an additional degree of freedom. As an example, for the specific case of the parabolic through, Ries and Spirkl (1995) proved that designing the contour using the caustic of the edge rays allows to noticeably increase the concentration ratio, as shown in Figure 13.1. With 100% ray collection efficiency within the acceptance angle $\pm\alpha$, the geometrical concentration is limited to $C_{MAX} = 1/\sin\alpha$. For a parabola with rim angle of $\pm 90°$ and also for 100% collection efficiency, the circular receiver only achieves $C/C_{MAX} = 1/\pi = 0.32$, while the optimum receiver provides $C/C_{MAX} = 0.47$.

13.2.2   Irradiance Uniformity on the Receiver

The irradiance distribution on the receiver produced by most concentrators is not uniform. The effect of this nonuniformity has not been critical in solar thermal systems, although it has some relevance in the development of Direct Steam Generation (DSG) systems (Goebel et al., 1996/1997). However, the nonuniform illumination of a photovoltaic cell may produce a dramatic decrease of its solar-to-electric power conversion efficiency. This is due to the Joule effect power losses in the cell series resistance R_S (if the series resistance were null, the efficiency will
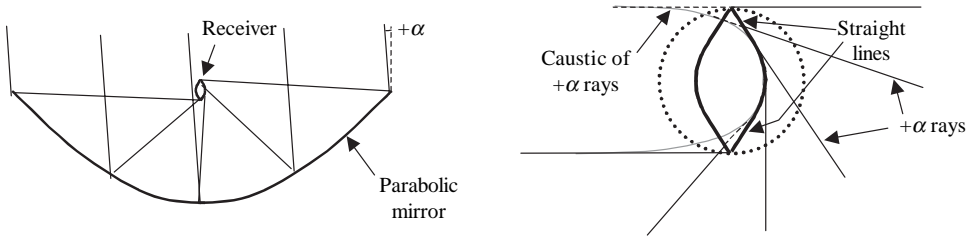
Figure 13.1   The concentration of a solar thermal parabolic trough (on the left) can be maximized by designing the contour of the receiver (on the right). The conventional circular receiver providing the same acceptance angle (for 100% ray collection) is shown for comparison purposes.
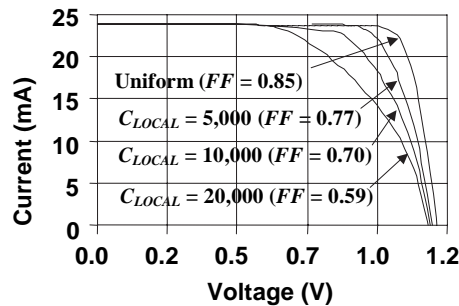
Figure 13.2   Simulation of the current-voltage characteristic of a GaAs solar cell under an average irradiance of 1,000 suns (1 sun = $1\,kW/m^2$) and different pillbox local concentration at the cell. *FF* denotes the cell fill factor.

always increase with concentration). The losses in the series resistance degrade the cell fill-factor *FF*, which is one parameter that is related with the cell efficiency by the formula

$$\eta = \frac{V_{OC}I_{SC}FF}{P_L} \tag{13.1}$$

where P_L is the solar power, V_{OC} is the open-circuit voltage, I_{SC} is the short-circuit current, and *FF* is the fill factor. The nonuniform illumination mainly affects the cell fill factor. As an example, Figure 13.2 shows the PSIPICE simulation with a discretized cell of the effect of illumination nonuniformity on the performance of a $1\,mm \times 1\,mm$ Gallium Arsenide (GaAs) solar cell (Álvarez, 2001). This cell was designed for optimum performance at a uniform irradiance of 1,000 suns (1 sun = $1\,kW/m^2$). The simulation considers an average irradiance along the total cell area of $1,000\,kW/m^2$ and a pillbox irradiance pattern with higher peak irradiance on a squared area at the cell center and null irradiance outside said area. When the cell is illuminated with a peak irradiance of 5 times the average, the cell conversion efficiency reduces approximately as the fill factor does, and then the simulation indicates that the efficiency decreases a $100(1 - 0.77/0.85) = 9.4\%$.

Both concentrator and cell designers should be concerned about this problem at the design stage and try to minimize its negative effects. The variation of the irradiance pattern with the mispointing angular error must also be considered. Therefore, two issues must be addressed:

1. How a given nonuniform irradiance distribution affects the cell efficiency. Note that analytical models rather than numerical simulations verb (as that in Figure 13.2) because the formers give much more information to the designers about where and how much the light can be concentrated.
2. How to design concentrators producing the desired irradiance patterns, independently of the sun position within the concentrator acceptance angle.

Discussing issue 1, the effect of the nonuniform illumination depends on both the physical parameters of the solar cell and on the irradiance distribution. No general analytical model to quantify this effect is available yet. As an illustrative example, in the case in which the front metal grid contribution to the series resistance R_S is negligible, the cell voltage near the maximum power point can be estimated with the following approximate formula (Benítez and Miñano, 2003):

$$V \approx V_0 - \frac{kT}{e}\ln\left(\int_0^\infty f(C)\exp\left(\frac{C}{C_0}\right)dC\right) \quad C_0 = \frac{kT/e}{R_S I_{sc,1sun}} \tag{13.2}$$

where V_0 is the cell voltage when the series resistance is null, kT/e is the thermal voltage (about $29\,mV$ at usual cell operating temperatures), $f(C)$ is the probability density function of the irradiance C, and $I_{sc,1sun}$ is the cell short circuit current under 1 sun irradiance. The parameter C_0 coincides approximately with the concentration level at which the cell efficiency is maximum. The integrand in Eq. (13.2) indicates the relative importance of the different concentration intervals $(C, C + dC)$ on the cell performance degradation, pointing out that the dependence on irradiance is strongly nonlinear due to the exponential function. Consequently, even if a small fraction of power is highly concentrated, it can produce the cell degradation dominating the integral in Eq. (13.2).

A general and simple rule of thumb that is generally used says that an irradiance distribution with a peak concentration doubling the average usually produces an affordable decrease of conversion efficiency. This rule is only inaccurate when the average irradiance is several times higher than the paramenter C_0. For instance, following with the example of the cells fulfilling Eq. (13.2), let us consider the two extreme irradiance distributions: (1) the pill-box type distribution (half the cell is not illuminated and the other half is illuminated with double the average irradiance), which has maximum standard deviation; and (2) the distribution with a very small area with the peak distribution, and the rest of the area is illuminated slightly below the average, which is the minimum (null, at the limit) standard deviation case. According to Eq. (13.2), the voltage difference near the maximum power point between cases (1) and (2) is

$$V(2) - V(1) \approx \frac{kT}{e}\ln\left(\frac{1}{2}\left(1 + \exp\left(\frac{<C>}{C_0}\right)\right)\right) \tag{13.3}$$

For a high concentration GaAs cell, the 10% relative difference between the efficiencies of cases (1) and (2), approximately coincides with $V(2) - V(1) = 100\,mV$. This implies $<C> \approx 4C_0$. Since presently high concentration GaAs cells have $C_0 \approx 500$ suns, this limits the 10% guaranteed accuracy of the rule to $<C> \approx 2{,}000$ suns.

It can be proved (Benítez and Miñano, 2003) that, for a given average concentration, the irradiance distribution maximizing Eq. (13.2) is uniform. However, when the grid series resistance is not negligible, it is clear that the irradiance pattern that produces the maximum conversion efficiency on conventional

concentrator cells is nonuniform (Benítez and Mohedano, 1999). The maximum efficiency of this optimum irradiance pattern is close to that obtained for the uniform illumination in practice. Therefore, the efficiency increase is not the main interest of this optimal nonuniform irradiance distribution but the fact that it guides the concentrator designers to where and how much can they concentrate the sun light.

In order to address now issue (2), let us consider the uniform illumination distribution as the goal. If only the uniformity is required for a given concentration factor, the solution to this design problem is well known and was described in Chapter 7. However, in the photovoltaic application, especially due to the nonlinear effects when series connected cell are illuminated differently (see Section 13.2.3), the concentrator is desired to have an acceptance angle α substantially greater than the sun angular radius α_S (typically, $\alpha \approx \pm 1°$, while $\alpha_S \approx \pm 0.26°$). The uniformity is then required for the sun placed anywhere inside the acceptance angle, which is a more complex design problem (the solutions in Chapter 7 coincide with case in which $\alpha = \alpha_S$), especially because in high-concentration systems ($<C> >1,000$), the value $\alpha \approx \pm 1°$ approaches the thermodynamic limit (and thus the maximum illumination angle of the cell, β, comes closer to $\pm 90°$, typically $\beta \approx \pm 60°–75°$).

There are two methods in classical optics that potentially can achieve this insensitivity to the source position. The two methods, used for instance in condenser designs in projection optics, are the light-pipe homogenizer and the Kohler illuminator (commonly called *integrator*) (Cassarly, 2001).

The light pipe homogenizer uses the kaleidoscopic effect created in the multiple reflections inside a light pipe, which can be hollow with metallic reflection or solid with total internal reflections (TIR). This strategy have been proposed several times in photovoltaics (Fevermann and Gordon, 2001; Jenkins, 2001; O'Gallagher and Winston, 2001; Ries, Gordon, and Laxen, 1997), essentially attaching the cell to the light pipe exit and placing the light pipe entry as the received of a conventional concentrator. It can potentially achieve (with the proper design of the pipe walls and length) good illumination on a squared light pipe exit with the sun in any position within the acceptance angle. For achieving high-illumination angles β, the design can include a final concentration stage by reducing the light pipe cross section near the exit. However, this approach has not been proven yet to lead to practical photovoltaic systems (and no company has commercialized it as a product yet).

On the other hand, the integrating concentrator consists of two imaging optical elements (primary and secondary) with positive focal length (that is, producing a real image of an object at infinity, as a magnifying glass does). The secondary is placed at the focal plane of the primary and the secondary images the primary on the cell. This configuration makes it that the primary images of the sun on the secondary aperture, and thus the secondary contour, defines the acceptance angle of the concentrator. As the primary is uniformly illuminated by the sun, the irradiance distribution is also uniform, and the illuminated area will have the contour of the primary and will remain unchanged when the sun moves within the acceptance angle (equivalently when the sun image moves within the secondary aperture). If the primary is tailored in square shape, the cells will be uniformly illuminated in a squared area. The squared aperture is usually the preferred contour to tessellate the plane when making the modules, while the
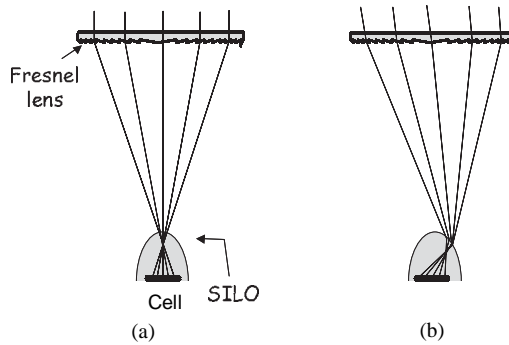
Figure 13.3   The "Sandia concept 90" proved that the cell could be nearly uniformly illuminated on a square (with a squared Fresnel lens) for any position of the sun in the acceptance angle. (a) Normal incidence. (b) Incidence near the acceptance angle.

squared illuminated area on the cell is also usually preferred because it fits the cell's shape.

Integrator optics in PV was first proposed (James, 1989) by Sandia Labs in the late 1980s, and it was commercialized later by Alpha Solarco. Now, high-concentration SMS concentrators (see Section 13.4.3) are including this strategy for achieving good uniformity and improving tolerances.

Sandia Labs' approach used a Fresnel lens as primary and a single-surface imaging lens (called SILO, from SIngLe Optical surface) that encapsulates the cell as secondary, as illustrated in Figure 13.3.

This simple configuration is excellent for getting sufficient acceptance angle α and highly uniform illumination, but it is limited to low concentrations because it cannot get high angles β. Imaging secondaries achieving high β (high numerical aperture, in the imaging nomenclature) are, to the present, impractical. Classical solutions, which would be similar to high-power microscopes objectives, need many lenses and would achieve $\beta \approx 60°$. Another simpler solution that nearly achieves $\beta = 90°$ is the RX concentrator (Benítez and Miñano, 1997; Miñano, Benítez, and Gonzalez, 1995) (see Figure 13.4). Although the Lens+RX integrator is still not practical, it is theoretically interesting because shows that the optimum photovoltaic concentrator performance (squared aperture concentrator, acceptance angle α several times larger than the sun radius α_S, isotropic illumination of the cell ($\beta = 90°$), squared uniform cell irradiance independently of the sun position within the acceptance) is nearly attainable.

As an example, for $\alpha = 1°$, this optimum performance concentrator will get a geometrical concentration $C_g = 7{,}387\times$, which is the thermodynamic concentration limit for that acceptance angle. Note that since all rays reaching the cell come from the rays within the cone of angular radius α, no rays outside this cone are collected by the optimum concentrator.

Of course, in practice, the optimum photovoltaic concentrator performance might be not desired. For instance, as already mentioned, the high reflectivity of nontextured cells for glazing angles may make the isotropical illumination useless. As another example, illuminating a squared area inside the cell is perfect for back-contacted solar cells (such as SunPower's or Ammnix' cells), but it may be not so perfect for front-contacted cells, for which an inactive area is needed to make the
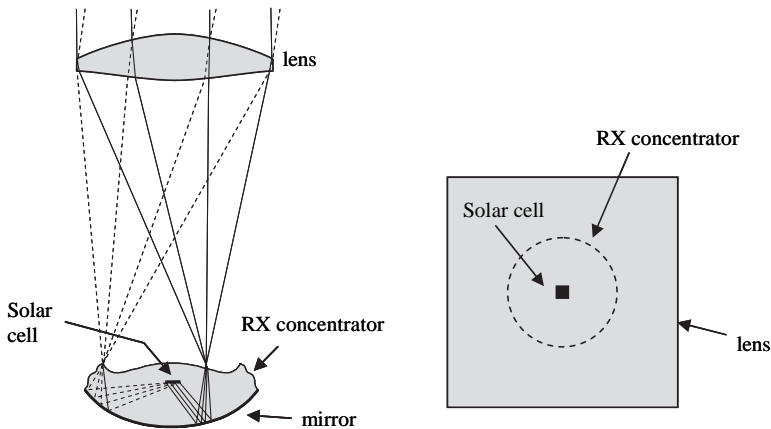
Figure 13.4   The use of an RX concentrator as the imaging secondary in the integrator (not shown to scale on the left) with a double aspheric imaging primary shows that it is theoretically possible to come very close to optimum photovoltaic concentrator performance.

front contacts (breaking the squared shape active area restriction). Finally, for medium concentration systems (let's say, $C_g = 100$) the aforementioned optimum performance would imply an ultrawide acceptance angle $\alpha = 8.6°$. It seems logical that over a certain acceptance angle, there must be no cost benefits due to the relaxation of accuracies (and $8.6°$ seems to be over such a threshold). If this is the case, coming close to the optimum performance seems to be unnecessary for this medium concentration level.

The present challenge in the optical design for high-concentration photovoltaic systems is concentrators that approach optimum performance and at the same time are efficient and suitable for low-cost mass production.

13.2.3   Dispersion of the Optical Efficiency Corresponding to Different Receivers

In solar thermal concentrating systems—for instance, in parabolic trough technology—the optical efficiency along the receiving tube does not need to be uniform because the system performance depends on the cumulative solar power cast along the receiver.

However, in general this is not the case in photovoltaic concentration systems. The level of degradation of performance due to the dispersion of the solar power cast by the different solar cells depends on the electrical interconnection configuration. The extreme cases are the all-parallel connected cells, whose performance degradation is unimportant, and the all-series connected cells, whose performance is much worse when nonequally illuminated.

Consider a set of N solar cells illuminated by nonperfect concentrators that produce dispersion of the solar power cast by each individual cell of the set. Assume that the cells are identical (thus, when illuminated equally and independently, all the individual cells present the same fill-factor FF, short-circuit current I_{SC}, and open-circuit voltage V_{OC}). Since I_{SC} is proportional to the power cast by the cell

with very good approximation, instead of referring to the dispersion of the solar power cast, we can directly refer to the dispersion of I_{SC} instead.

Let us consider the effect of illuminating the cell set in the two extreme cases (all cells are parallel connected or all cells are series connected). Referring to the parameters of Eq. (13.1) applied to the cell sets, if all cells are equally illuminated ($I_{SC,k} = I_{SC}$, $1 \leq k \leq N$), we get

$$I_{SC,series} = I_{SC} \quad V_{OC,series} = NV_{OC} \quad FF_{series} = FF$$
$$I_{SC,parallel} = NI_{SC} \quad V_{OC,parallel} = V_{OC} \quad FF_{parallel} = FF \tag{13.4}$$

where V_{OC} and FF are open-circuit voltage and fill-factor of all the (equally illuminated) individual cells. Therefore, according to Eq. (13.1), both series and parallel sets of cells get the same efficiency when uniformly illuminated.

However, when there is dispersion in the illumination, it is obtained that (Luqve, Lorenzo, and Ruiz, 1980)

$$I_{SC,series} = \min\{I_{SC,k} 1 \leq k \leq N\} \quad V_{OC,series} = N <V_{OC,k}> \quad FF_{series}‰ < <FF_k>$$
$$I_{SC,parallel} = N<I_{SC,k}> \quad V_{OC,parallel} \approx <V_{OC,k}> \quad FF_{parallel} \approx <FF_k> \tag{13.5}$$

Although the fill-factor when cells are series connected is higher than when they are parallel connected, the dominant difference with the equally illuminated case is the decrease of the short circuit current in the series connection case, and in a (pessimistic) first approximation, it can be said

$$\frac{\eta_{series}}{\eta_{parallel}} \approx \frac{I_{SC,series}}{I_{SC,parallel}} \tag{13.6}$$

Quantitatively, the degradation depends on the statistical distribution of the short circuit current of the differently illuminated cells, which depends on the dispersion of different concentrators, their alignment, and so forth. Due to the fact that the optical efficiency is usually maximum at normal incidence, misalignments introduce an asymmetry that makes that asymmetric distributions (far from Gaussian) are usually expected. As an illustrative example of this asymmetry, let us consider the case of the EUCLIDES™ concentrator, which was developed jointly by the Spanish Solar Energy Institute of the Technical University of Madrid (IES-UPM), the British company BP solar and the Spanish Institute of Technology and Renewable Energies (ITER). The biggest photovoltaic concentration plant in the world, of 480 kW, was made in Canary Islands with the EUCLIDES™ technology. This concentrator tracks the sun with a single north-south axis, and it is composed 14 units of a parabolic trough 84 m long (composed 140 individual parabolic mirrors, as shown in Figure 13.5) that concentrate the sunlight onto silicon solar cells, with a geometrical concentration of 38×. The interconnection configuration was based on 1,380 series connected cells grouped in 138 modules.

The structural support of the EUCLIDES™ is given by an equilateral triangular beam. This beam suffers flexion and torsion due to the system weight and the wind loads, which causes pointing errors along the concentrator trough. The losses associated they're beam deformations are not relevant for the annual energy production (because with below 1%; see Arboiro, 1997), but the instantaneous losses, which can be noticeable, are useful to illustrate the effect of the series connection of the cells. Figure 13.6 shows the results of a simulation of the instantaneous performance of the EUCLIDES™ if installed in Madrid for the worst case
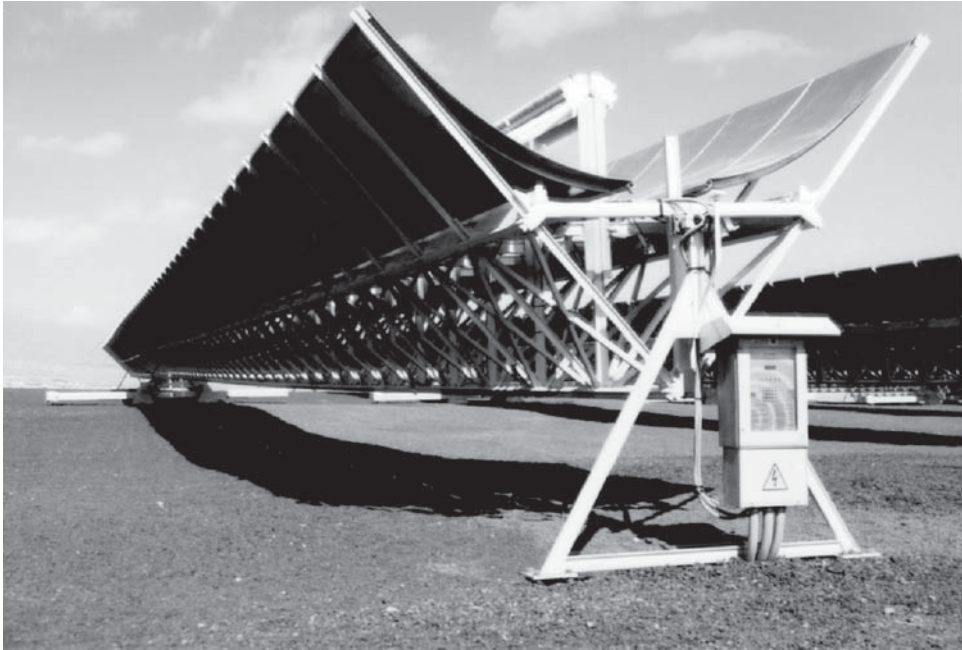
Figure 13.5   EUCLIDES™ photovoltaic concentrator plant installed in Tenerife, Spain. (Courtesy of ITER, BP, and IES-UPM)

(the time of the year when the beam deformation and the sun position make the pointing errors maximum). Figure 13.6a shows the photocurrent I_L of the cells along the concentrators, as a function of the cell position and for several wind speed values. The photocurrent of each cell I_L is proportional to the optical efficiency of the mirror that illuminates the cell. Figure 13.6b shows the losses introduced by the nonlinear effects of the series connection as a function of the wind speed. Note that this graph shows only the losses associated with the series connection (it does not include the decrease of the average of the optical efficiency). This means if the EUCLIDES™ were solar thermal, its corresponding value in Figure 13.6b would be 100% independently on the wind speed. Note that for the case of no wind, although there is up to ±5% of photocurrent dispersion, only a 2% loss should be expected. However, if the wind has a speed of 30 km/h, the minimum photocurrent reaches the 62% of the nominal value, and the series connection efficiency drops to 79%. In practice, this low value would be pessimistic because each one of the 138 modules has a bypass diode, which keeps the series connection efficiency at 83%. These bypass diodes (not mentioned before) introduce another nonlinear effect to provide a way for the current to flow skipping the low photocurrent modules.

This difference between solar thermal and photovoltaics is important because effects to basic concepts not always clearly identified. For instance, in solar thermal it is usual to use the concept of *effective sun.* This consists of analyzing a real system as perfect but transferring its different imperfections (mirror profile and scattering, tracking errors, structural misalignments, etc.) to the sun shape (Rabl, 1985). This model, which is correct for solar thermal, it is not correct for
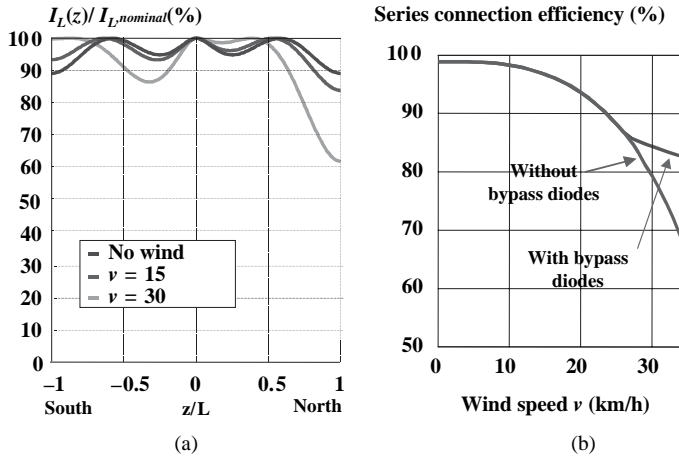
Figure 13.6   Results of the selected simulation of the EUCLIDES$^{\text{IM}}$ concentrator instantaneous performance in the worst case. (a) Photocurrent of the cells depending on the position along the concentrator and for several selected wind speed values. (b) Losses due to the nonlinear effects of the series connection.

photovoltaics when the aforementioned nonlinear effects of series connection take place.

Another direct consequence of all this is the definition of the *acceptance angle*. Photovoltaic concentrators with series-connected cells cannot be used effectively for the angular positions of the sun where the optical efficiency for some individual cells is low, whereas in solar thermal systems this can be allowed because the optical efficiency is averaged along the receiver. For example, consider the aforementioned EUCLIDES$^{\text{IM}}$ parabolic trough. Let us assume that each mirror has an angular collection curve that can be approximated by a Gaussian curve with a 50% transmission angle of ±1°. It is easy to obtain from the Gaussian curve formula that the 90% transmission angle is ±0.4°. In order to avoid a significant system performance degradation, standard deviation of the mirror positioning and structural deformation errors should be kept below 1° for the thermal case (specifically, if these errors are Gaussian distributed, a 5% degradation implies a standard deviation of 0.65°). However, in the photovoltaic case, probably a better criteria would be to keep the mirror positioning and structural deformation errors below ±0.4° during 95% of the operation time (if these errors are Gaussian distributed, this implies a standard deviation of 1°/4.4 = 0.23°, to be compared with 0.65° for the solar thermal). This is the reason why in photovoltaic concentrators the acceptance angle is usually defined at 90% transmission, whereas the 50% transmission angle is used as the definition of acceptance angle in solar thermal systems.

13.3   NONIMAGING CONCENTRATORS FOR SOLAR THERMAL APPLICATIONS

Table 13.1 shows various solar thermal concentrator types ordered by their geometrical concentration ratio (Cg). In this regard, it is worth pointing out that the

Table 13.1.   Implications of availability of solar flux over a range of 1–100,000 suns.

Flux (Suns)	Conditions	Applications
1–2	Fixed concentrator	Cooling and heating
2–100	Nontracking/tracking (linear focus)	Power generation (cooling and heating)
500–10,000	Tracking tower/dish	Power generation
20,000–50,000	Solar furnace	Materials, lasers space propulsion
70,000–100,000	Speciality solar furnace	Materials, lasers, experiments

Solar Geometry
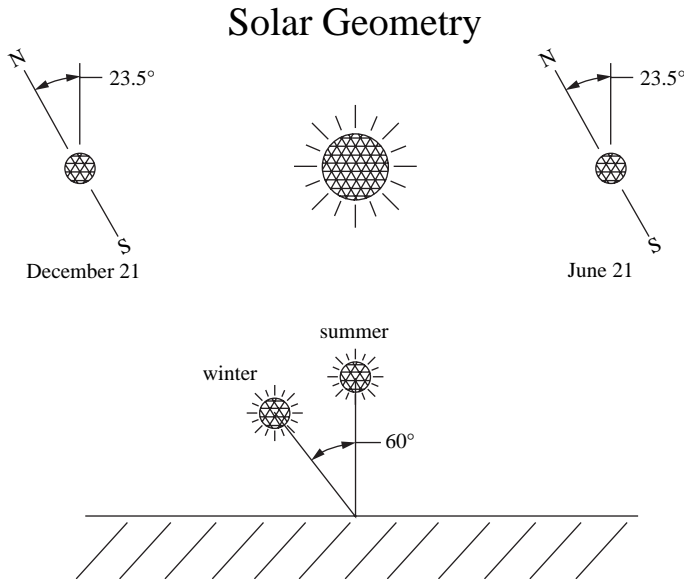
Figure 13.7   Tilt of Earth's axis to the ecliptic plane produces about 60° change in solar elevation over the year.

concept of concentration ratio is frequently misapplied. As is clear from the preceding discussion, this ratio divides the entrance aperture area by the area of the absorber. For example, the Cg for a trough of diameter D with tubular absorber of radius r is $D/2\pi r$, and not $D/2r$. The latter definition is sometimes quoted for parabolic troughs—perhaps because it gives a higher number—but it is not the thermodynamically correct definition. We will discuss various concentrator regimes with Cg as the organizing principle. As already noted, we are concerned with energy efficient concentrators only. Systems with low throughput are not particularly useful for solar applications.

13.3.1   Stationary Concentrators ($Cg < 2$)

This category may well be the most important of all because of the practical advantages enjoyed by fixed solar systems. Recall that the elevation angle of the sun varies by over 6° over the course of the year, a consequence of the tilt of Earth's axis of 23° to the plane of the ecliptic (Figure 13.7).
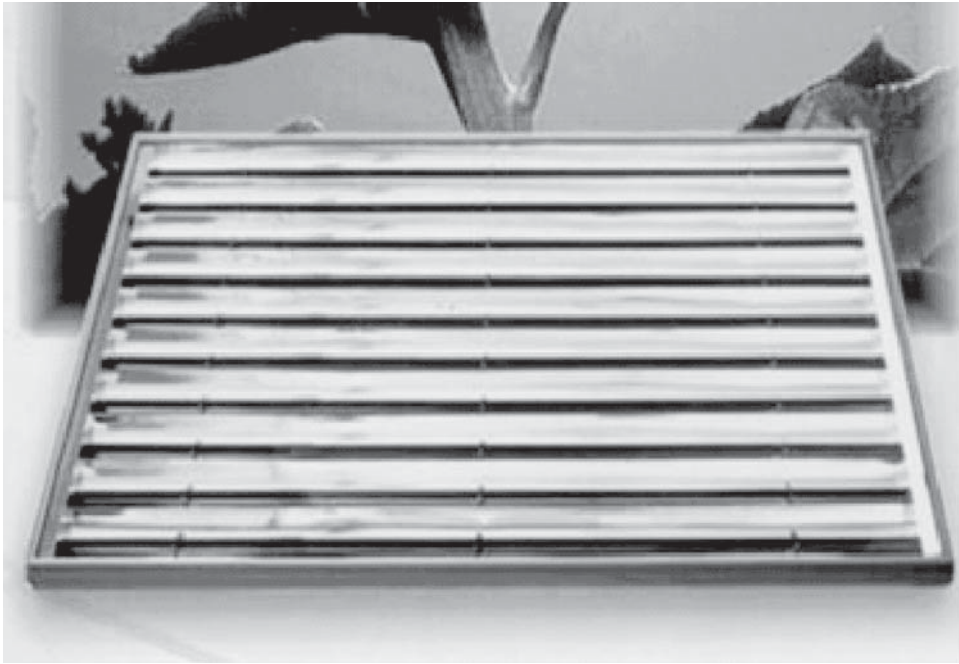
Figure 13.8   CPC "flat plate" collector for midtemperature.

Therefore a fixed concentrator would have to accommodate this annual excursion in elevation on top of the diurnal east-west variation of nearly 180°. Clearly, the diurnal variation indicates a 2D (trough) geometry aligned east-west. But what of the excursion in elevation? At one time, the concept of a fixed concentrator was considered an oxymoron. After all, the Cg of a parabolic trough is $1/\pi \sin\theta$ for a tubular receiver. So a parabolic trough aligned east-west that accepts, say, 60° in elevation would have a concentration ratio $Cg = 2/\pi$. Therefore, such an arrangement actually deconcentrates. To be sure, the geometry of the involute was known to the ancient Greeks, as was the geometry of the parabola. The involute in principle would have a concentration ratio approaching 1 but, in practice, with a necessary space or gap between absorber and reflector, Cg falls short of 1. With the advent of nonimaging optics and the recognition of the sine law of concentration the situation changed dramatically. Fixed solar concentrators were developed and applied to thermal uses where the temperature requirements exceeded that of the flat plate collector. A number of CPC collectors for midtemperature use are now manufactured (Figures 13.8 and 13.9). A particularly attractive use is thermal driven absorption cooling. In this application, rooftop solar collectors convert solar radiation that would otherwise warm up the building into heat at temperatures, say, 115°C to 180°C (Figure 13.10). The low end of the temperature range can drive a "single effect" absorption chiller, which can be characterized as quasi-static. They show a coefficient of performance (COP) of about 0.7–0.8. COP is the ratio of heat removed by the chiller divided by heat supplied to the chiller. Thermodynamically, this ratio can exceed 1. The upper end is sufficient to drive a "double-effect" chiller with a COP of about 1.1–1.2. The heat drives an absorption chiller that actively

Figure 13.9   Vacuum tube CPC collector for mid- to high temperature.

cools the building, a double benefit. At the low end, the solar collector's elements can be nonevacuated, whereas at the upper end, vacuum insulation is needed. When triple effect machines become commercially available, with driving temperatures in the 200°C range, solar cooling systems with stationary concentrators will be even more efficient. One can look forward to the day when the COP compensates for collector losses so that the amount of heat removed matches the heat collected.

13.3.2   Adjustable Concentrators (Cg = 2–10)

Narrowing the acceptance angle from 30° to, say, 6° still allows for stationary operation at least during the day. The required frequency of adjustments goes up rapidly with Cg. Thus, Cg = 3 needs only biannual adjustment, whereas Cg = 10 requires almost daily updates. Thus, systems of this kind can be characterized as quasi-static. Such collectors and systems have been successfully operated for thermal and PV applications on both prototype and commercial scales. An industrial process heat installation in Israel using seasonally adjusted nonevacuated CPCs is shown in Figure 13.11. A project for heating in a very cold ambient used a Cg = 3 nonevacuated CPC. This was a school building on a Navajo reservation in Breadsprings, New Mexico, USA, that is over 2,000 m elevation near the
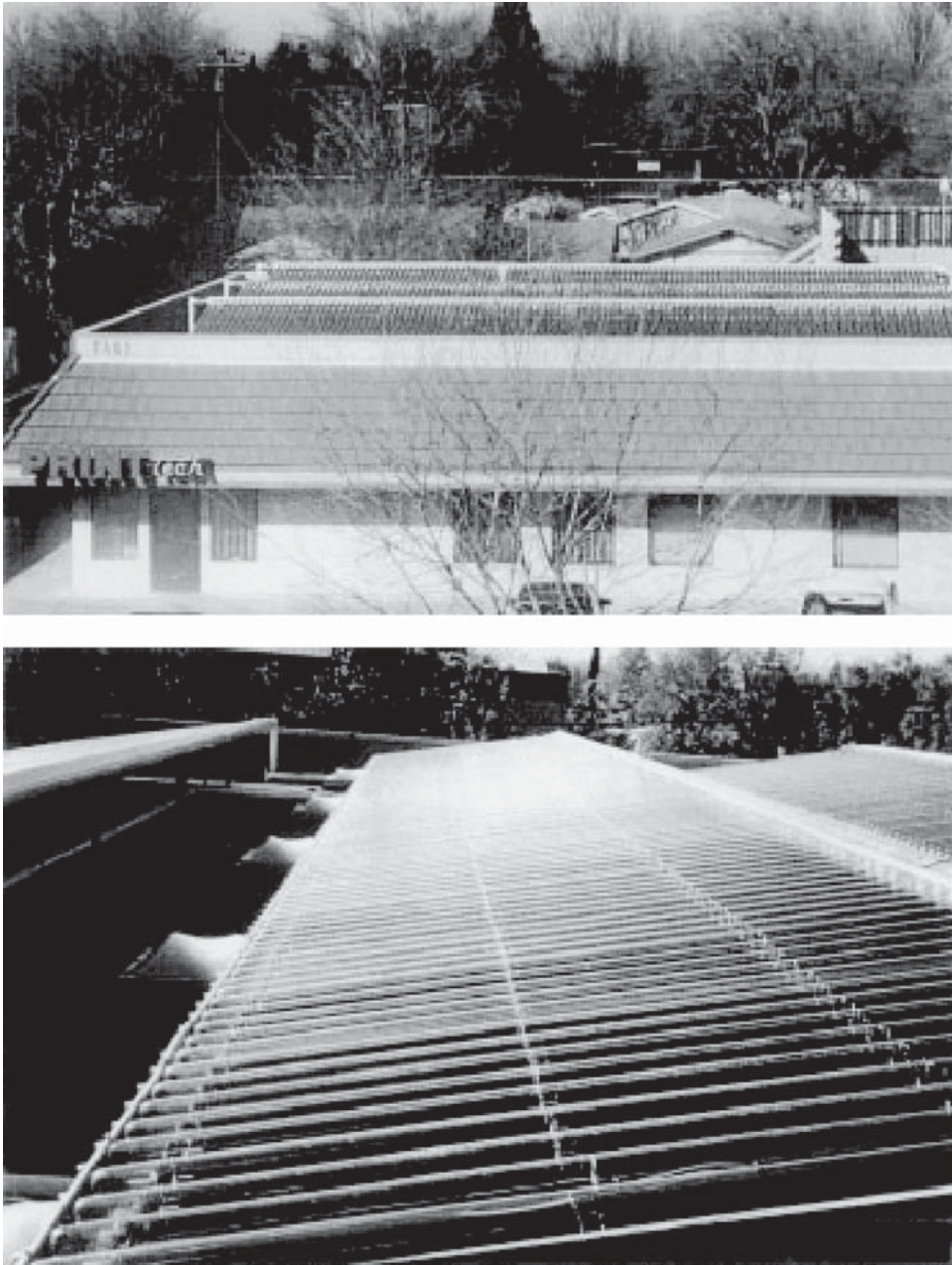
Figure 13.10   Solar cooling using evacuated CPCs driving double-effect absorption chiller.
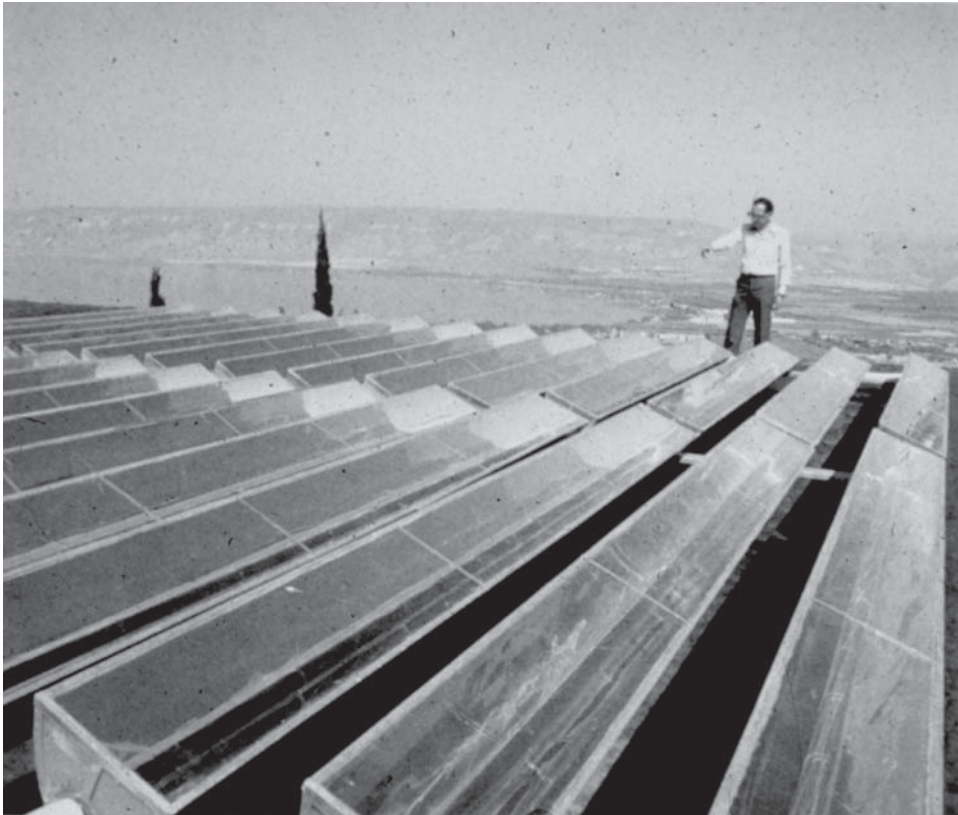
Figure 13.11   CPC array (nonevacuated) for process heat.

Continental Divide, which required only winter heating so that the collectors could be stationary and oriented for winter use. Figure 13.12a shows a schematic diagram of the collector, and Figure 13.12b shows the collector array on the roof.

13.3.3   Tracking Concentrators, One Axis (Cg = 15–70)

In this category, the most widely deployed type of solar concentrator is the parabolic trough, which is not a nonimaging solution. A good example is the solar power plant at Kramer Junction, California, USA (Figure 13.13). The modern parabolic trough technology was developed in the United States in the 1970s and early 1980s with extensive testing and demonstration at Sandia Laboratory in New Mexico. The technology was commercialized by the LUZ company in Israel, now succeeded by the Solel Corporation. The parabolic mirrors are back-silvered glass for high reflectivity and durability. The receiver heat collecting element (HCE) is stainless steel tubing coated with a highly selective metallo-ceramic (cermet) material and housed in an evacuated borosilicate glass tube. The glass is chemically treated to reduce surface (Fresnel) reflections. The working fluid is heated to 400°C so that the power block is a conventional steam turbine. The backup fuel to the solar heat

(a)

(b)
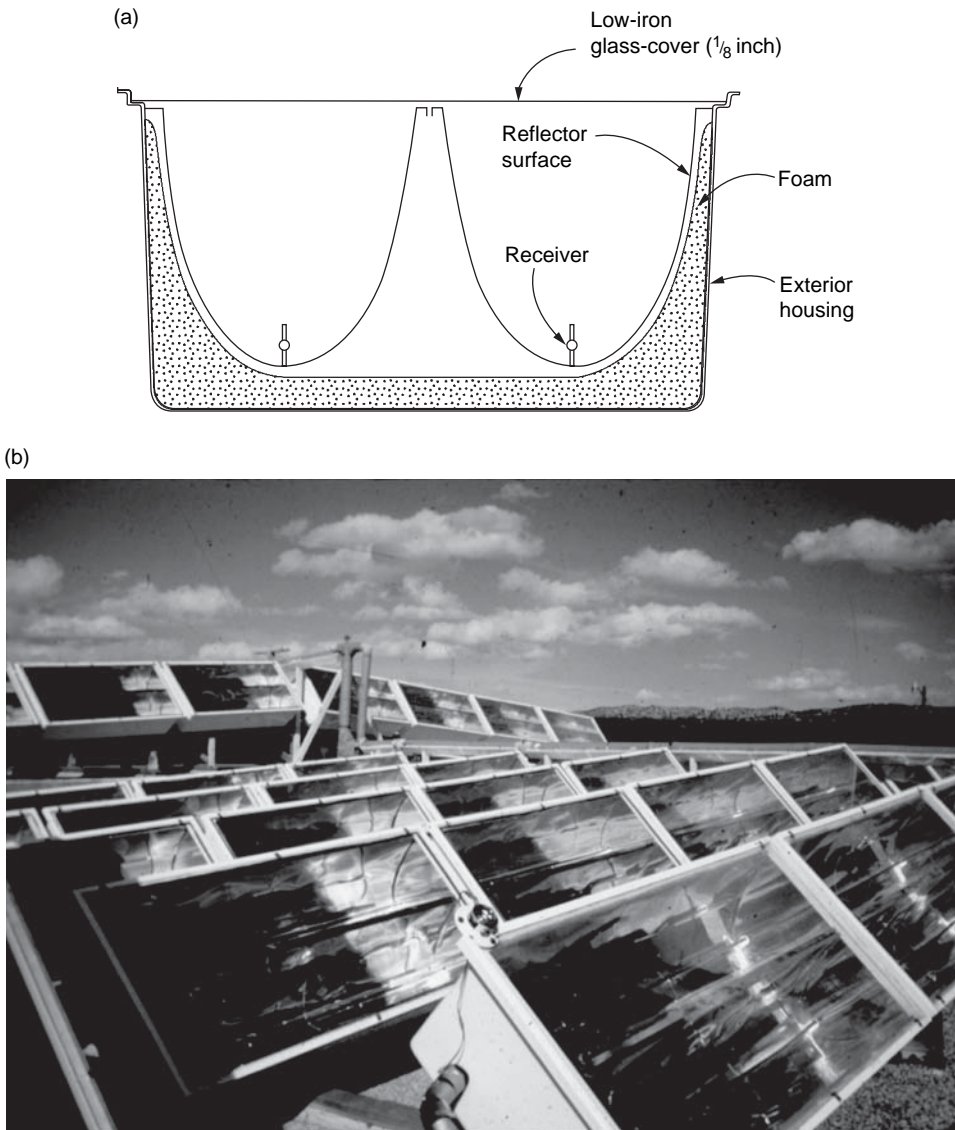
Figure 13.12   CPC (nonevacuated) collectors for winter heating and hot water.

input is natural gas. This system represents the largest-scale implementation by far of solar electric power generation.

Nonimaging design approaches in this concentration range have been mainly focused on the design of CPC-type secondary concentrators for the parabolic. This has been done for two reasons: (1) increasing the concentration (which can increase the conversion efficiency by raising the operation temperature of the absorber tube) and (2) obtaining a more uniform irradiance distribution on the absorber (which reduces the thermomechanical stress in the tube wall, especially in direct-steam generation systems; see Goebel et al., 1996/1997).

Figure 13.13   Large array of parabolic troughs for electric power generation.

Extensive work has been done considering the parabolic primary as an apparent Lambertian source (Collares-Pereira and Mendes, 1995; Collares-Pereira, Gordon, and Rabl, 1991; Mills, 1995). When the whole primary mirror is considered as a single Lambertian source, the achievable concentration is low when the primary f-number is decreased (which is desired for improving the mechanical design of the supporting structure). This occurs because the actual bundle impinging on the parabola within $\pm\alpha$, after the reflection, is far from performing as a Lambertian source (its étendue is significantly smaller). In order to obtain designs for $f/0.5$ parabolic primaries (i.e., rim angles $\phi = \pm 90°$), several solutions based on the division of the primary mirror in sections considered as Lambertian sources have been developed. Figure 13.14 shows two examples of these secondary designs, due to Mills (1995) and Collares et al. (1991). Neglecting the mirror reflectivity losses, the design of Figure 13.14b has 100% ray collection efficiency within the acceptance angle and gets a concentration of $C/C_{MAX} = 0.63$, being $C_{MAX} = 1/\sin\alpha$, while the parabola gets $C/C_{MAX} = 1/\pi = 0.32$. On the hand, the design of Figure 13.14a has 98% ray collection efficiency and gets a concentration of $C/C_{MAX} = 0.80$, while the parabola gets $C/C_{MAX} = 0.37$ for the same ray collection efficiency.

Later, Ries and Spirkl (1996) designed a CPC-type secondary, eliminating the restriction of assuming the Lambertian-primary approximation—that is, considering the actual bundle of rays reflected by the primary (see Figure 13.15a). This design is similar to that of the TIR mirror of the dielectric-filled CPC with circular aperture (Ning, Winston, and O'Gallagher, 1987), described in Chapter 5. Ries and Spirkl's design improved the simplicity of the aforementioned previous designs by Mills and Collares et al. and gets $C/C_{MAX} = 0.68$ for 100% ray collection efficiency.
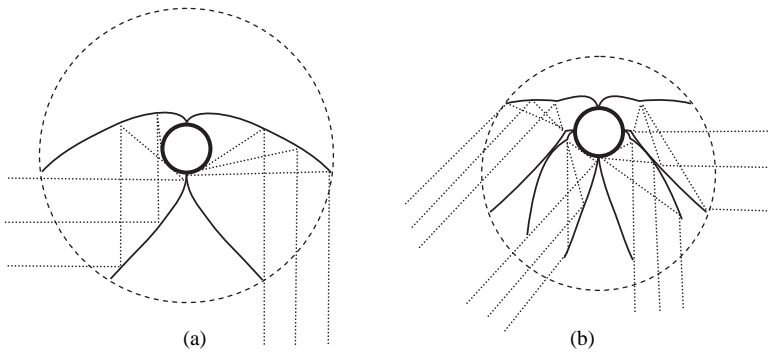
Figure 13.14   (a) Mills (1995) and (b) Collares et al. (1991) designed CPC-type secondaries for the Lambertian bundles coming from several primary sections.
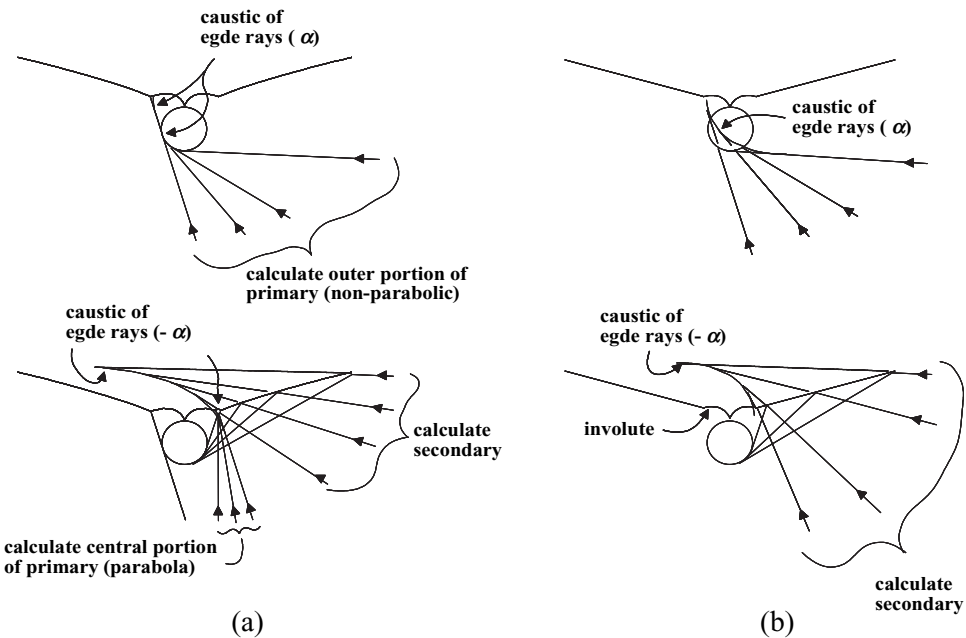
Figure 13.15   Secondary profile and caustics of the edge rays reflected by the primary when the primary is (a) fixed to a parabola and only the secondary is designed, and (b) is designed simultaneously with the secondary.

Assuming that the primary mirror contour is fixed to a parabola is also an unnecessary restriction, which limits the performance of the two-stage collectors. Benítez et al. (1997) applied the SMS design method (see Chapter 8) to redesign that of Ries and Spirkl to show that the simultaneous unrestricted design of primary and CPC-like secondary noticeably improved the optical performance. The comparison numbers of these two qualitatively identical designs given in Table 13.2. As a result, the SMS design for comparison, both secondaries and the caustic of the edge rays reflected on the primary, are shown in Figure 13.15.

Table 13.2.   Comparison of the edge rays and caustics of Benítez et al.'s (1997) concentrator and Ries and Spirkl's concentrator (1996). In the former the primary is designed simultaneously with the secondary, while in the latter the primary is constrained to have parabolic profile.

Rim angle ϕ (degs)	C/C_{MAX}		Secondary shade
	SMS	Parabola + TERC	
90°	0.74	0.68	2.2%
120°	0.67	0.50	1.7%
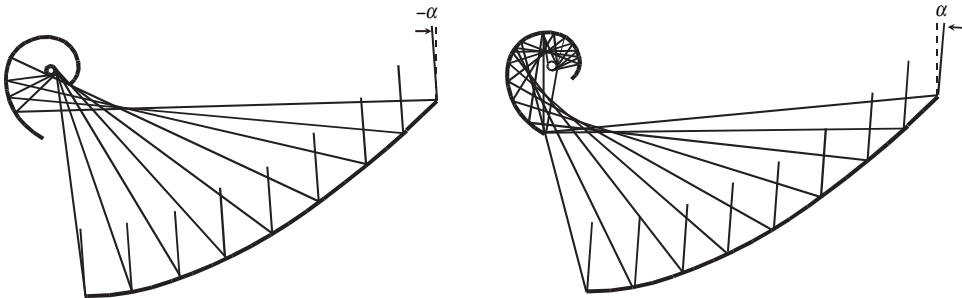
Figure 13.16   Edge rays used in the Snail concentrator design.

All the aforementioned secondary designs are flow-line mirrors (i.e., mirrors that guide the flow lines as a funnel). This type of mirrors has been called "non-sequential" mirrors in Chapter 8 and Appendix B. The flow-line mirrors have the feature that they have to be located in contact (or very close) to the absorber tube to achieve high concentration. This forces designers to locate the secondary inside the evacuated tube, which may be unpractical from the manufacturing point of view.

The SMS method of Chapter 8 was also used to design two contactless non-imaging solar thermal collectors, called Snail and Helmet concentrators, which use sequential mirrors instead of CPC-like mirrors (Benítez et al., 1997). These designs have a sizeable gap between the optics and the absorber, which has practical interest, as explained following. As in the case of the design of Figure 13.15b, in the Snail and Helmet concentrators the primary and secondary mirrors are calculated simultaneously. Since the primary is not forced to be parabolic, there are again more degrees of freedom available to achieve an improved performance.

The cross section of primary and secondary (this is not to scale) of the Snail concentrator is shown in Figure 13.14. Its name (and the Helmet's) is due to the shape of its secondary mirror. Figure 13.16 also shows some of the edge rays used in the design process. A detailed explanation of the design procedure for the Snail is given in Benítez et al., 1997.

The Helmet concentrator cross-sectional diagram is shown in Figure 13.17. On the contrary to the Snail, the Helmet is symmetric as the conventional parabolic trough. Further details about the design procedure can be found in Benítez (1998). Figure 13.18 shows the cross sections of two selected designs.
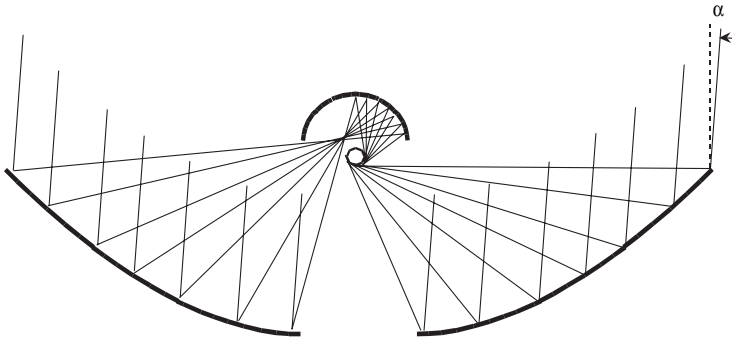
Figure 13.17   Edge rays used in the Helmet concentrator design (secondary not to scale).
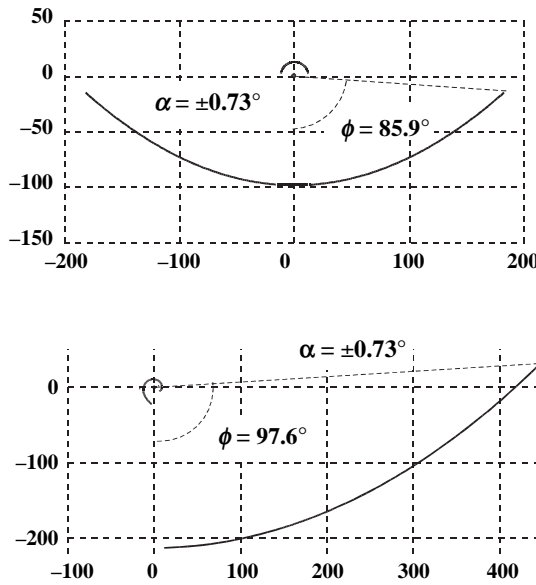
Figure 13.18   Cross section of selected Snail and Helmet concentrators, designed for an acceptance angle $\alpha = \pm 0.73°$. The angle ϕ is the upper rim angle. Other geometrical parameters are absorber radius = 1, (Helmet) C_G = 54.6, Gap/absorber radius = 11.1, (Snail) C_G = 69.3, Gap/absorber radius = 6.8.

The Helmet concentrator fulfills the condition that the average number of reflections on the secondary <n> is the lower bound attainable for two-stage collectors and for its design parameters—that is, the same rim angles and geometrical concentration. This lower bond can be expressed (for α small) as

$$\langle n \rangle \geq \eta_H - \eta_P \tag{13.7}$$

where η_H and η_P are the collection efficiencies for $\rho = 1$ of the Helmet concentrator and the parabolic concentrator with the same rim angles and geometrical concentration. The equality, reached by the Helmet, holds when (1) all the straight lines that link every point P of the primary with the absorber coincide with
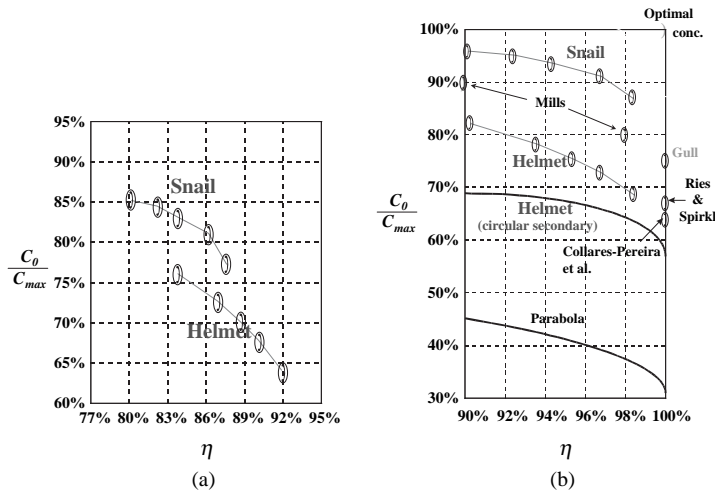
Figure 13.19   The performance of the concentrators can be compared when represented in the $\eta - C_0$ plane. The mirror reflectivity is assumed to be (left) $\rho = 0.95$, (right) $\rho = 1$.

trajectories of rays reflected at P, and (2) the remaining collected rays reflected at P suffer only a second reflection on the secondary before reaching the absorber.

In order to compare the performance of Snail and Helmet concentrators with the designs presented previously, let us represent their corresponding points of the $\eta - C_0$ plane, where η is the collection efficiency, defined as the fraction of the incident power that reaches the absorber, and the optical (or flux) concentration C_0, which is the ratio of the average irradiance on the absorber to the irradiance at the entry aperture and can be calculated as $C_0 = \eta C_G$. Figure 13.19a shows the performance points of several Snail and Helmet concentrators designed for $\alpha = \pm 0.73°$ in the $\eta - C_0$ plane, calculated by ray tracing assuming a uniform brightness distribution within $\pm \alpha$. A reflectivity $\rho = 0.95$ for the primary and secondary mirrors has been assumed. The C_0 axis is normalized to the thermodynamic limit of concentration, $C_{max} = 1/\sin \alpha = 78.5$.

The Snail concentrators get higher flux concentrations for the same collection efficiencies than the Helmet concentrators. If a higher reflectivity of the secondary is assumed (today available as the multilayer interferential mirrors of 3M), the efficiency and concentration of Snail concentrators increase more than those of the Helmet concentrators due to their different average number of reflections <n> on their secondaries: 1.2–1.4 in the Snails and 0.3–0.5 in the Helmets. Figure 13.19b shows the $\eta - C_0$ plane in the limit case when $\rho = 1$ (not only for the secondaries but also for the primaries). The performance points of the other two-stage concentrators of Collares-Pereira et al. (1991), Mills (1995), Collares-Pereira and Mendes (1995), Ries and Spirkl (1996) and Benítez et al. (1997) (the last one, labelled as Gull) are also represented in this figure for comparison purposes. These concentrators are not shown in Figure 13.19a because the values of η and C_0 with reflection losses have not been reported for some of them. The calculation of η considers as entry aperture only the nonshaded portion of the primary mirror in all cases.
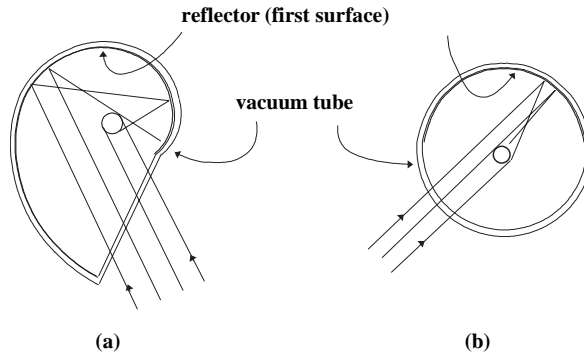
Figure 13.20   Example of integrated design of secondary concentrators and evacuated tube for (a) the Snail concentrator and (b) the Helmet concentrator.

The best concentrator in the $\eta - C_0$ plane of Figure 13.19b would be an optimal concentrator (not yet available in practice) which has $\eta = 100\%$ and $C_0 = C_{max}$. The best parabolic troughs (with rim angle $\phi = \pm 90$ degrees) are also represented in the figure.

The Snail and Helmet concentrators are as yet the only known concentrators with a sizeable gap between the optics and the absorber that are efficient and approach the optical concentration limit. The other two-stage concentrator designs need the secondary to be very close to the absorber in order to be efficient, and this implies that the secondary must be located inside the evacuated tube. Due to the high temperature of the absorber, this might be complex guaranteeing low heat-conduction losses through the secondary mirror supporting points. On the contrary, secondary mirrors placed outside the evacuated tube or integrated with it (as done with stationary evacuated CPC collectors) are possible when there is a sizeable gap between absorber and optics, as shown in Figure 13.20.

In the case of the Helmet concentrator, the integrated secondary option is especially attractive in terms of its simplicity: If the secondary can be approximated by an arc of circumference, it would be sufficient to metallize a conventional glass tube. In this approximation, the parameters to optimize are the ordinate of the center of the circumference z_C and its radius R'. Figure 13.21 shows the curves associated with the Helmet concentrators with circular secondaries for various values of R', with $z_C = R$. This figure also shows the points corresponding to the five Helmet concentrators in Figure 13.19 and the curve associated with the parabolic trough with rim angle $\pm 90°$, which are the references for comparison. The calculations made assume a reflection coefficient of $\rho = 100\%$. As it can be seen, for an efficiency $\eta = 96.8\%$, the circular secondary with radius $R' = 6$, R achieves a concentration 1.92 times that of the 100% efficient parabola and 1.56 times that of the parabola with the same collection efficiency. If a radius of $R_S = 11\ R$, is assumed for the secondary factors of 2.07 and 1.68 are obtained, respectively. The same comparison carried out for the Helmet concentrator with respect to the parabola gives values of 2.29 and 1.86.

The Helmet concentrator is symmetric like the conventional parabolic trough and has a rim angle close to $\pm 90°$, which is of interest for good mechanical stability of the collector. In contrast, the Snail concentrator is not symmetric, although
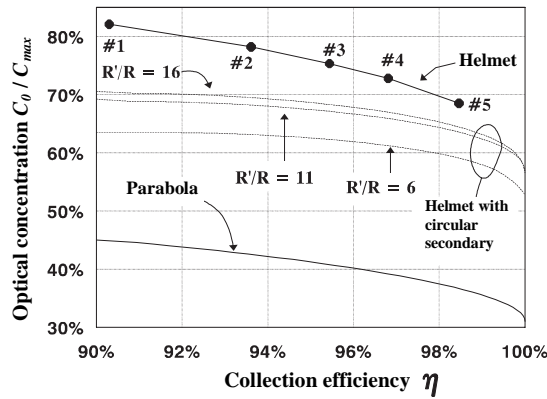
Figure 13.21   Results of the analysis of the approximation of the secondary of the Helmet concentrator by an arc of circumference.
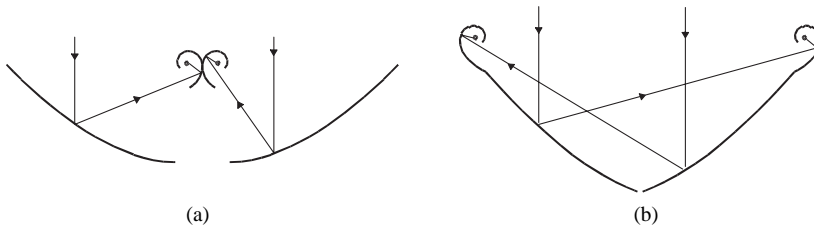
Figure 13.22   Symmetric two-Snail systems: (a) parabolic trough type and (b) Single-Mirror Two-Stage (SMTS) type.

the upper rim angle of the primary is also high (90°–105°). The design of asymmetric two-stage concentrators for tubular absorbers is not new. If a symmetric system is preferred, it can be obtained by placing the two asymmetric designs (with their respective receivers) "back to back." Another way of obtaining a symmetric collector is shown in Figure 13.22. An advantage of this configuration is that it does not need a specific supporting structure for the absorber and the secondary mirror. This strategy was first introduced in the SMTS (Single-Mirror Two-Stage) photovoltaic concentrator, which will be presented in Section 13.4.2.

13.3.4   Tracking Receiver, Stationary Reflector

A useful hybrid configuration that retains many of the advantages of static systems while maintaining the relatively high Cg of tracking troughs combines a fixed stationary reflector with a tracking absorber. Although a number of such configurations have been successfully demonstrated in the past, we will describe one promising system currently under development. The fixed primary reflector is a cylinder. By symmetry, rays of any elevation angle are equally (but poorly) concentrated into a region bounded by two caustics. The concentrated flux zone moves with incident direction so that the receiver has to track. By itself, this configuration would not have a high enough Cg to be effective. But when combined with a
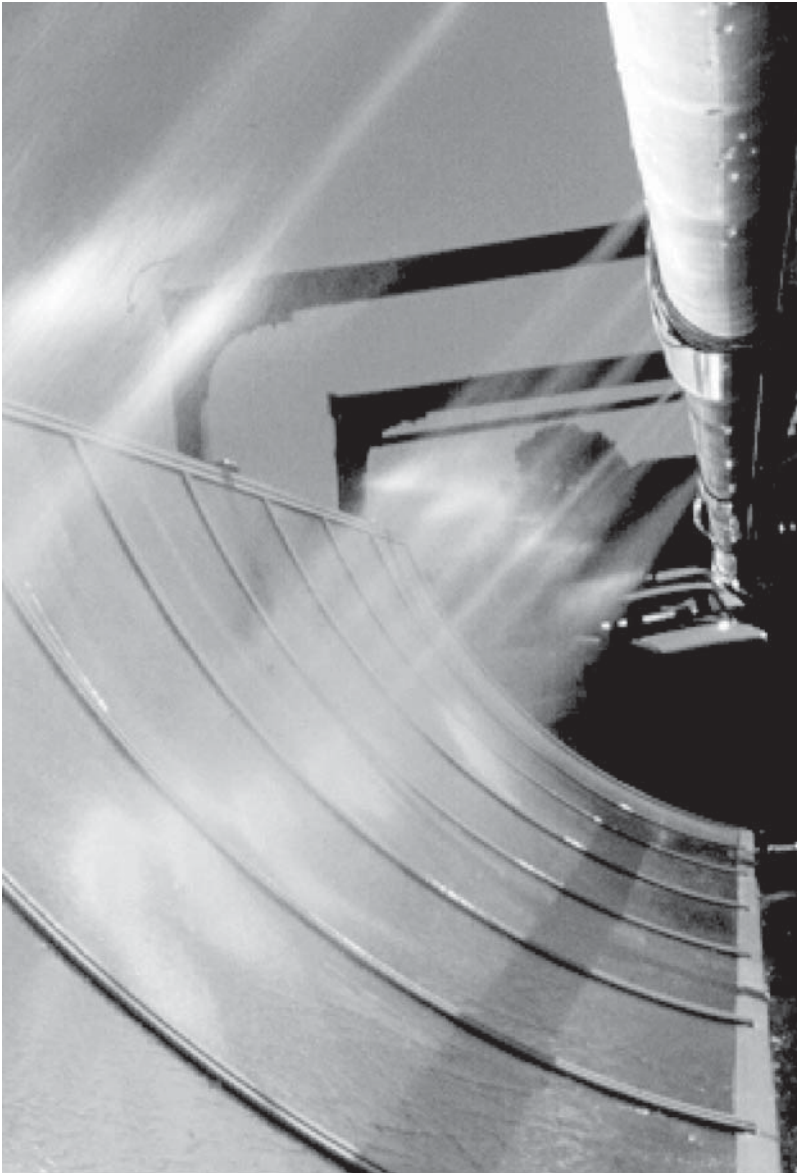
Figure 13.23   Stationary primary with tracking nonimaging secondary solar concentrator for high temperature.

secondary concentrator, the overall Cg can rival and even exceed the tracking parabolic trough. The angular distribution presented to the secondary concentrator is highly nonuniform and varies significantly across the aperture. Therefore, this is a case for the tailored edge-ray method of nonimaging concentrator design. Such systems are under development for mid- to high-temperature applications. A prototype with $Cg = 40$ and evacuated receiver was tested at Sandia (Figure 13.23) and showed a heat loss coefficient (the efficiency drops only 3% over a span of
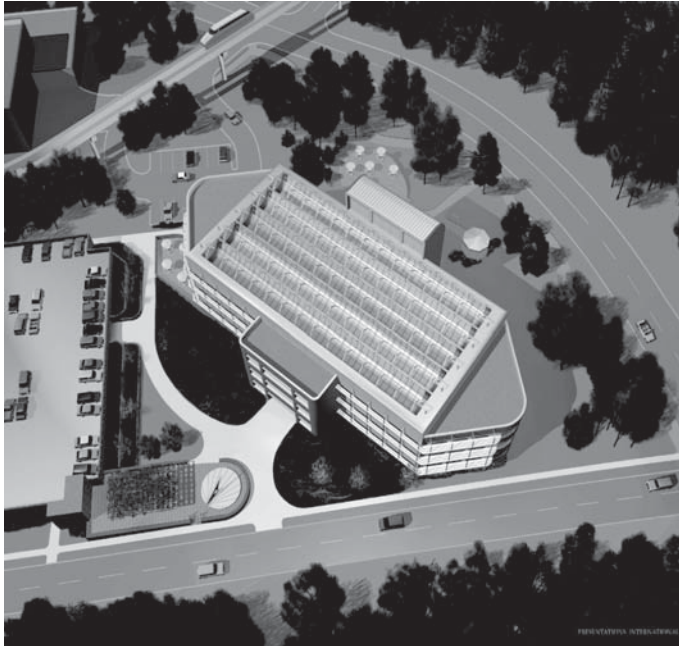
Figure 13.24   A building roof concept for stationary primary with tracking secondary solar thermal system.

300°C), which is even lower than tracking parabolic troughs with evacuated receivers. There are many advantages to having the primary reflector fixed. One of the principal benefits is the ability to make the primary reflector part of the roof structure of a building (Figure 13.24). Another is the potentially lower cost associated with a fixed primary reflector.

13.3.5   Tracking Concentrators, Two-Axis (Cg = 50–1,000)

Common types of 2D concentrators are parabolic dishes and Fresnel lenses. Parabolic dishes for driving high-temperature engines such as Stirling and various turbines have been under development for some time. The use of secondary concentrators has been proposed to improve performance by increasing either Cg, or intercept, or reducing tracking and tolerance requirements. For practical reasons, experiments have generally retrofitted existing dishes with secondary concentrators, rather than taking the preferred path of designing both primary and secondary elements together. The flow-line design lends itself to this retrofit mode because the secondary reflector is positioned forward of the focal plane of the primary. For thermal applications, concentration is necessary to achieve high temperature.

Since the concentration for a focusing single-stage reflecting concentrator maximizes for values of f near 45°, most primaries designed for use without secondaries have f-number F/D ratios of between 0.5 and 0.7. Reflecting secondaries in retrofit applications with these primaries such as illustrated in Figure 13.25
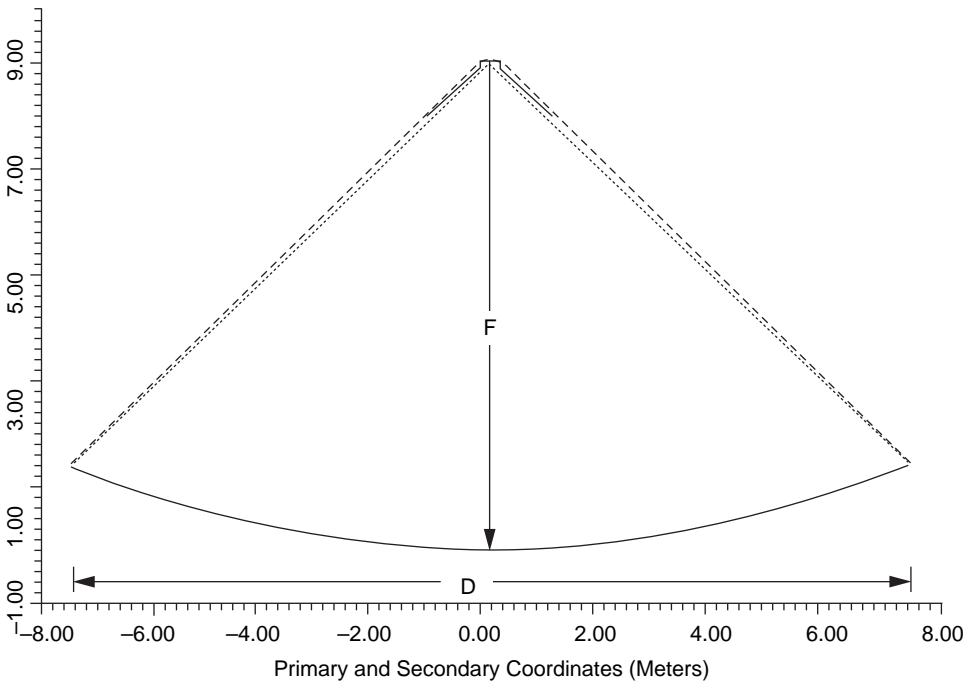
Figure 13.25   Dish primary with flow-line "trumpet" secondary.

are limited to concentrations between 1.5 and 2.0. However, even this moderate improvement can be quite important in increasing system performance. For example, if the root-mean-squared (rms) slope errors on the primary are about 5 milliradians, the corresponding "thermodynamic limit" for the concentration of sunlight in a medium with index of refraction $n = 1$ (i.e., in air) is about 1,600×, and the best single stage concentrator will be able to achieve only about one-quarter of this, or about 400×. A trumpet concentrator can increase this effective concentration readily to about 750×.

A project is currently being carried out at the University of Chicago group to design a practical trumpet secondary concentrator to be used in combination with a faceted membrane primary concentrator for dish-Stirling applications (Bean and Diver, 1992). This is a retrofit design for a dish that was originally designed as a single stage, and so it cannot attain the full power of a fully optimized two-stage concentrator. The configuration is shown in Figure 13.26. Nevertheless, preliminary ray-trace studies show that the addition of a small trumpet to the receiver aperture will allow a reduced aperture size and corresponding lower thermal losses at the operating temperature of 675°C. In addition the secondary will significantly increase tolerances for tracking error and surface slope errors on the faceted membrane primary. The optical design has been completed, and a water-cooled proto-type has been constructed and tested under the supervision of Sandia National Laboratories. This relatively small step marks the beginning of the incorporation of these devices into practical commercial solar thermal concentrators and is a first step toward the eventual utilization of their full potential.
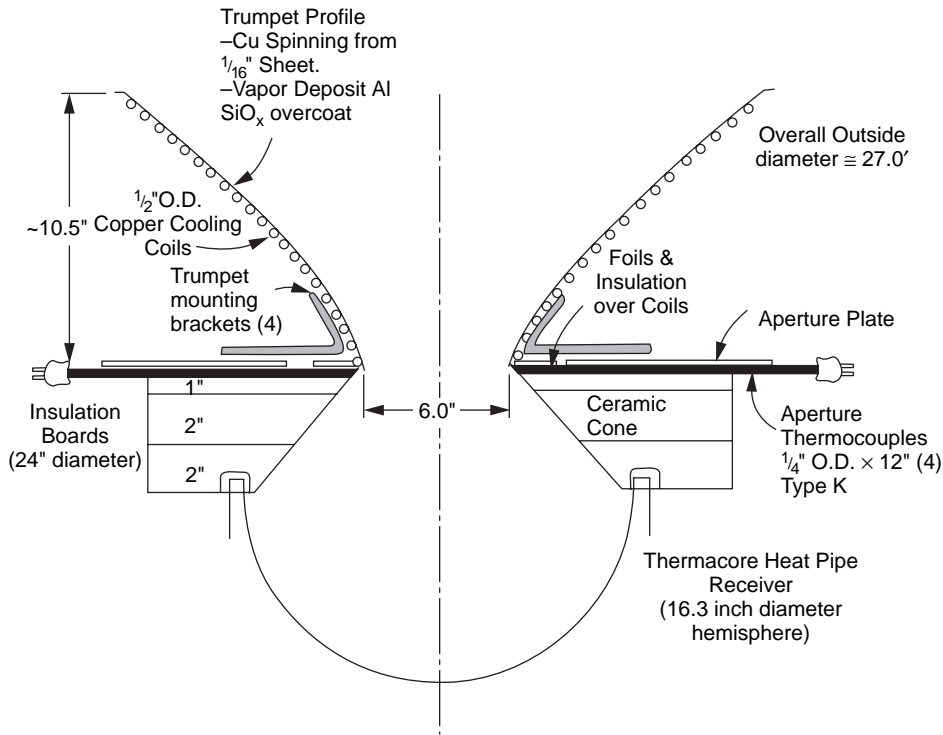
Figure 13.26   Trumpet secondary with thermal receiver.

The basic advantages of nonimaging concentrators have to do with their ability to achieve the highest possible concentration within a given angular field of view. The practical effects of this limit are very important but are often misunderstood. One can combine the entire source broadening mechanisms into a single effective Gaussian source distribution as follows:

$$S(\theta) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(\frac{-\theta^2}{2\delta^2}\right) \tag{13.8}$$

where θ is the incidence angle relative to the concentrator pointing axis and

$$\delta^2 = (\sigma_{\text{sun}})^2 + (2\sigma_{\text{slope}})^2 + (\sigma_{\text{spec}})^2 + (\sigma_{\text{tracking}})^2 \tag{13.9}$$

is the root-mean-squared (rms) deviation characterizing the combined effective source distribution and σ_{sun}, σ_{slope}, σ_{spec}, and σ_{tracking} are the effective rms deviations characterizing the sun shape, primary slope error, specularity, and tracking errors, respectively. The factor of two in the slope error term comes about because an error in the slope results in a change in the angle of a reflected ray by twice the amount of the slope error, as can be seen from Figure 13.27 (for further details, see Chapter 14.). In principle, if the primary mirror surface slope errors could be reduced to
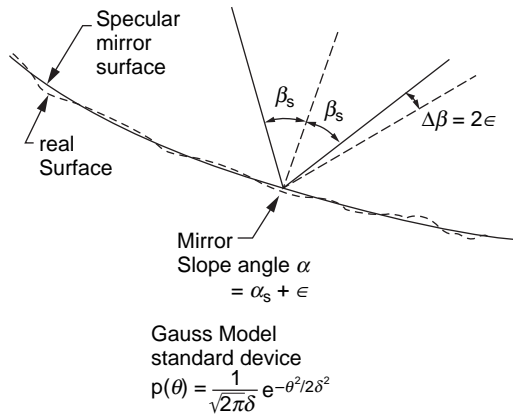
Figure 13.27   Effect of mirror slope errors. (For further details, see Chapter 14.)
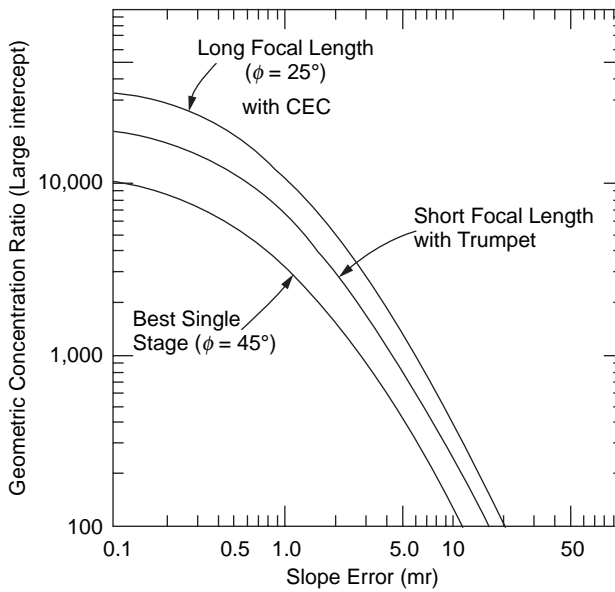
Figure 13.28   Concentration gain for secondary in conjunction with primary element.

zero, C max could be somewhat over 46,000. This is the "thermodynamic limit" for the concentration of sunlight in a medium with index of refraction $n = 1$.

For practical concentrators this slope error reduces the achievable concentration for a given slope error that can be achieved with a respectable intercept to values well below the limit as illustrated for single- and two-stage systems in Figure 13.28. Spherical primaries with a nonimaging secondary can exceed the performance of a single-stage paraboloid if long focal length configurations are employed as shown in Figure 13.29.
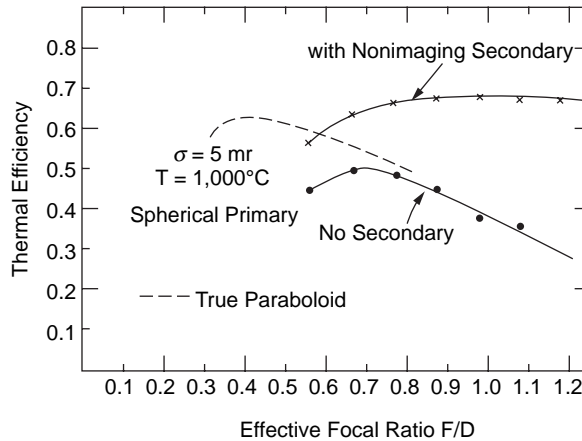
Figure 13.29   Efficiency gain for secondary in conjunction with primary element.

Table 13.3.   The Solar Two project.

A six-year, $52 million utility/industry/government cost-shared project to validate
  advanced molten-salt solar power tower technology by retrofitting Solar One.

Objectives:
- To stimulate the design, construction, and operation of the first commercial plants
- To validate the technical characteristics of a molten salt plant
- To improve the accuracy of economic projections
- To collect, evaluate, and distribute information to foster interest in commercialization
- To stimulate formation of a commercialization consortium

13.3.6   Central Receiver (Cg = 500–1,000)

The central receiver or "solar power tower" concept has been developed and demon-
strated over many years. Proposed by Soviet scientists in the 1960s, the concept
was adopted in a U.S. program in the 1970s as a promising path for power
generation (Vant-Hull and Hildebrandt, 1976). A good example is the very large-
scale demonstration *Solar Two*, a 10 MW facility in the California Mojave desert.
The objectives of this large-scale demonstration of central receiver technology are
shown in Table 13.3. A distinctive feature of Solar Two compared to its predeces-
sor Solar One was the use of molten salt as the heat transfer and storage medium
(Figure 13.30). A schematic diagram is shown in Figure 13.31. The heliostat field
directs the incident solar radiation to a molten salt receiver on top of a tower
(Figure 13.32). The heated salt can be used to drive a turbine or be stored in tanks.
After successfully meeting its demonstration objectives, the plant was decommis-
sioned in early 1999.

An innovative development of the central receiver concept was adding a relay
reflector to position the concentrated solar flux close to the ground (Rabl, 1976).

Figure 13.30   "Solar Two" central receiver plant.

This has the obvious advantage of allowing the power block also to be on the ground. In this configuration, the use of secondary concentrators is indicated to attain high flux at the receiver. This has been done in a large-scale demonstration at the Weizmann Institute of Science in Rehoveth, Israel. A schematic diagram of the concept is shown in Figure 13.33. The demonstration is being carried out at the Weizmann Solar Tower (Figure 13.34) (Yogev, 1997). Nonimaging final stage concentrators (Figure 13.35) convey highly concentrated solar flux to a receiver

Molten Salt Technology

Advanced molten salt technology with storage offers
- Solar/EPGS decoupling
- Load following capabilities
- Capacity factors to 60%
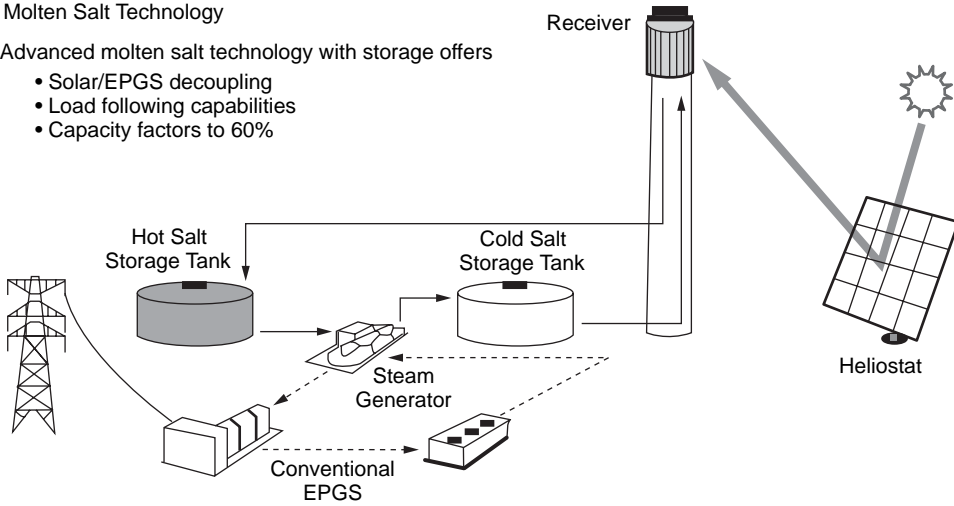
Figure 13.31   Schematic of central receiver solar power plant.
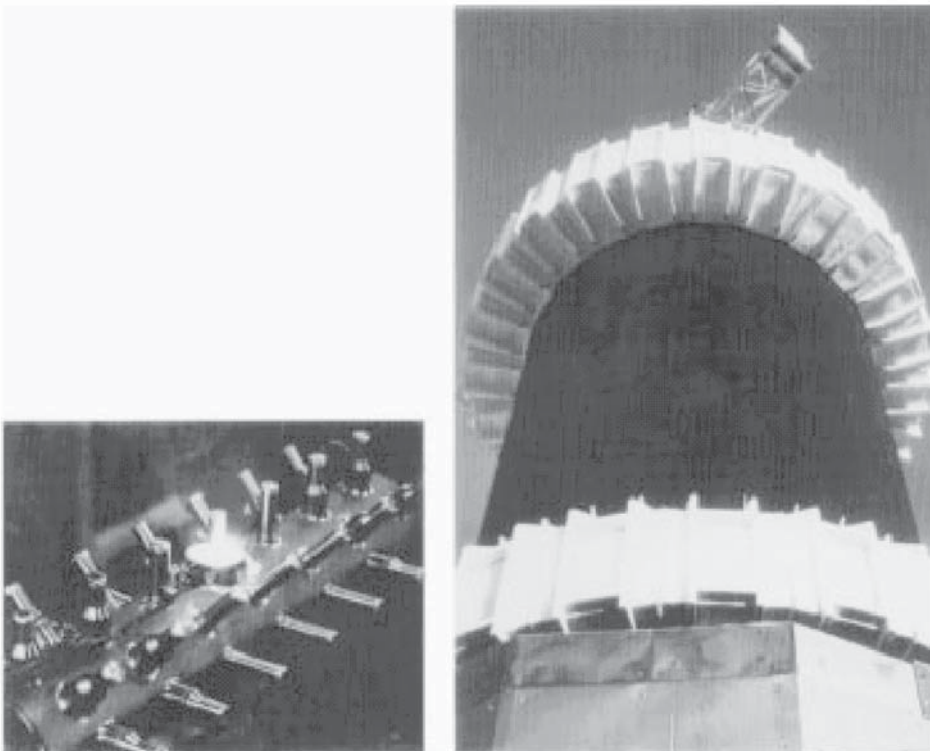
Figure 13.32   Molten salt receiver on top of tower.
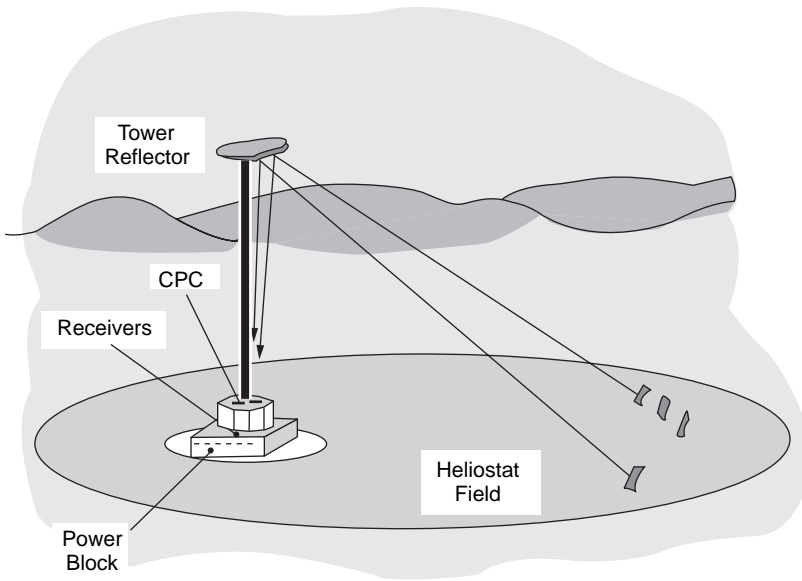
Figure 13.33   Schematic of Weizmann "beam down" central receiver plant.
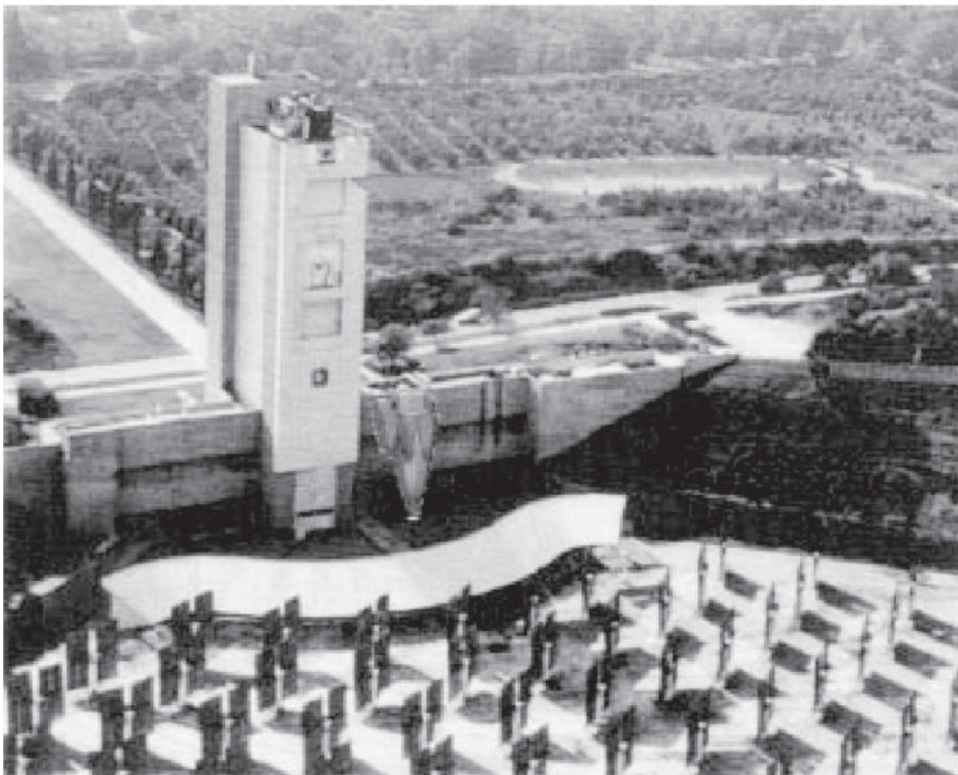
Figure 13.34   Weizmann solar tower with ground-based CPC receiver depicted to the right of tower.
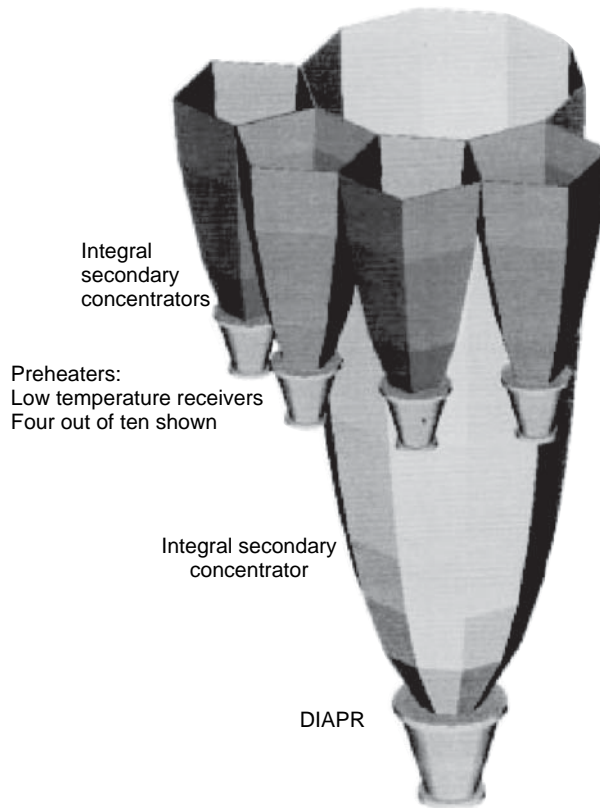
Figure 13.35   The CPC receiver for the Weizmann "beam down" central receiver.

and heat exchanger (Figure 13.36). A test of the receiver components at the solar tower is shown in Figure 13.37.

13.4   SMS CONCENTRATORS FOR PHOTOVOLTAIC APPLICATIONS

The progress in solar cells' development since their invention has been huge. The investment needed has also been important. Present cells are much more efficient and the technology much more advanced. Next-generation approaches attempt to further cell development, increasing the cell sophistication if needed.

However, the photovoltaic concentrator development has not accompanied such progress. This is obvious in the fact that most photovoltaic concentrator systems have been based either in the parabola or in the Fresnel lens (see Figure 13.38). The parabola has been recognized since Menaechmus (380–320 B.C.) and its optical properties since Apolonius (262–190 B.C.), and Galileo (1564–1642) used lenses for making the first refractive telescope. The thinning of such lenses to avoid absorption was proposed by Georges de Buffon in 1748 (and later optimized by Fresnel in 1822).
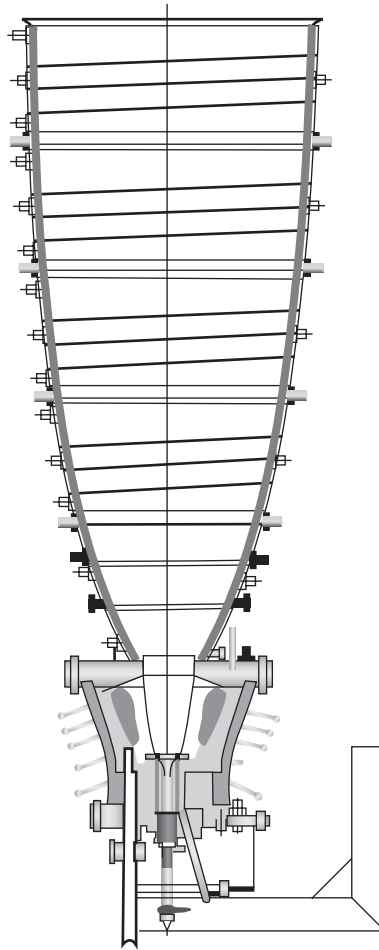
Figure 13.36   Detail of the Weizmann CPC and DIAPR receiver.

The parabola and the Fresnel lens perform far from the theoretical concentration-acceptance angle limits. The question is: Are they good enough, or do we need to come closer to the limit? At low concentration, maybe. But the present high-concentration trend, needed to reduce the effect in the global cost of the expensive high efficiency cells, suggests that parabola and Fresnel lens are not good enough.

Nonimaging optics came into the photovoltaic field as an option for low (static) concentration (Winston, 1974) or as a CPC-type secondary optical element for higher concentration photovoltaics (Rabl and Winston, 1976). The first full nonimaging device for nonstatic concentration was a nonimaging linear Fresnel lens with curved aperture (Kritchman et al., 1979).

Among all these nonimaging systems, only the addition of the CPC-type secondary to a classical primary could work for very high-concentration systems (>1,000×) with sufficient acceptance angle ($\alpha > 1°$) (Ning, Winston, and O'Gallagher, 1987). However, apart from the lack of illumination homogeneity and the practical problems related with the flow-line mirrors, this solution is necessarily
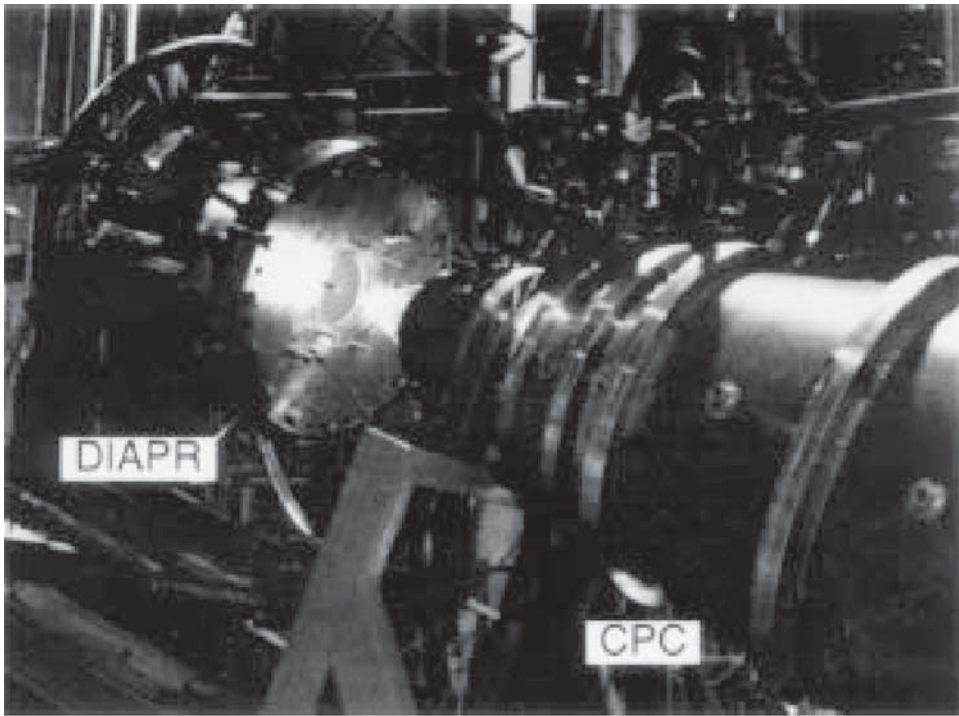
Figure 13.37   Weizmann receiver under test.
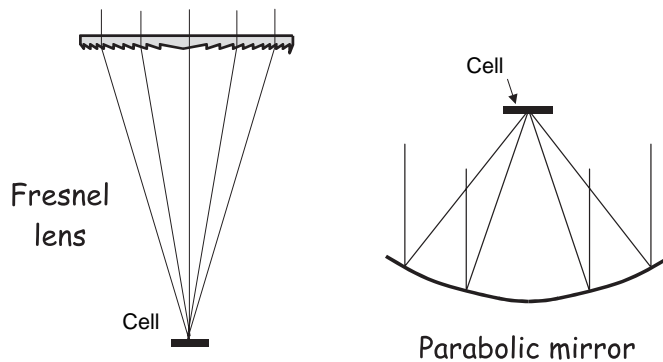
Figure 13.38   Fresnel lenses and parabolic reflectors are the classical concentrators in photovoltaics.

incompact (depth H to entry aperture diameter D ratio $H/D \approx 1.5$ for Fresnel lens, and $H/D \approx 1$ for parabolic mirror). This is due to the design approach, in which the secondary and primary are designed separately, and thus the secondary does not consider the actual rays exiting the primary but considers the primary as a Lambertian source (Welford and Winston, 1980), as indicated before in Section 13.3.3 for the solar thermal case. The smaller H/D ratio, the lower accuracy of this approximation, and thus the concentration-acceptance angle product is reduced.

The highest-performance compact concentrators need to design primary and secondary simultaneously, as is done with the SMS method. This method has pro-

duced new photovoltaic concentrator families during the last decade, which are presented in the next sections.

13.4.1   Static Concentrators for Bifacial Photovoltaic Solar Cells

The interest of static concentration in photovoltaics (PV) resides on several aspects. First, the concentration reduces the area of the solar cell, which is by far the most expensive component of a flat PV module, and thus cost reductions seem possible. Second, as the cost of the module components is more distributed in concentration systems, they are less affected by the fluctuations of the market price of silicon wafers. And third, when compared with tracked concentrators, the static ones do not need higher maintenance than flat modules to operate.

The PV static concentrators are usually designed for bifacial solar cells. The bifaciality doubles the possible ray trajectories that can reach the cell, which implies that the concentration limit is also doubled. This type of cell is not practical in high concentration systems (as tracking ones) because there is no inactive cell face to which to fix a heat-sink, and the dissipation becomes more difficult. The low concentration required is compatible with the design for an acceptance angle of $\alpha = \pm 30°$ (typically), which is needed to have a reasonable intercept factor (over 80% typically) when the collector is placed facing south and pointed to the ecliptic.

The concentration limit can also be increased if the cell is surrounded by an optically dense media. If its refractive index is n, then the concentration limit multiplies by n^2, so $n = 1.5$ implies a factor of 2.25 (slightly higher than that of the bifaciality). However, except for a few concentrator designs (Bowden, 1993; Yoshioka et al., 1995), most approaches have linear symmetry (i.e., that of a trough), which leads to a lower theoretical limit for the attainable concentration. The factor 2.25 reduces to ~1.73 (as seen in Chapter 5, section 5.2). This concentration limit reduction is possibly compensated with the simpler manufacturing of linear symmetric devices by low-cost extrusion techniques.

The SMS design method (see Chapter 8) was applied to the design of two static linear nonimaging designs for bifacial solar cells (Benítez et al., 1999). It has been done under the framework of the "Venetian Store" project, partially funded by the European Union via the JOULE program. This project is led by the Spanish PV manufacturer ISOFOTON (which has already a bifacial cell technology) and aims to develop a commercial static concentration product for building integration at competitive cost with the aesthetic of a Venetian store, whose slats will be the concentrators.

These two SMS designs are compact and have a performance close to the thermodynamic limit in 2D geometry. The concentrators are named RXI_F and the RX-RXI, according to the nomenclature introduced in Chapter 8. In this nomenclature, each concentrator is named with a succession of letters indicating the order and type of incidence on the optical surfaces that the transmitted bundle encounters on its way from source to receiver. The following symbols are used: R = refraction, X = sequential metallic reflection, I = sequential total internal reflection, X_F = nonsequential metallic reflection, and I_F = nonsequential total internal reflection (the subscript F refers to the coincidence of the nonsequential mirror with a flow line).

The RXI_F concentrator is made by a dielectric piece of refractive index n with one mirrored face (see Figure 13.39). The rays transmitted by the RXI_F are
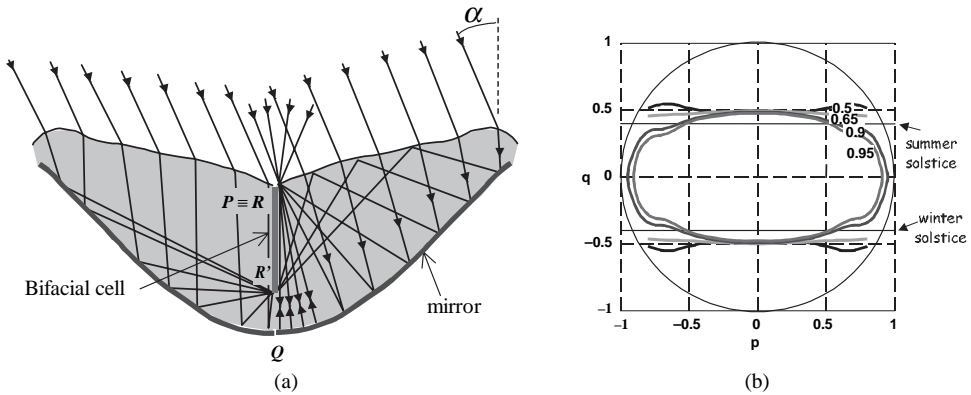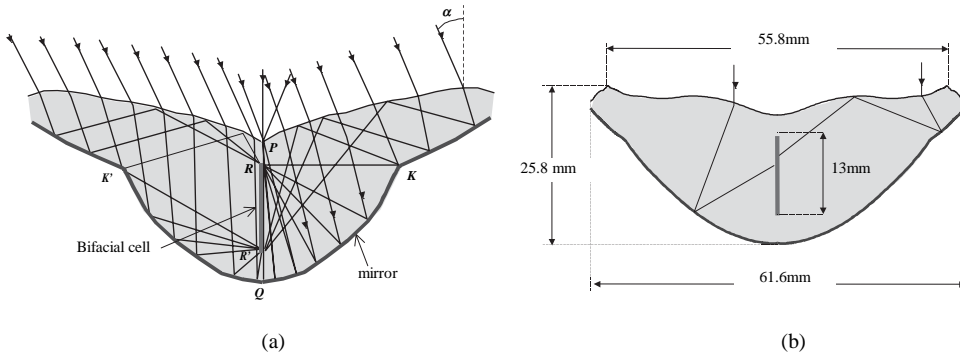
Figure 13.39   (a) Assignation of the edge rays used for the calculation of the optical surfaces in the RXI$_F$ concentrator. (b) Directional collection efficiency for concentrator 3 in Table 13.4 (neglecting the optical losses).

refracted on the upper refractive surface (R), reflected on the back mirror (X) and may reach the cell directly or after being reflected again (one or more times) by total internal reflection (TIR) on the upper refractive surface (I$_F$). Then, the upper aspheric has a double function: It acts as a refractive surface for the incident light and as a nonsequential mirror by total internal reflection for the rays after reaching the cell. Thus, the edge R of the cell is in contact with the refractive surface at P, as usual in nonsequential mirrors.

The edge rays of the input bundle that are used in the design are the rays that form an angle of $+\alpha$ with the vertical direction (let us call these rays i(+)) and those passing through the point P of the upper refractive surface. By the symmetry of the concentrator, the symmetric rays are also designed trivially. Figure 13.39 shows the assignation of the edge rays used for the calculation of the optical surfaces. The $+\alpha$ rays impinging on the left half of the concentrator are focused on the lower edge of the cell R' with one refraction and one metallic reflection. The rays passing through P are focused on R and then, as $P \equiv Q$, the left portion of the mirror is a circumference. The assignation of the $+\alpha$ rays impinging on the right half of the concentrator is more complex (Figure 13.39b). The rays impinging on the inner most portion of the upper refractive surface are focused on R with one refraction and one metallic reflection. The rays impinging on the outer portion are focused on R' with one refraction, one metallic reflection, and one total internal reflection. In the transition between both portions, the rays are sent after one refraction and one metallic reflection toward the TIR mirror with an incidence angle equal to the critical angle $\sin^{-1}(1/n)$ (as in the DTIR concentrator design, presented in Chapter 5, section 5.2).

The parameters of this design are (1) the acceptance angle α, (2) the refractive index n, (3) the cell width, (4) the distance from R' to Q, and (5) the geometrical concentration C_G. The geometrical concentration (since the whole input ray bundle is collected) is bounded by the 2D limit $C_{MAX} = 2n/\sin\alpha$. Because the total internal reflection condition must be fulfilled, the attainable concentration is slightly lower than this limit and can be estimated as $C'_{MAX} \cong (1 + \cos\beta)n/\sin\alpha$, where the angle $\beta = \sin^{-1}(1/n) - \tan^{-1}(\sin\alpha/(n - \cos\alpha))$. For $\alpha = 30°$ and $n = 1.5$, the former expres-

Table 13.4. Some geometrical characteristics of RXI_F designs with $\alpha = \pm 30°$ and $n = 1.473$ ($C_{MAX} = 5.89$).

RXI_F concentrator	#1	#2
C_G/C_{MAX}	0.933	0.815
Gap size (QR'/RR')	0.475	0.396
Aspect ratio (height/width)	0.472	0.438

Table 13.5. Cell temperature drop to ambient for the RXI_F #2. Operating conditions: Irradiance = 800 W/m², wind speed = 2.3 m/s (front) – 0 m/s (back), cell efficiency = 10%.

	Δ Temperature (in °C)	
Illumination profile	Uniform	Punctual
RXI_F, Glass	32.6	33.8
RXI_F, Plastic	80.2	80.7
Flat module	23.0	—

sion leads to $\beta = 3.5°$, and C'_{MAX} is 99.8% of as C_{MAX}. The higher the concentration, the less the compactness. Table 13.4 shows some geometrical characteristics of two selected RXI_F concentrator designs. Concentrator #2 is the one in Figure 13.39a, and its 3D directional collection efficiency is shown in Figure 13.39b.

The complete technical design of the concentrator must take into account not only optical aspects, but also electrical and thermal ones. Cell performance is affected by the nonuniformity of the illumination and the operation temperature. These aspects compromise the design, the manufacturing technology of the concentrator, and its scale. For instance, for the RXI_F design #2, with a scale such that the cell is 13 mm wide, the thermal compromise excludes plastics as dielectric materials. The reason is that the vertical position of the cell inside the dielectric material makes the heat removal more difficult than in a flat module. The finite differences solution of the problem of calculating the distribution of temperature inside the dielectric provides the data shown in Table 13.5.

It shows that the cell temperature under uniform illumination would be 9.6°C higher than that of a flat module under the same operating conditions. It also has been found that the nonuniform illumination does not increase appreciably the average cell temperature (1.2°C for a 56× pillbox-type illumination at the cell bottom). This is due to the lateral heat conduction along the cell width, which is good enough at these concentration scale levels.

The RXI_F has by design a sizable gap between the bottom of cell (point R') and the mirror. This situation is similar to that of the top edge of the CPC for bifacial vertical receiver. However, as opposed to this CPC case, the RXI_F gap has the property that no collected ray passes through it. Therefore, the metal tabs used for interconnecting the cells can be placed vertically at the gap without shading.

In the photovoltaic case, gaps are important to ensure the environmental protection of the cell. Unfortunately, providing a gap for the top of the cell in the RXI_F produces optical losses (as in the case of the bottom of the cell in the bifacial CPC). However, this is solved with the other aforementioned SMS design, the RX-RXI

Figure 13.40   (a) Assignment of the edge rays used for the calculation of the optical surfaces in the RX-RXI concentrator. (b) Practical RX-RXI concentrator with no kink designed for a cell 13 mm wide, $\alpha \pm 30°$, $n = 1.473$, $C_g = 4.3\times$ (0.728 C_{max}).

concentrator, which has nonshading sizable gaps for both edges of the cell (see Figure 13.40). Because no nonsequential mirror is used at this time, both edges R and R' of the cell are not in contact with the optical surfaces.

The RX-RXI concentrator is also made by a dielectric piece with one mirrored face. Conceptually, it is a combination of the well-known RX and RXI concentrators (see Chapter 8). The point K (and K'), where the mirror has a kink, separates the RX and RXI portions. The rays transmitted by the central portion (which is the RX one) are refracted on the upper surface (R) and reflected on the back mirror (X) toward the cell. On the other hand, the rays transmitted by the outer portion (which is the RXI one) reach the cell after being reflected once more by total internal reflection on the upper surface (I). Then, as in the RXI_F, the upper aspheric has again a double function, as a refractive surface and as a sequential mirror by total internal reflection. The input ray bundle and its edge rays used in the design are the same as those of the RXI_F.

Figure 13.40a shows the assignation of the edge rays used for the calculation of the optical surfaces. The parameters of this design are (1) the acceptance angle α, (2) the refractive index n, (3) the cell width, (4) the distance from R' to Q, (5) the distance from R to P, and (6) the geometrical concentration C_G (which is bounded by the limit $C_{MAX} = 2n/\sin \alpha$). Figure 13.40b shows a slightly modified RX-RXI design that provides no kink at the center, which is an interesting feature for the practical realization.

13.4.2   Single-Mirror, Two-Stage, One-Axis Concentrators

Although the higher concentration seems to lead to the lower cost, the necessity of greater initial investments suggests that the medium concentration option can be a good first step with low risk. There are several commercial products that operate in the low concentration range ($C \approx 10$–40). They use modified one-sun silicon solar cells, already under fabrication, which permits the penetration in the market with moderate volume of production. The optic of these systems consists of linear Fresnel lenses. Because their focal length varies with the sun declination (due to the apparent increase of the refractive index of the lens with declination; see Appendix C), one axis tracking is only used for concentrations below 15. A different approach is the EUCLIDES™ system (shown in Figure 13.5), which is
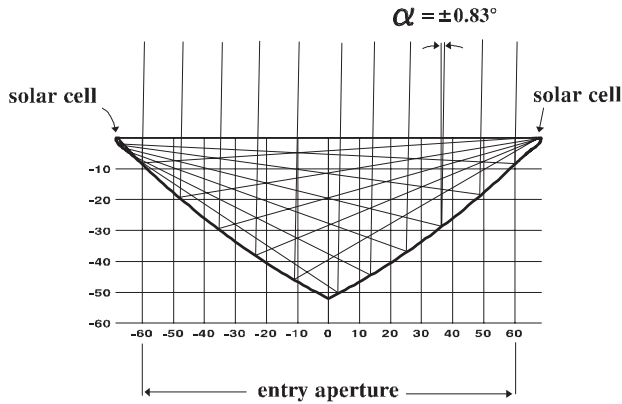
Figure 13.41   SMTS concentrator designed with $\alpha = \pm 0.83°$ and geometrical concentration 60×.

the only one that uses mirrors and does one axis tracking (horizontal and N-S oriented). Its geometrical concentration is 38×.

The Single-Mirror Two-Stage concentrator (SMTS) is also a mirror-based linear system (as the EUCLIDES™) for the medium concentration range. It was designed with the SMS design method (Benítez et al., 1995), and is made up of two mirrors and two receivers, as shown in Figure 13.41. Although the SMTS has only one mirror per receiver, it is actually a two-stage concentrator: one portion of each mirror acts as a sequential primary (X) for a receiver, while the rest is a non-sequential secondary (X_F) for the other receiver. Thus, the SMTS is a combination of the two symmetric XX_F designs. It gets a concentration of 60× with only one axis tracking.

The simplicity of the SMTS makes it interesting for low-cost applications, such as photovoltaic concentration. Moreover, the position of the receivers with respect to the optic makes it unnecessary to add a specific support structure for the solar cells and heat sinks (used for cooling the cells), and the latter may even have a structural function. At the University of Reading (UK) a prototype of an SMTS concentrator was developed jointly with the IES-UPM with an aperture of $1.5\,m^2$, a geometrical concentration of $Cg = 30×$, and an acceptance of $\alpha = \pm 1.5$ degrees, as part of a JOULE project financed by the EU (shown in Figure 13.42).

The Dielectric Single-Mirror Two-Stage concentrator (DSMTS) can be considered an advanced version of the SMTS (Benítez et al., 1997). It can work with one axis tracking and a geometrical concentration as high as 100× . Its design is also based on the SMS method.

Figure 13.43 shows the cross section of a DSMTS concentrator for $Cg = 30×$ and some representative sun rays have been drawn. Like the SMTS concentrator, it has a single mirror per cell row, which acts as a primary for one of the cell rows and as a secondary for the other row. Additionally, a linear dielectric secondary in contact with the cells increases the concentration-acceptance product by a factor around 1.5. Figure 13.44a shows the cross section of the dielectric piece with a bigger scale. It has three faces: one optically inactive, another refractive, and the last one reflective.

When the sun rays impinge on the right mirror, they are reflected toward the cell row on the left. Then there are three possible paths for the rays before

Figure 13.42   SMTS concentrator for $Cg = 30\times$ and an acceptance angle $\pm1.5°$.
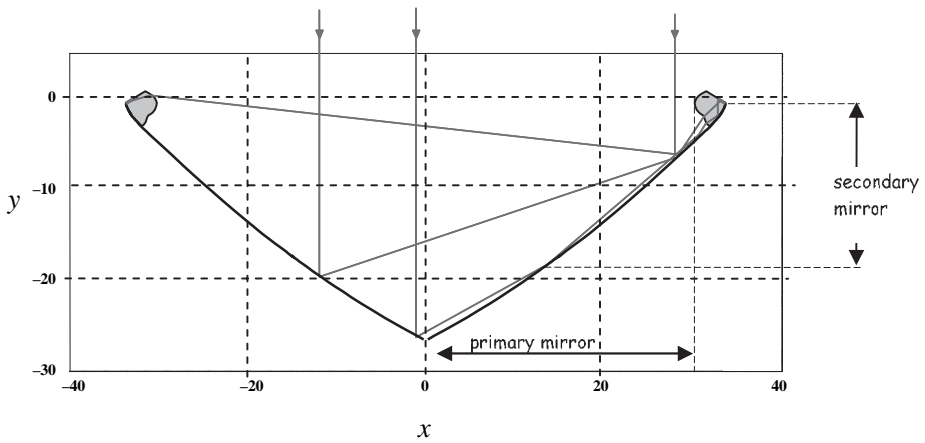
Figure 13.43   DSMTS concentrator designed with $\alpha = \pm2.37°$ and geometrical concentration $30\times$ (unit length = receiver width).
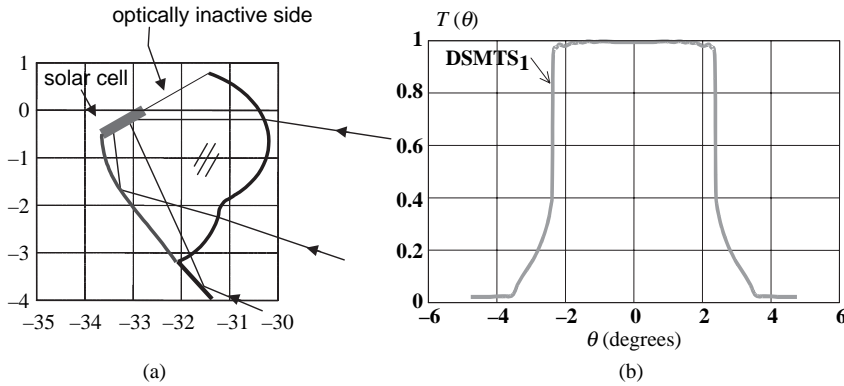
Figure 13.44   (a) Dielectric secondary that covers the cell row (unit length = receiver width) and (b) relative optical 2D transmission curve for the DSMTS of Figure 13.43.

reaching the cells (see Figure 13.44a): (1) directly, after being refracted by the lens, (2) with a refraction on the lens followed by a reflection on the mirror, and (3) with a reflection on the mirror followed by one refraction on the lens.

The design method of the DSMTS is two-dimensional. This means that if polar mounting one-axis tracking is used, the concentrator is designed for a given declination angle δ—that is, for a day within the solstices. As the only refractive element in a DSMTS works with high local spread of the ray bundles, the concentrator will present the same efficiency and nearly equal acceptance angle for all declinations within the solstices ($|\delta| < 23.5°$), as shown following.

The DSMTS shown in Figure 13.43 and Figure 13.44a has been designed for the equinoxes ($\delta = 0°$) with a geometrical concentration of $Cg = 30\times$. The resulting 90% transmission acceptance angle is $\alpha = \pm2.4°$ (see Figure 13.44b). This angle is very wide compared to other designs with the same concentration: The SMTS had $\alpha = \pm1.5°$, and the EUCLIDES^{TM} parabolic trough has $\alpha = \pm0.7°$ for this concentration factor. This means that the manufacturing, mounting, and pointing tolerances of the DSMTS are much more relaxed.

On the other hand, it is also possible to design the DSMTS for higher concentration with reasonable acceptance angle. For instance, for $Cg = 100\times$ and dielectric refractive index 1.5 and acceptance angle $\alpha = \pm0.72°$. This concentration is 2.6 times that of a parabolic trough with the same acceptance angle, a single cell row, and rim angle $\pm45°$ (this rim angle has the greatest concentration).

Since the acceptance angle of the 100× DSMTS is smaller than that of the 30× DSMTS, it is more sensitive to the apparent increase of the refractive index of the secondary lens. The continuous curve in Figure 13.45a represents the acceptance angle of the 100× DSMTS designed for $\delta = 0$, as a function of the sun declination angle δ. Since it has been designed for the equinox, the acceptance decreases slightly to a minimum value (0.64°) at the solstices.

As there is more irradiation at solstices than at equinoxes, it would be interesting to have a higher acceptance at solstices. This can be done by designing for an intermediate declination between the solstices ($\delta = \pm23.5°$) and equinoxes ($\delta = 0°$). A value $\delta = 15°$ has been used to design another 100× DSMTS. Figure 13.45a shows the acceptance declination angles' dependence for this case. Its angular transmission at the equinox and the solstice is represented in Figure 13.45b.
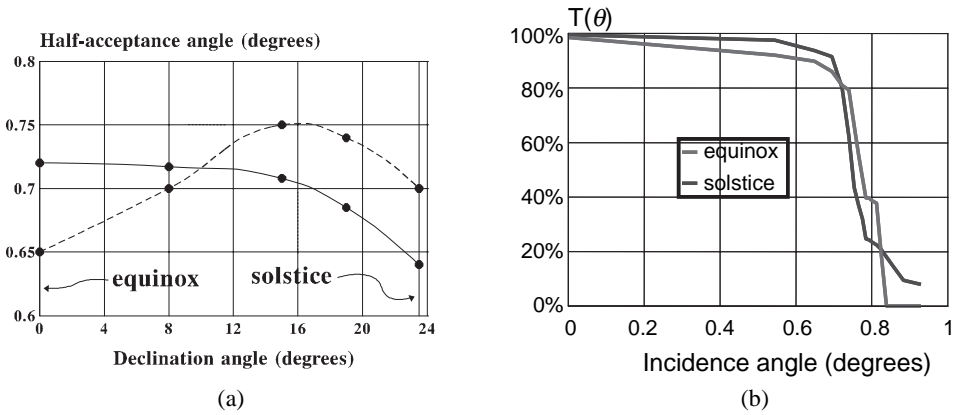
Figure 13.45   (a) Acceptance angle versus declination angle for two 100× DSTMSs with polar axis tracking, designed for $\delta = 0°$ (continuous line) and $\delta = 15°$ (dashed line). (b) Angular transmission at the equinoxes and the solstices of the 100× DSMTS designed for polar axis tracking and $\delta = 15°$
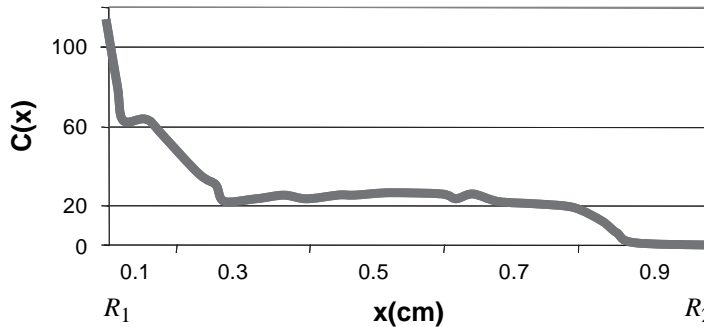
Figure 13.46   Irradiance distribution along the 0.9 cm of the cell width at the summer solstice when the sun is centered within the acceptance (input irradiance = 800 W/m²).

With respect to the illumination uniformity, the DSMTS is not specifically designed for producing good uniformity. As an example, the illumination profile near the solstice when the sun is centered in the acceptance angle is the one shown in Figure 13.46. The input irradiance is assumed to be 800 W/m² (i.e., 0.8 suns), and thus the resulting modeled averaged cell irradiance is 20.9 suns. The cell had a single connection busbar at $x = 0$. Although the peak irradiance is over 100 suns, since it is close to the busbar, the simulated cell efficiency at ambient temperature only drops from 18.6% (uniform illumination) to 18.2%.

IES-UPM manufactured and measured one prototype of the DSMTS with $Cg = 30×$. Here we will include a summary of the results, and the details can be found in Mohedano (2001). Instead of using the design in profile, shown in Figure 13.43, which would require making a mirror coating of the reflective face of the dielectric secondary, the design was slightly modified to guarantee that total internal reflection is guaranteed in that incidence (as it is done in the design of the
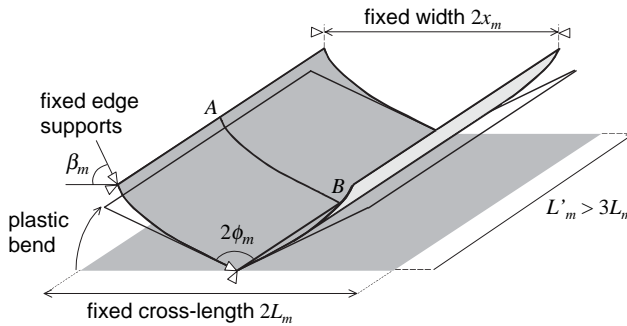
Figure 13.47   The DSMTS mirror can be conformed just as the elastic curve is achieved when an aluminum sheet of fixed cross length $2L_m$ is bent by a given angle ($2\phi_m$) and the slope at the sheet edges (β_m) and the width of the system ($2x_m$) subsequently fixed.

DTIR concentrator; see Chapter 5), and thus the mirror coating was not needed. As a consequence, for the same concentration factor, the acceptance angle reduced from $\alpha = \pm 2.4°$ to $\alpha = \pm 2.0°$, which is still very wide.

The manufacture of the dielectric secondary of this prototype was done by direct machining on PMMA (for production, the glass-extrusion technology is preferred). On the contrary, the fabrication of the mirror was done using a potential mass-production technique, which is interesting for its simplicity. It consists of conforming the mirror by (see Figure 13.47) (1) taking an aluminum sheet of a specified size (specifically, its cross-sectional length $2L_m$) lamination on it a reflective film, (2) folding it at the center at a specified angle $2\phi_m$, and (3) fixing the edges at a specified distance $2x_m$ with a slope β_m angle.

For given parameters L_m, ϕ_m, x_m, and β_m, the sheet profile adopted by elastic deformation can be calculated by solving numerically the elastic curve differential equation of mechanics (Timoshenko, 1983). These parameters can be optimized to obtain the elastic mirror profile that has minimum standard deviation of the slope error σ with respect to the theoretical profile. For evaluating the slope errors, the section of the mirror that works only as a secondary was excluded (see Figure 13.43), which is much more tolerant to errors due to the higher local angular spread of the impinging bundle (see Chapter 14). The rest of the parameters of the mirror (length of the cylinder L'_m and sheet thickness T) did not affect significantly the elastic curve. In the case of the thickness, it must be large enough to avoid the gravitational or normal wind losses to modify the profile significantly. The calculations showed that, in the prototype, $T > 1\,mm$ for $2L_m = 71.7\,cm$ was sufficient.

After the optimization, the resulting values of L_m, ϕ_m, x_m, and β_m were close to the theoretical ones (but not exactly the same). The resulting optimized mirror for the 30× SMTS design had a surprisingly low standard deviation of the slope error, $\sigma = 0.014°$ (see Figure 13.48a). Ray tracing on the optimized elastic curve makes it possible to predict the performance of the prototype. The angular transmission curves (considering no optical losses and a perfect dielectric secondary) of the selected 30× DSMTS concentrator with the designed mirror and the calculated optimal elastic one are shown in Figure 13.48b. The acceptance angle at 90% transmission reduced from $\alpha = \pm 2.0°$ to $\alpha = \pm 1.9°$. This low standard deviation suggests that this technology may even be suitable for the aforementioned 100× DSMTS (when applied to its mirror, $\sigma = 0.033°$ is obtained).
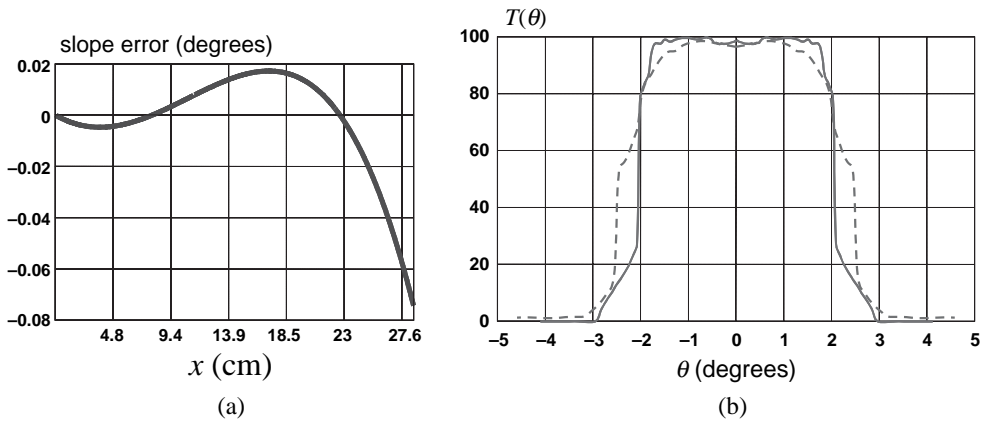
Figure 13.48   (a) Slope error along the section of the mirror that works as primary. (b) Angular transmission of the DSMTS when ray traced on the design mirror profile (continuous line) and on the calculated elastic mirror profile (dashed line).
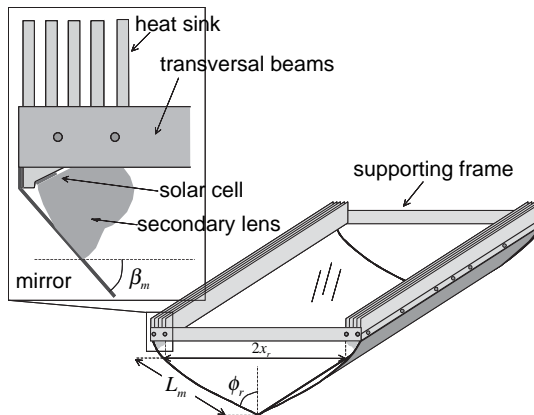

Figure 13.49   Diagram of the DSMTS prototype. In the DSMTS concentrator, as in the SMTS, the heat sinks can act as structural beams.

Apart from the simplicity of the mirror elastically conforming to the desired shape, as in the SMTS, the DSMTS has the additional advantage of a simple supporting structure because the heat sinks can act as structural beams, as shown in Figure 13.49 and Figure 13.50.

The geometrical transmission of the DSMTS prototype versus incidence angle and versus position at the entry aperture was carried out with a laser beam. By integration over the spatial coordinate, the angular transmission curve shown in Figure 13.50a was obtained. The maximum transmission was performed at normal incidence, with 98% of the entry aperture collecting the laser beam light. The acceptance angle at 90% transmission reduced from the designed value $\alpha = \pm2.0°$ to $\alpha = \pm1.6°$. The discrepancies between the calculated elastic curve model were due to the inaccuracies in the actual fixing of the optimized parameters L_m, ϕ_m, x_m, and β_m.
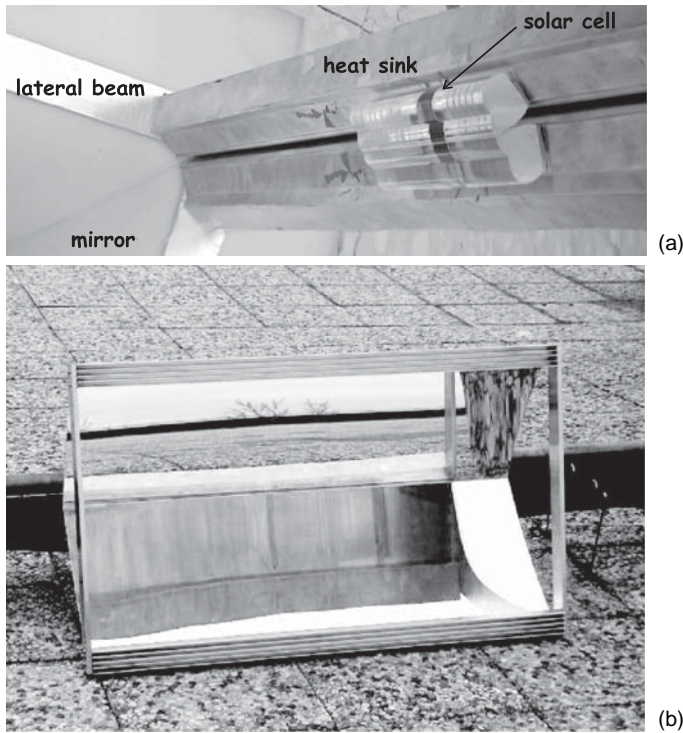
Figure 13.50   DSMTS prototype designed for $Cg = 30\times$ and $\alpha = \pm2$, (a) TIR secondary and (b) with mirror conformed by elastic bending.

The optical efficiency of the DSMTS was also measured outdoors. The current gain with respect to the same cell with the flat panel encapsulation was 26.4. Therefore, the optical efficiency was $\eta_{opt} = 26.4/30 = 0.88$. The modeling of the different causes of optical losses fit well with measurement, and their contributions are detailed in Figure 13.50b.

13.4.3   High-Concentration Photovoltaic Designs

In the case of high-concentration systems, the present trend goes toward two research and development lines: single large optic dishes for illuminating a PV module and many *miniconcentrators* (concentrator entry aperture diameter = 30–85 mm) for illuminating small cells (1–2 mm). Up to now the SMS designs have been applied to specific miniconcentrator designs, and we will refer only to that line.

In order to make efficient modules, the miniconcentrators are usually truncated as squares for tessellation. In these designs, the acceptance angle remains unchanged after truncation, and then the acceptance angle, concentration product is reduced to the factor $2/\pi = 64\%$. This reduction would be smaller if the hexagonal concentrators were used, but square tessellation is usually considered more aesthetic because it is free from edge effects.

The interest in miniconcentrators comes from the fact that the small cell technology is very close (even simpler) to that of the LED's, which is very much

(a)   (b)
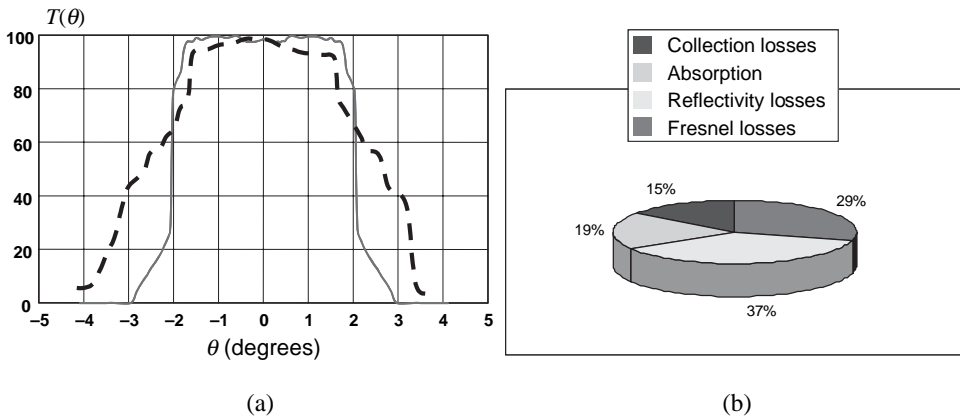
Figure 13.51   (a) Angular transmission curve of the DSMTS prototype measured with a laser. The maximum transmission was 98%. (b) Modeled relative contributions of the optical losses to the measured optical efficiency of the DSMTS prototype.
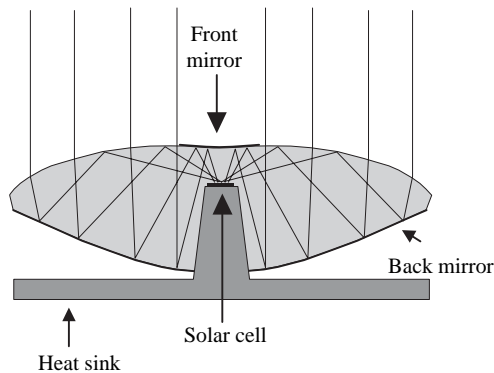
Figure 13.52   Cross section of photovoltaic RXI concentrator.

developed and highly automated. As an example, highly efficient $1\,mm^2$ GaAs cells for 1,000 suns operation have been demonstrated (Algora et al., 2001). The power to be dissipated is around 0.7 W. On the other hand, the Luxeon III LED of Lumileds has a high-flux blue chip of also $1\,mm^2$, and its package was designed to passively dissipate about 4 W. Therefore, the thermal problem is higher in LEDs than in high-concentration photovoltaics.

The first proposed photovoltaic miniconcentrator was the RXI (presented in Chapter 8). In the framework of the European JOULE project entitled *Ultra Compact High-Flux GaAs Cell Photovoltaic Concentrator* (HERCULES), project, the IES-UPM developed an RXI and manufactured prototypes by injection molding of PMMA (Álvarez et al., 1999).

The RXI design, whose cross section is shown in Figure 13.52, had a geometrical concentration $Cg = 1{,}256\times$, circular entry aperture (40 mm diameter), square GaAs cell ($1 \times 1\,mm^2$), cell illumination up to $\beta = \pm 70°$. The ray-traced 90% transmission acceptance angle was $\alpha = \pm 1.8°$.
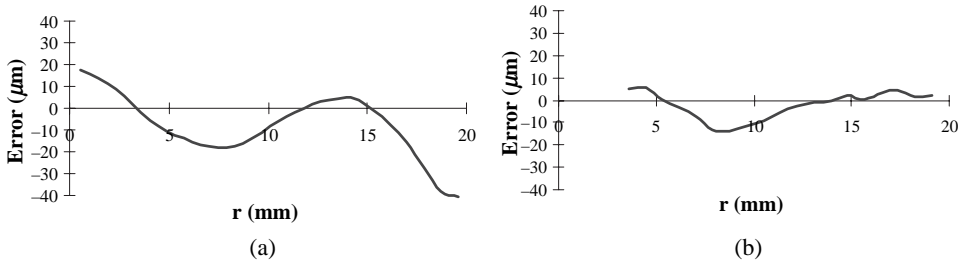
Figure 13.53 (a) Top and (b) bottom surface profile error of one injected RXI piece.
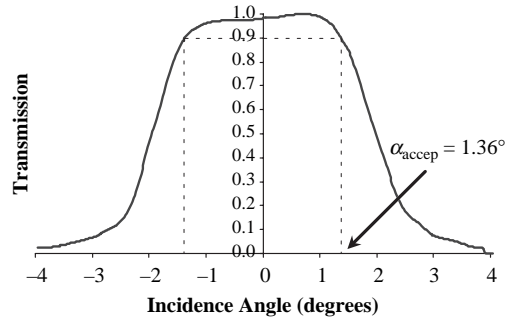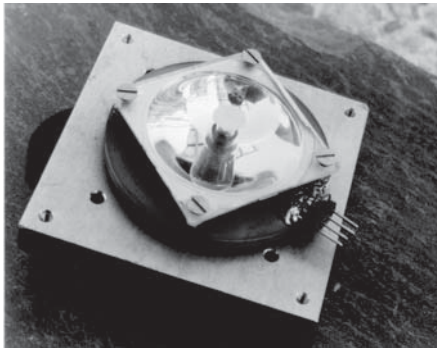
Figure 13.54 (a) Photovoltaic RXI prototype, (b) measured angular transmission curve of RXI at 890 nm.

RXI prototypes were made by injection molding of PMMA. The mold is made by a diamond turning technique, with an accuracy of $\pm 1\,\mu$m. The profiles of the concentrator parts were measured with a contact profilometer. The profile error, as the difference between theoretical and manufactured profile along the concentrator radius, can be seen in Figure 13.53. Metallization of the lower surface and front mirror of RXI was made by evaporation of silver. After this, the metallization is protected with a coat of painting. To encapsulate the receptor in the RXI, we use a silicone rubber (Sylgard 184, Dow Corning), which glues the GaAs cell to the concentrator.

In order to characterize the prototypes, the optical efficiency and angular transmission curve were measured. The measurement of angular transmission was made with a source of 890 nm wavelength, using a collimated beam created with a lens and an infrared LED. Figure 13.54 shows the angular transmission curve of an RXI concentrator obtained from this measurement. The measured acceptance angle was $\alpha = \pm 1.36°$, and the 50% transmission angle was 1.9°.

The current gain (ratio of current with concentrator to current without it) measured was 1,095, and from it an optical efficiency $\eta_{\text{opt}} = 87.2\%$ was deduced. This was estimated taking into account the dependence of the noticeable reflectivity of the GaAs with the incidence angle (see Álvarez, 1999; for further details).

The RXI concentrator had excellent performance in terms of optical efficiency and acceptance angle, but its illumination uniformity was not good (unless its acceptance angle was reduced, which was undesirable). In the prototype, the cell
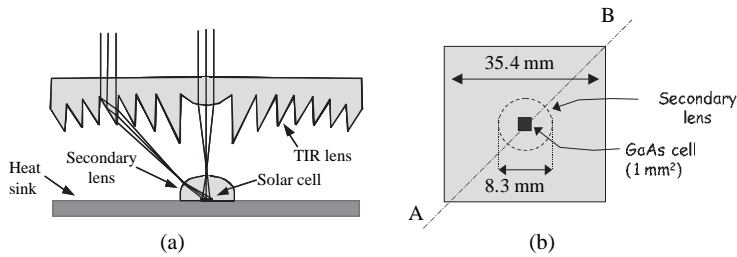
Figure 13.55   (a) Diagonal cross section of the TIR-R concentrator. (b) Top view of the *Cg* = 1,250× design with entry aperture truncated as a square (for posterior tessellation).

fill factor under the RXI concentrator was *FF* = 0.77, to be compared with the *FF* = 0.85 obtained at the same concentration level under uniform illumination with a flash lamp.

In order to improve the RXI, its design strategy could be modified. However, other practical limitations for the photovoltaic application (curved top surface, nonprotected total internal reflection, need of high reflectivity low-cost mirror) suggested changing the development into a different concept. Thus, the TIR-R concentrator was proposed in the continuing European project INFLATCON (shown in Figure 13.55).

The TIR lens was invented by Fresnel (Stevenson, 1835) as a revolutionary optics for lighthouses at the beginning of the 19th century. More recently, TIR lenses have proven to be excellent devices for illumination applications (Medreder, Parkyn, and Pelka, 1997), with good exit aperture uniformity with the inclusion of a second optical element (called a mushroom lens), whose geometry coincides qualitatively with the TIR-R concentrator, although their design was not suitable for getting good photovoltaic performance.

Using the SMS method to the TIR-R configuration (Álvarez et al., 2001) made it possible to achieve high-concentration-acceptance angle products, good illumination uniformity, and sufficiently large aspheric facets (this is important for keeping high efficiency, due to the possible vertex rounding when manufactured).

The theoretical characteristics of the squared aperture 1,250× TIR-R design, with 2° tilt angle of the vertical facets, are acceptance angle $\alpha = \pm 1.3°$, optical efficiency (at $\lambda = 630\,\text{nm}$) $\eta_{\text{opt}} = 82.5\%$, and irradiance on cell surface at normal incidence (@850 W/m²) <2,300 suns.

The inclusion of integration strategies in the TIR-R design is permitting to get closer to the challenge of design for high-concentration photovoltaic systems that have high efficiency, a sufficiently wide acceptance angle, sufficient illumination uniformity, and suitable low-cost mass production.

13.5   DEMONSTRATION AND MEASUREMENT OF ULTRA-HIGH SOLAR FLUXES (C_g UP TO 100,000)

There has been very rapid progress from the first ultra-high flux measurements conducted on the roof of the high-energy physics building on the University of

Table 13.6.   Summary of Ultra-High Flux Measurements.

Date (Ref)	Location	Secondary	Measured Flux (suns)	Total power
February 1988 (Gleckman, 1988 Gleckman et al., 1989)	Chicago	Lens–oil filled silver vessel ($n = 1.53$)	$56,000 \pm 5,000$	~44 W
March 1989 (Cooke et al., 1990)	Chicago	Solid sapphire DTIRC ($n = 1.76$)	$84,000 \pm 3,500$	~72 W
July–August 1990 (O'Gallagher et al., 1991)	NREL (Golden, CO)	Water-cooled reflecting silver CPC—air-filled ($n = 1.0$)	$22,000 \pm 1,000$	~3.5 kW
March 1994 (Jenkins et al., 1996)	NREL (Golden, CO)	Fused silica (quartz) DTIRC with "extractor" tip ($n = 1.46$)	$50,000 \pm 2,000$	~900 W

Chicago in 1988, to the recent experimental investigation of potential laser pumping and materials processing experiments carried out at the National Renewable Energy Laboratory High-Flux Solar Furnace and the Weizmann Institute Solar Tower (Yogev, 1997). The progression of demonstration experiments is summarized in Table 13.6. We shall describe each of these experiments in order, beginning with the small-scale experiments at Chicago and leading up to the experimental demonstration of the attainment of 50,000 suns at a power level approaching 1 kW. However, before discussing the individual experiments a few general features of the measurement procedures are presented.

The measurement of ultra-high fluxes may seem straightforward in principle, but it is difficult to implement in practice. To measure intense solar fluxes, a measurement of the power passing through an aperture of known area or incident on a target surface of known area is first required. The flux is then determined by the ratio of the measured power, P, to the corresponding area, A. Even when the high fluxes are generated in air, the intense thermal input to most semiconductor detection devices can very quickly raise their temperature above their operating limits. In addition, when the flux is generated inside a high-index medium, the measurement technique has to provide an accurate power measurement taking account of the different indices of refraction between the medium and that of the detector. For these reasons all of the measurements described in this work rely on fluid-based calorimetric measurements. The calorimeter provides a medium for absorbing the thermal energy and maintaining relatively low operating temperatures. Measurements of temperature differences or temperature increases then allow a quantitative measurement of the power delivered to the fluid.

For some of our measurements, a "ballistic calorimeter" approach was used in which the fluid was exposed to the high flux for a measured time interval, and the temperature rate of rise was measured. An internal electric heater was used to calibrate the response of the calorimeter fluid to a known power input. Where the operating scale permitted flow-through, calorimetry was used. In these cases the

power was delivered to an absorbing cavity whose walls were surrounded and cooled by a measured flow of water. The temperature difference between the inlet and outlet of the cooling water together with the flow rate provides a continuous measure of the steady state power, P, delivered.

Coupling of the calorimeter to the higher-index concentrating media requires an even more elaborate setup. Because the high flux is inside a high-index medium, it is necessary to get the light to the walls of the calorimeter so that it can be absorbed. For the small-scale experiments (power levels 100–200 W) we used an oil-filled ballistic calorimetric as will be described further following. In these cases, the exit aperture of the secondary concentrators is optically coupled to an indexed-matched window into the calorimeter filled with an index-matched calorimetric fluid. The oil-filled calorimeter, its window, the optical couplant, and secondary concentrator act as one optical element. The light, after some small Fresnel reflection losses, is absorbed by the calorimeter wall. The heating of the oil inside the calorimeter determines the power passing through the concentrator's exit aperture, and the concentration is given by P/IA.

In all the geometries an accurate determination of the effective area is made, and the measured flux is then given by $F = P/A$. The net "concentration factor," C, is then given by the ratio of F to the solar insolation level I measured by a normal incidence pyronometer (NIP).

13.5.1   Exceeding the $n = 1$ Limit: Small-Scale Experiments

To explore the application of these concepts to the generation of very high-solar fluxes, laboratory scale versions of small two-stage systems were constructed at the University of Chicago and used to establish new records for the concentration of sunlight. The first of these systems used primary spherical mirror with a focal ratio $F/D = 2.0$ and rim angle $\phi = 11°$, and the entire configuration was mounted on an equatorial tracking system. The long focal length made the spherical aberrations from this primary negligible for this nonimaging application. The second system was actually a miniature furnace, in that a flat heliostat was used to track the sun and direct a parallel beam of unconcentrated sunlight toward a larger fixed primary also with a focal ratio of 2.0 and a rim angle of 11°. The system followed conventional design principles in that the target is centered on the axis of the primary. Both primaries produced a concentrated flux of about 2,000 kW/m^2 (2,000 suns) in their focal zones. High-index secondaries and appropriate calorimetry were then used to concentrate this sunlight manifold higher and to measure the resulting flux.

13.5.2   A New Record for the Concentration of Sunlight

For these very first ultra-high-flux demonstrations, the secondary used was a specially designed CPC-type vessel that was filled with an oil of high refractive index ($n = 1.53$). The vessel was made of very pure solid silver (obtained from a bullion dealer), and its interior walls were highly polished. The vessel was covered with a small lens that becomes a part of the secondary (its optical properties were used in applying the "edge-ray" design approach to determine the wall profile curve).
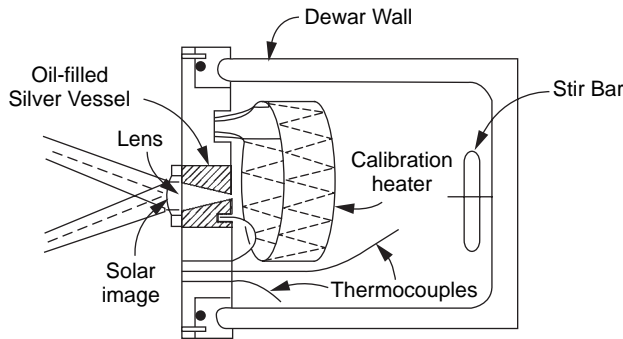
Figure 13.56   56,000 sun experiment using oil-filled secondary.

The exit aperture of the vessel forms the aperture through which the high flux must pass, and it opens directly into the volume of the calorimeter, as shown in Figure 13.56.

The ideal limit for this configuration is over 100,000, but reflection losses from the primary along with absorption losses in the secondary reduced the net expected concentration to about 60,000 suns. Another reason for a small shortfall is that we measure insolation with a normal incidence pyranometer, or NIP, which has a rather "loose" 5° (full-angle) acceptance cone that admits much of the circumsolar radiation, while our concentrating optics is more tightly limited to the solar disc. The "ballistic calorimeter" measured a flux concentration 56,000 1/–5,000 times the ambient terrestrial sunlight intensity (56,000 suns) (Gleckman, 1988; Gleckman, O'Gallagher, and Winston, 1989). This exceeded previous high-solar-flux measurements by more than a factor of three and provided the very first demonstration of the capability of nonimaging optics in generating such high solar power levels.

13.5.3   Brighter than the Sun

In order to achieve even higher fluxes a second small-scale demonstration was carried out using the heliostat and a slightly larger primary. In this case the primary and secondary calorimeter assembly were able to be held fixed. The secondary was formed from sapphire with an index of refraction of $n = 1.72$, permitting even higher flux concentrations to be attained than with the first experiments with oil. The side wall profile was shaped so that all radiation inside the acceptance angle fulfilled the condition for total internal reflection (no additional metal coating was necessary). However, in this case the secondary was solid and had to be coupled into the calorimeter through an index matching (sapphire) window. A special oil compound was formulated to have an index that precisely matched that of the sapphire window and secondary, and this was used to fill the calorimeter assembly shown in Figure 13.57. These experiments produced a flux of 84,000 suns (Cooke et al., 1990). These measurements paved the way for a series of low-power experiments on laser pumping using this heliostat and primary configuration (Cooke, 1992).
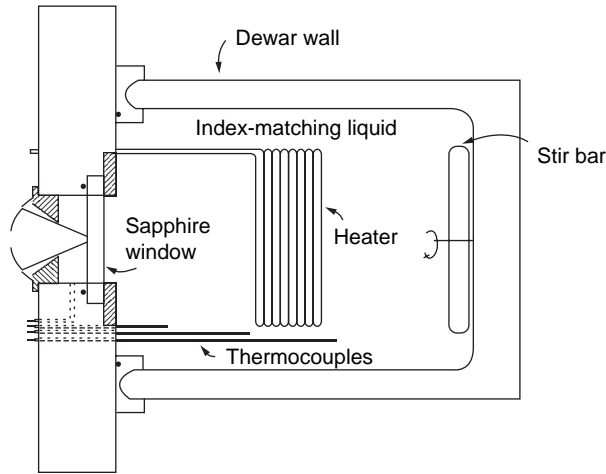
Figure 13.57   84,000 sun experiment using sapphire secondary.

13.5.4   Large-Scale Experiments at the NREL High-Flux Solar Furnace

The ultra-high-flux concentrators just described are low-power systems (approximately 100 W total) intended for concept demonstration and small-scale applications. Following through with a long-standing desire to explore the development of these techniques for larger-scale, higher-power applications, the National Renewable Energy Laboratory (NREL) designed and constructed a scaled-up solar concentrating furnace facility specifically intended to take advantage of nonimaging optics (Lewandowski et al., 1991). This high-flux solar furnace (HFSR) concept uses a modified long focal length design and is capable of delivering up to 10 kW to the focal zone and in particular to the entrance aperture of a secondary concentrator.

    An artist's rendition of the furnace concept is shown in Figure 13.58. The design uses a faceted primary with a long focal length so that when combined with a nonimaging secondary element, the resulting configuration can approach the allowed limit with a high intercept. This is very different from conventional short focal length furnaces, since it is designed to be able to take advantage of the use of optimized nonimaging secondaries and approach the thermodynamic limit. As can be clearly seen in Figure 13.59, aside from the long focal length, the other optically significant difference between the design of the NREL primary and the conventional furnaces (made possible by the requirements of the nonimaging design) is that the incident and target directions are off-axis. The major advantage of this configuration is that the secondary concentrator and target, as well as any associated equipment (e.g., cooling, data acquisition, etc.) do not shadow the primary. In these high-concentration geometries an optimum reflecting design is a CPC with an acceptance angle close to, but slightly larger than, the primary rim angle, truncated to about 80% of its full height and with design entrance aperture diameter sized to intercept as much focal plane energy as possible. An optimum high-index secondary is a DTIRC with an acceptance angle about equal to the primary rim angle and a spherical or aspheric front surface.
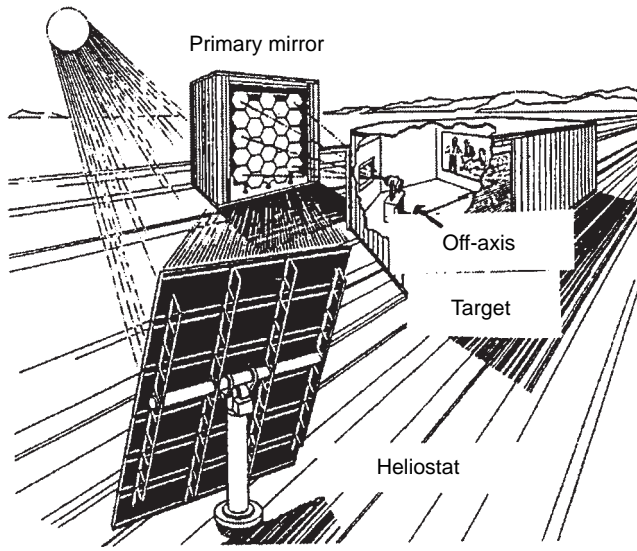
Figure 13.58   Schematic of NREL solar furnace.
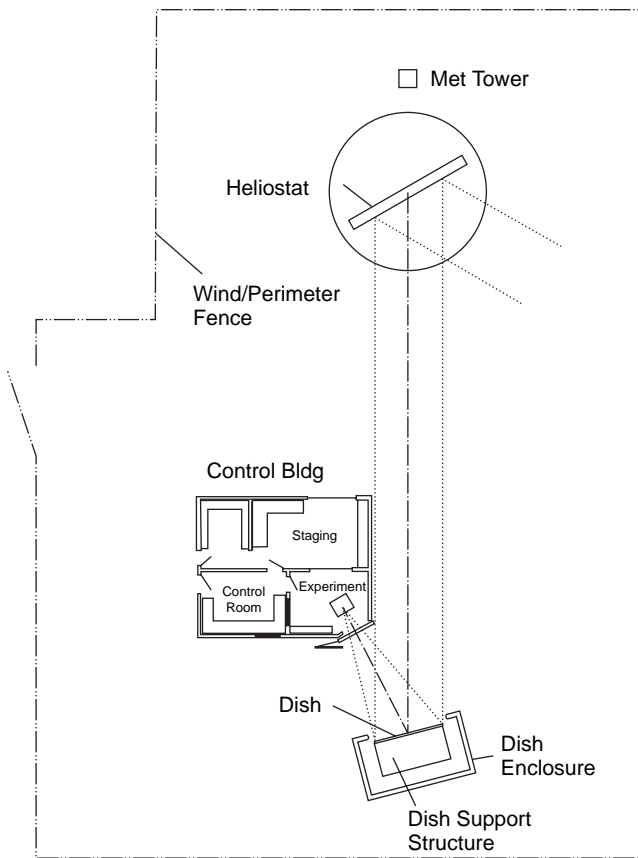
Figure 13.59   Diagram of NREL solar furnace laboratory.

13.5.5   Measurement of High Flux at Kilowatt Levels in Air ($n = 1.0$)

In order to demonstrate the potential of the furnace for conducting measurements and experiments with high solar flux in air, the Chicago University group designed and constructed an optimized CPC reflecting secondary concentrator for use with the HFSR primary. In collaboration with the NREL staff, we tested two prototype units during the summer of 1990. On several occasions, flux concentrations at multikilowatt levels in excess of 22,000 suns were achieved, surpassing the previous values achieved in air ($n = 1.0$) by more than 20%. The secondary CPC units have continued to be used at NREL for a variety of materials processing and chemical reactor high-flux experiments.

The remainder of this section is concerned with the design and fabrication of these CPC secondaries and the measurement and analysis of the high-flux performance. A ray-trace model for analyzing the characteristics of nonimaging secondary elements to be used with the NREL facility was developed. The inputs to this model are "ray files" generated by NREL. Each of these is a set of rays chosen by a Monte-Carlo procedure to characterize the focal plane spatial and directional distribution for a particular set of primary concentrator parameters. For these first experiments it was decided to use a simple reflecting CPC-type secondary, since these are easy to cool and the concentrated flux can be used in a low index ($n > 1$) target. An analysis of the optical throughput and net concentration ratio of a large number of different CPCs covering a wide range of design parameters was carried out by tracing 5,000 or 10,000 rays through each. The peak and average flux concentration and total power delivered were investigated as a function of CPC design acceptance angle, entrance aperture intercept factor, degree of truncation, and secondary surface reflectivity. In addition, information on the position and directional distribution of concentrated solar flux in the target plane at the CPC exit aperture was generated. The optimization procedure is shown in Figure 13.60. Each point represents the predicted average flux concentration ratio at the exit aperture of a particular CPC plotted versus the corresponding total power delivered. The shape of each CPC is specified completely by three parameters, the aperture radius, r_{ent}, the design acceptance angle f_c, and the truncation fraction, f_t. The effects of all reflection losses are included in the ray-trace calculations. For purposes of this calculation, a conservative value of $\rho = 0.91$ was used for the secondary reflectivity. Points for several families of CPCs are shown. The NREL primary consists of 23 facets, each with a spherical contour and hexagonal perimeter, arranged in a roughly rectangular array, and each is individually aligned to point toward the same off-axis design aiming point. The maximum convergence angle f at this aiming point for the outer edges of the corner facets is 17.3°, while 19 of the 23 facets are contained within a cone of half angle 15°. The resulting flux distribution at the secondary entrance plane is approximately rotationally symmetric with 90% of the energy lying within a circle 10 cm in diameter. The maximum secondary geometric concentration ratio for the reflecting ($n = 1.0$) CPC decreases with increasing ϕ_c. For a fixed-acceptance angle, the average flux concentration in the entrance aperture decreases with increasing radius (and therefore decreases in the exit aperture), while the total intercepted power increases, as can be seen from the dashed curves in Figure 13.60. The points represented by the (solid squares) are for untruncated (full-height) CPCs of the indicated values
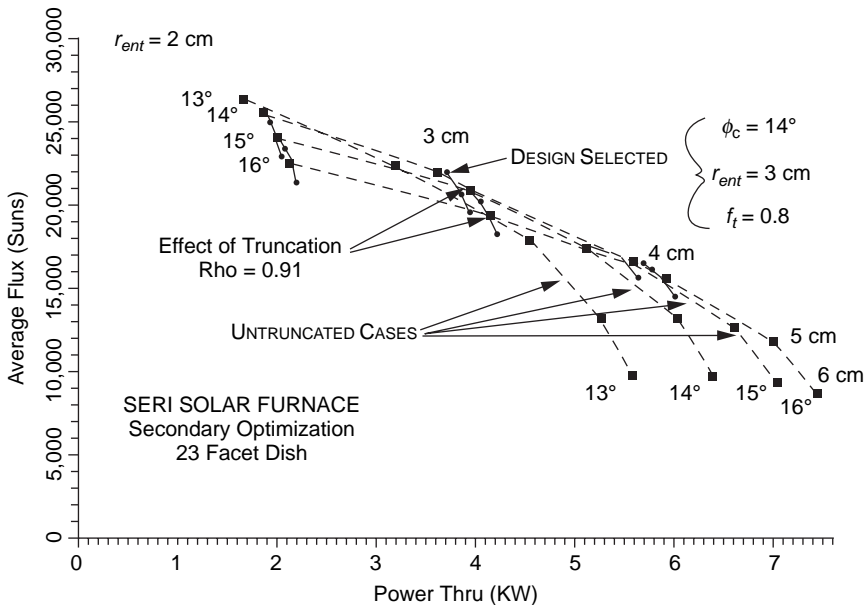
Figure 13.60   Design considerations of secondary for high-flux solar furnace at NREL (previously called SERI for Solar Energy Research Institute).

of ϕ_c and r_{ent}. At a fixed small-entrance aperture (e.g., r_{ent} = 2 cm) the average concentration achieved increases rapidly with decreasing ϕ_c at a relatively small sacrifice in power, while for large apertures, the net concentration is not strongly dependent on acceptance angle, but the total power is, as ϕ_c increases to "see" the entire primary. Since one goal of these early experiments was to demonstrate flux levels higher than ever achieved previously in air, and at the same time to deliver these levels at a multikilowatt level over an appreciable area, the design selected represented a compromise between the two extremes of high flux–low power and low flux–high power. We chose a CPC with an entrance aperture radius of 3 cm and a design acceptance angle of ϕ_c = 14°.

Finally, the effect of truncation is illustrated in Figure 13.60 for six cases (ϕ_c = 14° and 15°, and r_{ent} = 2, 3, and 4 cm) by the points labeled with asterisks. As each CPC is truncated to a fraction f_t of its full height, the profile was rescaled to keep the entrance aperture fixed. Thus, the exit aperture increases accordingly, and the achieved flux eventually decreases. For illustration, a fully developed (untruncated) 14° CPC profile with an entrance aperture radius of 3 cm is shown in Figure 13.61a alongside a CPC profile with the same acceptance angle and aperture but truncated to 0.8× its full height. For each case values of f_t from 1.0 down to 0.5 are plotted in the Figure. For small truncations, the effect is negligible, and in some cases it is more than offset by a decrease in the net reflection loss with respect to the untruncated cases. Based on this analysis, the final secondary design selected has an acceptance angle ϕ_c = 14° and was truncated to 80% of its full height (see Figure 13.60b), corresponding to a geometric concentration ratio Cg = 16.4. The entrance and exit aperture diameters are 6.0 cm and 1.48 cm, respectively. The combined optical models for the furnace and secondary predict that,

$f_t = 1.0$
(Untruncated)

$r_{exit} = 0.730$ cm
$c = 17.1$

(a)

$f_t = 0.8$
Prototype
Design

$r_{exit} = 0.735$ cm
$c = 16.7$

(b)
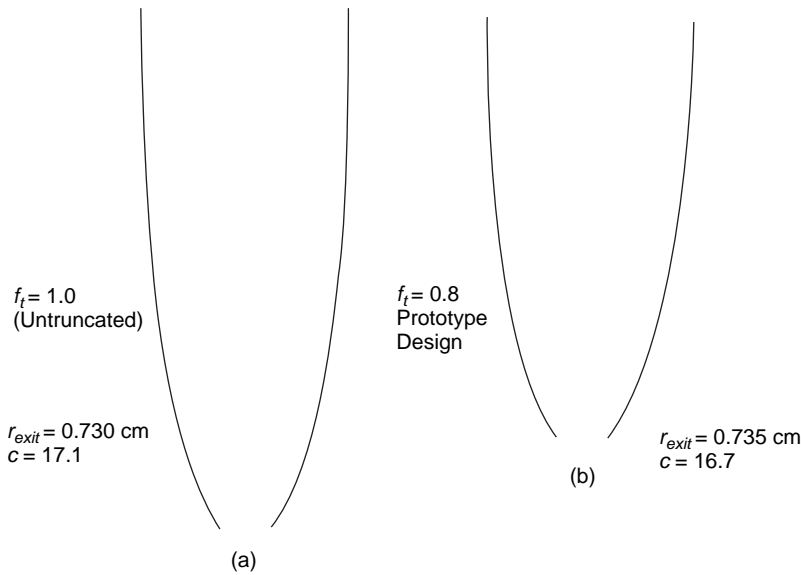
Figure 13.61   CPC design for NREL 20,000 sun experiment.

when deployed in the focal plane of the SERI high-flux solar furnace, an average concentration in excess of 20,000 suns at a net power of about 3.5 kW will be delivered at the exit aperture. Two mechanically identical water cooled CPCs with the profile of Figure 13.61b were constructed.

The material selected for the concentrator body was an alloy of tellurium and copper that has thermal properties close to those of pure copper (high conductivity and heat capacity) but is much more readily machined. A cross-section drawing is shown in Figure 13.62a. The concentrator profile was machined out of the inside of a solid cylinder. Spiral grooves were cut around the outside surface to channel the cooling water. The whole inside piece was then inserted into a separately machined outer jacket—also made of the tellurium-copper alloy—and the two pieces were soldered together. In order to provide clearance for the CPC exit aperture to be fitted into the flux measuring calorimeter (see following), the lower 12 mm of the concentrator wall just above the exit aperture could not be directly coupled to the cooling jacket. The wall in this portion was left thick to permit heat to be conducted away from the concentrated flux region expected at the exit aperture. After assembly and soldering, each unit was leak tested, and the inner surface was given a preliminary polish.

The entire unit was then plated, first with nickel, then silver, and the inner surface was again polished. Finally, a protective chromate coating was applied over the inner silver surface. The appearance of the reflecting surface for each of the two units was less specular than desired, although extra care was taken with the second unit with some success. Quantitative data on the specular and total reflectivity of these surfaces is not available, although values of ρ_2 between 0.90 and 0.95 are typical. To measure directly the radiant energy emerging from the CPC exit aperture, a small cold water calorimeter was designed and built. A cross section (not to scale) is shown in Figure 13.62b. The basic design consists of a

WATER COOLED CPC   WATER COOLED CALORIMETER
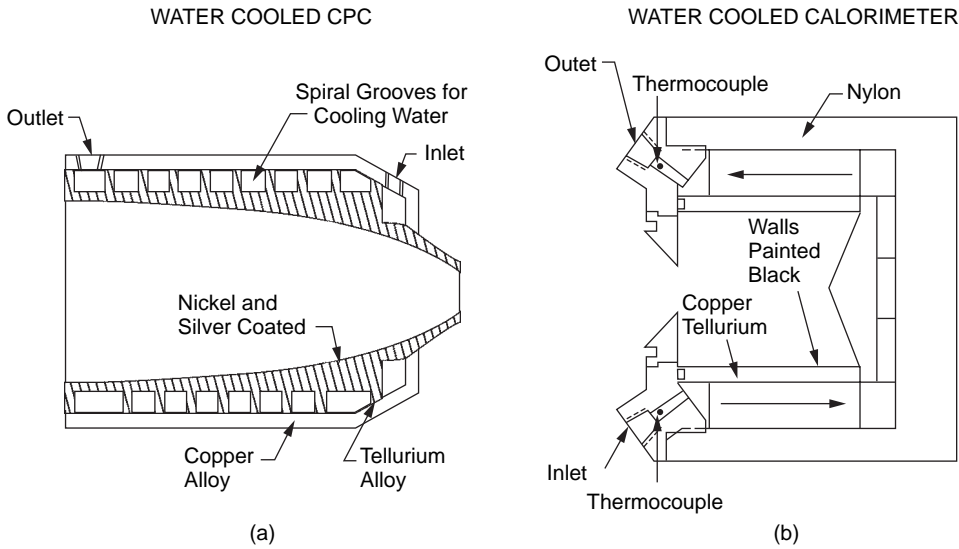
Figure 13.62   CPC used for NREL 20,000 sun experiment.

cavity absorber whose walls are made of the same tellurium-copper material as the CPC. Grooved fins were machined in its outer surface to distribute cooling water evenly around the outsides and bottom of the cavity walls, and the whole piece was inserted into an insulating cylindrical container made of Teflon. The components were held together by an O-ring seal and insulating cover, machined so that the CPC exit aperture was flush with and precisely matched to the cavity opening. The inside of the cavity was painted black with Pyromark paint ($\alpha = \varepsilon = 0.95$), and from the aspect ratio of the cavity the effective absorptivity of the cavity aperture is calculated to be 0.998. The power absorbed within the cavity, P_{cal}, is determined simply by recording ΔT, the difference between the inlet and outlet temperatures of the cooling water measured with Type K immersion thermocouples connected in a differential configuration and monitoring the volumetric flow rate $\dot{V}$ of the water. Then

$$P_{\text{cal}} = \mu \dot{V} c_p \Delta T \qquad (13.10)$$

where μ is the mass per unit volume of water. The calorimeter was calibrated with a 1 kW lamp designed to fit tightly inside the cavity. When carefully insulated and operated in equilibrium, the efficiency of the calorimeter was reliably and reproducibly measured to be $\varepsilon = 0.99$.

The high-flux measurement procedure was quite straightforward. The CPC calorimeter assembly was mounted with its aperture and axis positioned at their nominal (design) values. The calorimeter flow rate and DT were monitored while an attenuating shutter was opened. During and after the time to reach equilibrium (1–2 minutes), a measure of the net observed concentration defined as

$$C_{obs} = \frac{P_{cal}/\left(\pi r_{exit}^2\right)}{I_{DN}} \qquad (13.11)$$

was calculated and continuously monitored. Here I_{DN} is the direct normal component of the insolation as measured by a normal incident pyronometer (NIP). For instance, a value of $P_{cal} = 3.5\,\text{kW}$ with $r_{exit} = 0.74\,\text{cm}$ and I_{DN} 51,000 W/m2 corresponds to $C_{obs} = 20{,}345$ (suns).

Measurements with the first unit fell somewhat short of predictions, although they did represent a new record for the concentration of sunlight in air ($n = 1$). On June 26 and 27, values of $C_{obs} = 17{,}000$ to over 18,000 suns were obtained for several time intervals of several minutes or more, exceeding previous highest values of 16,000 suns. Later in July, regular readings over 19,000 suns were found after moving the secondary back and forth along the optic axis to find the highest value. The major reason for the lower than predicted values with this first proto-type unit is thought to be nonspecularity of the CPC reflecting surface due to insufficient polishing of the reflector substrate and some characteristics of the nickel undercoat. Other possible factors contributing are source broadening due to circumsolar radiation and short time-scale tracking instabilities caused by wind. The greater care taken in polishing the second unit paid dividends. This CPC produced values over 22,000 suns. Although these are very slightly higher than ray-trace predictions, based on nominal heliostat, primary, and secondary reflectivity, this result more than met expectations.

13.5.6   Measurement of High Flux at High Power Levels in a Refractive Medium ($n > 1.0$)

After the successful demonstration of a high-power high-flux concentrator in air at the solar furnace, it was desired to repeat the high-index demonstration at this new larger scale. Because of the inherent optical errors in a larger system, it was recognized that flux levels equal to those of the small-scale measurements could not be reproduced at the same intercept and concentration levels. However, the goal of generating and measuring a flux of 50,000 suns at near kilowatt levels was adopted as a means of demonstrating the technology. Of course, for these experiments a secondary DTIRC would have to be used, and an optimization procedure similar to that carried out for the reflecting CPC in air and illustrated in Figure 13.60 was carried out by tracing rays from the primary through a large number of individual trial DTIRC secondary designs. In all cases the material for the secondary was fused silica and the front surface arc angle was 50°. The results of this optimization are shown in Figure 13.63, and the design selected had an entrance radius of 1.5 cm and an acceptance angle of 12°. The initial experiments attempted to use a scaled-up version of the "ballistic calorimeters" used at Chicago for the successful small-scale experiments. (See Figures 13.56 and 13.57.) The secondary/calorimeter assembly used at the HFSR is shown in Figure 13.64. Several unsuccessful attempts were made to measure the high flux produced by the furnace and this secondary. The problem had to do with the inability to find an optical couplant material that would serve to couple the highly concentrated sunlight from the tip of the secondary to the index-matched fused silica window (and from there into the oil-filled calorimeter). The couplant invariably overheated and burned, ruining the coupling and often destroying the tip of the secondary and damaging the window in the process. It is felt that the inability to scale up from the low-power measurements has to do with the fact that the heat dissipation surface area grows only linearly with the linear dimension while the total power
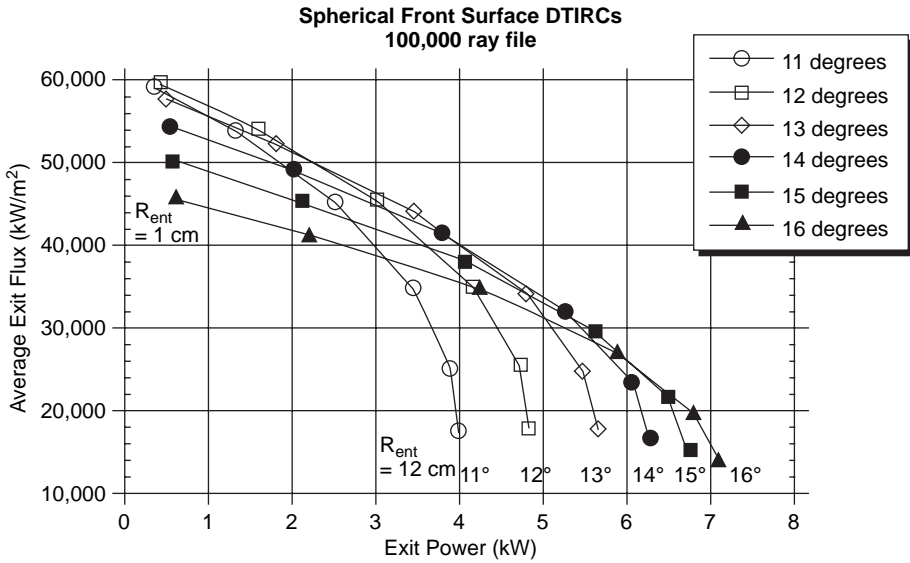
Figure 13.63 Design considerations for dielectric secondary used in NREL high-flux experiment.
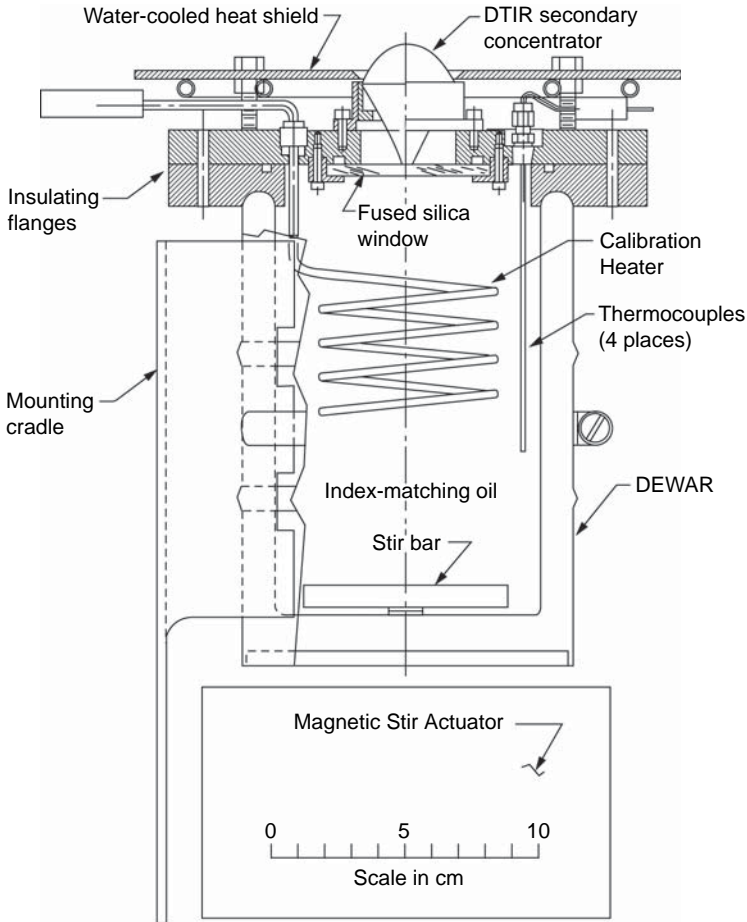

Figure 13.64 NREL high-flux experiment using dielectric secondary.

is growing as the square of the linear dimension, so there is therefore a fundamental problem with this approach.

Because of these difficulties in measuring high fluxes in large-scale high-index media, new techniques were developed (Jenkins et al., 1995; Karni et al., 1994; 1995) to extract light from higher-index materials into lower-index ones. A new way to measure the flux passing through a large dielectric aperture was developed from research at the Weizmann Institute of Science (WIS) in Israel and at the University of Chicago. These investigations studied the use of long extensions attached to the exit apertures of secondary nonimaging concentrators to extract light into air. A relatively simple device was developed that extracts light into air from a borosilicate medium ($n = 1.46$) with an efficiency greater than 99%. New DTIRC secondaries were fabricated with such "extractor tips." These secondaries were then combined with the cold water calorimeter developed for the air measurements (Figure 13.62b) as shown in Figure 13.65, and a successful measurement of the power delivered through a small aperture surrounding the "waist" of the extractor tip was carried out. A photograph of the DTIRC with extractor tip is shown in Figure 13.66.

How does an extractor work? Well, an efficient extractor works by the multiple reflections of light rays at incrementally increasing angles along its length, causing each ray to eventually fail TIR and refract into a lower-index medium. In two dimensions, the model for how an extractor works is easy to visualize and is shown in Figure 13.67a. A triangular extractor with a small wedge angle β forces light rays hitting one surface with angle δ_n to deflect slightly so that the next reflection occurs with angle

$$\delta_{n+1} = \delta_n - \beta \tag{13.12}$$
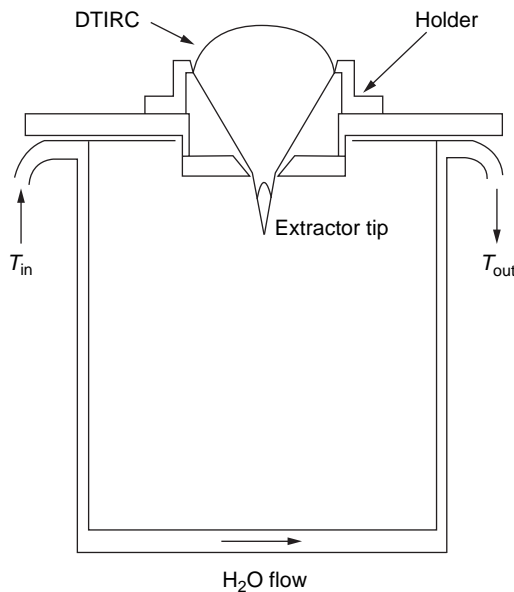
Figure 13.65   Schematic of NREL 50,000 sun experiment.

If the magnitude of the angle made with the normal to the triangle edge θ_n is less than the critical angle θ_c ($\theta_c = \arcsin(1/n)$), TIR fails and most of the light refracts out. About 5% of each ray Fresnel reflects under this condition.

But as β gets smaller (the extractor is lengthened) the average number of reflections that fail TIR increases, causing the efficiency to approach 100% as the extractor is lengthened. Designing of a three-dimensional extractor, however, is not so simple. The 3D analogue to the triangle wedge in two dimensions is a cone. This shape does not have 100% extraction efficiency as the length becomes very large. Because axially symmetric systems conserve the skew invariant of light rays upon reflection, some rays can spiral around inside the cone and never fail TIR. More information explaining skew invariance is given in Appendix C. To visualize this effect, we depict the light reflecting around a circle as shown in Figure 13.67b. The angle each reflection makes with the normal to the surface of the circle is conserved, and the skew invariant of a reflected ray is

$$h = r \sin\theta_n \tag{13.13}$$
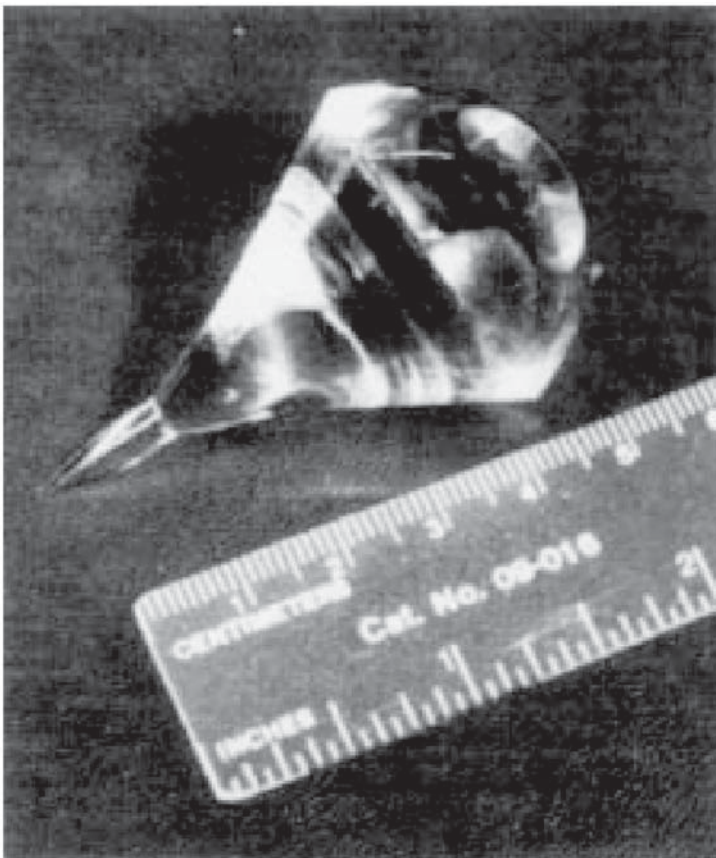
Figure 13.66   Photo of fused silica secondary with extractor tip.
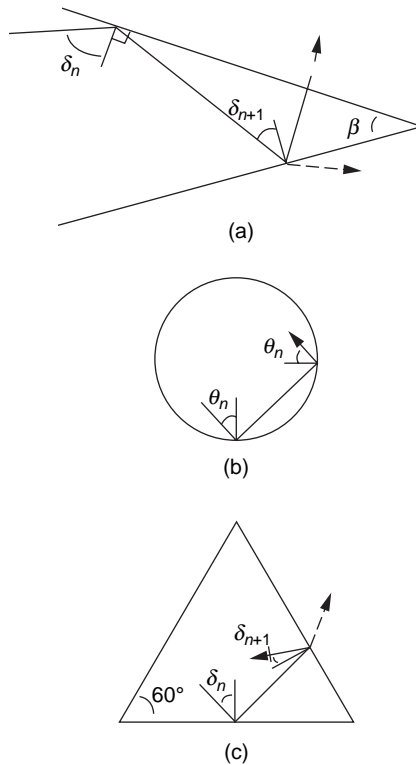
Figure 13.67   Design considerations for the extractor tip.

where r is the circle radius and θ_n is the angle made with the normal to the circle. Rays with $h > R \sin \theta_c$ cannot exit the circle, as reflection always fails TIR ($\theta_n > \theta_c$). In a cone, the skew invariant is conserved, but the cross-sectional radius decreases along the length. Keeping h constant increases the value θ_n. It is easily shown that rays with $h > R \sin \theta_c$ never fail TIR at any position inside the cone. The performance efficiency tested via a Monte Carlo ray-tracing algorithm indicates that the maximum extraction efficiency of an axially symmetric conical extension is under 90%. Such a loss of 10% of the output power is very undesirable, so a modification is needed.

It turns out that by taking an extension consisting of a long cylinder and shaving off three planes, one breaks the axial symmetry, and there is no limit on extraction. This extractor acts as a cone near the entrance and as a three-sided pyramid at its tip. Inside a three-sided pyramid, the cross-section is an equilateral triangle, as shown in Figure 13.67c. Rays inside this triangle, unlike those inside a circle, cannot bounce around indefinitely without failing TIR. Since all rays eventually refract out, the maximum extraction efficiency approaches 100%. Cross sections with higher-degree polygons, such as squares and hexagons, also break the axial symmetry that limits the maximum extraction efficiency. These may also be able to reach 100% extraction efficiency, but they are inferior to an extractor with triangular cross section as ray traces indicate that it has the highest efficiency for a given extractor length. The efficiency of the three-sided extractor
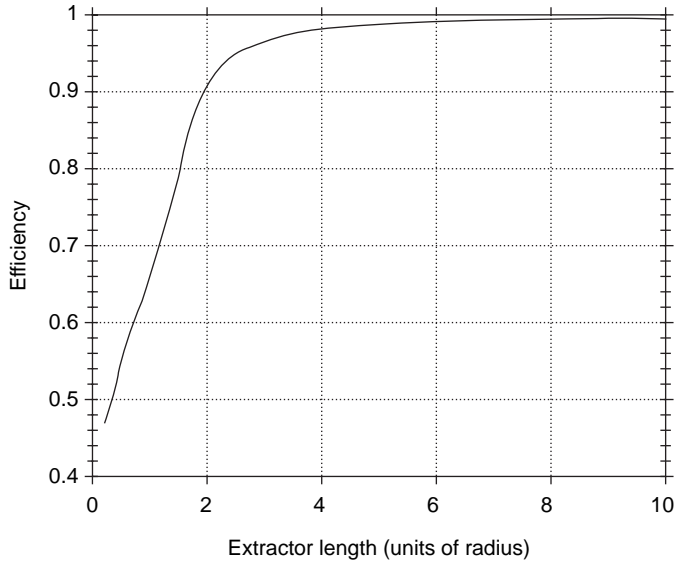
Figure 13.68   Optimizing the extractor tip.

for $n = 1.46$ as a function of the length L relative to the exit radius R is shown in Figure 13.68. Making the pyramid as long as possible for a given cylinder length maximizes the extractor efficiency. For the experiment measuring a solar concentration of 50,000, a $L = 15$ mm long and $R = 2.3$ mm extractor gives $L/R > 6$ and more than 99% efficiency. The secondary concentrator and extractor tip is shown in Figure 13.66. This design is very robust in that the tolerance on manufacturing the planar surfaces is fairly high. Any plane can be tilted by fairly large angles (up to 10°) with little effect on extraction efficiency. Implications and applications of such devices on future work is explained later in this chapter.

13.6 APPLICATIONS USING HIGHLY CONCENTRATED SUNLIGHT

Experiments performed at the Weizmann Institute of Science Solar Tower (Krupkin, Kagan, and Yoger, 1993; Yoger, 1997) and at the National Renewable Energy Laboratory's (NREL) high-flux solar furnace (HFSF) have demonstrated the effectiveness of using concentrated sunlight and advanced nonimaging secondaries to pump lasers and produce fullerenes (potentially useful new forms of molecular carbon). The recently developed techniques that allow more flexibility in design (Jenkins and Winston, 1996) have been used to develop two new configurations, each of which couples the high solar flux available at the HFSF to unusually shaped targets that impose unusual constraints. The first was a Nd:YAG laser crystal rod, 1 cm in diameter and 13 cm long. The second was a 6 mm diameter graphite rod used as feed-stock in an experimental fullerene production reactor. The laser-pumping secondary is designed to deliver the solar flux around the sides

of the barrel of the laser rod, which is perpendicular to the optic axis of the concentrator, while the fullerene secondary develops the high flux on the tip of one end of the graphite rod, which could be fed into the bottom of the reactor along the axis of the concentrator. Both the laser crystal rod and the fullerene reactor require active water cooling.

13.6.1   Solar-Pumped Lasers

Soon after the first laser was built, people imagined machines that would convert sunlight directly into laser radiation. Simple solar lasers have been around for almost 30 years (Young, 1965). One main problem preventing widespread use of sunlight-pumped lasers is that the power levels and efficiencies obtained are very low. The power of sunlight hitting a typical square meter of Earth's surface is relatively low, and concentrating enough sunlight to achieve lasing threshold can be difficult. Recent advances in solid-state crystals have been shown to increase the performance of commercial lasers greatly (by reducing the size of a given power system), and these allow much higher-efficiency solar-pumped lasers to be built. The main constraint is concentrating sunlight sufficiently to surpass lasing threshold. The measurement of ultra-high fluxes limits how effectively sunlight can pump a laser and what materials could possibly lase.

The most common solid-state lasing material is neodynium-doped yttrium aluminum garnet (Nd:YAG), and many experiments have pumped it using sunlight. Other less common crystals, such as Nd:GSGG, have been used to increase the laser efficiency. The overlap of the solar spectrum with Nd:YAG is about 14%, and less common (and costlier) lasing materials increase this value significantly. A list of solar lasing experiments to date is given in Table 13.7. As can be seen, many different types of crystals have been used, and a wide range of solar configurations from one-stage to two-stage systems have been employed. There are two methods

Table 13.7.   Summary of solar laser experiments.

Reference (year)	Crystal Type	Output Power	Collecting Area (m^2)	P_o/A (W/m^2)
High-power solar systems				
Arashi et al. (1984)	Nd:YAG	18	78.5	0.3
Weksler and Shwarz (1988)	Nd:YAG	60	38.5	1.6
	Nd:Cr:GSGG	100	38.5	2.6
Benmair et al. (1990)	Er,Tm,Ho:YAG	65	38.5	1.7
Jenkins (1996)	Nd:YAG	57	12.1	4.7
Multiple crystal lasers				
Krupkin et al. (1993)	Nd:YAG	500	660*	0.8
Small-scale systems (end-pumped systems)				
Falk et al. (1975)	Nd:YAG	4.5	0.29	15.5
	Nd:Cr:YAG	5	0.29	17.2
Cooke (1992)	Nd:YAG	1.7	0.41	4.1
	Nd:Cr:GSGG	3.2	0.41	7.8

*Estimated by number of heliostat mirrors in secondary's field of view (12) × area of each mirror (54.7 m^2).

for pumping lasers using sunlight: end pumping and side pumping. Figure 13.69 shows the two techniques. There are advantages and disadvantages in using both ways. The end-pumping scheme can be used on very small scales, but the pump light enters the laser crystal entirely from one end. This requires fluxes above 40,000 suns to lase Nd:YAG. Previous high-flux measurements (Jenkins et al., 1995) show that this system cannot be scaled up indefinitely, as the optical coupling between dielectric surfaces degrades when exposed to ultra-high solar fluxes at the kilowatt power scale. An end-pumping scheme has a maximum output power to about 5 W. The side-pumping scheme doesn't face this limitation because the lasing medium is excited by sunlight entering over a much larger aperture (the side walls of the laser crystal). This lowers the threshold concentration, and the output power can be scaled by increasing the laser rod's length. The disadvantage is that the diameter of the laser crystal must be at least 0.5 cm to get high efficiency. If the mean absorption length of sunlight is large compared with the diameter of the laser crystal, less sunlight is absorbed, and output power and efficiency is reduced. The disadvantage of side-pumping is that large-scale solar furnaces (matching the flux region size to the size of the laser crystal) are required, while the advantage is that these systems can be scaled up greatly, and output powers above 500 W have been reported (Krupkin et al., 1993). The most advanced work on solar pumped lasers has been done at the Weizmann Institute Solar Tower.
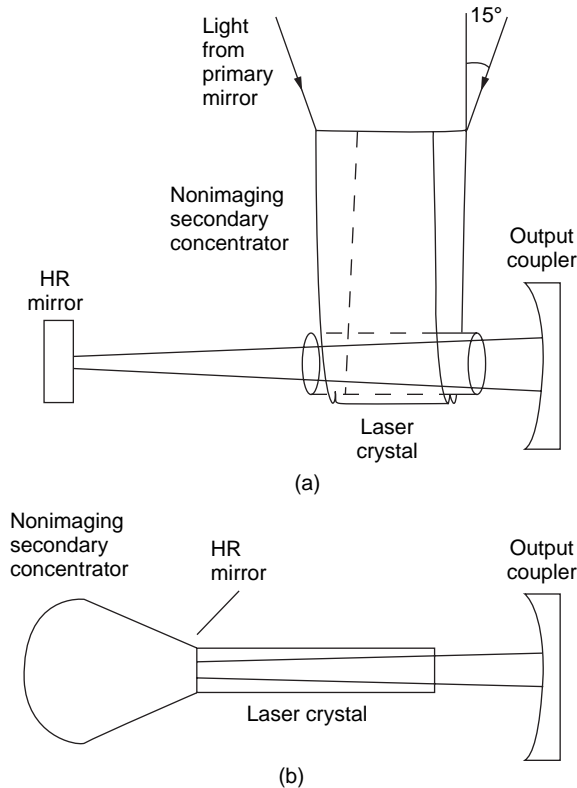
Figure 13.69   Schematic of solar laser showing (a) side pumping and (b) end pumping.
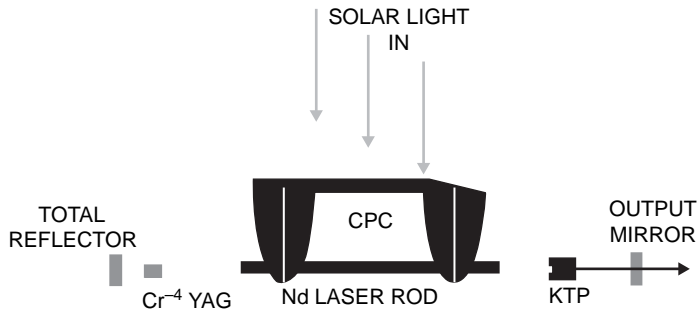
Figure 13.70   Side-pumped solar laser.
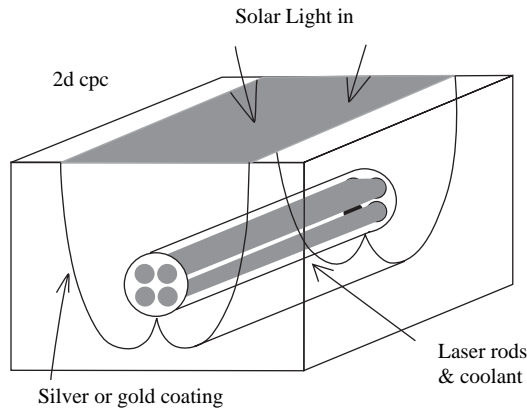
Figure 13.71   Weizmann solar-pumped lasers.
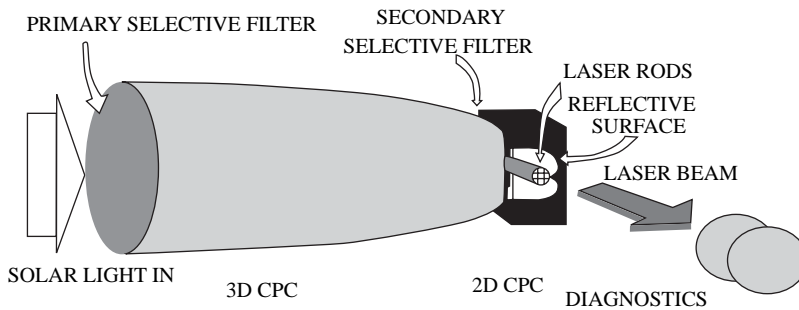
Figure 13.72   Weizmann solar-pumped laser using CPCs in tandem.

Figure 13.70 shows a schematic view of a side-pumped solar laser, and Figure 13.71 is a detailed view of how the solar energy is coupled to the laser crystal. The overall experiment, which uses both 3D and 2D CPCs, is shown in Figure 13.72.

As can be seen from Table 13.7, the efficiencies of present solar lasers are fairly low when compared with other solar processes, such as PV-electric generation. A

square meter of primary mirror can presently be converted into about 5 W of laser output. With a solar insolation level of 1,000 W/m2, this is a conversion efficiency of 0.5%. This is not the appropriate standard of comparison because the cost of industrial laser power is much more expensive than that of electricity. The important factor determining solar laser use will depend on cost/benefit analysis of a given application.

This is where forecasting becomes problematic, as applications for solar lasers are somewhat speculative and futuristic. Current research suggests that space-based applications hold the most promise for solar lasers (Brauch et al., 1991; 1992). In space, the solar insolation level, the input pump source, is more intense and much more stable. This in turn allows more stable lasing configurations that increase laser output efficiency and brightness (and building an electrically pumped laser is difficult). On Earth, high-brightness lasers may require the resonator mirrors to move in real time as the solar insolation varies. Also, there are partially and totally cloudy days (and nights) on Earth. This will prevents round-the-clock operation of a solar-pumped laser and decreases potential for widespread commercial use, as electrical-pumped lasers are a well-developed technology and are becoming cheaper all the time. In space, the generation of electrical power to pump a laser is not a trivial endeavor. The competition in space to direct pumping of solar lasers is the use of photovoltaic cells to generate electricity for a commercial laser. Directly pumped solar lasers compete very well with energy-inefficient flash-lamp-pumped lasers. These commercial lasers are impractical to power by photovoltaic (PV) cell arrays and have a low brightness (10–15 times the diffraction limit) in general. However, recently, high-power diode-pumped solid-state lasers are being produced that have very high electrical to laser light conversion efficiencies and beam divergences at 1.01 times the diffraction (minimum allowed) limit (Gitin and Reingrube, 1995).

This system is the main competition to directly solar-pumped lasers. Additional advantages are that any material can be pumped, and the lasing technology is very well developed. Conceptually, powering these lasers with PV panels is not very difficult to imagine. For PV cells converting sunlight into electricity with around a 10% efficiency and with a conversion of electricity into laser power of 5%, one gets the same efficiency as our directly pumped system (0.5%). Using different materials may increase the efficiency of either direct- or diode-pumped solar lasers, leaving the question of which type of laser to use for future research. Applications of space-based lasers include industrial processing materials at a lunar base, power transmission between satellites, and communications in space.

13.7 SOLAR PROCESSING OF MATERIALS

The properties of concentrated solar flux make it very amenable to any industrial process that involves the heating of materials or surfaces. Sunlight has a very wide wavelength band from 0.3 mm to 2.5 mm—much greater than that found in most other industrial light sources, such as flash lamps or lasers.

Ideally (no atmospheric or optical system losses), concentrated sunlight can cause materials to reach equilibrium temperatures equal to those found at the sun's surface (around 6,000 K). Since, for most applications, heating materials need
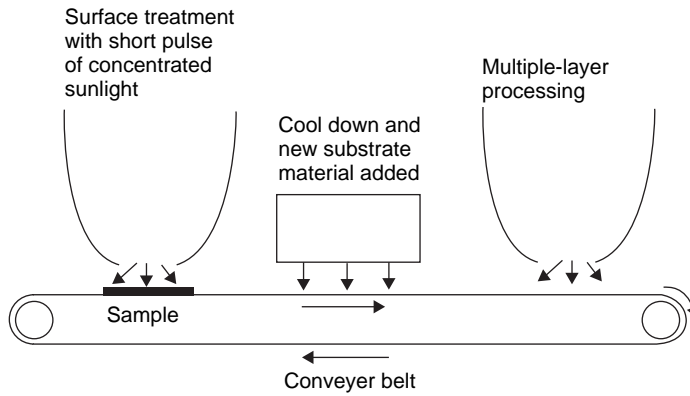
Figure 13.73   Schematic of a high-solar-flux materials process.

significantly lower temperatures (or else vaporization occurs), sunlight could possibly be used in a vast array of industrial processes. Ultra-high fluxes achieve the highest temperatures and test the limits of what can be done with sunlight. Samples with any absorption bands in the solar spectrum can be heated with solar flux, and the rate of heating can be controlled. Laser light (with its extremely narrow line widths) can have much lower overall absorption). Coupling in the cost of producing laser light to direct power from sunlight, there is potential for sunlight to replace laser processing in many areas. By varying the concentration levels, the heating rate and final equilibrium temperature can be adjusted. The limiting factor on heating is how much sunlight is absorbed over the entire solar spectrum, the emissivity of the sample, and maximum concentration of sunlight that a solar furnace can produce. Additionally, some processes, such as producing fullerenes (Chibante et al., 1993; Fields et al., 1993), require that parts of the solar spectrum be removed. Filtering of sunlight can be done before it is concentrated, preventing damage to the filters when exposed to ultra-high solar fluxes. Ultra-high fluxes probe the limits of what solar processing methods can cover. With short-duration heating of a bulk material (1–30 s), nonequilibrium processes are studied. During short heating pulses the bulk material's temperature barely rises, while the surface temperature increases greatly. Thus, the surface is modified, while the material's bulk properties are unchanged. By setting down solar-treated surface layers one after another, one sees the vast potential for industrial applications. Industrial processing of materials is depicted in Figure 13.73.

An excellent summary of the processes that are being researched currently using highly concentrated sunlight is available in the article by Pitts and Lewandowski (1993); Pitts et al., (1993); and Stanley, Fields, and Pitts (1990). By using the high-flux solar furnace (HFSF) at NREL, solar-induced surface transformation of materials (SISTM) includes the following research highlights: surface-hardening of steel, cladding (bonding) of thin surface layers to cheaper bulk materials, and rapid thermal annealing to remove defects in silicon wafers (used as semiconductors and thin-film solar cells). Table 13.8 lists various processes that could use ultra-high solar fluxes for materials processing.

Table 13.8.   Processes that use ultra-high solar fluxes for materials processing.

Process	Explanation
Transformation hardening of steels	Hardens surface while keeping bulk properties unchanged. Thin layers, such as those found on knife blades, could be produced with relatively little heat.
Cladding (bonding) of materials	Similar to gold-plating materials.*
Self-propagating high-temperature synthesis (SHS)	Thermal pulse starts exothermic reaction and propagates throughout reaction mixture.
Plasma- and flame-sprayed coatings	Solar processing may replace use of lasers to reduce cracking and imperfections of these coatings.
Modifying films produced by physical vapor deposition (PVD)	Thin layers are placed on substrates. Solar process creates desired coating by solid-state diffusion and reacting or by melting and liquid diffusion and reacting.
Chemical vapor deposition (CVD)	Solar systems heat surfaces to help the reactive formation of materials procede.
Rapid thermal annealing	Ion implantation very important for semiconductors. Fast solar pulses of 1–$10\,s$ can anneal out defects during implantation, and current research studies the production of high-efficiency thin-film solar cells†
Metalorganic deposition	Similar to CVD. Heated surface aids in creation of materials.
Joining materials together	Welding of mismatched joints may be helped by selective application of sunlight.
Space applications (more details in this chapter)	All of the preceding and processes in zero gravity may allow extremely pure samples to be created.

*See Tsuo et al. (1996).
†See Pitts and Lewandowski (1993) and Rawers et al. (1994).

13.8   SOLAR THERMAL APPLICATIONS OF HIGH-INDEX SECONDARIES

This is another possible use of high-index secondaries other than end-pumping solid-state laser crystals in generating high temperatures to drive gas turbines. Research is in progress at the Weizmann Institute of Science in Israel (Karni et al., 1995) on the coupling of dielectric secondaries with gas turbines. They use a six-sided extractor tip (Figure 13.74) to let solar flux out of the high-index secondary and into the high-temperature gas environment of the turbine ($n = 1$). By sealing it to the electric generation system, it acts as the window to the turbine and as a novel optical device to extract the solar flux from the borosilicate secondary into the low-index air environment where flowing gases are heated to temperatures of 1,000°–1,350°C. Their thermal system design of the directly irradiated annular pressurized receiver (DIAPR) requires solar concentrations of 5,000–10,000 to minimize reradiation losses when operating at such high
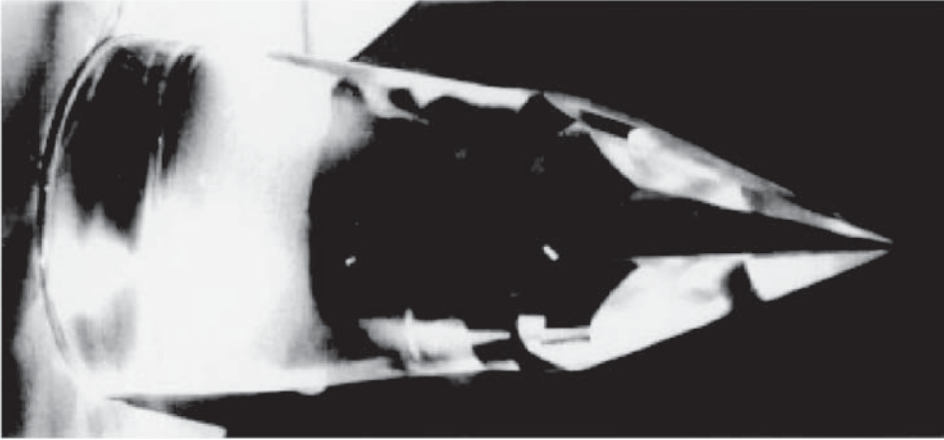
Figure 13.74   Weizmann quartz (fused silica) secondary with extractor tip.
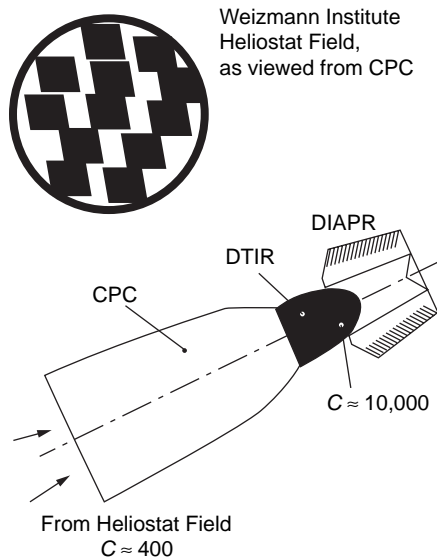
Figure 13.75   Weizmann receiver using dielectric secondary.

temperatures (Figure 13.75). A summary of their system is given by Karni et al. (1995). More detailed information on the extractor can be obtained from Karni et al. (1994). The use of high-index concentrators with TIR reflection conditions (no losses from reflection that occur on metal surfaces) minimize optical losses in the system and increases the concentration limit by n^2. In general, the use of a higher-index material to form the aperture of a gas turbine will not produce higher operating temperatures.
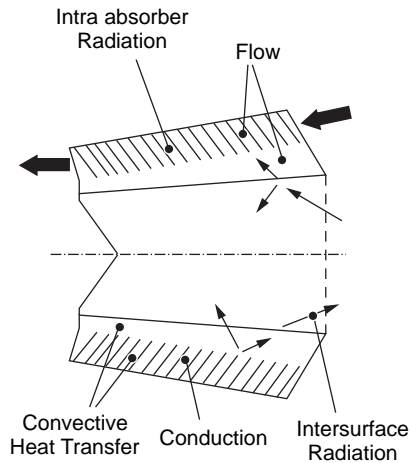
Figure 13.76   Detail of Weizmann high-flux receiver.

This is because the reradiation is proportional to n^2 times the area of the secondary exit aperture, which normally cancels the n^2 gain in concentration obtained by using higher-index materials. All gas turbines operate at temperatures much less than that of the sun (6,000 K), and therefore the spectrum of reradiation is considerably red-shifted into the infrared (IR) region. By applying selective coatings onto the dielectric concentrator and extractor optics, one may prevent IR radiation from reradiating, and it may be possible to reduce heat losses inside high-temperature electrical generation systems. The various thermal transfer paths in the high-temperature receiver are shown in Figure 13.76.

13.9   SOLAR THERMAL PROPULSION IN SPACE

Solar thermal propulsion systems in space will require very high temperatures to generate necessary levels of thrust by the direct solar heating and resulting expansion and expulsion of the propellent material. The optimal material is hydrogen gas because this lowest-mass molecule has the highest thrust to mass ratio. The generation of such temperatures, in the range 1,400°–2,200°C, will in turn require very high levels of solar flux concentration. In practice, to attain such levels it may be useful and perhaps even necessary to incorporate some form of ideal or near ideal nonimaging concentrator. In particular, recent work indicates that very substantial gains in performance can be generated through the incorporation of a properly designed nonimaging secondary concentrator combined in a two-stage configuration with a focusing primary concentrator. Two strategies for final stage concentration are shown in Figure 13.77. The all-dielectric option has superior thermal performance but is more demanding on materials. Figure 13.78 shows a conceptual diagram of a solar-powered hydrogen turbine using an all-dielectric concentrator with an extractor. The considerations in evaluating the two design
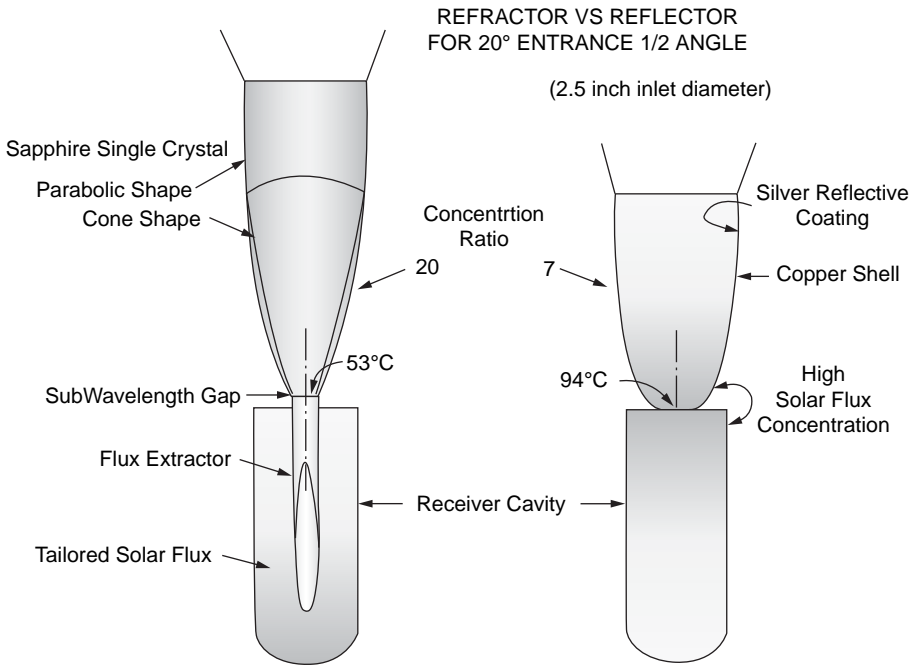
Figure 13.77   Final stage concentration.
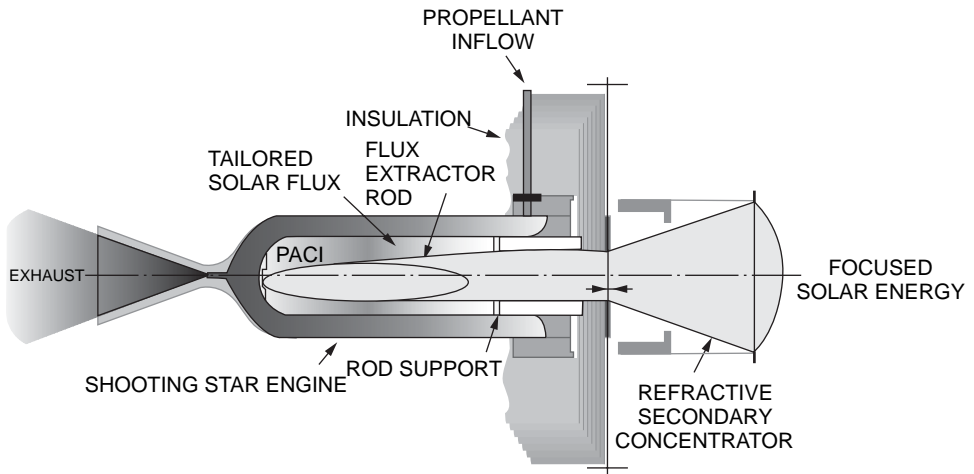
Figure 13.78   NASA concept for hydrogen propulsion using high solar flux.

Table 13.9. Refractive secondary concentrator overview.

Background	Challenges
As the need for achieving super-high temperatures (2500°K and above) in solar heat receivers has developed, so also has the need for utilizing secondary concentrators. Secondary concentrators refocus the already highly concentrated solar energy provided by the primary collector, thereby significantly reducing the heat receiver light entrance aperture and the resulting radiation reflux loss from the cavity. *Benefits* The use of solid single-crystal optically clear materials for the secondary concentrator offer a number of advantages over a hollow reflective secondary: • Higher thermal efficiency (~95% vs ~80%) • No special cooling required • Blocks cavity outgassing by sealing the cavity • Provides for flux tailoring via flux extractor • Potential reduction of IR reflux loss from cavity	The technical challenges of designing and fabricating refractive secondary concentrator/flux extractors are many: • Optical materials that can survive the environment (high temperature, hydrogen atmosphere, etc.) • Coatings for enhancing optical and thermal performance • Concentrator/extractor interface gap that reduces conductive losses • Hardware attachments that can survive launch loads • Material and fabrication sources *Applications* • Solar thermal propulsion (AITP, Shooting Star) • Solar thermal power systems • Solar bimodal system (e.g., ISUS) • Space commercialization (e.g, solar furnaces)

options are summarized in Table 13.9. As of this writing, NASA has adopted the more conservative option of a metal reflecting CPC (albeit the metal is a more exotic iridium than copper) as the final stage concentrator in a solar-driven hydrogen turbine.

REFERENCES

Algora, C., Ortiz, E., Rey-Stolle, I., Díaz, V., Peña, R., Andreev, V. *et al.* (2001). A GaAs solar cell with an efficiency of 26.2% at 1,000 suns and 25.0% at 2,000 suns. *IEEE Transact. on Elec. Devices* **48**(5), 840–844.

Álvarez, J. L. (2001). Doctoral Thesis, Universidad Politecnica de Madrid.

Álvarez, J. L., Hernández, M., Benítez, P., and Miñano, J. C. (1999). RXI concentrator for 1,000× photovoltaic energy conversion. *Nonimaging Optics: Maximum Efficiency Light Transfer V* (Winston, R., ed.). Proceedings of SPIE, Vol. 3781, pp. 30–37.

Álvarez, J. L., Hernández, M., Benítez, P., and Miñano, J. C. (2001). TIR-R concentrator: A new compact high-gain SMS design. In *Nonimaging Optics: Maximum Efficiency Light Transfer VI* (Winston, R., ed.). Proc, SPIE, Vol. 4446, pp. 32–42.

Arboiro, J. C. (1997). Doctoral Thesis, Universidad Politécnica de Madrid.

Bean, J., and Diver, R. (1992). The CPG 5-kWe dish-stirling development program, Paper No. 929181. *Proceedings of the 27th IECEC*, San Diego, CA.

Benítez, P. (1998). Doctoral Thesis. ETSI Telecominicación, UPM, Madrid, pp. 23–53.

Benítez, P., and Miñano, J. C. (1997). Ultrahigh-numerical-aperture imaging concentrator. *J. Opt. Soc. Am. A.* **14**, 1988–1997.

Benítez, P., and Miñano, J. C. (2003). Concentrator optics for next generation photovoltaics. In *Next Generation Photovoltaics*. Institute of Physics Publishing. (Eds. A., Martí and Luque, A.)

Benítez, P., and Mohedano, R. (1999). Optimum irradiance distribution for photovoltaic solar energy conversion. *Appl. Physics Letters* **74**(17), 2543–2545.

Benítez, P., García, R., and Miñano, J. C. (1997). Contactless efficient two-stage solar concentrators for tubular absorber. *App. Opt.* **36**, 7119–7128.

Benítez, P., Hernández, M., Mohedano, R., Miñano, J. C., and Muñoz, F. (1999). New nonimaging static concentrators for bifacial photovoltaic solar cells. *Nonimaging Optics: Maximum Efficiency Light Transfer V* (Winston, R., ed.). Proceedings of SPIE, Vol. 3781, pp. 22–29.

Benítez, P., Miñano, J. C., García, R., and Mohedano, R. (1997). Contactless two-stage solar concentrators for tubular absorber. In *Nonimaging Optics: Maximum Efficiency Light Transfer IV* (Winston, R., ed.). Proceedings of SPIE, Vol. 3139, pp. 205–216.

Benítez, P., Miñano, J. C., and González, J. C. (1995). Single-mirror two-stage concentrator with high acceptance angle for one axis tracking PV systems. *Proceedings of the 13th European Photovoltaic Solar Energy Conference*. pp. 2406–2409.

Benítez, P., Mohedano, R., and Miñano, J. C. (1997). DSMTS: A novel linear PV concentrator. *26th IEEE Photovoltaic Specialists Conference*.

Benítez, P., Mohedano, R., Miñano, J. C., García, R., and González, J. C. (1997). Design of CPC-like reflectors within the Simultaneous Multiple Surface design method. In *Nonimaging Optics: Maximum Efficiency Light Transfer IV* (Winston, R., ed.). Proceedings of SPIE, Vol. 3139, pp. 205–216.

Bowden, S. (1993). High Efficiency Photovoltaic Roof Tile with Static Concentrator. 23rd IEEE Photovoltaic Specialist Conference, pp. 1068–1072.

Brauch, U., Muckenschnabel, J., Opower, H., and Wittwer, W. (1991). Solar-pumped solid state lasers for space to space power transmission. *Space Power* **10**(3–4), 285–294.

Brauch, U., Muckenschnabel, J., Thompson, G., Bernstein, H., Yogev, A. *et al.* (1992). Influence of operating temperature on the power, divergence, and stress-induced birefringence in solar-pumped solid state lasers. *Optical Engineering* **31**, 1072–1078.

Cassarly, W. (2001). Nonimaging optics: Concentration and illumination. In *Handbook of Optics*, 2nd ed. McGraw-Hill, New York, pp. 2.23–2.42.

Chibante, L., Thess, A., Alford, J., Diener, M., and Smalley, R. (1993). Solar generation of the fullerenes. *J. Phys. Chem.* **97**(34), 8696–8700.

Collares-Pereira, M., and Mendes, J. F. (1995). Novel second stage solar concentrator for parabolic troughs. In *Nonimaging Optics: Maximum Efficiency Light Transfer III* (Winston, R., ed.). Proceedings of SPIE, Vol. 2538, pp. 125–130.

Collares-Pereira, M., Gordon, J. M., and Rabl, A. (1991). High concentration two-stage optics for parabolic trough solar collectors with tubular absorber and large rim angle. *Solar Energy* **47**(6), 457–466.

Cooke, D. (1992). Sun-pumped laser: Revisiting an old problem using nonimaging optics. *Applied Optics* **31**, 7541–7546.

Cooke, D., Gleckman, P., Krebs, H., O'Gallagher, J., Sagie, D., and Winston, R., (1990). Brighter than the sun. *Nature* **346**, 802.

Feuermann, D., and Gordon, J. M. (2001). High-concentration photovoltaic designs based on miniature parabolic dishes. In *Nonimaging Optics: Maximum Efficiency Light Transfer VI* (Winston, R., ed.). Proceedings of SPIE, Vol. 4446, pp. 43–51.

Fields, C., Pitts, R., Hale, M., Bingham, C., Lewandowski, A. *et al.* (1993). Formation of fullerenes in highly concentrated solar flux. *J. Phys. Chem.* **97**(34), 8701–8702.

Gitin, M., and Reingrube, J. (1995). Diode-Pumped Solid-State Lasers Show Bright Future, Ind. Laser Rev. **10**, 8–10.

Gleckman, P. (1988). Achievement of ultrahigh solar concentration with potential for efficient laser pumping. *Applied Optics* **27**, 4385–4391.

Gleckman, P., O'Gallagher, J., and Winston, R. (1989). Concentration of sunlight to solar-surface levels using non-imaging optics. *Nature* **339**(18), 198–200.

Goebel, O., Geskes, P., Geyer, M., Hemecke, K., and Herrmann, U. (1996/1997). Forschungs verbund sonnenenergie, Themen 96/97, pp. 71–74.

James, L. W. (1989). Use of imaging refractive secondaries in photovoltaic concentrators. SAND89-7029, Alburquerque, New Mexico.

Jenkins, D. G. (2001). High-uniformity solar concentrators for photovoltaic systems. In *Nonimaging Optics: Maximum Efficiency Light Transfer VI* (Winston, R., ed.). Proceedings of SPIE, Vol. 4446, pp. 52–59.

Jenkins, D., and Winston, R. (1996). Integral design method of nonimaging optics. *J. Opt. Soc. Am. A.* **13**, 2106–2116.

Jenkins, D., Winston, R., O'Gallagher, J., and Lewandowski, A. (1995). Uses of ultrahigh solar flux. *Proceedings of American Solar Energy Society Conference*, Minneapolis, pp. 130–135.

Karni, J., Ries, H., Segal, A., Krupkin, V., and Yogev, A. (1994). Delivery of Radiation for a Transparent Medium. Israel Patent 109,336 and international patent application PCT/US95/04915.

Karni, J., Ries, H., Segal, A., Krupkin, V., and Yogev, A. (1995). The DIAPR: A High-Pressure, High-Temperature Solar Receive. *International Solar Energy Conference*, Hawaii, pp. 591–596.

Kritchman, E. M. *et al.* (1979). Efficient Fresnel lens for solar concentrator. *Solar Energy* **22**, 119–123.

Krupkin, V., Kagan, Y., and Yogev, A. (1993). Nonimaging optics and solar laser pumping at the Weizmann Institute. Nonimaging optics: Maximum Efficiency Light Transfer II. Proceedings of SPIE, Vol. 2016, pp. 50–60.

Lewandowski, A., Bingham, C., O'Gallagher, J., Winston, R., and Sagie, D. (1991). *One and Two Stage Flux Measurements at the SERI High Flux Solar Furnace*. Proceedings ASME-JSES-JSME International Solar Energy Conference, Reno, NE, pp. 329–335.

Luque, A., Lorenzo, E., and Ruiz, J. M. (1980). Connection losses in photovoltaic arrays. *Solar Energy* **25**, 171–178.

Medvedev, V., Parkyn, W. A., and Pelka, D. G. (1997). Uniform high-efficiency condenser for projection systems. In *Nonimaging Optics: Maximum Efficiency Light Transfer IV* (Winston, R., ed.). Proceedings of SPIE, Vol. 3139, pp. 122–134.

Mills, D. R. (1995). Two-stage collectors approaching maximal concentration. *Solar Energy* **54**, 41–47.

Miñano, J. C., Benítez, P., and Gonzalez, J. C. (1995). RX: A nonimaging concentrator. *Appl. Opt.* **34**, 2226–2235.

Mohedano, R., Benítez, P., Miñano, J. C., Tercero, F., and Lobato, P. (2001). Design, construction and measurements of a dielectric-single mirror two stage (DSMYS) photovoltaic concentrator. *Nonimaging Optics: Maximum Efficiency Light Transfer VI* (Winston, R., ed.). Proceedings of SPIE, Vol. 4446, pp. 20–31.

Ning, X., Winston, R., and O'Gallagher, J. (1987). Dielectric totally internal reflecting concentrators. *Appl. Opt.* **26**(2), 300–305.

O'Gallagher, J. J., and Winston, R. (2001). Nonimaging solar concentrator with near-uniform irradiance for photovoltaic arrays. In *Nonimaging Optics: Maximum Efficiency Light Transfer VI* (Winston, R., ed.). Proceedings of SPIE, Vol. 4446, pp. 60–64.

Pitts, R., and Lewandowski, A. (1993). Deposition of diamond-like carbon films and other materials processes using a solar furnace. *Materials Technology* **8**, 237.

Pitts, R., Tracy, E., Shinton, Y., and Fields, C. (1993). Applications of solar energy to surface modification processes. *Critical Reviews in Surface Chemistry* **2**(4), 247–269.

Rabl, A. (1976). Tower reflector for solar power plants. *Solar Energy* **18**, 269–271.

Rabl, A. (1985). *Solar Thermal Collectors and their Applications*. Oxford University Press, Oxford.

Rabl, A., and Winston, R. (1976). Ideal concentrators for finite sources and restricted exit angles. *Appl. Opt.* **15**(11), 2880–2883.

Ries, H., and Spirkl, W. (1995). Caustic and its use in designing optimal absorver shapes for 2D concentrators. Proceedings of SPIE, Vol. 2538.

Ries, H., and Spirkl, W. (1996). Nonimaging secondary concentrators for large rim angle parabolic troughs with tubular absorber. *Appl. Opt.* **35**, 2242–2245.

Ries, H., Gordon, J. M., and Laxen, M. (1997). High-flux photovoltaic solar concentrators with kaleidoscope-based optical designs. *Solar energy* **60**(1), 11–16.

Stanley, J., Fields, C., and Pitts, R. (1990). Surface treating with sunbeams. *Advanced Materials and Processes* **138**(6), 16.

Stevenson, A. (1835). *Report to the Committee of the Commissioners of the Northern Lights, appointed to take into consideration the subject of illuminating the lighthouses by means of lenses*, Elton Engineering Books.

Timoshenko, S. (1983). *Strength of Materials*. 3rd Ed. Krieger Publishing Co, New York.

Vant-Hull, L. L., and Hildebrandt, A. F. (1976). Solar thermal power system based on optical transmission. *Solar Energy* **18**, 31.

Welford, W. T., and Winston, R. (1980). Design of nonimaging concentrators as second stages in tandem with image forming first-stage concentrators. *Appl. Opt.* **19**(3), 347–351.

Winston, R. (1974). Principles of solar concentrators of a novel design. *Solar Energy* **16**, 89–94.

Yogev, A. (1997). Large scale solar projects at the Weizmann Institute. Proceedings of SPIE, Vol. 3139, pp. 186–193.

Yoshioka K. et al. (1995). Performance Simulation of a Three-Dimensional Lens for a Photovoltaic Static Concentrator, *13th European Photovoltaic Solar Energy Conference*. pp. 2371–2376.

Young, C. (1965). A sun-pumped CW 1-W laser. *Applied Optics* **5**, 993–997.

<p align="center">14</p>

MANUFACTURING TOLERANCES

14.1   INTRODUCTION

When a series of concentrators is manufactured, they necessarily differ from the theoretical design model in several aspects. In the first place, the optical surfaces have microscopic roughness, which causes some scattering of the incident light. Second, the surfaces profiles have contour errors—that is, they do not fit perfectly the theoretical ones. This results in slope variations between the two and thus angular deviations of the exiting rays. Third, the relative positions of source, receiver, and optical surfaces involve a certain tolerance. All of these errors affect the performance of the concentrators, and for each application they should be kept below a certain level to guarantee a specified quality.

The effect of errors is well known in the case of parabolic reflectors (Butler and Pettit, 1977; Lof and Duffie, 1963; Rabl, 1985; Winter, Sizmann, and Vant-Hull, 1990). Moreover, there exists an analytical framework that provides functional relationships between the parameters involved. Within this theory, concentrator errors are characterized by the probability density function (pdf) of the angle of deviation of the reflected rays. The angular transmission of the real parabola (i.e., with errors) is calculated as the convolution of this pdf with the theoretical angular transmission (i.e., without errors). It is also possible, in the two-dimensional case, to transfer the mirror errors to the source (the sun, in solar applications): The parabola can be considered error-free, but the source must be widened through the convolution with the pdf. However, for concentrators with two or more optical surfaces, no analytical theory that quantifies the effect of errors is available, and manufacturing tolerances are therefore studied by means of ray-tracing calculations.

In Section 14.2 we present a 2D-analytical error model for multiple optical surface concentrators. The optical surfaces may be refractive surfaces, sequential mirrors (such as a conventional parabola), or nonsequential mirrors (i.e., CPC-like mirrors). The key to the model is the error transfer between optical surfaces: The errors of one surface can be transferred to another one, leaving the first surface apparently error-free and increasing the errors of the second one. Through this transfer process it is possible, following the hypothesis of the model, to transfer

<p align="center">395</p>

the errors of all the optical surfaces up to the entry aperture and combine them there.

Using this model the complete concentrator can be studied at the entry aperture with a single probability density function, as in the case of the parabolic reflector, and therefore the errors can be transferred from the concentrator to the source. The model can be employed for predicting the behavior of the average concentrator from a manufactured series (if certain parameters of the process are known) or for estimating the behavior of a specific concentrator of the series.

In Section 14.4 we define the concentrator error multiplier M, which is a single number that characterizes the concentrator for error calculations, when the errors at all surfaces are equal. The utility of this number, which depends only on the concentrator optics and geometry, is its principal characteristic: It is used for multiplying the slope error at the surfaces, which gives as a result an angle of deviation associated with each concentrator. The convolution of the pdf of this global angle of deviation with the theoretical angular transmission of the concentrator is the transmission of the concentrator with errors. In the well-known case of the parabolic reflector M equals 2.

Finally, Section 14.5 looks at the problem of specifying the maximum slope error acceptable for a given application. As will be seen, the maximum acceptable error depends on the error multiplier M but not solely on it.

14.2   MODEL OF REAL CONCENTRATORS

The model presented in this section is two-dimensional—that is, only objects contained in a plane are considered. Section 14.3.5 deals with the problem of extending this theory to three-dimensional concentrators with rotation symmetry. This study concentrates on contour errors of optical surfaces, although other errors, such as scattering, could easily be included.

14.3   CONTOUR ERROR MODEL

When a ray impinges on an optical surface with contour errors, it is found that (1) the point of incidence is different from the expected one, and (2) the slope of the surface at the point of incidence does not coincide with the theoretical value (see Figure 14.1). Are position error (a) and slope error (b) of equal importance? The general opinion is that, for optical purposes, position error is irrelevant, and only slope errors should be considered (Rabl, 1985). The problem with regard to demonstrating this statement is How do we compare the two types of error? In a specific case, the comparison can be made by means of ray-tracing. However, this solution does not quantify the sensitivity of the concentrator to position error, nor does it provide information on the parameters that, in general, affect this sensitivity.

Let us prove that the position error can be transformed into an apparent slope error, which is directly comparable with the real slope error. First, we'll consider that the optical surface is a refractive surface θ on which a wavefront Σ impinges. Let r be the ray associated with the front Σ that would impinge at point P of θ if this were error-free, and let R_I and R_C be the respective radii of curvature of Σ and
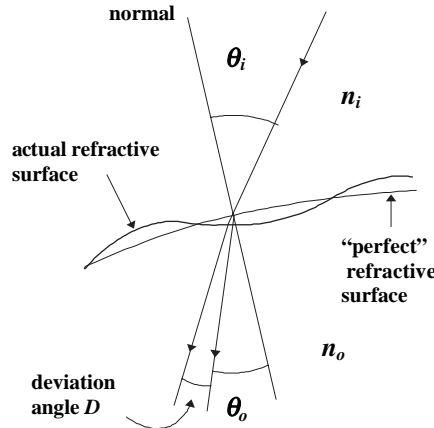
Figure 14.1   Contour errors of a refractive surface (not to scale).

(a)   (b)
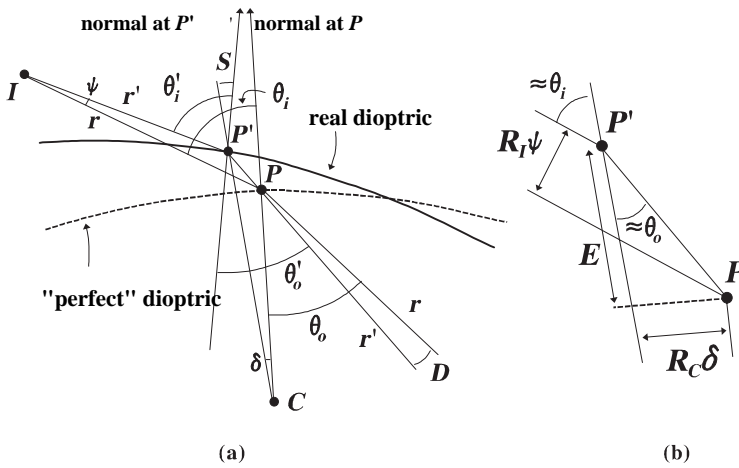
Figure 14.2   (a) The ray that appears to refract at point P with a deviation angle D actually impinges at P'. (b) Detail of the zone of incidence in the real refractive surface and the theoretical one.

θ at P. Figure 14.2 shows the points I and C, center of curvature of Σ and θ at P, which are therefore distant from P by the values R_I and R_C, respectively.

Due to the contour errors of the refractive surface θ, the real ray that emerges from point P is not the theoretical ray r, but the ray r' that is refracted at P' and with which r forms the deviation angle D [see Figure 14.2a]. The contour error at P' is manifested as a displacement E with respect to the normal to the theoretical profile and an angular difference S between its tangents. It is easy to see that E and S are related, in mathematical terms, as $S(s) \approx dE(s)/ds$, s being the arc-length parameter of the theoretical profile.

The following relationships between angles are deduced from Figure 14.2:

$$\theta_i' - \theta_i = S - \delta + \psi$$
$$\theta_o' - \theta_o = S - \delta - D \tag{14.1}$$

As Snell's law is fulfilled at P and P', we have

$$\left.\begin{array}{l} n_i \sin\theta_i = n_o \sin\theta_o \\ n_i \sin\theta_i' = n_o \sin\theta_o' \end{array}\right\} \Rightarrow n_i \sin\left(\frac{\theta_i' - \theta_i}{2}\right)\cos\left(\frac{\theta_i' + \theta_i}{2}\right) =$$

$$n_i \sin\left(\frac{\theta_o' - \theta_o}{2}\right)\cos\left(\frac{\theta_o' + \theta_o}{2}\right) \Rightarrow \frac{\theta_o' - \theta_o}{\theta_i' - \theta_i} \approx \frac{n_i \cos\theta_i}{n_o \cos\theta_o} = \frac{\tan\theta_o}{\tan\theta_i} \tag{14.2}$$

having obtained the final expression through the application of Snell's law. From Figure 14.2b, where the central area of Figure 14.2a is represented in enlarged form, approximating the arc length for that of the chord in the theoretical profile, it is deduced that

$$\left.\begin{array}{l} R_C\delta \approx E\tan\theta \\ R_i\psi \approx (E/\cos\theta_o)\sin(\theta_i - \theta_o) \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \delta \approx \dfrac{E\tan\theta_o}{R_C} \\[2mm] \psi \approx \dfrac{E(\sin\theta_i - \cos\theta\tan\theta_o)}{R_i} \end{array}\right. \tag{14.3}$$

Combining Eqs. (14.1), (14.2), and (14.3), we obtain finally

$$D = \left(1 - \frac{\tan\theta_o}{\tan\theta_i}\right)\left[S - E\tan\theta_o\left(\frac{1}{R_C} + \frac{\cos\theta_i}{R_i}\right)\right] \tag{14.4}$$

The first factor in Eq. (14.4) will be called m, the error multiplier. The first term of the second factor is S, the slope error at P'. The second term can be interpreted as an apparent slope error, associated with the position error E, of value

$$S_E = -E\tan\theta_o\left(\frac{1}{R_C} + \frac{\cos\theta_i}{R_i}\right) \tag{14.5}$$

The definition of S_E permits the comparison of the effects of the slope and position errors. Assuming a deterministic sinusoidal error $E = E\sin(2\pi s/\lambda) - s$ is the arc length parameter of the theoretical profile, it is possible to calculate the spatial wavelength λ_C for which the magnitudes of the two terms are equal—that is, $|S| = |S_E|$. Taking into account that $S = dE/ds$, we have

$$\lambda_C = 2\pi\left[\tan\theta_o\left(\frac{1}{R_C} + \frac{\cos\theta_i}{R_i}\right)\right]^{-1} \tag{14.6}$$

If the wavelength λ associated with a contour error fulfils $\lambda \ll \lambda_C$, then (and only then) the slope error will be dominant with respect to the position error. Thus, Eq. (14.6) quantifies when the position error is negligible, an approximation that is usually given only vague justification.

By way of an example, we might consider a parabolic concentrator with focal distance f and rim angle $\pm\phi_M$. Given a point P of the parabola, if we call ϕ the angle formed by the straight line passing through P and the focus with the vertical, it is easy to deduce that the radius of curvature R_C of the parabola at P is $R_C = 2f/\cos^3(\phi/2)$. As for normal incidence $R_i \to \infty$ and $\theta_o = \phi/2$, we obtain from Eq. (14.6) that the wavelength $\lambda_C = 4\pi f/(\sin(\phi/2)\cos^2(\phi/2))$, which is the minimum so that $\phi = 2\arctan(1/2^{1/2}) = 70.5°$, where $\lambda_C = 32.6\ f$. Thus, the position errors would not be negligible with respect to the slope errors if the error wavelength

were greater than $\approx 30\ f/10 = 3\ f$, with the maximum position error being around $\phi = 70.5°$. Note that the mirror length, for $\phi_M = \pm 90°$, is $L \approx 4.6\ f$.

The extreme case $\lambda \rightarrow \infty$, in which the error is converted into a linear one, is useful for studying tolerance with regard to errors of mirror positioning with respect to the receiver. As we shall see following, the usual hypothesis that the slope error is dominant is not valid in this case.

For wavelength errors λ greater than λ_C, the expression Eq. (14.6) would say the opposite to the usual hypothesis—that is, that the slope error is negligible with respect to the position error. However, for values of $\lambda > 4L$ (for $\phi_M = \pm 90°$, this implies $\lambda > 18\ f$), the calculation carried out with Eq. (14.6) is not correct because there is an edge effect: Since in the mirror there is "no room" for a quarter of a wavelength of the error, the maxima of slope and position used for obtaining Eq. (14.6) do not coexist in the mirror. Let us calculate, for example, the ratio between the slope and position errors when $\lambda \rightarrow \infty$, in the most favorable case for the slope error to dominate (maximum at $s = 0$). At this extreme $E \rightarrow E_o(2\pi L/\lambda)$ and $S \rightarrow E_o(2\pi/\lambda)$, from which we deduce that $S_E/S \rightarrow L\sin(\phi/2)\cos^2(\phi/2)/2f$. This expression is also maximum for $\phi = 70.5°$ so that $S_E/S < 0.24\ L/f$. For example, in the case $\phi_M = \pm 90°$ it is found that $S_E/S < 1.1$—that is, that the position error is not negligible with respect to the slope error even in this case of maximum slope error.

This tool is used for quantifying the maximum position error acceptable for the "general opinion" to be correct. In any case, even though position error may be considerable, its transformation into slope error permits the study of the optical surface as though it had no position errors, at the cost of increasing (apparently) its slope errors. For this reason, from now on we shall consider that only slope errors are present.

Summarizing, a small slope error S produces angular deviation D with respect to the theoretical exiting ray with value $D = mS$, where m is the error multiplier, is given by

$$m = \begin{cases} 2\ (mirror) \\ 1 - \dfrac{\tan(\theta_o)}{\tan(\theta_i)}\ (dioptric) \end{cases} \tag{14.7}$$

where θ_i is the angle of incidence of the ray and θ_o the theoretical angle of refraction, both calculated using the normal to the theoretical surface (see Figure 14.1). Observe that the refractive surface formula can also be used for the mirror simply by making $\theta_o = -\theta_i$. The quotient $\tan(\theta_o)/\tan(\theta_i)$ tends to n_i/n_o when $\theta_i \rightarrow 0$ (n_i and n_o are the refractive indices at the entry and exit sides, respectively).

When an optical surface is manufactured, the slope errors S that it presents at each point may take different values with certain probability. From a mathematical point of view, the slope errors S constitute a stochastic process, which has at a point Q of the surface an ensemble average $E[S(Q)]$ and a standard deviation $\sigma_{S(Q)}$, and whose probability density function (pdf) is $f_{s(Q)}(\theta)$. The errors at different points of the same surface may be correlated, with a self-correlation function $R_S(Q, Q')$.

The deviation angle D is also a stochastic process, a function of the slope error and, in the case of a refractive surface, dependent on the angle of incidence θ_I. Its pdf is calculated as (Rios, 1977)

$$f_{D(Q,\theta_I)}(\theta) = \frac{1}{|m(Q, \theta_I)|} f_{S(Q)}(\theta/m(Q, \theta_I)) \tag{14.8}$$

Therefore, the ensemble average of D is $E[D(Q, \theta_I)] = m(Q, \theta_I) E[S(Q)]$ and its standard deviation $\sigma_{D(Q, \theta_I)} = m(Q, \theta_I)\sigma_{S(Q)}$.

14.3.1   ERROR TRANSFER

Consider an error-free concentrator with acceptance angle $\pm\alpha$. Assume that this concentrator is composed of sequential optical surfaces (refractive surfaces or mirrors) and nonsequential mirrors, which fulfill that no more than two edge rays impinge on each point of a sequential surface and only one on each point of a nonsequential mirror (excepting its edge points). An additional condition is that the number of incidences of the edge rays on the concentrator must be finite (which excludes, for instance, the flow-line concentrator (Winston and Welford, 1979)).

The local half-acceptance angle $\beta(Q)$ of the concentrator bundle at a given point Q of an optical surface is given by half the angle formed by the two edge rays r_{QE} and $r_{QE'}$ impinging at Q in the case of sequential optical surfaces, and, in the case of nonsequential mirrors, by the angle formed by the edge ray r_{QE} impinging at Q and the tangent to the mirror at Q (see Figure 14.3).

Suppose now that said concentrator has contour errors and that there may be correlation of errors between points of one optical surface but that the errors at points of different optical surfaces are independent. These contour errors cause angular deviations of ray trajectories as indicated in Section 14.3. This implies that there may be rays inside the acceptance at Q that are not collected and rays outside it that reach the receiver. In the case of sequential optical surfaces, we shall call neighbors of the edge rays $r_{QE}(r_{QE'})$ those rays exiting from Q and forming an angle with $r_{QE}(r_{QE'})$ smaller than the angle formed with $r_{QE'}$ (r_{QE}). There exists a ray that is a neighbor of r_{QE} and $r_{QE'}$: that which coincides with the ray trajectory that bisects the bundle at Q, which forms an angle $\beta(Q)$ with both edge rays, an angle we shall refer to as $D_M(Q)$. In the case of nonsequential mirrors, the edge-ray neighbors of r_{QE} are those exiting from Q which, like r_{QE}, are not subject to a second reflection on the nonsequential mirror. For this case, let us call $D_M(Q)$ the maximum angle formed by r_{QE} with its neighbors within the acceptance at Q (i.e., with the first ray that impinges on the nonsequential mirror a second time).

For the correct application of the error transfer method explained following, the deviation angle D at Q should be smaller than $D_M(Q)$ (Condition I). If this
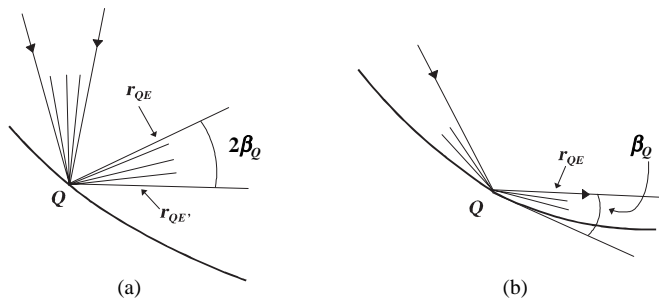
Figure 14.3   Local acceptance angle β_Q at the point Q (a) for a sequential optical surface and (b) for a nonsequential mirror.

condition is fulfilled, the study can be made independently for the edge-ray neighborhoods of $+\alpha$ and $-\alpha$. There is a second limitation to the size of D (Condition II), which will be specified later.

Figure 14.4 shows two optical surfaces, labeled #1 and #2 (refractive surfaces in the figure, i.e., the concentrator is a lens), and a ray trajectory exiting from the point P of surface #2. As the point of incidence of this ray on surface #1 is Q, let us call this ray r_{PQ}. In the example of this figure, r_{PQ} belongs to the bundle of collected rays so that it would reach the receiver if the surfaces were error-free. The figure also shows the trajectory of r_{PQ} after the refraction at Q when no errors are present and the edge rays of the bundle at P and Q. The ray r_{PQ} is a neighbor of the edge rays labeled r_{PE} and r_{QE}.

Let us call $S(QQ)$ the slope error of surface #1 at Q. In accordance with Section 14.3, the slope error $S(QQ)$ that causes the exiting direction of r_{PQ}, after refraction at Q, to be deviated with an angle $D(QQ) = m(Q, \theta_{IQ})S(QQ)$, where θ_{IQ} is the incidence angle of r_{PQ} at Q. The ray r_{PQ} will reach the receiver if its trajectory still coincides with that of a collected ray. The collected directions are characterized by the pillbox transmission function $t(Q, \theta)$, which is one between the edge-ray trajectories at Q and zero outside (θ is the angle with respect to the vertical at Q).

$$t(Q, \theta) = \begin{cases} 1 & \text{if } \theta(r_{QE}) < \theta < \theta(r_{QE'}) \\ 0 & \text{in the rest} \end{cases} \tag{14.9}$$

The direction in which the ray PQ is refracted is a random variable, a function of the deviation angle, which is given by $\theta_D = \theta(r_{QE}) + D(QQ)$. Thus, the transmission at Q is a random variable (discrete, which takes a value 0 or 1), due to its dependence on $D(QQ)$ on substituting the previous expression in Eq. (14.9). The ensemble average of the transmission of r_{PQ}—which coincides with the probability $P[t(Q, \theta_D) = 1]$—is given by the convolution (Rios, 1977)
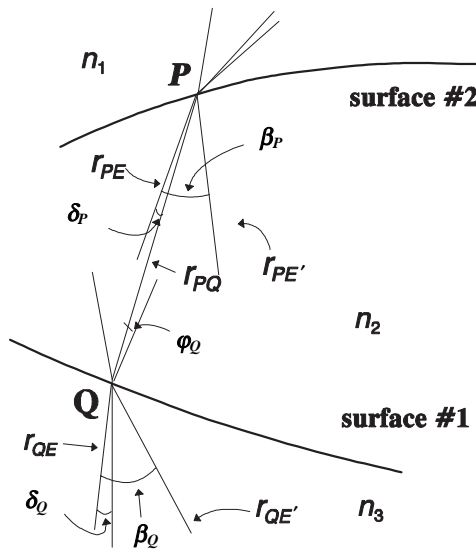
Figure 14.4   Error transfer of the slope error at Q to the point P.
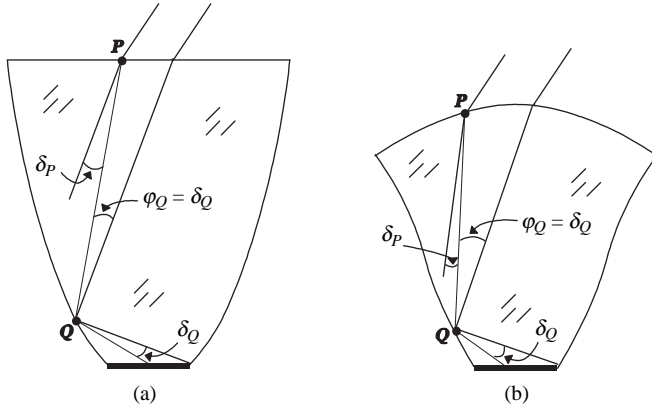
Figure 14.5 Example of error transfer in (a) a dielectric-filled CPC, and (b) a DTIRC.

$$E[t(Q, \theta_D)] = t(Q, \theta(r_{PQ})) \otimes f_{D(QQ)}(\theta(r_{PQ})) = \int_{-\infty}^{\infty} t(Q, \theta_D) f_{D(QQ)}(\theta_D - \theta(r_{PQ})) d\theta_D$$

$$= \int_{-\infty}^{\delta_Q} f_{D(QQ)}(D) dD \qquad (14.10)$$

where $\delta_Q = \theta(r_{QE}) - \theta(r_{PQ})$, as indicated in Figure 14.4. The final integral indicates that, as r_{PQ} is a neighbor of r_{QE} and Condition I is fulfilled, the convolution can be made with a steplike transmission function $u(\theta, r_{QE})$, which has no step at $r_{QE'}$. In the knowledge that, as shown in Figure 14.4, $\delta_P = \theta'(r_{PE}) - \theta'(r_{PQ})$, θ' being the angle with respect to the vertical at P, with the help of the linear change of variables $D' = D(\delta_P/\delta_Q)$ in Eq. (14.10), we obtain

$$E[t(Q, \theta_D)] = \int_{-\infty}^{\delta_Q} \left\{ \frac{\delta_Q}{\delta_P} f_{D(QQ)}\left(\frac{\delta_Q}{\delta_P} D' \right) \right\} dD' = \int_{-\infty}^{\delta_Q} f_{D(PQ)}(D') dD' \qquad (14.11)$$

the new random variable being (Rios, 1977)

$$D(PQ) = \frac{\delta_P}{\delta_Q} D(QQ) \qquad (14.12)$$

This means that the deviation angle at Q can be studied as an apparent deviation angle at P, given by Eq. (14.12). The preceding change of variables is well defined if $\delta_Q \to 0$ when $\delta_P \to 0$ and Condition I is fulfilled for $D(QQ)$ at Q (if Condition I is not fulfilled, the change of variables is still valid if $\delta_P/\delta_Q \approx \delta_P'/\delta_Q'$, where δ_P' and δ_Q' are the angles formed by r_{PQ} with $r_{EP'}$ and $r_{EQ'}$, respectively).

Figure 14.5 shows two well-known concentrators (Welford and Winston, 1989), a dielectric-filled CPC and a DTIRC, with which we shall illustrate the error transfer process. Assume that both concentrators have slope errors $S(QQ)$ at their respective points Q, and let us transfer the errors to the points P. It is easy to see that for the CPC $\delta_Q = \delta_P$ and therefore, in accordance with Eq. (14.12), the apparent deviation angle at PQ due to the error at Q is $D(PQ) = D(QQ) = 2S(QQ)$. In the case of point Q of the DTIRC, the converging refractive surface causes $\delta_Q > \delta_P$,

so the apparent deviation is $D(PQ) < 2S(QQ)$. Thus, the slope error at Q "seen" from P is smaller in the DTIRC than in the CPC.

The next step consists of combining the deviation angle $D(PP)$ produced by the slope error at P, $S(PP)$, with the apparent deviation angle $D(PQ)$ produced by $S(QQ)$. The total angular deviation at P is

$$D(P) = D(PP) + D(PQ) \tag{14.13}$$

However, the calculation of the pdf of $D(P)$ from Eq. (14.13) is not direct because $D(PP)$ and $D(PQ)$ are not, in principle, independent random variables (even with $S(PP)$ and $S(QQ)$ being independent). The reason is that, for a given point P, the deviation angle $D(PQ)$ differs, in general, for each ray r_{PQ} because it depends on point Q through δ_P/δ_Q and $D(QQ)$ in Eq. (14.12). Mathematically, $D(PQ)$ is said to be a random variable conditional on the ray r_{PQ}, which can be expressed as $D(PQ/r_{PQ})$. This implies that the final integral in Eq. (14.11) is not in fact a convolution integral between the pdf of $D(PQ/r_{PQ})$ and the steplike function $u(\theta, r_{PE})$, because this pdf varies with r_{PQ}.

In order to make $D(PP)$ and $D(PQ)$ independent, or what amounts to the same, to make the integral at P a convolution integral, we must impose a second condition on the deviation angles. At the limit $\delta_P \to 0$, the ray r_{PQ} tends to the edge ray r_{PE}, which has an apparent deviation angle $D(PQ/r_{PE})$. Let us consider the direction r_{PQ} obtained by deviating the ray r_{PE} with the angle $D(PQ/r_{PE})$. Condition II on $D(PQ/r_{PE})$ states the following: it has to be small enough to make $D(PQ/r_{PQ}) \approx D(PQ/r_{PE})$. If this condition is fulfilled, the integral Eq. (14.11) represents the convolution between the pdf of $D(PQ/r_{PE})$ and the steplike function $u(\theta, r_{PE})$.

Thus, provided that Condition II is fulfilled, the total deviation angle at P for all the neighbor rays of r_{PE} is calculated as (for the neighbor rays of $r_{PE'}$ an analogous expression holds)

$$D(P/r_{PE}) = D(PP) + D(PQ/r_{PE}) \tag{14.14}$$

where the random variables $D(PP)$ and $D(PQ/r_{PE})$ are independent (assuming independence between $S(PP)$ and $S(QQ)$). The pdf of total deviation angle for the neighbor rays of r_{PE} is also given by the convolution (Rios, 1977)

$$f_{D(P/r_{PE})}(\theta) = f_{D(PP)}(\theta) \otimes f_{D(PQ/r_{PE})}(\theta) \tag{14.15}$$

The process of transfer of the slope error at Q to the apparent angular deviation at P does not require $D(PQ/r_{PE})$ to fulfill Condition I. However, fulfillment of the condition may lend the transfer a more physical sense (unnecessary from the mathematical point of view), since the slope error at Q will be equivalent to a slope error at P for the ray r_{PE}, with a value

$$S(PQ/r_{PE}) = \frac{\delta_P}{\delta_Q} \frac{m(Q, \theta_{IQ})}{m(P, \theta_{IP})} S(QQ) \tag{14.16}$$

where δ_P and δ_Q must be small enough to fulfill Conditions I and II. The total error is $S(P/r_{PE}) = S(PP) + S(PQ/r_{PE})$. In the case of sequential optical surfaces, the ratio δ_P/δ_Q can usually be approximated by β_P/β_Q. Observe that if $m(Q)/\beta_Q \ll m(P)/\beta_P$, the errors at Q are strongly attenuated when transferred to P.

The calculations of the transfer process can also be made with the angle ϕ_Q (shown in Figure 14.4) instead of the angle δ_Q. These two angles are related by the refraction law $n_2 \sin(\theta_{IQ} + \phi_Q) = n_3 \sin(\theta_{RQ} + \delta_Q)$, which implies $\delta_Q \approx \phi_Q \tan \theta_{RQ}/\tan \theta_{IQ}$.

The ratio δ_P/ϕ_Q, if calculated at the limit $\delta_P \to 0$, coincides with QC/PC, where C is the point of the caustic of the edge rays at r_{PE} (as is proven in Benítez, 1998).

All of the preceding formulas given for the two-refractive surface case can be used if one or both surfaces are mirrors, making only $\theta_R = -\theta_I$.

The process of error transfer described can be applied to multiple surface concentrators, going from the receiver to the entry aperture along the edge rays that emerge from the edges of the source (each one separately), accumulating the errors of all the optical surfaces in between. The accumulated errors should fulfill Conditions I and II so that precision is not lost. An edge-ray tracing can be used to calculate the required parameters. The final formulas for the complete transfer process for each point x of the entry aperture, to be applied separately to each edge ray that passes through x, are

$$f_{D_{eff}(x)}(\theta) = \overset{N}{\underset{k=1}{\otimes}} f_{D_{k,eff}(x)}(\theta) \quad , \quad D_{k,eff}(x) = D_k \frac{\delta_{N+1}}{\delta_k} \tag{14.17}$$

where the indices k indicate the intersection of the edge ray that passes through the point x of the aperture with the surface k ($k = 1, 2, \ldots, N$, numbered from the receiver to the aperture). The surface $N + 1$ is the concentrator entry aperture. The number N may be different for each of the two edge rays that pass through x, as, for example, in the case of the SMTS concentrator presented in Chapter 13.

The ensemble average and standard deviation of the deviation angle at the point x for each edge ray is thus given by

$$E[D_{eff}(x)] = \sum_{k=1}^{N} E[D_{k,eff}(x)] = \sum_{k=1}^{N} E[D_k] \frac{\delta_{N+1}}{\delta_k}$$

$$\sigma_{D_{eff}(x)} = \sqrt{\sum_{k=1}^{N} \sigma_{D_{k,eff}(x)}^2} = \sqrt{\sum_{k=1}^{N} \sigma_{D_k}^2 \left(\frac{\delta_{N+1}}{\delta_k}\right)^2} \tag{14.18}$$

14.3.2   Angular Transmission

In Section 14.3.1 we obtained, under certain hypotheses, the two pdfs of the effective deviation angles for each point x of the entry aperture, associated with each edge ray that passes through x. Let us suppose that the concentrator has been designed for an infinite source with angular size $\pm\alpha$. Let $f_{Deff(x,+)}$ and $f_{Deff(x,-)}$ be the pdfs associated with the rays impinging on x with angles $+\alpha$ and $-\alpha$, respectively. The next step should be to obtain a global pdf for the whole concentrator. The ensemble average of the angular transmission $T(\theta)$ of the concentrator when errors are present is calculated with the integrals

$$E[T(\theta)] = \frac{1}{L} \int_{x=0}^{L} f_{D_{eff}(x,+)}(\theta) \otimes t_o(\theta, x)dx \quad , \quad \theta > 0$$

$$E[T(\theta)] = \frac{1}{L} \int_{x=0}^{L} f_{D_{eff}(x,-)}(\theta) \otimes t_o(\theta, x)dx \quad , \quad \theta < 0 \tag{14.19}$$

where $t_o(\theta, x)$ is the pulselike transmission function of the error-free concentrator at x (which is equal to one between the edge rays and zero outside). The integral can be called $t(\theta, x)$, the transmission function of x without errors.

If Condition (I) is fulfilled, the ensemble average of $T(\theta)$ does not depend on the correlation between points of the same optical surface.[1] On the other hand, the standard deviation of the values of $T(\theta)$ (which provides information about the best and worst concentrators) does depend on such a correlation, although it is not calculated in this study.

If the concentrator and the error parameters are symmetrical with respect to $x = L/2$, the resulting expected transmission will be symmetrical with respect to $\theta = 0$. Thus, only one of the integrals Eq. (14.19) need be studied.

The calculations in Eq. (14.19) can be simplified if the angular transmission of the error-free concentrator $T_o(\theta)$ is of pillbox type (i.e., one between $\pm\alpha$ and zero outside), since the function $t_o(\theta, x)$ will be independent of x. Thus, it can be removed from the integrals Eq. (14.19), leading to the global pdfs of the deviation angle

$$f_{\overline{D}_{(+)}}(\theta) = \frac{1}{L} \int_{x=0}^{L} f_{D_{eff}(x,+)}(\theta)dx \quad \text{,,} \quad \theta > 0$$

$$f_{\overline{D}_{(-)}}(\theta) = \frac{1}{L} \int_{x=0}^{L} f_{D_{eff}(x,-)}(\theta)dx \quad \text{,,} \quad \theta < 0$$

(14.20)

If, on the other hand, $t_o(\theta, x)$ depends on x (so that $T_o(\theta)$ is not stepped) but $f_{D_{eff}(x,\pm)}$ does not, then $f_{D_{eff}(x,\pm)}$ is the global pdf and can be removed from the integrals Eq. (14.19). In both cases of independence of x, the expected angular transmission is calculated as

$$E[T_E(\theta)] = f_{\overline{D}_{(+)}}(\theta) \otimes T_o(\theta) \quad \text{,,} \quad \theta > 0$$

$$E[T_E(\theta)] = f_{\overline{D}_{(-)}}(\theta) \otimes T_o(\theta) \quad \text{,,} \quad \theta < 0$$

(14.21)

When both $t_o(\theta, x)$ and $f_{D_{eff}(x,\pm)}$ depend on x, the exact formula for calculating the ensemble average of the transmission with errors is Eq. (14.19), and the process given by Eqs. (14.20) and (14.21) is an approximation. Although strictly speaking no global deviation angle pdf exists in this case, Eq. (14.20) can be considered as an estimation.

Note that, since $T_o(\theta)$ and $f_{D(\pm)}$ are the respective averages of $t_o(\theta, x)$ and $f_{D_{eff}(x,\pm)}$ along the aperture, the process Eqs. (14.20) and (14.21) consists in calculating these averages separately and then convolving them, while in the exact procedure Eq. (14.19) the convolution is carried out at each point and the mean of the result calculated. If the functions $t_o(\theta, x)$ and $f_{D_{eff}(x,\pm)}$ vary greatly with x, and do so in a similar (i.e., parallel) way, the approximate procedure will be pessimistic, since in the exact calculation there is a compensation of errors and acceptance that is not present in the approximation: where the error is large (small) the acceptance follows suit. However, if the functions $t_o(\theta, x)$ and $f_{D_{eff}(x,\pm)}$ vary greatly with x, but in opposite ways, the approximate procedure will be optimistic, since the exact transmission is situated in the worst case: Where there is most acceptance there is also most error.

---

[1]   In nonsequential mirrors with a strong correlation of errors between their points this affirmation may not hold, as it is possible that rays subjected to two reflections (which we do not consider as "neighbors" of the edge rays) are not collected if these reflections are strongly correlated.

Observe that, in general, the integration Eq. (14.20) does not conserve the pdf shape. For example, if $\sigma_{Deff(x,\pm)}$ depends on x, the resulting global pdf is not Gaussian, although $f_{Deff(x,\pm)}$ is Gaussian for all x.

If the ensemble average of the slope errors is zero, it is easy to prove from Eq. (14.20) that the standard deviation of the global deviation angle is

$$\sigma_{\bar{D}_{(\pm)}} = \sqrt{\frac{1}{L} \int_{x=0}^{L} \sigma^2_{\bar{D}_{eff}(x,\pm)} dx} \tag{14.22}$$

Moreover, if the concentrator and the error parameters are symmetrical with respect to $x = L/2$, then $\sigma_{\bar{D}} \equiv \sigma_{\bar{D}(+)} = \sigma_{\bar{D}(-)}$, which can be calculated as

$$\sigma_{\bar{D}} = \sqrt{\frac{2}{L} \int_{x=0}^{L/2} \frac{\sigma^2_{\bar{D}_{eff}(x,+)} + \sigma^2_{\bar{D}_{eff}(x,+)}}{2} dx} \tag{14.23}$$

It is possible to give a breakdown of the contributions of each surface to the global standard deviations Eq. (14.22). The global deviation angle associated with surface k (used in Eq. (14.17)) has a standard deviation

$$\sigma_{\bar{D}_{k(\pm)}} = \sqrt{\frac{1}{L} \int_{x=0}^{L} \sigma^2_{\bar{D}_{k,eff}(x,\pm)} dx} \tag{14.24}$$

and fulfills

$$\sigma_{\bar{D}_{(\pm)}} = \sqrt{\sum_{k=1}^{N(\pm)} \sigma^2_{\bar{D}_{k(\pm)}}} \tag{14.25}$$

14.3.3   Acceptance Angle

Once $E[T(\theta)]$ has been computed, it is possible to calculate the ensemble average of the acceptance angle with errors, $E(\alpha)$, defined, for example, at 95% of the normal incidence transmission in $E[T(\theta)]$. This value may be different for $\theta > 0$ and $\theta < 0$ in an asymmetric concentrator (or one with asymmetric errors).

However, it is possible to estimate $E(\alpha)$ without calculating $E[T(\theta)]$ when the errors have zero mean. Consider a symmetric concentrator and let us assume the following simplifying approximations: (1) the global deviation angle pdf is Gaussian, and (2) the acceptance reduction due to errors for a nonstepped error-free angular transmission is equal to that for a stepped one. In this situation, if the error-free concentrator has an acceptance angle α_o, $E(\alpha)$ is the number verifying

$$0.95 = \int_{-\infty}^{\alpha} \frac{1}{\sigma_{\bar{D}} \sqrt{2\pi}} e^{\frac{(x-E(\alpha))^2}{2\sigma^2_{\bar{D}}}} \tag{14.26}$$

This leads to the approximate value (Rios, 1977)

$$E(\alpha) \approx \alpha - 1.65\sigma_{\bar{D}} \tag{14.27}$$

If the acceptance angle with errors is defined at 90% of the normal incidence transmission, the coefficient 1.65 in Eq. (14.27) must be replaced by 2.

14.3.4   Validation of the Model

The errors model just presented is strictly valid when its hypotheses are fulfilled. It is not unusual that, for instance, Condition II fails at the concentrator edges. However, as the failure of the model is limited to small portions of the aperture, precision may still be high.

There are concentrators, such as the parabolic mirror-secondary CPC combination, where the condition $\delta_Q \to 0$ when $\delta_P \to 0$ is not fulfilled at any point of its aperture. In this case, the model as presented is not applicable unless the secondary errors are negligible.

Let us check the validity of the model for the case of the XR concentrator (Miñano and González, 1992) shown in Figure 14.6 by means of a 2D ray-tracing. The design parameters of this XR are (1) 2D-geometrical concentration $C_g = 34\times$, (2) acceptance angle $\alpha = \pm 2.48°$, and (3) secondary refractive index $n = 1.5$.

For this example, suppose that the slope errors of each surface have equal standard deviations at all their points. Consider also that the slope errors are 10 times greater in the refractive surface than in the mirror—that is, σ_{S2} (mirror) = 0.1 σ_{S1} (refractive surface). Expression Eq. (14.17) was used to transfer the errors to the entry aperture and calculate the standard deviation of effective deviation angle for each surface and for the whole concentrator at every point of the aperture (i.e., standard deviations of $D_{k,eff}(x)$ and $D_{eff}(x)$). This calculation was made for only one edge ray, given the symmetry, and the approximation $\delta_P/\delta_Q \approx \beta_P/\beta_Q$ was used. Observe that, as $\beta_P \ll \beta_Q$ (and $m(Q) < m(P)$ also), the tolerance to errors of the refractive surface is far greater than that of the mirror. Figure 14.7a shows the result of this process, normalized to the value σ_{S2}.

The effective deviation angle of the refractive surface is smaller than that of the mirror up to $x \approx 0.8$ $L/2$ ($x = 0$ corresponds to the symmetry axis). From this point to the aperture rim, the effective deviation of the refractive surface increases rapidly due to the decrease of the local acceptance β_Q in the external portion of the refractive surface.

The standard deviation of the global deviation angle, calculated with Eq. (14.24), is similar for the two surfaces. The mirror has $2\sigma_{S2}$ and the refractive surface $1.58\sigma_{S2}$ ($2.55\sigma_{S2}$ total) in accordance with Eq. (14.25), although the refractive surface errors are 10 times higher.
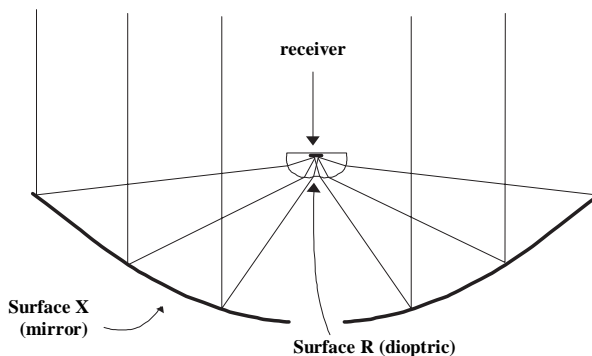
Figure 14.6   XR concentrator with a 2D-geometrical concentration 34× and acceptance angle ±2.48°.
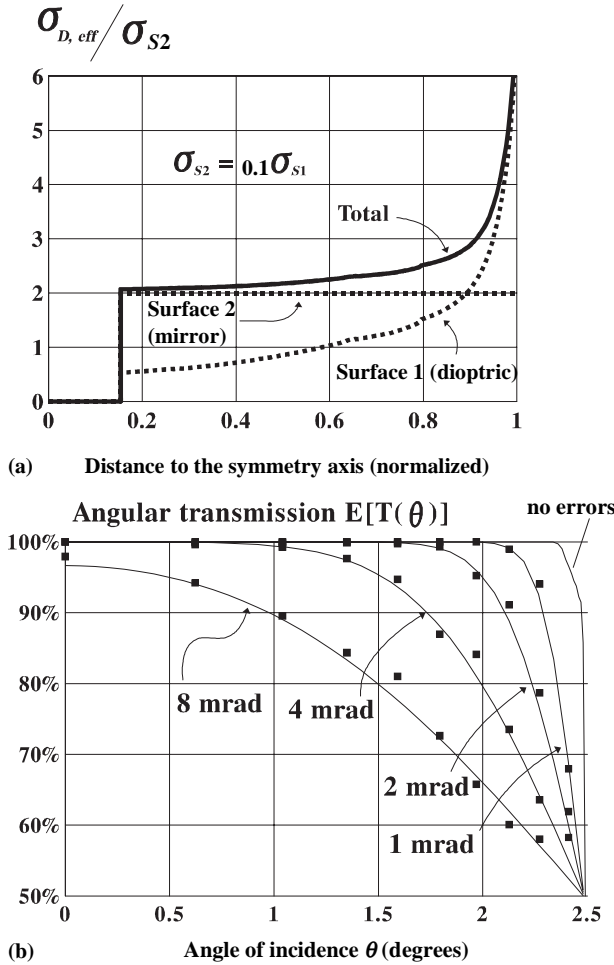
Figure 14.7   (a) Standard deviation of the effective deviation angles for each surface of the XR and the complete concentrator as a function of the distance to symmetry axis. (b) Ensemble average of the angular transmission of the XR predicted by the model (full curves) and by ray-tracing calculations (broken curves). The labels indicate the different values of the mirror slope error, which are 10% of those of the refractive surface in all cases ($\sigma_{S2} = 0.1$ σ_{S1}).

Assuming slope errors with Gaussian distributions and zero mean values, we calculated with the model the ensemble average of the angular transmission $E[T(\theta)]$ for four different values of σ_{S2}. The result of the model's predictions is shown in Figure 14.7b with full curves. This figure also shows as points the values obtained with a 2D-ray-tracing program, which fit well with the model even for slope errors as high as $\sigma_{S2} = 8$ mrads. Observe that, as the ray-tracing program generates random numbers for surface slope for each incidence, the slope errors within each surface are totally uncorrelated (i.e., S is "white noise"). It is easily seen that in this situation the standard deviation of the angular transmission with error $\sigma_{T(\theta)}$ is null, which implies that the ray-tracing generates (if a sufficiently

high number of rays are traced) the expected value of the transmission $E[T(\theta)]$, which is independent of such correlation.

14.3.5 Example of Application: Predictions for a Specific Concentrator of a Manufactured Series

The error model presented can predict the behavior of the average concentrator of a series, if the stochastic processes of the slope errors S of the optical surfaces of the manufactured concentrator are known. This data can be obtained through the measurement of these errors at each point of the surfaces of a sufficiently large sample of the series.

Estimation of the behavior of a specific concentrator of the series can also be made by applying the model, considering that the stochastic processes S_k are deterministic and that their values coincide at each point with the measured value of that concentrator, $S_{c,k}(Q)$. This means that the pdf of S_k is given by

$$f_{S_K(Q)}(\theta) = \delta(\theta - S_{c,k}(Q)) \tag{14.28}$$

where $\delta(\cdot)$ represents Dirac's delta function. Thus, $E[S_k(Q)] = S_{c,k}(Q)$ and $\sigma_{S,k}(Q) = 0$ for all points Q. From Eqs. (14.16) to (14.18) it is easily demonstrated that the apparent deviation angle at all entry points, $D_{k,eff}$, caused by the error S_k, is also deterministic: It has a mean value $D_{c,k}$ and null standard deviation. From the condition $\sigma_{Dk,eff}(x) = 0$ and Eq. (14.20) it is deduced that

$$E[\overline{D}_k] = \frac{1}{L}\int_0^L D_{c,k}(x)dx \equiv \langle D_{c,k}\rangle \quad \sigma_{\overline{D}_k}^2 = \frac{1}{L}\int_0^L D_{c,k}^2(x)dx - \langle D_{c,k}\rangle^2 \equiv \sigma_{Dck}^2 \tag{14.29}$$

$\sigma_{Dc,k}$ being the standard deviation of the function $D_{c,k}$ for the variable x. This result can also be obtained from another more general one: when $D_{k,eff}(x)$ is deterministic, the total/global deviation angle associated with the surface k coincides, as a random variable, with that which is obtained on applying the function $D_{k,eff}(x)$ to the random variable X, uniform in the interval $[0, L]$. The pdf of the global deviation angle can therefore be calculated by applying the fundamental theorem of functions of random variables (Rios, 1977).

The combination of the errors associated with the different surfaces can be estimated by applying Eq. (14.25), which will be more precise the less the correlation between the errors of the different surfaces (i.e., the smaller their correlation coefficient, Rios, 1977).

14.3.6 Manufacturing Errors in Three-Dimensional Concentrators

The model proposed is 2D, so its predictions and analysis may not be precise in practice, given that real concentrators are three-dimensional. In the case of linear concentrators, the model takes into account only the rays perpendicular to the translation direction, whereas in that of rotational concentrators only meridian rays are considered.

For rotational concentrators it is possible to improve the estimation given by the model by taking into account that in rotation symmetry the points of the profiles in 2D farthest away from the axis give rise to "more" points on the 3D

surfaces (proportional to r, distance from the axis). This can be achieved by calculating the integrals of the model that calculate the averages at the aperture $(0 < x < L)$, extended to the 3D aperture (in $0 < r < R$, changing dx for $2\pi r\, dr$).

Given that the processes of mechanization of the surfaces produce mainly contour errors with rotation symmetry, the ensemble average of the angular transmission should be calculated using, instead of Eq. (14.21), the integral

$$E[T(p, q)] \approx \frac{1}{2\pi} \int_{t=-\infty}^{t=+\infty} f_{\bar{D}}(t) \int_{\phi=0}^{\phi=2\pi} T(p.1\cos\phi, q - t\,\text{sen}\,\phi)\,d\phi\,dt \qquad (14.30)$$

14.4   THE CONCENTRATOR ERROR MULTIPLIER

Consider the case in which the slope errors of the points of all surfaces of a concentrator have zero mean value and the same standard deviation σ_S. If the concentrator is symmetrical, the concentrator error multiplier M is defined as the ratio

$$M = \frac{\sigma_D}{\sigma_S} \qquad (14.31)$$

where σ_D is the standard deviation of the concentrator global pdf calculated in Eq. (14.22). If the concentrator is not symmetrical, two M values should, in general, be defined: one for $\theta > 0$ and another for $\theta < 0$.

The error multiplier M is dependent only on concentrator optics and geometry, and in accordance with Eqs. (14.8), (14.12), (14.18), and (14.22), M is given by

$$M = \sqrt{\frac{1}{L} \int_{x=0}^{L} \sum_{k=1}^{N} m_k^2 \left(\frac{\delta_{N+1}}{\delta_k} \right)^2} \qquad (14.32)$$

As in the case of sequential optical surfaces, usually $\delta_{N+1}/\delta_k \approx \alpha/\beta_k$, it is deduced from Eq. (14.32) that if $m_k/\beta_k \ll m_j/\beta_j$ for $0 \le x \le L$, the errors of surface k do not contribute to the concentrator sensitivity (since they are negligible with respect to the errors of surface j). Therefore, the tolerance to errors of said surface k is far greater than that of surface j. This effect occurs, for instance, in the XR concentrator, as seen in Section 14.3.4.

The concentrator error multiplier, although a single number, contains much information about the concentrator errors. Figure 14.8 shows the M values for several selected concentrator designs. Both the parabola and the CPC have $M = 2$. The calculation with Eq. (14.32) in these cases is direct: $N = 1$, and for all $x\, \delta_2 = \delta_1$ and $m_1 = 2$, so that $M = 2$.

The dielectric-filled CPC with moderate acceptance ($\tan \alpha \approx \alpha$) has $M = 3.04$, which is approximately $n = 1.5$ times that of the empty CPC. This factor of n coincides with the reduction of the acceptance angle of the bundle after refraction on the flat refractive surface. The remaining 0.04 is due to the refractive surface errors. The calculation of M is also easy: $N = 2$, and for all $x\ \delta_2 = \delta_1$, $\delta_3 \approx n\ \delta_2$, $m_1 = 2$, $m_2 \approx 1 - 1/n$. Therefore, $M = ((2n)^2 + (1 - 1/n)\ n)^2)^{1/2} = 3.04$.

When a secondary is added to the parabola (or in the case of the XR), the M factor is not increased because the secondary errors (if working with far greater acceptances than the primary), when transferred to the primary, are negligible

Concentrator error multiplier
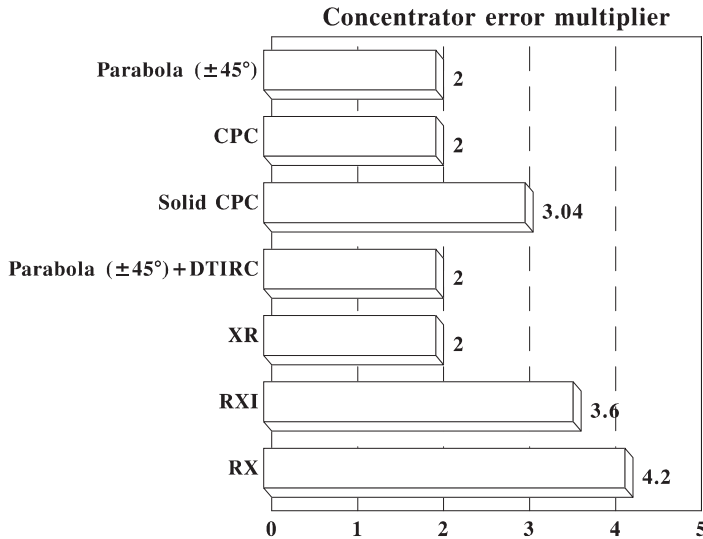
Figure 14.8   Concentrator error multipliers of several selected concentrators.

(this was already known for the CPC secondaries (Rabl, 1976) but is found to be general).

The selected RXI and RX concentrators have $M = 3.6$ and $M = 4.2$, respectively. It seems surprising that, although the collected rays undergo one more reflection in the RXI than in the RX, the M value of the former is lower. This effect, which occurs because the local acceptances in their respective mirrors are markedly different, is another clear exemplification of the fact that not only the optics but also the geometry of the concentrator is important with regard to errors.

14.5   SENSITIVITY TO ERRORS

The error multiplier M does not alone define the error sensitivity of a concentrator, since the acceptance angle α_o of the error-free concentrator also plays a role. For instance, in the example from the previous section, the CPC increases its M value by a factor $n = 1.5$ on being filled with dielectric media. However, its acceptance α_o also increases by the same factor so that the resulting acceptance with errors $E(\alpha)$ can be expected to improve proportionally to n, even though M has increased.

The parameter to compare must be this resulting acceptance angle $E(\alpha)$, which depends on the theoretical acceptance angle α_o and on M. It can be easily estimated from Eqs. (14.27) and (14.31) as

$$E(\alpha) \approx \alpha - 1.65 M \sigma \tag{14.33}$$

where all concentrator surfaces are assumed to have slope errors with standard deviation σ.

As an example of the practical interest of the parameter M, let us calculate the maximum slope error σ_{max} that can be tolerated for a given application. If the minimum acceptance angle required is α_{min}, according to Eq. (14.33)

maximum slope error (mrad)

Parabola (±45°)	26.4
CPC	56.7
Solid CPC	57.2
Parabola (±45°)+DTIRC	68
XR	86.9
RXI	47.6
RX	40.9

0   20   40   60   80   100

(a)

maximum slope error (mrad)

Parabola (±45°)	
CPC	2.2
Solid CPC	3.4
Parabola (±45°)+DTIRC	3.4
XR	4.9
RXI	2.6
RX	2.3

0   1   2   3   4   5   6

(b)
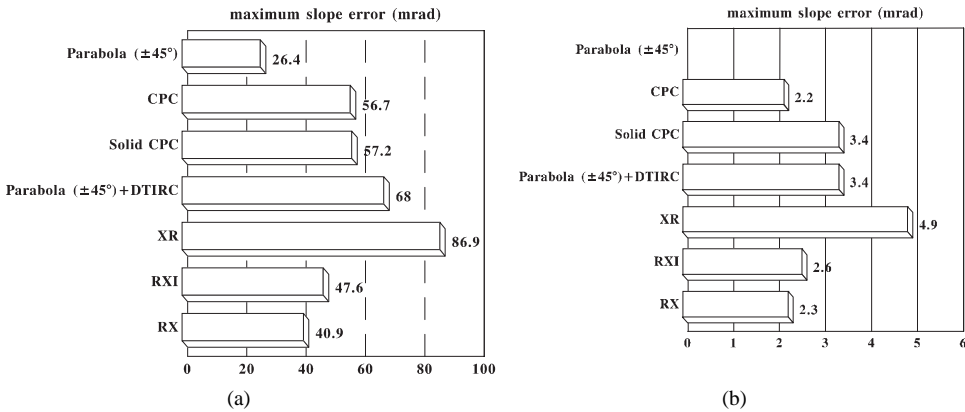
Figure 14.9   Specification of the maximum slope errors that can be allowed to obtain an acceptance angle of ±0.75 degrees for several selected concentrators. The 2D-geometrical concentration is (a) $Cg = 5x$, and (b) $50x$.

$$\sigma_{\max} \approx \frac{\alpha - \alpha_{\min}}{1.65M} \tag{14.34}$$

provided that $\alpha_o > \alpha_{\min}$. If $\alpha_o < \alpha_{\min}$, the concentrator could not be used in this application, since it would not reach the minimum acceptance even without errors.

Figure 14.9 shows the specification of maximum errors admissible for several selected concentrators for two different geometrical concentrations if the resulting acceptance must be at least $\alpha_{\min} = \pm0.75°$. The values of $\sigma_{\max}$ depend on C_g through α_o. The M values calculated in the previous section have been considered valid for both cases (M does not vary much with concentration).

The parabolic mirror with concentration $C_g = 5\times$ can be used for this application if manufactured with slope errors below ≈22 mrads. However, the parabola cannot be used for $C_g = 50\times$ because the acceptance without errors is already inferior to ±0.75 degrees. Observe that a concentrator may be more sensitive to errors than another one for a given concentration but less sensitive for a different concentration value. This is the case of the RXI and the CPC in the preceding example.

14.6   CONCLUSIONS

This chapter has presented an error model for real concentrators that allows, under certain hypotheses, the prediction of the (average) behavior of concentrators after manufacture if the tolerances of the technology employed are known. This prediction can be made in terms, for example, of the angular transmission curve. From this model, moreover, design rules can be inferred in order to avoid solutions sensitive to errors or with critical portions of optical surfaces.

The key to the model is error transfer between surfaces, a process that consists in transforming the system composed of two surfaces with errors into an equivalent system in which errors are transferred from one surface to the other, eliminating the errors of one of them and apparently increasing the errors of the other. Through the transfer process the errors of all the concentrator surfaces are

transferred to the surface of the entry aperture or even to the source (leaving the concentrator error-free).

A noteworthy result of this model is that the behavior of each concentrator is characterized by a single number: its error multiplier M. In consequence, it is possible to classify concentrators on the basis of M in terms of errors in order to obtain a degree of intrinsic sensitivity to manufacturing errors for each design.

Moreover, the number M has practical applications that are both simple and very useful. For example, multiplied by the standard deviation of the slope error of the optical surfaces of the concentrator, it gives directly the standard deviation that must be added (quadratically) to the angular width of the source for transferring to it the errors of the concentrator. It is also possible to estimate from M the maximum slope errors admissible for a given application, simply by carrying out a subtraction and a division, as shown in equation Eq. (14.34).

REFERENCES

Benitez, P. (1998). PhD. UPM, Madrid.

Butler, B. L., and Pettit, R. B. (1977). Optical evaluation techniques for reflecting solar concentrators. *Opt. Apl. To Sol. Ener. Conv*, SPIE, Vol. **114**, 43–49.

Lof, G. O G., and Duffie, J. A. (1963). Optimization of focusing solar collector design. *J. Eng. Power*, 221–228.

Miñano, J. C., and González, J. C. (1992). New method of design of nonimaging concentrators. *Appl. Opt*. **31**, 3051–3060.

Rabl. A. (1976). Comparison of solar concentrators. *Solar Energy* **18**, 93–111.

Rabl. A. (1985). *Active Solar Collectors and Their Applications*. Oxford University, Oxford.

Rios, S. (1977). *Métodos estadisticos*. Ediciones del Castillo, Madrid.

Welford, W. T., and Winston, R. (1989). *High Collection Nonimaging Optics*. Academic Press, New York.

Winston, R., and Welford, W. T. (1979). Geometrical vector flux and some new nonimaging concentrators. *J. Opt. Soc. Am. A.* **69**, 532–536.

Winter, C.-J., Sizmann, R. L., and Vant-Hull, L. L. (1990). *Solar Power Plants.* Springer Verlag, Berlin.

Appendix A

DERIVATION AND EXPLANATION OF THE ÉTENDUE INVARIANT, INCLUDING THE DYNAMICAL ANALOGY; DERIVATION OF THE SKEW INVARIANT

A.1 THE GENERALIZED ÉTENDUE

In Section 2.7 we introduced the invariant

$$n^2 dx\, dy\, dL\, dM \tag{A.1}$$

The meaning of this was as follows: Let any ray be traced through an optical system, and let rectangular coordinate axes be set up in the entry and exit spaces in arbitrary orientations. Also, let the ray meet the x, y plane in the entry space in (x, y), and let its direction cosines be (L, M, N), and similarly for the exit space. Then for any nearby ray with coordinates $(x + dx, y + dy, L + dL, M + dM)$ we have

$$n'^2 dx'\, dy'\, dL'\, dM' = n^2 dx\, dy\, dL\, dM \tag{A.2}$$

We shall prove in Section A.2 that this result is true for any optical system and for any choice of the directions of the axes. But first we discuss some different ways of interpreting this result. We put $p = nL$ and $q = nM$, and we treat (x, y, p, q) as coordinates in a four-dimensional space. Let U be any enclosed volume in this space. Then U is given simply by

$$\int dU = \iiiint dx\, dy\, dp\, dq \tag{A.3}$$

taken over the volume. The result of Eq. (A.2) means that if we let a ray sweep out the boundary of this volume in the four-dimensional entry space, it will sweep

out an equal volume in four-dimensional exit space or, clearly, in any intermediate space of the optical system. There is a useful analogy between this space and the multidimensional phase space of Hamiltonian mechanics, so we shall call it simply phase space. Thus, our result is that if a ray sweeps out a certain volume of phase space in the entry region, it sweeps out the same total volume in the exit region, or, more concisely, phase space volume is conserved. Of course, the shape of the volume will change from entry to exit space or even if the origin of coordinates is shifted parallel to the z direction. However, it can easily be shown that rotating the axes about their origin does not change the phase space volume, so it is a physical significant invariant.

Another picture that is particularly valuable for the purposes of this book may be obtained by supposing that an area in the x, y plane (in real space, not phase space) is covered with uniformly closely spaced points and that through each point a large number of rays is drawn. These rays are to be in different directions so that their direction cosines increase by uniform small increments over certain ranges of L and M. The rays so drawn then represent "ray points" uniformly spaced throughout a certain volume of phase space, and our theorem can then be stated in the following form: The density of rays in phase space is invariant through the optical system. If the four-volume U of Eq. (A.3) has been defined by means of physical components in the optical system—that is, an aperture stop and a field stop—it is known as the throughput or étendue of the system.

We can see from the ray-point representation that the étendue is a measure of the power that can be transmitted through the optical system from a uniformly bright source of sufficient extent, in the geometrical optics approximation. Here, of course, in addition to neglecting interference and diffraction effects, we are assuming no losses in the system due to reflections, absorption, or scattering. The concept of étendue is applied in comparing the properties of many different kinds of optical instruments, although usually only a simplified version appropriate to axially symmetric systems is used.

A.2   PROOF OF THE GENERALIZED ÉTENDUE THEOREM

Equation (A.2) has been proved in many different ways. In particular, the analogy from Hamiltonian mechanics can be used to give a simple proof (see, e.g., Marcuse, 1972; Winston, 1970). However, since we are here concerned with applications to geometrical optics, it seems appropriate to give a proof based directly on optical principles. We shall discuss the mechanical analogies in some detail in section A.3.

We follow closely the method of Welford (1974), which makes use of the point-eikonal, or characteristic function, of Hamilton. This function is defined as follows. We take arbitrary cartesian coordinate systems in entry and exit spaces of the optical system under discussion. Let P and P' be any two points in the entry and exit spaces, respectively, and let them be in the x, y planes of their respective coordinate system[1]. Then the eikonal V is defined as the optical path length from P to

---

[1]   This is in no way essential to the definition of the eikonal; it merely simplifies the present calculation.

P' along the physically possible ray joining them. In general, one and only one ray passes through P and P', but if there is more than one, then V is multivalued. Thus, with the preceding restriction V is a function of x, y, x', and y'. Let the direction cosines of the ray in the two spaces be (L, M, N) and (L', M', N'); then the fundamental property of the eikonal can be stated as follows:

$$\partial V/\partial x = -nL, \quad \partial V/\partial y = -nM$$
$$\partial V/\partial x' = -n'L' \quad \partial V/\partial y' = -n'M' \tag{A.4}$$

This property is proved in many texts on geometrical optics (e.g., Born and Wolf, 1975; Welford, 1974). To prove our theorem we differentiate Eq. (A.4) again, and we obtain, using the notations of Eq. (A.3) and using subscripts for partial derivatives,

$$dp = -V_{xx}dx - V_{xy}dy - V_{xx'}dx' - V_{xy'}dy'$$
$$dq = -V_{yx}dx - V_{yy}dy - V_{yx'}dx' - V_{yy'}dy'$$
$$dp' = -V_{x'x}dx + V_{x'y}dy + V_{x'x'}dx' + V_{x'y'}dy'$$
$$dq' = -V_{y'x}dx + V_{y'y}dy + V_{y'x'}dx' + V_{y'y'}dy' \tag{A.5}$$

We next rearrange these terms and put the equations in matrix form

$$\begin{pmatrix} V_{xx'} & V_{xy'} & 0 & 0 \\ V_{yx'} & V_{yy'} & 0 & 0 \\ V_{x'x'} & V_{x'y'} & -1 & 0 \\ V_{y'x'} & V_{y'y'} & 0 & -1 \end{pmatrix}\begin{pmatrix} dx' \\ dy' \\ dp' \\ dq' \end{pmatrix} = \begin{pmatrix} -V_{xx} & -V_{xy} & -1 & 0 \\ -V_{yx} & -V_{yy} & 0 & -1 \\ -V_{x'x} & -V_{x'y} & 0 & 0 \\ -V_{y'x} & V_{y'y} & 0 & 0 \end{pmatrix}\begin{pmatrix} dx \\ dy \\ dp \\ dq \end{pmatrix} \tag{A.6}$$

If we denote the two matrices by B and A and the column vectors by M and M', this equation takes the form

$$BM' = AM \tag{A.7}$$

and multiplying through by the inverse of B,

$$M' = B^{-1}AM \tag{A.8}$$

This matrix equation can be expanded, and we get together with three similar equations

$$dx' = \frac{\partial x'}{\partial x}dx + \frac{\partial x'}{\partial y}dy + \frac{\partial x'}{\partial p}dp + \frac{\partial x'}{\partial q}dq \tag{A.9}$$

It can be seen that the determinant of the matrix $B^{-1}A$ is the Jacobian

$$\det(B^{-1}A) = \frac{\partial(x', y', p', q')}{\partial(x, y, p, q)} \tag{A.10}$$

which transforms the differential four-volume $dx\ dy\ dp\ dq$—that is, we have

$$dx'dy'dp'dq' = \frac{\partial(x', y', p', q')}{\partial(x, y, p, q)}dx\ dy\ dp\ dq \tag{A.11}$$

Our result will be proven if we can show that the Jacobian has the value unity. But the determinant of matrix B has the value

$$V_{xx'}V_{yy'} - V_{xy'}V_{yx'} \tag{A.12}$$

and that of matrix A has the same value, since $V_{xy'} = V_{y'x}$, and so forth. Also, the determinant of the product of two square matrices is the product of their determinants, so that

$$\det(B^{-1}) = (V_{xx'}V_{yy'} - V_{xy'}V_{yx'})^{-1} \tag{A.13}$$

Thus, $\det(B^{-1} A) = 1$ and Eq. (A.13) yields our theorem, Eq. (A.2).

A.3   THE MECHANICAL ANALOGIES AND LIOUVILLE'S THEOREM

In this section we shall indicate the analogies used to identify our theorem of the invariance of U, the étendue, with Liouville's theorem in statistical mechanics.

Fermat's principle, on which all of geometrical optics can be based, can be stated in the form

$$\delta \int_{P_1}^{P_2} n(x, y, z)ds = 0 \tag{A.14}$$

where ds is an element of the ray path from P_1 to P_2. This can be written in the form

$$\delta \int_{P_1}^{P_2} \mathcal{L}(x, y, z, \dot{x}, \dot{y})dz = 0 \tag{A.15}$$

where

$$\mathcal{L}(x, y, z, \dot{x}, \dot{y}) = n(x, y, z)\sqrt{(1 + \dot{x}^2 + \dot{y}^2)} \tag{A.16}$$

and the dots denote differentiation with respect to z. Also, we define

$$p = \frac{n\dot{x}}{\sqrt{1 + \dot{x}^2 + \dot{y}^2}} \quad q = \frac{n\dot{y}}{\sqrt{1 + \dot{x}^2 + \dot{y}^2}} \tag{A.17}$$

The analogy is to regard L as the Lagrangian function of a mechanical system in which x and y are two generalized coordinates, p and q are the corresponding generalized momenta, and z corresponds to the time axis. On this basis the ordinary development of mechanics can be carried out, such as by Luneburg (1964, Article 18), by solving the variational problem of Eq. (A.15). The Hamiltonian is found to have the value

$$\mathcal{H} = -\sqrt{n^2 - p^2 - q^2} \tag{A.18}$$

In other words, it is $-nN$ where N is the z-direction cosine, and, of course, p and q as just defined are respectively equal to nL and nM. The phase space for this system has the four coordinates (x, y, p, q), and Liouville's theorem in statistical mechanics can be invoked immediately to state that phase space volume is conserved.

However, the meaning of Liouville's theorem in mechanics is rather different from the theorem of conservation of étendue. Liouville's theorem is essentially statistical in nature, and it refers to the evolution in time of an ensemble of mechanical systems of identical properties but with different initial conditions. Each

system is represented by a single point in phase space, and the theorem states that the average density of points in phase space is constant in time. An example would be the molecules of a perfect classical gas in equilibrium in a container. Each point in phase space, which in this example has $2N$ dimensions, where N is the number of molecules, represents one of an ensemble of identical containers, an ensemble large enough to permit taking a statistical average of the density of representative points. Liouville's theorem states that if all the containers remain in equilibrium, the average density of points remains constant.

Another example would be focused beams of charged particles, as in a particle accelerator. Here we can regard one pulse of particles as constituting one realization of the ensemble and therefore one point in phase space. The statistical averaging is carried out over the random positions and momenta of the particles entering the focusing system from pulse to pulse.

The theorem has been applied to many different physical systems but the essential point is that it makes a statistical statement about the average density of points in phase space, whereas the throughput or étendue theorem is deterministic in nature. Thus, although for convenience we may call the throughput theorem "Liouville's theorem," it is desirable to remember that there is a fundamental difference in meaning.

A.4 CONVENTIONAL PHOTOMETRY AND THE ÉTENDUE

In discussing the action of concentrators we use the notion of a beam of radiation in which a certain cross section $dx \, dy$ is uniformly filled with rays covering uniformly the direction cosine solid angle $dL \, dM$. Note the useful relation $dL \, dM = N \, d\Omega$. Then an ideal concentrator takes all the rays from a source of finite area within a certain range of direction cosines and delivers them at the exit aperture to emerge over a solid angle 2π.

These ideas relate easily to particle beams, but it may not be quite clear how they relate to classical photometric ideas. In classical photometry we have a particular kind of ideal source called Lambertian for which the flux or power radiated per unit solid angle per unit projected area of the source is constant over all directions. Many physical sources approximate closely to the Lambertian condition and an ideal blackbody radiator must be Lambertian, since the radiation in a blackbody cavity is isotropic. It is easy to show that if a source radiates the same flux per unit area of the source and per unit element of direction cosine space, then it is Lambertian. Thus, we see than an ideal concentrator will take radiation from a Lambertian source over a certain solid angle and deliver it so as to make the exit aperture appear to be a complete Lambertian radiator over solid angle 2π.

REFERENCES

Born, M., and Wolf, E. (1975). "Principles of Optics," 5th Ed. Pergamon, Oxford, England.

Marcuse, D. (1972). "Light Transmission Optics." Van Nostrand-Reinhold, Princeton, New Jersey.

Welford, W. T. (1974). "Aberrations of the Symmetrical Optical System." Academic Press, New York.

Winston, R. (1970). Light collection within the framework of geometrical optics. *J. Opt. Soc. Am.* **60**, 245–247.

Appendix B

THE EDGE-RAY THEOREM

B.1 INTRODUCTION

This theorem provides the key for the synthesis methods of concentrators. It has been used as a powerful tool since the first designs of "nonimaging concentrators" in the 1960s, although its first proof was given by Miñano much later (Miñano, 1985; 1986). Since then, other proofs and formulations have been proposed by Davis (1994), Ries & Rabl (1994), and Benítez et al. (1997).

For simplicity of the explanation, we will consider here only the theorem in 2D geometry, partially following the approaches in Miñano (1985) and Benítez et al. (1997). The reader interested in the 3D treatment can consult Miñano (1986). First, in Section B.2, the theorem is formulated for the case of continuous inhomogeneous refractive index media. The theorem in presence of sequential optical surfaces and nonsequential mirrors (see Chapter 8 for this classification) is considered in Sections B.3 and B.4, respectively. Finally, in Section B.5 the modifications to the theorem introduced by the presence of slope discontinuities in sequential optical surfaces is considered, and its application to an enjoyable example, due to P. Davis (1994), that seems to constitute an offense against the edge-ray theorem at first sight.

B.2 THE CONTINUOUS CASE

Consider a certain two-parametric input bundle described at the plane $y = y_i$ by the points contained in the region M_i of the space x-y-p, and assume that we must design an optical system such that all the rays contained in the bundle M_i are cast into the exit bundle described at the plane $y = y_o$ by the region M_o, if reversed; also vice versa: all the rays of M_o if reversed, are cast into M_i. Consequently the trajectories $\Im(M_i)$ and $\Im(M_o)$ must be the same rod in x-y-p space. Consider now the one-parametric bundle of rays described by the boundary of region M_i (denoted by ∂M_i) and the one-parametric bundle described by ∂M_o. Since both bundles are one-parametric, $\Im(\partial M_i)$ and $\Im(\partial M_o)$ will be two surfaces. The theorem states that (under certain conditions that will be shown later) if $\Im(\partial M_i) = \Im(\partial M_o)$, then $\Im(M_i) = \Im(M_o)$.
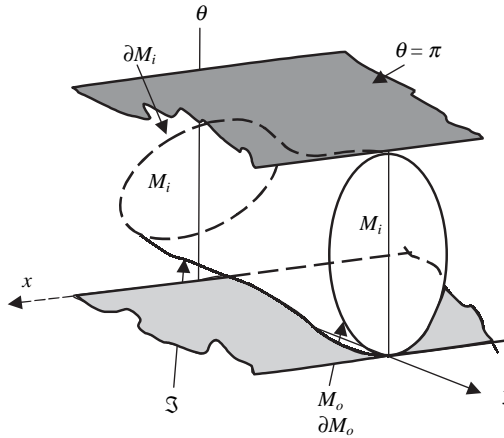
Figure B.1   The edge-ray theorem using coordinates x-y-θ.

The inverse theorem, also considered in the references, has no application for the design and will not be discussed here.

In order to prove this theorem, it is essential to consider the ray trajectories in the Hamiltonian formulation (see Section 6.1) as solutions of the following system of 1st order differential equations:

$$\frac{dx}{dt} = H_p \quad \frac{dp}{dt} = -H_x$$
$$\frac{dy}{dt} = H_q \quad \frac{dq}{dt} = -H_y \tag{B.1}$$

where the Hamiltonian $H = p^2 + q^2 + r^2 - n^2(x, y, z)$ and t is a parameter along the ray. This parameter is $t = s/(2^{1/2}n) + k$, being s is the arc length of the trajectory. The constant k can be set arbitrarily to zero. The solution must be consistent with $H = 0$.

Before proving the theorem, let us change the variables in the system of equations Eq. (B.1). The new variables are x, y, θ, μ, where x, y are equal to the old Cartesian coordinates of a plane and θ, μ are related to p, q by the equations

$$p = n\cos\theta + \mu \quad q = n\sin\theta + \mu \tag{B.2}$$

The proof of the theorem is clearer using the new variables, and its conclusions can easily be extrapolated to the variables x, y, p, q. In these variables, the defined regions are represented as shown in Figure B.1.

The trajectories of the rays are the solutions of the system Eq. (B.1) consistent with $H = 0$ that is, consistent with $n^2 = p^2 + q^2$. Using the new set of variables, the trajectories of the rays remain as the solutions of another system of equations involving x, y, θ, μ consistent with $\mu = 0$. This equation is necessarily a particular integral of the new system of equations, since $H = $ constant was a first integral of system Eq. (B.1). Since $\mu = 0$, then the variable θ can be viewed as the angle formed by the ray and a line parallel to the x-axis.

Introducing the new variables in the system Eq. (B.1) and taking into account that $\mu = 0$ the following system is obtained:

$$dx/dt = n\cos\theta$$
$$dy/dt = n\sin\theta \quad\quad\quad (B.3)$$
$$d\theta/dt = n_x \sin\theta + n_y \cos\theta$$

The first two equations are obtained directly. Equations $dp/dt = nn_x$ and $dq/dt = nn_y$ from system Eq. (B.1) lead to the same equations (the last one of system Eq. (B.3)) when $\mu = 0$ and $d\mu/dt = 0$. Note that

$$\frac{dp}{dt} = \frac{dn}{dt}\cos\theta \frac{d\theta}{dt} + \frac{d\mu}{dt} \quad\quad\quad (B.4)$$

and that

$$\frac{dn}{dt} = n_x \frac{dx}{dt} + n_y \frac{dy}{dt} = nn_x \cos\theta + nny\sin\theta \quad\quad\quad (B.5)$$

Assume that the conditions of existence and uniqueness of the solutions of system Eq. (B.1) are fulfilled. Then, there is only one solution crossing a given point (x_1, y_1, p_1, q_1) of this space x-y-p-q. Since the trajectories of the rays fulfill $n^2 = p^2 + q^2$, there are only two rays crossing a given pint x_1, y_1, p_1 of this space x-y-p: one of them with $q_1 = (n^2 - p_1^2)^{1/2}$ and the other with $q_1 = -(n^2 - p_1^2)^{1/2}$; that is, one of them proceeds toward increasing y and the other toward decreasing y (excluding the case $q_1 = 0$). Now consider only the rays with $q_1 \geq 0$. In this case there is only one ray trajectory in the x-y-p space crossing a given point x_1, y_1, p_1. The relationship between p and θ $(p = n\cos\theta)$ establishes a point-to-point mapping between the region U of the space x-y-θ defined by $0 \leq \theta \leq 180°$ and the region U' of the space x-y-p defined by $n^2 \geq p^2$. We assume that this mapping is continuous from U to U' and vice versa (for that n must be a continuous function), so it becomes a homeomorphism.

If only the rays with $q \geq 0$ are considered in x-y-p space then any trajectory in U is transformed by the mapping into another trajectory in U', and vice versa— that is, the mapping relates the solutions of the system of Eq. (B.3) with those of the system of equations Eq. (B.1) together with $H = 0$.

We now give a proof of the theorem. Let M_i and M_o be two (simply connected) sets in the planes $y = y_i$ and $y = y_o$ (∂M_i and ∂M_o are the boundaries of these sets). Let $\Re$ be an open set in the x-y plane, including the trajectories of the rays of the bundle ∂M_i (or ∂M_o, as we have assumed that the concentrator is built so that $\Im(\partial M_i) = \Im(\partial M_o)$), and let U be the region of the x-y-θ space such that (x, y, θ) belongs to U if (x, y) belongs to $\Re$.

The hypotheses of the theorem are as follows:

a. The conditions for existence and uniqueness of the solutions of the system of Eq. (B.3) are fulfilled in U, as are the conditions for the continuity of these solutions with respect to the initial conditions x_1, y_1, θ_1.
b. $\Im(\partial M_i) = \Im(\partial M_o)$. This surface will be called $\Im$ for simplicity.
c. $\Im$ is contained in the region $0 \leq \theta \leq \pi$ (the angle in radians).
d. $\Im$ is bounded.
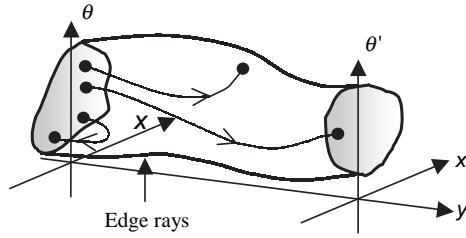
Figure B.2   Trajectories considered in the proof of the edge-ray theorem.

From the first hypothesis, and particularly from the uniqueness of the solutions, it is derived that the trajectories of two rays, in the space x-y-θ, cannot have common points.

The equation of the surface $\Im$ can be written in parametric form as

$$x = x(x_1(\tau), y_i, \theta_1(\tau), t)$$
$$y = y(x_1(\tau), y_i, \theta_1(\tau), t) \qquad\qquad (B.6)$$
$$\theta = \theta(x_1(\tau), y_i, \theta_1(\tau), t)$$

where the parameters are t and τ. The functions $x_1(\tau)$ and $\theta_1(\tau)$ give the values of x_1 and θ_1 for the points of ∂M_i, so these functions are periodical, and give the same values when τ is increased by one period. The parameter t is the one appearing in Eq. (B.3). Without loss of generality, we can set $t = 0$ at the plane $y = y_i$. It varies along the ray trajectory from $t = 0$ when the ray departs from ∂M_i to the value corresponding to a point of ∂M_o. The functions $x(x_1, y_1, \theta_1, t)$, $y(x_1, y_1, \theta_1, t)$ and $\theta(x_1, y_1, \theta_1, t)$ are the solutions of Eq. (B.3) and x_1, y_1, θ_1 are the initial conditions. Note that these functions are continuous with respect to all their parameters. By choosing the appropriate parameter τ a homeomorphism can be established between the surface $\Im$ and a tube where ∂M_e and ∂M_r are the boundaries of $\Im$. So we can conclude that $\Im$ together with M_e and M_r encloses a certain region of the x-y-θ space that will be called B.

Consider a ray described by the point p_i that belongs to the interior points of M_i (Int (M_i)). Four trajectories of this ray in the x-y-θ space can be considered (see Figure B.2) as follows:

1.   The trajectory intercepts the surface $\Im$.
2.   It intercepts the surface M_i and leaves B through it.
3.   The trajectory never leaves B.
4.   The trajectory intercepts a point of Int (M_o).

It is obvious that the trajectory departs from p_i. Option (1) is not possible because $\Im$ represents trajectories of rays, and these cannot cross in U. Option (2) is also impossible: The value of θ at the point p_i fulfills necessarily $0 < \theta < \pi$ (note that if θ were zero or π for a point of Int (M_i), due to the definition of the interior of a set, then the value of θ for some point of ∂M_i, which belongs to $\Im$, would no fulfill hypothesis ($\Im$)) and so the trajectory of the ray cannot be tangent to the plane $y = y_i$. Then this trajectory enters into B. If the trajectory comes back to M_i, it is necessary that θ takes, at some point of the trajectory, values zero or π (note that θ is continuous along the trajectory)—that is, it is necessary that the trajectory intercepts the planes $\theta = 0$ or $\theta = \pi$ before it reaches M_i and these planes are out-

side B or coincide at some point with $\Im$. Then the trajectory must exit B before reaching M_i, which is contradictory to (2).

Assume that the trajectory of the ray p_i never exits B—in other words, the trajectory is inside B for $t \in (0, \infty)$, or in terms of the arc length along the trajectory, $s = t/(2^{1/2}n)$:

$$\lim_{s \to \infty} y(s) = \int_{s=0}^{s=\infty} \sin\theta(s)\,ds < y_o \tag{B.7}$$

Since $0 < \theta(s) < \pi$, $0 < \sin\theta(s) < 1$, and thus, Eq. (B.7) implies that

$$\lim_{s \to \infty} \sin\theta(s) = 0 \Rightarrow \lim_{s \to \infty} \theta(s) = 0 \text{ or } \pi \tag{B.8}$$

But then

$$\lim_{s \to \infty} x(s) = \int_{s=0}^{s=\infty} \cos\theta(s)\,ds = \begin{cases} +\infty \text{ if } \theta(\infty) = 0 \\ -\infty \text{ if } \theta(\infty) = \pi \end{cases} \tag{B.9}$$

which contradicts that $\Im$ is bounded (see condition (4)).

Summarizing these last results, the only possibility for the trajectory of p_i is to intercept a point of M_o, and so $\Im(\text{Int}(M_i))$ belongs to $\Im((M_o))$. A similar reasoning with respect to a ray p_o of $\text{Int}(M_o)$ establishes that $\Im(\text{Int}(M_i))$, and so $\Im(\text{Int}(M_i))$ = $\Im(\text{Int}(M_o))$. Since $\Im(\partial M_i) = \Im(\partial M_o)$ by hypothesis, then $\Im(M_i) = \Im(M_o)$.

The homeomorphism between the region U of the space x-y-θ and the region U' of the space x-y-p allows us to establish easily the edge-ray theorem in the space x-y-p. Effectively, if a ray intersects (or doesn't) a surface in U, it will intersect or not the corresponding surface in U'. Also note that all the rays of M_i and M_o have $q \geq 0$. The theorem expressed in the variables x-y-p has the following hypotheses:

1. The conditions for existence and uniqueness of the solutions of the system of equations Eq. (B.1) are fulfilled in U (U' is formed by the points x, y, p, q, such that x, y belongs to $\Re$) as well as the conditions for the continuity of these solutions with respect to the initial conditions.
2. $\Im(\partial M_i) = \Im(\partial M_o)$. This surface is called $\Im$.
3. The rays of $\Im$ have $q \geq 0$ all along their trajectories, as do the rays of M_i and M_o at $y = y_i$ and at $y = y_o$, respectively 4.$\Im$ is bounded.

The fulfillment of these assumptions implies, as before, that $\Im(M_i) = \Im(M_o)$.

The conditions mentioned in hypothesis (a) (or (a′)) are (Hurewicz, 1978) that the right-hand side of Eq. (B.3) (or Eq. (B.1)) are continuous with respect to x, y, θ, t (or x, y, p, q, t) and that they satisfy locally the Lipschitz condition with respect to x, y, θ (or x, y, p, q). These conditions can be substituted by requiring that n, their first, and their second partial derivatives are continuous in $\Re$ (these last conditions are sufficient but not necessary).*

The application of the theorem of conservation of the étendue implies that the areas of M_i and M_o must be the same. However, it seems that M_i and M_o can be

---

* All these conditions are sufficient conditions but not necessary. (As an example, the Luneburg lens, see Appendix F, does not fulfill the uniqueness of the ray trajectories). The formulation of the strictly necessary conditions still consitutes an open problem.

chosen such that their areas are different. This is not true. By means of the Stokes theorem (Abraham and Marsden, 1978; Arnold, 1975; Godbillon, 1969) applied to the conservation of the étendue theorem, it can be concluded that the condition $\Im(\partial M_i) = \Im(\partial M_o)$ also implies that the areas of M_i and M_o must be equal.

This proof of the edge-ray theorem can be widened to input and exit aperture shapes different than straight lines by using the coordinate system $i\text{-}j$ defined by the flow lines and the orthogonal lines. The entry and exit aperture would be contained in two lines $j =$ constant ($j = j_i$ and $j = j_o$, respectively).

The conditions of existence, uniqueness, and continuity of the solutions with respect to the initial conditions are now fulfilled if the right-hand side of equations Eq. (B.6) are continuous with respect to i, j, u, v, t and fulfill the Lipschitz condition with respect to i, j, u, v. Condition (c') must now be replaced by $v \geq 0$. This implies that the optical direction cosine of the rays with respect to the j lines (lines with $i =$ constant), bv, is not negative. Note that $b = |\text{grad } j|$ Practical concentrators use reflectors and discontinuous refractive index distributions (lenses), neither of which is considered in the theorem just mentioned. The following sections will deal with these cases.

B.3   THE SEQUENTIAL SURFACE CASE

Once we've seen the continuous refractive index media case, consider the discontinuous case in which sequential optical mirrors or refractive surfaces are used. Sequential surfaces (as defined in Section 8.13) are defined as those on which all rays of the transmitted simply connected bundle impinge once. First, consider the line R of the $x\text{-}y$ as a 2D refractive surface. As in the former section, assume first that all rays have $q \geq 0$ (i.e., $0 \leq \theta \leq \pi$). Without loss of generality, the origin of the parameter t can be set at the points of the surface R for all the rays.

Since the whole bundle is refracted at that surface, we can identify the bundle before and after such incidence, which we will denote as M^- and M^+, respectively. Consider the points of the space $x\text{-}y\text{-}\theta$ of the bundle just before the refraction at every point $A = (x, y)$ of incidence on the surface R, which will be denoted as $x(t = 0^-), y(t = 0^-)$. $\theta(t = 0^-)$, and also the points of the space $x\text{-}y\text{-}\theta$ of the bundle just after the refraction, $x(t = 0^+), y(t = 0^+), \theta(t = 0^+)$.

The edge ray theorem for the continuous media case of the former section applies for $t < 0$ and also for $t > 0$. Then, the proof of the theorem for this discontinuous case is reduced to proof that if $\Im(\partial M^-) = \Im(\partial M^+)$, then $\Im(M^-) = \Im(M^+)$.

In general, the ray trajectories in the phase space become discontinuous at every point A (see Figure B.3). It will be sufficient to prove that the mapping of the points of M^- and M^+ is a homeomorphism.* This is immediate because the mapping between M^- and M^+ is given by

$$x(0^+) = x(0^-)$$
$$y(0^+) = y(0^-)$$
$$\theta(0^+) = \theta_A + \sin^{-1}\left(\frac{n_A(0^{-1})}{n_A(0^+)} \sin(\theta(0^{-1}) - \theta_A) \right)$$

(B.10)

---

* We must exclude the case in which there is glancing incidence on a refractive surface from the medium of lower refractive index or incidence at the critical angle from the medium of higher refractive index. In such incidences the trajectories of the rays are not uniquely determined.
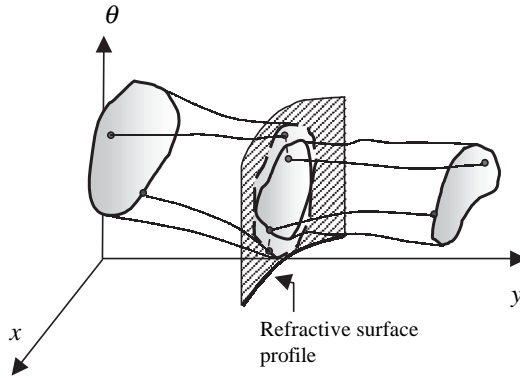
Figure B.3   The ray trajectories in the x-y-θ are discontinuous at the cylinder whose base is a sequential refractive surface. The blue dots belong to the trajectory of an edge ray and the red dot to an interior ray.

where n_A is the (discontinuous) refractive index at A, and θ_A is the angle of the normal vector to R at A with respect to the x-axis.

The first two equations in Eq. (B.10) indicate that ray trajectory in the x-y space is continuous. If the surface is refractive, the third surface is the Snell's law at A. Since by hypothesis $0 \leq \theta \leq \pi$, the third equation is a homeomorphism.

The proof for the case of a reflective surface is analogous (in fact, setting $n_A^+ = -n_A^-$ we obtain the reflection law). For the analogous formulation, all the rays after the reflection must have $0 \leq \theta \leq \pi$ or $\pi \leq \theta \leq 2\pi$. Nevertheless, as in the continuous case, the generalization for both refractive and reflective sequential surfaces can easily be extended to other input and exit aperture shapes by using the coordinate system i-j defined by the flow lines and the orthogonal lines. The entry and exit aperture would be contained in two lines $j = $ constant ($j = j_i$ and $j = j_o$, respectively).

B.4   THE FLOW-LINE MIRROR CASE

As seen above, in the case of the SMS designs without flow-line mirrors (also called nonsequential mirrors in Chapter 8), the edge-ray theorem states that

$$\Im(\partial M_i) = \Im(\partial M_o) \Rightarrow \Im(M_i) = \Im(M_o) \tag{B.11}$$

In order to include nonsequential mirrors we must define other bundles of edge rays and reformulate the theorem. Let us consider nonsequential mirrors of the type shown in Figure 3.2a, in which the bundle M that is reflected on the mirror is also represented.

This mirror fulfills the following:

1. The rays of M that pass through the point A are edge rays and form a connected bundle at A. The angle formed between the tangent to the mirror at A and the ray r_a is smaller than that which it forms with r_a'. The rays of M at A situated between r_a and its symmetric with respect to the tangent are the edge rays ∂M_A, which do not extend beyond A (that is, they are edge rays only before reaching A).
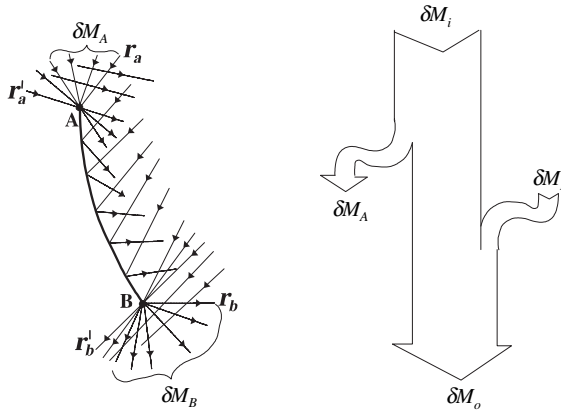
Figure B.4   Edge rays of the bundle reflected in a flow-line mirror.

2. The rays of M that pass through the point B are edge rays and form a con-
    nected bundle at B. The angle formed between the tangent to the mirror at B
    and the ray r_b is smaller than that which it forms with r_b'. The rays of M at B
    situated between r_b and its symmetric with respect to the tangent are the edge
    rays ∂M_B, which begin at B (that is, they are edge rays only after passing
    through B).
3. All edge rays impinging between A and B undergo a single reflection on the
    mirror.

In the case of Figure B.4a no more than one edge ray impinges at each point of
the mirror between A and B. This is not a condition for the formulation of the
theorem (although in practice it may be desirable).

   If a concentrator is composed of sequential optical surfaces and a nonsequen-
tial mirror of the type just described, the edge-ray theorem can be expressed as

$$\left.\begin{array}{l} \Im(\partial M_A) \cap \Im(\partial M_B) = \varnothing \\ \Im(\partial M_i) \cup \Im(\partial M_B) = \Im(\partial M_o) \cup \Im(\partial M_A) \end{array}\right\} \Rightarrow \Im(M_i) = \Im(M_o) \qquad \text{(B.12)}$$

   Thus, the design of such a concentrator is achieved by coupling (1) the edge
rays of ∂M_A with the edge rays of the entry bundle ∂M_i, (2) the edge rays of ∂M_B
with the edge rays of the exit bundle ∂M_o, and (3) the rest of the rays of ∂M_A with
the rest of the rays of ∂M_B. Figure B.4b shows an explanatory diagram of this assig-
nation of edge rays.

   The demonstration of the edge-ray theorem can be obtained under the same
hypothesis as the continuous case. Now the region B is bounded by the surface of
the edge ray trajectories and the cylinder obtained by the translation of the flow-
line mirror profile parallel to the θ axis (see Figure B.5). The edge rays ∂M_A and
∂M_B are represented as two vertical lines and belongs to the boundary of B. The
ray trajectories are discontinuous on the flow-line mirror, suffering a vertical jump
in the phase space x-y-θ as in the case of the sequential refraction in Figure B.3,
given by the mapping
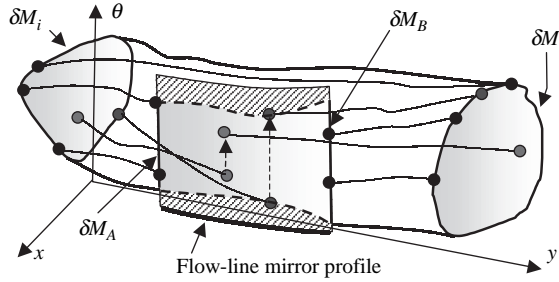
Figure B.5   Edge-ray theorem in presence of a flow-line mirror. The ray trajectories in the x-y-θ are discontinuous at the cylinder whose base is flow-line surface. The blue dots belong to the trajectory of an edge ray and the red dot to an interior ray.

$$x(0^+) = x(0^-)$$
$$y(0^+) = y(0^-) \qquad\qquad (B.13)$$
$$\theta(0^+) = 2\theta_m - \theta(0^-)$$

where θ_m is the mirror direction at the point given by the two first equations. (The last equation in Eq. (B.11) is the reflection law). The proof of the theorem in this case relies on the fact that any ray in B reflected at a point A of the mirror transforms into another ray, which is also in B. This is guaranteed by Eq. (B.11) along with the local symmetry of the bundle with respect to the flow line, from which it is deduced that

$$\theta_m \leq \theta(0^-) \leq \theta_{edge-ray}(0^-) \Rightarrow \theta_{edge-ray}(0^+) \leq \theta(0^+) \leq \theta_m \qquad (B.14)$$

This is illustrated in the phase space in Figure B.5.

B.5   GENERATION OF EDGE RAYS AT SLOPE DISCONTINUITIES

There are surfaces of practical interest for the nonimaging design that have slope discontinuities. Examples are found, for instance, in microlens arrays, Fresnel, and TIR lenses. To understand the effect of slope discontinuities in the edge ray theorem, it is useful to consider a succession of continuous slope surfaces, converging to the slope-discontinuous one. As an example, the case of a sequential refractive surface is shown in Figure B.6. Two ray families are represented, corresponding to the two edge-ray families that cross a sequential surface. It is deduced that the ray fan emitted from the discontinuity point P between the two rays refracted at each side of the discontinuity (r_a and r_b for the top edge-ray family, and r_c and r_d for the bottom one) are also edge rays that must be considered in the design.

Furthermore, we could distinguish between two cases, depending on the connection of the two ray fans emitted from P. The first case, shown in Figure B.7a, is that in which the slope discontinuity is small enough so r_b belongs to the ray fan r_c-r_d and r_c belongs to the ray fan r_a-r_b. In this situation, even the discontinuity is present, the bundle as a set in the phase space remains as a connected set. However, if the slope discontinuity is large enough compared with the bundle
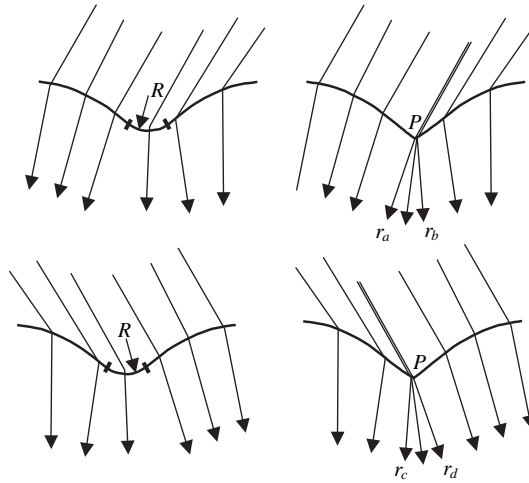
Figure B.6   Edge rays generation at the slope discontinuities on a surface. On the left, a continuous slope surface with a region of small curvature radius R. When R tends to zero, the surface converges to the discontinuous slope surface on the right.
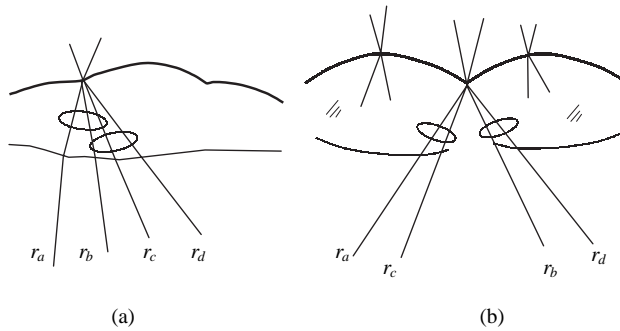
Figure B.7   Depending on the slope discontinuity size compared to the bundle angular spread at the discontinuity point, the discontinuity will (b) or will not (a) cause a disconnection of the bundle.

angular spread at the discontinuity point P, the two edge ray fans at P disconnects, and so does the bundle (see Figure B.7b).

B.6   OFFENCE AGAINST THE EDGE-RAY THEOREM

Consider the bundle M_i at $n = 1$ defined at the planes $y = y_i$ formed by the rays that point toward decreasing y-values and form with the y-axis less than $45°$. These rays can be described as those fulfilling $q < 0$ & $|p| < |q|$. The edge rays of this bundle (which has no spatial limitation) are the rays with $p = q$ and $p = -q$.

Analogously, consider the bundle M_o at $n = 1$ defined at the planes $y = y_o$ and also formed by the rays that point toward decreasing y-values and form with the y-axis less than $45°$. Also, the edge rays of this bundle (which has no spatial
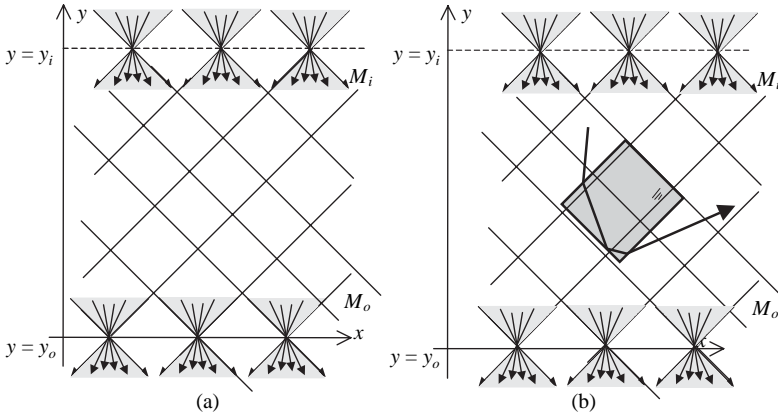
Figure B.8 (a) Extended bundle in 2D geometry formed by all the rays that progress downward and form with the y-axis less than $45°$. (b) The edge rays shown in (a) are not modified when a solid square of refractive index $n > 1$ is introduced. This seems to violate the edge ray theorem, applied to the transmission of the bundle from the plane $y = y_i$ to $y = y_o$.
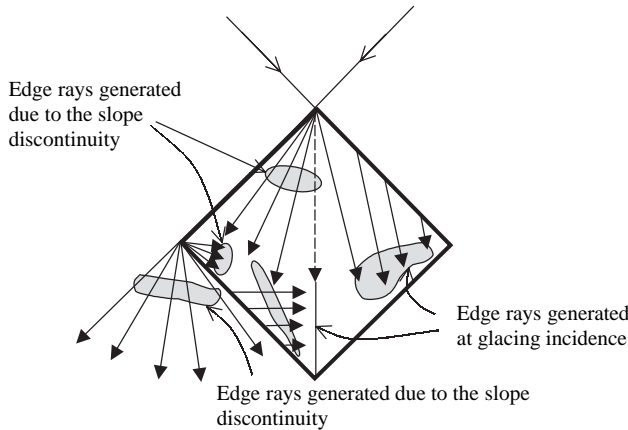
Figure B.9 Solution to the offence against the edge ray theorem.

limitation) are the rays with $p = q$ and $p = -q$. Trivially, the two bundles are coupled by just the free-space propagation in index $n = 1$ (i.e., with no optical system), as shown in Figure B.8a. The edge-ray theorem can be applied to this trivial case: Since $\Im(6M_i) = \Im(6M_o)$, then $\Im(M_i) = \Im(M_o)$.

Imagine now that we introduce a solid square region of refractive index $n > 1$ in this plane, as shown in Figure B.8b. Since the square faces are parallel and perpendicular to the edge rays shown in Figure B.8a, their trajectories remain unchanged. Therefore, it seems that we can also deduce, according to the edge-ray theorem, that the bundle coupling $\Im(M_i) = \Im(M_o)$ is also produced.

However, this is not true, because, as shown in Figure B.8b, there are interior rays of M_i that do not transform into rays of M_o (of course, and vice versa). The explanation of this apparent paradox is that the edge rays of the bundle in Figure B.8a are not the only rays of the bundle transmitted in Figure B.8b, because, as illustrated in Figure B.9:

1. The square corners are slope discontinuities and thus generate new edge rays.
2. We have "forgotten" in Figure B.8b to refract the glacing incidence rays (all along the square side). The generation of edge rays at glacing incidence is present in the Luneburg lens and was specifically used for the design of the ITC (Miñano, Ruiz, and Luque, 1983).

REFERENCES

Abraham, R., and Marsden, J. E. (1978). *Foundations of Mechanics.* Benjamin/Cummings, Reading, MA.

Arnold, V. (1975). *Classical Mechanics, Mathematics Methods*. Mir, Moscow.

Benítez, P., Mohedano, R., Miñano, J. C., García, R., and González, J. C. (1997). Design of CPC-like reflectors within the simultaneous multiple surface design method. In *Nonimaging Optics: Maximum Efficiency Light Transfer IV* (Wintson, R., ed.). Proceedings of SPIE, Vol. 3139, 19–28.

Davis, P. A. (1994). *Edge-ray principle of non-imaging optics*. *JOSA A*, Vol. 11, **4**, 1256–1259.

Godbillon, C. (1969). Géométric differentielle et mécanique analytique. Hermann, Paris.

Hurewicz, W. (1978). *Lectures on Ordinary Differential Equations*. Spanish translation. Rialp, Madrid.

Miñano, J. C. (1985). Two-dimensional nonimaging concentrators with inhomogeneous media: A new look. *JOSA A*, Vol. 2, **21**, 1826–1831.

Miñano, J. C. (1986). Design of three-dimensional nonimaging concentrators with inhomogeneous media. *J. Opt. Soc. Am. A*. **3**(9), 1345–1353.

Miñano, J. C., Ruiz, J. M., and Luque, A. (1983). Design of optimal and ideal 2D concentrators with the collector immersed in a dielectric tube. *Appl. Opt*. **22**, 3960–3965.

Ries, H., and Rabl, A. (1994). Edge-ray principle of non-imaging optics. *JOSA A*, Vol. 11, **10**, 2627–2632.

Appendix C

CONSERVATION OF SKEW AND LINEAR MOMENTUM

C.1 SKEW INVARIANT

The conservation of skew in rotational symmetric systems or the conservation of linear momentum in the linear symmetric ones comes easily in the Hamiltonian formulation. The Hamiltonian system is given by Eq. (6.16) for a system i_1, i_2, i_3 of orthogonal coordinates. For a rotationally symmetric system, let us assume that θ is the coordinate of the symmetry—that is, $\partial n/\partial \theta = 0$, where n is the refractive index distribution function. We will identify θ with i_2 of Eq. (6.16) . Then we have

$$\begin{aligned} \frac{di_1}{ds} &= \frac{\partial H}{\partial u_1} & \frac{du_1}{ds} &= -\frac{\partial H}{\partial i_1} \\ \frac{d\theta}{ds} &= \frac{\partial H}{\partial h} & \frac{dh}{ds} &= -\frac{\partial H}{\partial \theta} \\ \frac{di_3}{ds} &= \frac{\partial H}{\partial u_3} & \frac{du_3}{ds} &= -\frac{\partial H}{\partial i_3} \end{aligned} \qquad (C.1)$$

where the Hamiltonian function is

$$H \equiv u_1^2 a_1^2(i_1,i_3) + h^2 a_2^2(i_1,i_3) + u_3^2 a_3^2(i_1,i_3) - 1 \qquad (C.2)$$

a_1, a_2, and a_3 are, respectively, the modulus of the gradient of i_1, θ and i_3 over the refractive index n (i.e., $a_j = |\nabla i_j|/n$). The refractive index n is in general a function of i_1, i_3, and it is independent of θ. h is the conjugate variable of θ, which is usually called the skew.

Because i_1, θ, i_3 is a triorthogonal coordinate system, and because of the rotational symmetry, $\partial |\nabla i_1|/\partial \theta = \partial |\nabla i_3|/\partial \theta = 0$. Then, there is no dependence of a_1 a_3, on θ, as it is specified in Eq. (C.2). a_2, which is $a_2 = |\nabla\theta|/n$, can be written as $a_2 = 1/(\theta\, n)$, where θ is the distance of the point i_1, θ, i_3 to the axis of rotational symmetry—that is, θ is one of the cylindrical coordinates. θ is a function of i_1, i_3 solely.

Now, it is simple to see that h is an invariant parameter along a ray: one of the equations of the system given by Eq. (C.1) sets

$$\frac{dh}{ds} = -\frac{\partial H}{\partial \theta} = 0 \qquad (C.3)$$

C.2   LUNEBURG TREATMENT FOR SKEW RAYS

In mathematical terms it is said that h is a first integral of the Hamiltonian system. When a first integral of a Hamiltonian system is known, then a new Hamiltonian system with two fewer variables can be built up, which obviously simplifies the problems. Using this property Luneburg show that the new Hamiltonian system represents also an optical system in 2D geometry. With the Hamiltonian formulation that we are using this is simple to see:

As we have seen, the Hamiltonian function does not depend on θ.

$$H \equiv u_1^2 a_1^2(i_1, i_3) + \frac{h^2}{\rho^2(i_1, i_3)n^2(i_1, i_3)} + u_3^2 a_3^2(i_1, i_3) - 1 \tag{C.4}$$

For a given value of $h = h_0$, we can solve the system

$$\begin{aligned}
\frac{di_1}{ds} &= \frac{\partial H}{\partial u_1} & \frac{du_1}{ds} &= -\frac{\partial H}{\partial i_1} \\
\frac{di_3}{ds} &= \frac{\partial H}{\partial u_3} & \frac{du_3}{ds} &= -\frac{\partial H}{\partial i_3}
\end{aligned} \tag{C.5}$$

with the condition $H = 0$. The results are sets of four functions $i_1 = i_1(s)$, $i_3 = i_3(s)$, $u_1 = u_1(s)$, $u_3 = u_3(s)$; which give the projection of the phase space ray trajectories in the subspace i_1-i_2-u_1-u_3.

For instance, assume that i_1, i_3 are the cylindrical coordinates ρ, z, respectively. Then, the solution of the system Eq. (C.5) can provide us the functions $\rho = \rho(s)$, $z = z(s)$, (besides two other functions) that are the projection of the ray trajectories on the ρ, z coordinates.

To get the complete trajectories in the phase space we must add the equation $h = h_0$ and integrate

$$\frac{d\theta}{ds} = \left.\frac{\partial H}{\partial h}\right|_{h=h_o} = \frac{2h_0}{\rho^2(i_1, i_3)n^2(i_1, i_3)} \tag{C.6}$$

which can be integrated because the functions $i_1 = i_1(s)$, $i_3 = i_3(s)$ are known.

The system of Eq. (C.5) together with Eq. (C.4) describes a 2D optical system with configuration variables i_1, and i_3, and conjugate variables u_1 and u_3. The refractive index of this 2D optical system is $n_{eq}(i_1, i_3)$ given by

$$n_{eq}(i_1, i_3) \equiv \sqrt{n^2(i_1, i_3) - \frac{h_0^2}{\rho^2(i_1, i_3)}} \tag{C.7}$$

This can be easily verified by comparing the Hamiltonian equations of a 2D optical system with this refractive index taking into account that the solutions of a Hamiltonian system remains the same if we multiply the Hamiltonian function by a nonzero function (see Section 6.12.2).

Meridian rays are the ones whose trajectory in the space i_1, θ, i_3 is contained in a θ = constant plane—that is, $d\theta/ds = 0$. These meridian rays are those with $h = 0$ (see Eq. (C.1) and Eq. (C.2)). Then, $n_{eq} = n$ for meridian rays.

The Luneburg result can be summarized as follows: The projection of the trajectory of a skew ray (with $h = h_0$) in the θ, zi_1, i_3 coordinates coincides with the trajectory of a meridian ray if the refractive index distribution is modified as given by Eq. (C.7).

C.2.1   Optical Path Length

According to Eq. (6.19) the optical path length can be written as

$$L_{AB} = \int_B^A u_1 di_1 + h\,d\theta + u_3 di_3 = h(\theta_A - \theta_B) + \int_B^A u_1 di_1 + u_3 di_3 \qquad \text{(C.8)}$$

The second term of the last part of Eq. (C.8) is the optical path length ray (in the optical system with refractive index distribution $n_{eq}(i_1, i_3)$) of a meridian ray that coincides with the projection of the ray on the i_1, i_3 coordinates.

C.3   LINEAR MOMENTUM CONSERVATION

A similar reasoning as done for rotationally symmetric systems can be done for linear symmetric ones. The result equivalent to Eq. (C.3) gives

$$\frac{dp_3}{ds} = -\frac{\partial H}{\partial x_3} = 0 \qquad \text{(C.9)}$$

where p_3 is the optical direction cosine of a ray with respect the x_3-axis, which is assumed to be the axis of linear symmetry.

The linear momentum conservation can also be deduced from the skew conservation by considering the linear system as a limit case of a toroidal system (and thus a rotational symmetric system) that is infinitely far from the axis of rotational symmetry—that is, when $\rho \to \infty$ (Miñano, 1984).

The Luneburg treatment of skew rays (which will be called nonmeridian rays in linear symmetric systems) can also be applied here.

The projection of the trajectory of a nonmeridian ray (with $p_3 = p_0$) on a plane normal to the axis of linear symmetry coincides with the trajectory of a meridian ray if the refractive index distribution is modified as follows:

$$n_{eq}(i_1, i_2) \equiv \sqrt{n^2(i_1, i_2) - p_0^2} \qquad \text{(C.10)}$$

The optical path length can also be split in two parts, one of which is the optical path length (in the optical system with refractive index distribution $n_{eq}(i_1, i_2)$) of the meridian ray that coincides with the projection of the ray on the i_1, i_2 coordinates

$$L_{AB} = \int_B^A u_1 di_1 + u_2 di_2 + p_3 dx_3 = p_3(x_{3A} - x_{3B}) + \int_B^A u_1 di_1 + u_2 di_2 \qquad \text{(C.11)}$$

C.4   DESIGN OF CONCENTRATORS FOR NONMERIDIAN RAYS

With the exception of the direct 3D design method (like the Poisson brackets method), all other methods design a 2D device from which a 3D device is later generated by linear or rotational symmetry. In the final 3D device, this procedure only guarantees that the optical device couples the meridian edge rays at the entry and at the exit. With the aid of the preceding results we can do designs that couple
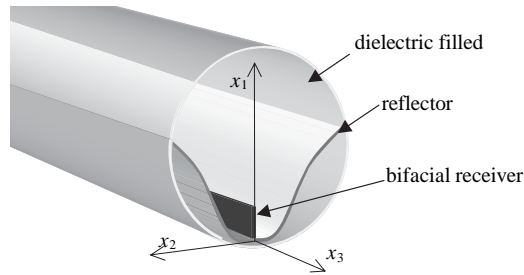
Figure C.1   Linear Integrated Tubular Concentrator. Designed according the section 5.6 for a bifacial solar cell as receiver (dielectric refractive index 1.5).
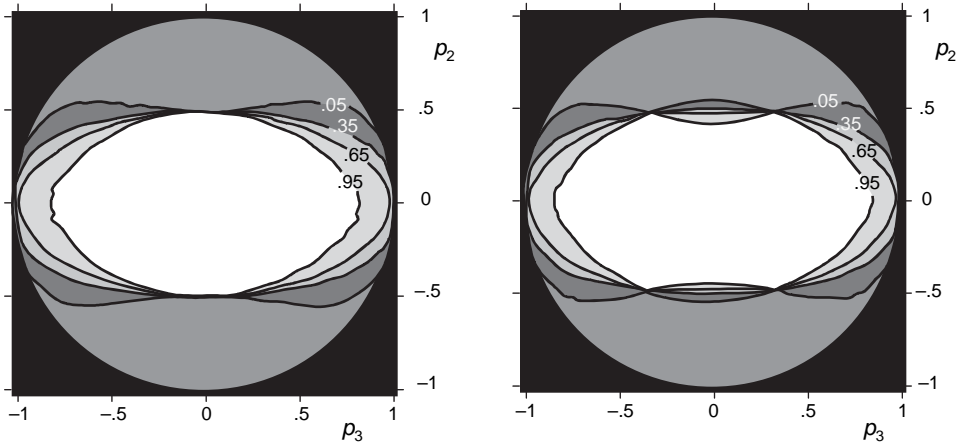
Figure C.2   Angular transmission for a linear systems whose cross-section is designed for meridian rays, i.e., for rays with $p_3 = 0$ rays (left and side). All the $p_3 = 0$ rays are collected when $p_2 < 0.5$ and all of the remaining $p_3 = 0$ rays are rejected. Thus, the angular transmission takes only values 0 or 1 for $p_3 = 0$. The right hand side shows the angular transmission of a linear system whose cross section is designed for $p_3 = 0.35$ rays.

nonmeridian edge rays instead of the meridian ones. The only thing we have to do is to take into account Eq. (C.7) or Eq. (C.10) when passing from the 2D design to the 3D one. In this way we can have, for instance, a rotational CPC designed for meridian rays ($h = 0$), a rotational CPC designed for skew rays of $h = h_0$, a linear CPC designed for meridian rays ($p_3 = 0$) or a linear CPC designed for nonmeridian rays of $p_3 = p_0$.

Figure C.1 shows a linear concentrator that was designed for $p_3 = 0.35$. This concentrator, called Integrated Tubular Concentrator (ITC) (Miñano, Ruiz, and Luque, 1983) is the first tailored design in the sense of Chapter 5. Figure C.2 shows the angular transmission function of two linear designs. The one corresponding to the figure on the left is a linear design whose cross-section is an ideal 2D design done for meridian rays—that is, $p_3 = 0$. Consequently the transmission curve $T(p_2, p_3 = 0)$ takes only two values: 0 or 1. The figure on the right side corresponds to a linear concentrator whose cross section is an ideal 2D design for $p_3 = \pm0.35$ rays and consequently the curve $T(p_2, p_3 = \pm0.35)$ takes only two values: 0 or 1.

REFERENCES

Miñano, J. C. (1984). Cylindrical concentrators as a limit case of toroidal concentrators. *Applied Optics* **23**, 2017–2020.

Miñano, J. C., and Ruiz, J. M. (1983). Design of optimal and ideal 2D concentrators with the collector immersed in a dielectric tube. *Applied Optics* **22**, 3960–3965.

Appendix D

CONSERVATION OF ÉTENDUE FOR TWO-PARAMETER BUNDLES OF RAYS

The conservation of étendue belongs to a set of integral invariants applicable to the solutions of any Hamiltonian system. Several authors discovered some of these invariants independently. Henri Poincaré (1957) recognized them as a part of a general statement. In this appendix we are going to see the applications to optics one of them, which we call conservation of étendue for two-parameter bundles of rays.

Different forms of this invariant have appeared in Geometrical Optics: The theorem of Malus and Dupin, the Lagrange invariant, the Herzberger's (1944, 1958) fundamental optical invariants (Herzberger, 1958) express totally or partially this invariant (Miñano, 1984).

We will show this invariant in Poincaré's way. Let's consider a two-parameter bundle of rays—that is, a set of rays such that each ray can be distinguished from others of the bundle by specifying the values of two parameters, which we will denote by τ_1 and τ_2. The trajectory of a ray of this bundle in the phase space x_1-x_2-x_3-t-k_1-k_2-k_3-ω will be given in parametric form by eight functions $x_1 = x_1(\tau_1, \tau_2, s)$, $x_2 = x_2(\tau_1, \tau_2, s)$, $x_3 = x_3(\tau_1, \tau_2, s)$, $t = t(\tau_1, \tau_2, s)$, $k_1 = k_1(\tau_1, \tau_2, s)$, $k_2 = k_2(\tau_1, \tau_2, s)$, $k_3 = k_3(\tau_1, \tau_2, s)$, $\omega = \omega(\tau_1, \tau_2, s)$, where s is the parameter along the ray curve (see Section for the variables definition). Let R be a hypersurface of the phase space x_1-x_2-x_3-t-k_1-k_2-k_3-ω that intercepts once and only once any phase space trajectory of rays of the bundle. Let $s(\tau_1, \tau_2)$ be the parameter s at the point of interception. Thus, the coordinates x_1, x_2, x_3, t, k_1, k_2, k_3, ω of the rays of the bundle at the surface R can be expressed as functions of τ_1 and τ_2. The conservation of étendue for two-parameter bundle of rays says that the differential form dE_{2P} is independent of the surface R provided R intercepts all the rays of the bundle once and only once

$$dE_{2P} = dx_1 dk_1 + dx_2 dk_2 + dx_3 dk_3 - dt\, d\omega \tag{D.1}$$

that is, dE_{2P} is constant along the bundle trajectory in the phase space (see Flanders, 1989, for a detailed explanation of differential forms). The minus sign of $dt d\omega$ comes from the fact that the conjugate variable of t is not ω but $-\omega$. The

variables x_1, x_2, x_3, t, k_1, k_2, k_3, $\dot\omega$ in Eq. (D.1) are the coordinates at the surface R. Thus, dE_{2P} can also be written as

$$dE_{2P} = \left(\frac{D(x_1,k_1)}{D(\tau_1,\tau_2)} + \frac{D(x_2,k_2)}{D(\tau_1,\tau_2)} + \frac{D(x_3,k_3)}{D(\tau_1,\tau_2)} - \frac{D(t,\omega)}{D(\tau_1,\tau_2)} \right) d\tau_1\, d\tau_2 \qquad (D.2)$$

where $D(x, k)/D(\tau_1, \tau_2)$ is the Jacobian. The conservation of dE_{2P} in Eq. (D.2) means that the term between brackets is independent of the surface R. These surfaces R can be such that $s(\tau_1, \tau_2) = $ constant. For this case we can write

$$\frac{d}{ds}\left(\frac{D(x_1,k_1)}{D(\tau_1,\tau_2)} + \frac{D(x_2,k_2)}{D(\tau_1,\tau_2)} + \frac{D(x_3,k_3)}{D(\tau_1,\tau_2)} - \frac{D(t,\omega)}{D(\tau_1,\tau_2)} \right) = 0 \qquad (D.3)$$

where now the variable s in the functions $x_1 = x_1(\tau_1, \tau_2, s)$, $x_2 = x_2(\tau_1, \tau_2, s)$, $x_3 = x_3(\tau_1, \tau_2, s)$, $t = t(\tau_1, \tau_2, s)$, $k_1 = k_1(\tau_1, \tau_2, s)$, $k_2 = k_2(\tau_1, \tau_2, s)$, $k_3 = k_3(\tau_1, \tau_2, s)$, $\omega = \omega(\tau_1, \tau_2, s)$ is not substituted by a function of τ_1, τ_2.

If a canonical transformation of the phase space x_1-x_2-x_3-t-k_1-k_2-k_3-ω in another phase space i_1-i_2-i_3-i_4-u_1-u_2-u_3-u_4 is done, Eq. (D.1) remains in a similar way

$$dE_{2P} = di_1 du_1 + di_2 du_2 + di_3 du_3 + di_4 du_4 \qquad (D.4)$$

When the optical media is time independent the number of independent variables of the Hamiltonian system can be reduced in two (see Section 6.12.2). Consequently, Eqs. (D.1) and (D.4) remain as

$$dE_{2P} = dx_1 dk_1 + dx_2 dk_2 + dx_3 dk_3 \qquad (D.5)$$

$$dE_{2P} = di_1 du_1 + di_2 du_2 + di_3 du_3 \qquad (D.6)$$

where now, i_1-i_2-i_3-u_1-u_2-u_3 are obtained from a canonical transformation of x_1-x_2-x_3-k_1-k_2-k_3.

The Lagrange invariant can be expressed as (Herzberger, 1993)

$$\frac{d}{ds}\left(\left(\frac{\partial x_1}{\partial \tau_1}, \frac{\partial x_2}{\partial \tau_1}, \frac{\partial x_3}{\partial \tau_1} \right) \cdot \left(\frac{\partial p_1}{\partial \tau_2}, \frac{\partial p_2}{\partial \tau_2}, \frac{\partial p_3}{\partial \tau_2} \right) - \left(\frac{\partial x_1}{\partial \tau_2}, \frac{\partial x_2}{\partial \tau_2}, \frac{\partial x_3}{\partial \tau_2} \right) \cdot \left(\frac{\partial p_1}{\partial \tau_1}, \frac{\partial p_2}{\partial \tau_1}, \frac{\partial p_3}{\partial \tau_1} \right) \right) = 0 \quad (D.7)$$

that is, the expression between brackets is constant along the ray trajectories (the dot denotes dot product). This expression between brackets is

$$\left(\frac{\partial x_1}{\partial \tau_1}, \frac{\partial x_2}{\partial \tau_1}, \frac{\partial x_3}{\partial \tau_1} \right) \cdot \left(\frac{\partial p_1}{\partial \tau_2}, \frac{\partial p_2}{\partial \tau_2}, \frac{\partial p_3}{\partial \tau_2} \right) - \left(\frac{\partial x_1}{\partial \tau_2}, \frac{\partial x_2}{\partial \tau_2}, \frac{\partial x_3}{\partial \tau_2} \right) \cdot \left(\frac{\partial p_1}{\partial \tau_1}, \frac{\partial p_2}{\partial \tau_1}, \frac{\partial p_3}{\partial \tau_1} \right)$$
$$= \frac{D(x_1,p_1)}{D(\tau_1,\tau_2)} + \frac{D(x_2,p_2)}{D(\tau_1,\tau_2)} + \frac{D(x_3,p_3)}{D(\tau_1,\tau_2)} \qquad (D.8)$$

Then, Eq. (D.7) is expressing that dE_{2P} (see Eq. (D.5)) is constant along the ray trajectories—that is, the Lagrange invariant is the same as the conservation of dE_{2P}.

A two-parameter bundle such that dE_{2P} is zero is a normal system of rays (Luneburg, 1964), which are also called orthotomic systems or normal congruences (Stavroudis, 1972)—that is, there is a family of surfaces normal to the rays (the wavefronts). The converse theorem is also true: A normal system of rays has zero dE_{2P}. From here it follows that the theorem of Malus-Dupin is equivalent to the conservation of the zero dE_{2P} in normal system of rays.

The usual expression of the Lagrange invariant is a line integral on a closed curve, and it refers only to normal systems of rays (Born and Wolf, 1993). This

integral invariant relative to closed curves can be transformed into the integral invariant obtained from Eq. (D.7) with the Stokes theorem.

D.1   CONDITIONS FOR ACHROMATIC DESIGNS

As an application of the preceding results, consider the two-parameter bundle of rays (parameters τ_1 and τ_2) defined by the equations (in this example we follow Minaño, Benítez, and Muñoz, 2001).

$$
\begin{aligned}
x_1 &= A \quad x_2 = 0 & t &= 0 \\
k_1 &= \tau_1 \quad k_2 = \sqrt{\left(\frac{\tau_2 n}{c_o}\right)^2 - \tau_1^2} & \omega &= \tau_2
\end{aligned}
\tag{D.9}
$$

These equations represent a pulse of light leaving the point $x_1 = A$, $x_2 = 0$ at $t = 0$ and traveling in the x_1-x_2 plane. Consequently, the coordinates z and r do not appear. Observe that the modulus of the wave vector k is $k = \omega n/c_o$ where n is the refractive index which depends on x_1, x_2, and ω. Then, the expression of k_2 in Eq. (D.9) ensures that the bundle fulfills $K = 0$, where K is the Hamiltonian function (see Section 6.12.2).

We want to find the conditions that n must fulfill for the pulse to reach a point $x_1 = A'$, $x_2 = D$ (image point) independently of the value of ω. We don't care about the value of t when the pulse rays reach at the image point. Thus, the bundle at the image point will fulfill

$$
\begin{aligned}
x_1 &= A' & x_2 &= D & t &= t(\tau_1, \tau_2) \\
k_1 &= k_1(\tau_1, \tau_2) & k_2 &= \sqrt{\left(\frac{\tau_2 n}{c_o}\right)^2 - k_1^2(\tau_1, \tau_2)} & \omega &= \tau_2
\end{aligned}
\tag{D.10}
$$

Since n does not depend on t, then ω is constant along the ray trajectories and thus $\omega = \tau_2$ for the entire trajectory.

Let us calculate dE_{2P} at the start and at the end of the trajectories with the aid of Eq. (D.2) and equate both results.

$$
\begin{vmatrix} \dfrac{\partial x_1}{\partial \tau_1} & \dfrac{\partial x_1}{\partial \tau_1} \\[2mm] \dfrac{\partial k_1}{\partial \tau_1} & \dfrac{\partial k_1}{\partial \tau_2} \end{vmatrix} + \begin{vmatrix} \dfrac{\partial x_2}{\partial \tau_1} & \dfrac{\partial x_2}{\partial \tau_2} \\[2mm] \dfrac{\partial k_2}{\partial \tau_1} & \dfrac{\partial k_2}{\partial \tau_2} \end{vmatrix} - \begin{vmatrix} \dfrac{\partial t}{\partial \tau_1} & \dfrac{\partial t}{\partial \tau_2} \\[2mm] \dfrac{\partial \omega}{\partial \tau_1} & \dfrac{\partial \omega}{\partial \tau_2} \end{vmatrix}
\tag{D.11}
$$

A simple calculation gives the following expression:

$$
\frac{\partial t(\tau_1, \tau_2)}{\partial \tau_1} = 0
\tag{D.12}
$$

Let p_1 be the direction cosine with respect x_1 of the pulse issuing at $t = 0$—that is, $p_1 = k_1/k$. Since $\tau_1 = k_1$ (at $t = 0$) and taking into account that $\tau_2 = \omega$ is kept constant in the derivative of Eq. (D.12), this equation can be written as

$$
\frac{\partial t(p_1, \omega)}{\partial p_1} = 0
\tag{D.13}
$$

With the aid of the equation of the Hamiltonian system in Eq. (6.4) giving the total derivative of t and the expression of the Hamiltonian function Eq. (6.8) we get (note that in our problem the trajectory is contained in the x_1-x_2 plane)

$$t = \int \frac{2\omega n(n + \omega n_\omega)}{c_o^2} ds \qquad (D.14)$$

where $n_\omega = \partial n / \partial \omega$. Let dl be the differential of length in the x_1-x_2 plane. Using again Eq. (6.4) and Eq. (6.8) we can get

$$dl = \sqrt{(dx_1)^2 + (dx_2)^2} = 2\sqrt{k_1^2 + k_2^2}\, ds = 2\frac{\omega n}{c_o} ds \qquad (D.15)$$

Thus, Eq. (D.14) can be written as

$$t = \frac{1}{c_o} \int_P^{P'} n\, dl + \frac{\omega}{c_o} \int_P^{P'} n_\omega dl \qquad (D.16)$$

where it has been taking into account that ω is constant along the trajectory, and thus it can go out of the integral. P and P' are, respectively, the points $x_1 = A$, $x_2 = 0$, and $x_1 = A'$, $x_2 = D$ of the x_1-x_2 plane. The first integral of Eq. (D.16) is the optical path length between P and P' over the light velocity in vacuum c_o. Let us call the second integral of Eq. (D.16) the dispersion path length between P and P'. This second integral is evaluated along the ray paths.

If the pulse trajectories issuing from the point $x_1 = A$, $x_2 = 0$ at $t = 0$ reach the coordinates $x_1 = A'$, $x_2 = D$ in the space x_1-x_2-t it is obvious that the trajectories projected in the x_1-x_2 plane are imaging the point P into the point P'. Since the principle of Fermat establishes that the optical path length between P and P' does not depend on the direction cosine of emission p_1, Eq. (D.13) with the aid of Eq. (D.16), can be written as

$$\int_P^{P'} n_\omega dl = \text{constant} \qquad (D.17)$$

which expresses that the dispersion path length from P to P' must be constant for all the ray paths. Eq. (D.17) should not be confused with a variational principle.

Eq. (17), which is a generalized form of the Conrady's formula (Welford, 1989), can be used to design achromatic lenses in a simple way.

D.1.1   Example of Achromatic Aplanatic Lens Doublet Design

For instance, consider the design of a lens focusing the rays from P at the point P' that is also achromatic in a neighborhood of $\omega = \omega_o$. For this purpose, two equations should be fulfilled:

$$\int_P^{P'} n(\omega_o) dl = \text{constant} \qquad (D.18)$$

$$\int_P^{P'} n_\omega(\omega_o) dl = \text{constant} \qquad (D.19)$$
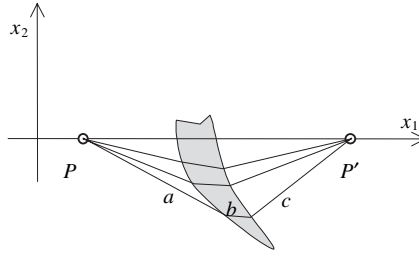
Figure D.1   Single achromatic lens focussing P onto P'. Their surfaces are not normal to P-P'.

To fulfill Eq. (D.18) a single refractive surface is enough (Cartesian ovals). A lens fulfilling both (Eqs. (D.18) and (D.19)) requires a minimum of two surfaces. Assume that the media outside the lens has $n = 1$ and $n_\omega = 0$, and let us call a, b, c to the three segments of the ray trajectory from P to P' (see Figure D.1). Then, Eqs. (D.18) and (D.19) can be written as

$$a + bn(\omega_0) + c = L$$
$$bn_\omega(\omega_0) = D$$

(D.20)

where L and D are two constants.

From Eq. (D.20) it is got $b = $ constant—that is, the ray path within the dielectric is constant (and also the ray path outside the dielectric). If besides these equations we require the lens surfaces to be normal to the axis P-P', then the problem has no solution (excepting trivial cases). Thus, we cannot get a rotational symmetric solution with continuous derivatives at the axis P-P'. Schultz (1988) got the same result for single achromatic lenses in which the word *achromatic* meant that the lens is sharp focusing for two values of the wavelength. This definition of achromatic lens and our definition are coincident when the two values of the wavelength converge to a single value. Nevertheless, we can get a nonsymmetric solution. For this, we need a third equation plus Eq. (D.20). This third equation is Snell's law at any of the two dielectric interfaces.

To solve the symmetry problem we can add another surface and another material—that is, we need a doublet. The new surface can be prescribed (not totally free), or it can be designed. In the last case we can add an additional condition in the design. This is the way we are going to solve the problem. This additional condition will be the sine condition (see Figure D.2). In this way the final design will be an achromatic and aplanatic aspheric doublet.

For a given value of ϕ, we have five unknowns to determine the points of the three surfaces crossed by the ray determined by ϕ. Note that the two coordinates of the point in surface A are linked once we fix ϕ. These five unknowns are solved with five equations: the sine condition, Eq. (D.18), Eq. (D.19), and Snell's law applied to two of the three surfaces A, B, C. Note that Snell's law must necessarily be fulfilled at the third surface because we made also use of Fermat's principle (Eq. (D.18)) for the full trajectory.

Figure D.3 shows the cross sections of two achromatic and aplanatic aspheric doublets. Both have been designed with aperture diameters and f-numbers equal to two commercial spherical achromatic doublets in order to compare the results.
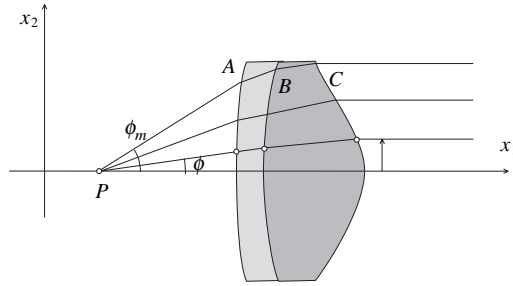
Figure D.2   Sine condition: The angle ϕ is proportional to the coordinate x_2 at the exit.

Ø 30 mm, *f/#* 1.66

Ø 6.25 mm, *f/#* 2.4
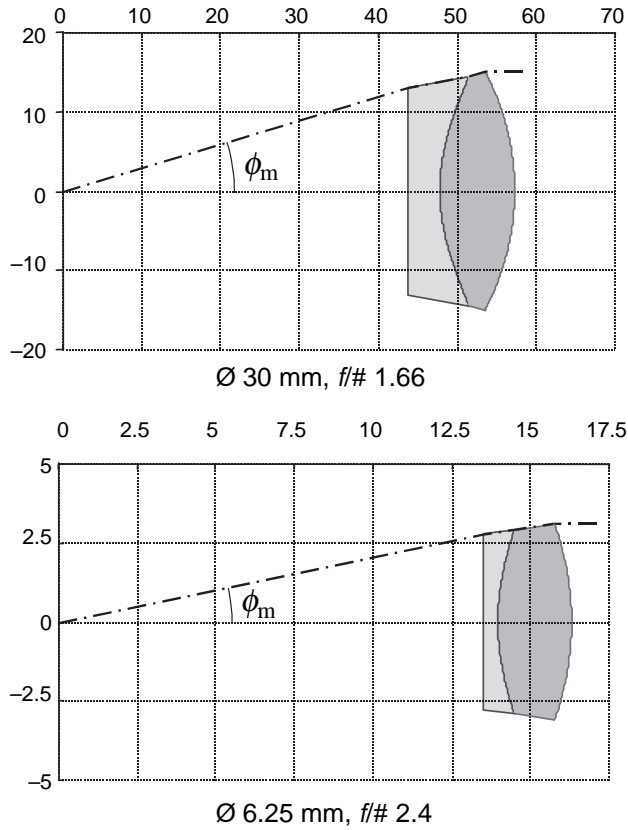
Figure D.3   Cross-section of the aplanatic achromatic aspheric doublets. Surface *A* is aspheric, although in the figure it looks flat.

The wavelength of design was chosen to $\lambda_0 = 486.13$ nm. The data of the materials used in the lenses are $\varnothing$ 30 mm, *f*/1.66.

1. Material between surface *A* and *B*: SF10 (Flint material) $n(\lambda_0) = 1.7465$; $n_\lambda(\lambda_0) = -0.2497\,\mu\mathrm{m}^{-1}$.
2. Material between surface *B* and *C*: BaFN10 (Crown material); $n(\lambda_0) = 1.6800$; $n_\lambda(\lambda_0) = -0.1326\,\mu\mathrm{m}^{-1}$. $\varnothing$6.25 mm, *f*/2.4
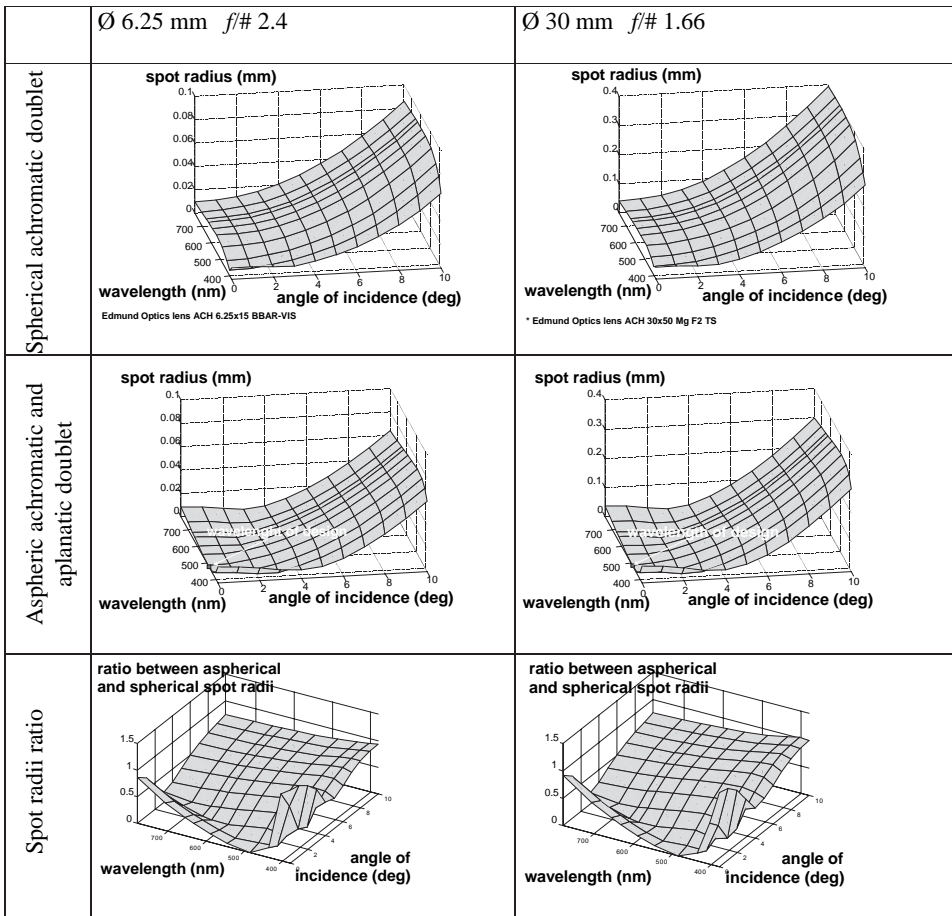
Figure D.4   Results of the comparison between spherical achromatic doublets and aspheric aplanatic achromatic doublets.

3. Material between surface A and B: SF14 (Flint material) $n(\lambda_0) = 1.7823$; $n_\lambda(\lambda_0)$ $= -0.2821 \, \mu m^{-1}$.

4. Material between surface B and C: BaF13 (Crown material); $n(\lambda_0) = 1.6794$; $n_\lambda(\lambda_0) = -0.1394 \, \mu m^{-1}$.

The geometrical (no diffractive effects) spot radius for different wavelengths and different incidence angles has been calculated by ray tracing. Here incidence angle means the angle formed by a bundle of parallel rays with the x_1-axis (for this calculation the rays proceed in reverse direction to the x_1-axis). Figure D.4 shows the results of the commercial spherical achromatic doublets compared with the achromatic and aplanatic aspheric doublets. It can be verified that the spot radius is smaller for all wavelengths and incidence angles studied. The spherical doublets look very similar to the aspheric ones. The comparison has been done on the basis that the angle ϕ_m (see Figure D.3) as well as the aperture diameter are the same for the spherical and aspheric lenses being compared. This means that the f/#, defined as the effective focal length over the aperture diameter, is the same.

Two different sets of $f/\#$ and aperture diameter $\varnothing$ have been used: $\varnothing = 30\,\text{mm}$, $f/1.66$, and $\varnothing = 6.25\,\text{mm}$, $f/2.4$.

D.2   CONDITIONS FOR CONSTANT FOCAL LENGTH IN LINEAR SYSTEMS

The same procedure shown here can be applied to the design of linear lenses whose focal length must be kept constant in a neighborhood of incidence angles around the design condition.

Consider a linear symmetric optical system such that x_3 is the axis of symmetry. Assume that we want to design such a linear system with the condition that the rays issuing from the straight-line $x_1 = 0$, $x_2 = A$ are focused on a straight line $x_1 = D$, $x_2 = A'$ (see Figure D.5). This problem is, from the mathematical point of view, equivalent to the achromatic design problem considered in Section D.1.

This linear optical system is invariant with respect to x_3, and we are only interested in the projection of the trajectories on the x_1-x_2. We require that these trajectory projections focus the rays leaving from $x_1 = 0$, $x_2 = A$ onto $x_1 = D$, $x_2 = A'$. In a similar way, the achromatic system of Section D.1 is invariant with respect to t and we are only interested in the x_1-x_2 trajectory (we don't mind about the time t).

Proceeding in the same way let us calculate dE_{2P} of the bundle of rays leaving a point in the object line. The coordinates x_1, x_2, x_3, p_1, p_2, p_3 of the rays of this bundle as a function of the two parameters τ_1 and τ_2 are

$$
\begin{aligned}
x_1 &= 0 & x_2 &= A & x_3 &= 0 \\
p_1 &= \sqrt{n^2 - \tau_1^2 - \tau_2^2} & p_2 &= \tau_1' & p_3 &= \tau_2
\end{aligned}
\tag{D.21}
$$

Because this is a time invariant system and we are not interested in the chromatic response, it is more advisable to use the Hamiltonian with the optical direction cosines as conjugate variables as in Eq. (6.10) (instead of the wave vector components which are just proportional to the optical direction cosines). The differential dE_{2P} can be written as

$$
dE_{2P} = \left(\frac{\omega}{c_o}\right)\{dx_1 dp_1 + dx_2 dp_2 + dx_3 dp_3\}
\tag{D.22}
$$
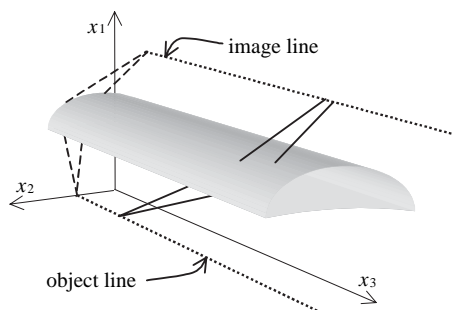
Figure D.5   Linear system that focuses the rays issuing from the object line into the image line.

The resulting $dE_{2P} = 0$ for the bundle given by Eq. (D.21). It should be also zero for the bundle described at the image line. This description is

$$x_1 = D \qquad\qquad x_2 = A' \qquad\quad x_3 = x_3(\tau_1, \tau_2)$$
$$p_1 = \sqrt{n^2 - p_2^2(\tau_1, \tau_2) - \tau_2^2} \quad p_2 = p_2(\tau_1, \tau_2) \quad p_3 = \tau_2 \tag{D.23}$$

Because $dE_{2P} = 0$ at the image line, then

$$\frac{\partial x_3(\tau_1, \tau_2)}{\partial \tau_1} = 0 \tag{D.24}$$

Similarly as done before, we will express Eq. (D.24) in terms of x_1, x_2. For this, consider the equation giving dx_3/ds in Eq. (6.9)

$$x_3 = 2\int p_3 ds \tag{D.25}$$

Now let us express ds as a function of the differential of length of the ray trajectory projected on the $x_1 - x_2$ plane

$$dl = \sqrt{(dx_1)^2 + (dx_2)^2} = 2\sqrt{p_1^2 + p_2^2}\, ds = 2\sqrt{n^2 - p_3^2}\, ds \tag{D.26}$$

$$x_3 = \int_P^{P'} \frac{p_3}{\sqrt{n^2 - p_3^2}}\, dl \tag{D.27}$$

where P and P' are, respectively, the points $x_1 = 0$, $x_2 = A$; and $x_1 = D$, $x_2 = A'$ of the $x_1 - x_2$ plane. Note that the integral in Eq. (D.27) is done along a trajectory on the $x_1 - x_2$ plane and that p_3 is constant along the trajectory of any ray in a linear system (because n only depends on x_1 and x_2 and not on x_3). These projected trajectories fulfill Fermat's principle with the equivalent refractive index distributions (see Eq. (C.10)); that is,

$$\int_P^{P'} \sqrt{n^2 - p_3^2}\, dl = \text{constant} \tag{D.28}$$

Eq. (D.24) expresses that the integral in Eq. (D.27) along the projection of the ray trajectories on the $x_1 - x_2$ plane must be constant for all the rays leaving from point P with the same value of p_3. When $p_3 = 0$ (i.e., rays contained in a plane normal to the axis of translational symmetry), then Eq. (D.27) is always fulfilled. If n is constant along the whole trajectory of rays (in the case of reflectors) then the integrand in Eq. (D.27) is constant, and this equation only expresses that the projected trajectories have a constant length. Which turns to be the same as Fermat's principle.

Combining Eqs. (D.27) and (D.28) it is found again that a single lens fulfilling them cannot be symmetric with respect the projection of the line joining P and P'. An analysis similar to those done in Sections 6.1 and 6.2 can also be done for rotationally symmetric systems.

REFERENCES

Born, M., and Wolf, E. (1993). *Principles of Optics*, 6th Ed. Pergamon, NY.
Flanders, H. (1989). *Differential Forms with Applications to Physical Sciences.* Dover, NY.

Herzberger, M. (1958). *Modern Geometrical Optics.* Wiley Interscience, NY.

Herzberger, M. (1994). *Mathematics and Geometrical Optics*, Suplementary Note III in Luneburg, R. K. (1964) *Mathematical Theory of Optics*, University of Califomia Press, Berkeley.

Miñano, J. C. (1984). Application of the conservation of étendue theorem for 2D subdomains of the phase space. *Applied Optics* **23**, 2021–2025.

Miñano, J. C., Benítez, P., and Muñoz, F. (2001). Application of the 2D étendue conservation to the design of achromatic aplanatic aspheric doublets. In *Non-imaging Optics: Maximum Efficiency Light Transfer VI* (Winston, R., ed.). Vol. **46**, 11–19.

Poincaré, H. (1957). *Les Méthodes Nouvelles de la Mécanique Céleste.* Dover, NY.

Schultz, G. (1988). Aspheric surfaces. In *Progress in Optics* (Wolf, E., ed.). North Holland, Amsterdam, pp. 351–416.

Stavroudis, O. N. (1972). The optics of rays, wavefronts, and caustics. Academic, New York, NY.

Welford, W. T. (1989). *Aberrations of Optical Systems.* Adam Hilger, Bristol.

Appendix E

PERFECT OFF-AXIS IMAGING

E.1 INTRODUCTION

Let us consider a conventional axially symmetric lens system acting as a concentrator for a source at infinity (size $\pm\beta$). The edge-ray theorem, when applied to this case, states that in order to obtain an ideal imaging concentrator, it must focus perfectly (i.e., on a single image point) the rays impinging on the entry aperture of the system forming an angle $+\beta$ with the optical axis.[1] The impossibility of such perfect imaging of an off-axis point in three dimensions (3D) by an axially symmetric optical system was the subject of a theorem proven in the late 1970s (Welford and Winston, 1979). It was restricted to optical systems with homogeneous media separated by a finite number of optical surfaces and considered as exceptions some trivial cases and systems with spherical symmetry (as the sphere with its aplanatic points).

This theorem, but reduced to two dimensions (2D), or equivalently, to the meridian rays of the axially symmetric system, caused some controversy. In Welford and Winston (1978) it was stated that there was no solution (excluding also some exceptions), but in their 1979 article, which corrected the one they wrote in 1978, they admitted the solution existed, although none was provided.

The Simultaneous Multiple Surface (SMS) design method (Miñano and Gonzalez, 1992) of Nonimaging Optics (explained in Chapter 8) came into being through the attempt to solve a conjecture, whose solution would also prove the feasibility of imaging two off-axis points in 2D. It is a well-known fact that with the incidence on a single refractive surface, it is possible to focus perfectly a wavefront on a point if it is verified that for each point of that surface there passes only a single ray of the bundle. These refractive surfaces are the well-known Cartesian ovals[1] (Stavroudis, 1972) (see Figure E.1a), whose equation is given by the constancy of the optical path length from the wave front to the point.

---

[1]   Note that this condition is sufficient but too restrictive from the nonimaging requirements. The sufficient and necessary condition in this example is that those $+\beta$ rays are focused along the receiver edge (i.e., astigmatism is an allowable aberration).
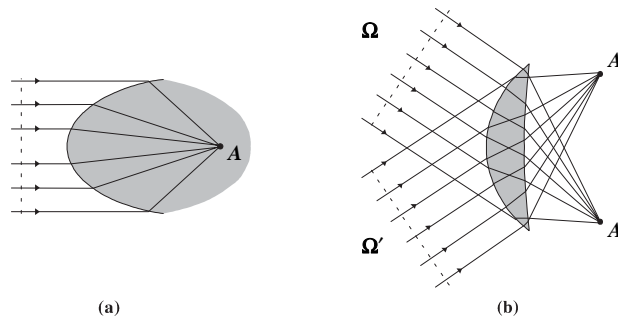
Figure E.1   (a) Cartesian oval that focuses a plane wave front on point A. (b) RR concentrator that focuses two plane wave fronts on two points, A and A^*.

The initial hypothesis of the SMS method was to affirm that it is possible with sequential incidences on two refractive surfaces to focus perfectly two wave fronts on two points, if it is verified that for each point of the two surfaces only one ray of each wave front passes. Although no formal demonstration of this hypothesis was offered, it was found to be true at least from a practical point of view, by direct calculation of the surfaces (Miñano and González, 1992). The uncertainty about the validity of this calculation as proof stems from the fact that, for the ray assignation chosen, the design procedure had no degrees of freedom to guarantee the focusing of some small fractions of the wave fronts near the optical axis. Nevertheless, an "a posteriori" ray tracing showed that these rays fit surprisingly well (demonstrating validity for that example in practice). The first solution found was an aspheric lens (in the current SMS nomenclature, RR concentrator, because of the two refracting surfaces) that made the edges of the two refracting surfaces coincide (González, 1989–1990). This lens was designed to focus on two points, A and A', two plane wave fronts, Ω and Ω', symmetric with respect to the mediatrix m of the segment AA^*, as shown in Figure E.1b. The symmetry of the problem permitted to find a symmetrical solution with regard to this mediatrix.

E.2   THE 2D CASE

However, a rigorous solution of the 2D conjecture can be obtained by means of the following assignation of rays. Consider the points P and Q in Figure E.2a located in the object and image spaces, respectively, and that the refractive index n of the lens is known. The object point P has not been placed at infinity (as in Figure E.1) for the sake of simplicity, though its location is not relevant for the proof anyway. Because of the symmetry of the lens, P' and Q', symmetric points of P and Q with respect to the optical axis, will be also perfect conjugates.

Let D_L and D_R be the left- and the right-hand sides refracting surfaces, respectively, that define the lens. The rays emerging from P undergo a first refraction on D_L and then a second refraction on D_R before impinging at Q. The procedure to calculate D_L and D_R consists of the following steps (see Figure E.2a):
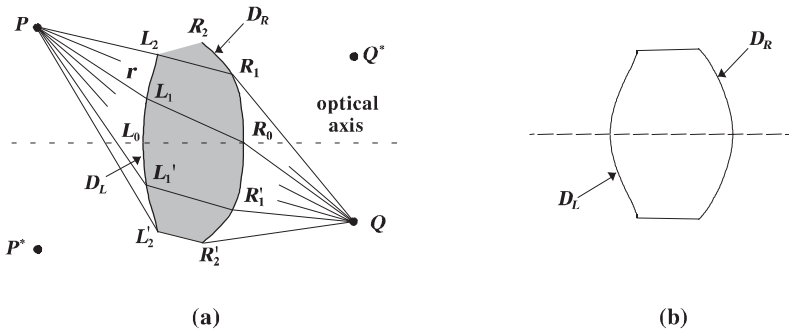
Figure E.2 (a) Design procedure of a symmetric lens that images perfectly P on Q in 2D. (b) Lens designed with $n = 1.5$, $PP^* = 3$, $QQ^* = 1.5$, dist$(PP^*, L_0) = 5.48$, dist$(QQ^*, R_0) = 2.25$, dist$(L_0, R_0) = 2.27$.

1. Choose the vertices L_0 and R_0 of curves D_L and D_R at the optical axis. For the symmetry and, furthermore, to avoid a kink at the axis of the lens, the normal vectors of the surfaces at L_0 and R_0 are chosen parallel to this axis.
2. Calculate the trajectory of the ray r that passes through Q and R_0 before the refraction at R_0. Choose the point L_1 of curve D_L on said trajectory. The normal at L_1 is found applying the condition that the trajectory of ray r must pass, before the refraction at L_1, through P. The optical path length from P to Q, $l(P, Q)$, is calculated along the whole trajectory of the ray r.
3. Choose the profile of D_L between L_0 and L_1 by interpolating a smooth curve between them (for instance, a polynomial of third degree), which verifies that its normal at L_0 and L_1 coincide with the ones found in steps 1 and 2. The symmetric curve with respect to the optical axis is also a portion of D_L and goes from L_0 to L'_1, the symmetric point of L_1.
4. Calculate the portion of D_R between R_0 and R'_1 with the condition that the rays emitted from P and refracted on D_L between L_1 and L'_1 must be focused on Q, after a second refraction on this curve. The solution to this problem is qualitatively similar to that of the Cartesian ovals and is given by the constancy of the optical path length from P to Q, calculated in step 1. The symmetric curve with respect to the optical axis is also a portion of D_R and goes from R_0 to R_1, the symmetric point of R'_1.
5. Calculate the portion of D_L between L_1 and L_2, with the condition that the rays emitted from P and refracted on this curve are refracted again on D_R between R_0 and R_1 toward Q. The dioptric D_L extends from L_2 to its symmetric point L'_2.
6. Calculate the portion of D_R between R'_1 and R'_2 with the condition that the rays emitted from P and refracted on D_L between L'_1 and L'_2 must be focused into Q after a second refraction on this curve. The dioptric D_R extends from R'_2 to its symmetric point R_2.
7. Repeating steps 5 and 6 new portions of both surfaces (not shown in Figure E.2a) can be calculated if a larger aperture is desired. Figure E.2b shows a solution designed with this procedure that adds two portions to the one depicted in Figure E.2a.
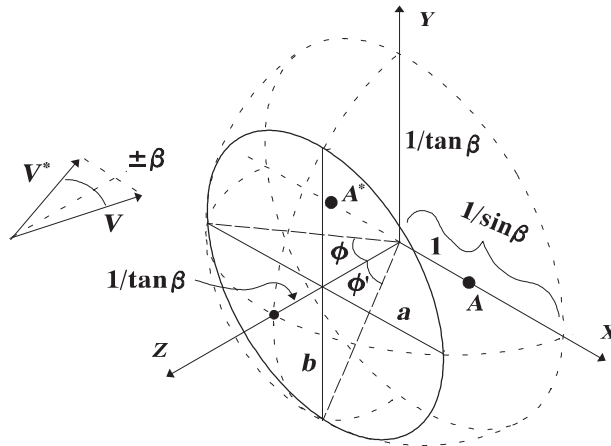
Figure E.3   The edge C_1 of the RR(3D) concentrator is an ellipse located in a plane $z =$ constant and contained in the ellipsoid of revolution with foci A and A^*.

E.3   THE 3D CASE

Returning to the three dimensional case, a new question arises: Is it possible to extend the initial conjecture of the SMS method to three dimensions? Do there exist two refractive surfaces that focus perfectly two wave fronts on two points in 3D geometry? As we shall see, the answer is once again yes (at least from a practical point of view), if it is verified, as in the 2D case, that for each point of the surfaces only one ray of each wave front passes. Notice that the surfaces cannot have rotational symmetry, as proven in Welford and Winston (1979).

Although a rigorous solution as in the 2D case could be constructed, we shall find two solutions analogous to the first found with the SMS method in 2D: two three-dimensional RR concentrators—from now on, RR(3D)—which verify that the edges of the two refractive surfaces coincide in a curve C. The RR(3D) is designed with the SMS method applied for the first time to 3D geometry.

Consider the Cartesian reference system shown in Figure E.3. We will denote p, q, and r to the direction cosines with respect to the X-, Y-, and Z-axes, respectively. Points A and A' have respective coordinates $(1, 0, 0)$ and $(-1, 0, 0)$ in this system. Consider also two plane wave fronts Ω and Ω^*, normal to the vectors $V = (\sin\beta,\ 0,\ -\cos\beta)$ and $V^* = (-\sin\beta,\ 0,\ -\cos\beta)$, respectively. The rays associated with Ω and Ω', which are contained in planes y = constant (since q is null) and form with the Z-axis angles $+\beta$ and $-\beta$, will be referred to as rays $e(+)$ and rays $e(-)$, respectively. Given that the lens will focus the rays $e(+)$ on A, the optical path length from Ω to A will be the same for all the rays, $l(\Omega, A)$. In a similar way, $l(\Omega', A')$ will be the optical path length common to all the rays $e(-)$. Moreover, given the symmetry of the data of the problem with respect to the planes $x = 0$ and $y = 0$, it is interesting to force the solution to have this same symmetry. This means that $l(\Omega, A) = l(\Omega', A')$.
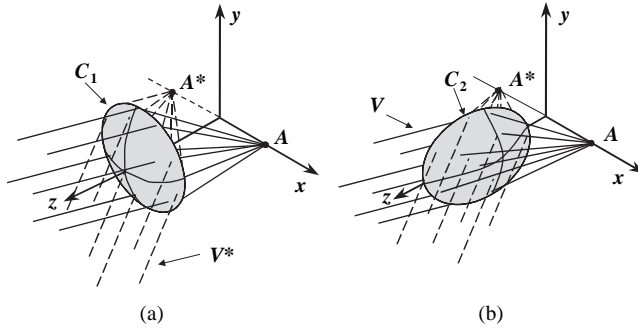
Figure E.4   Qualitative difference between the solutions with (a) edge C_1 and (b) edge C_2.

E.3.1   Calculation of the Edge

In the first place we shall calculate the curve C, the edge of the lens, applying the condition $\mathrm{l}(\Omega, A) = \mathrm{l}(\Omega', A') = l_C$ to the points of C. This is expressed as

$$d(X,A) + d(X,\Omega) = d(X,A^*) + d(X,\Omega^*) = l_C \tag{E.1}$$

There are two different curves, C_1 and C_2, fulfilling the equations (E.1). Let us consider first the case of curve C_1, which is contained in the ellipsoid of revolution with respect to the X-axis with foci A and A' (see Figure E.3) and a plane $z = z_C$:

$$x^2 \sin^2 \beta + (y^2 + z^2)\tan^2 \beta = 1 \quad z = \frac{2}{\sin 2\beta} - \frac{l_C}{\cos \beta} \equiv z_C \tag{E.2}$$

Whose intersection is the ellipse (contained in $z = z_C$)

$$\frac{x^2}{a^2} + \frac{y^2}{(a\cos\beta)^2} = 1 \tag{E.3}$$

where a, the semi-axis parallel to the X-axis, depends on z_C via $a^2 \sin^2\beta + z_C^2 \mathrm{tag}^2\beta = 1$, as deduced from the left equation in (E.2), making $y = 0$. The semi-axis of the ellipse parallel to the Y-axis is $b = a\cos\beta$. The ellipse (E.3) is also depicted in Figure E.3. The solution that has the curve C_1 as edge fulfills that both wave fronts, Ω and Ω', impinge first on a refractive surface on the left, D_L, and both emerge afterward from a refractive surface on the right, D_R, toward the object points A and A^* (see Figure E.4a).

The second curve fulfilling (1) is an ellipse in the plane $x = 0$, whose equation is

$$\frac{y^2}{\left(\dfrac{l_C}{\sin\beta}\right) - 1^2} + \frac{\left[z - \dfrac{l_C}{\sin\beta\tan\beta}\right]^2}{\dfrac{1}{\sin^2\beta}\left[\left(\dfrac{l_C}{\sin\beta}\right)^2 - 1\right]} = 1 \tag{E.4}$$

The solution that has the curve C_2 as edge fulfills that the wave front Ω impinges first on a refractive surface on the back, D_B, and emerge afterward from a refractive surface on the front, D_F, toward the object point A; while the wave front Ω'

impinges first on D_F and emerge afterward from D_B toward the object point A' (see Figure E.4b).

E.3.2   Design of the Lens

Let us consider first a lens with a curve C_1 as edge. The input parameters of the design of the RR(3D) concentrator are (a) the refractive index n, (b) the angle β, and (c) the coordinate z_C. The curve C_1 is known as soon as z_C and β are fixed. The steps to follow for calculating these two surfaces are as follows:

a. Take a finite number of points P_i ($i = 1, 2, \ldots, N$) along the curve C_1 in the quadrant $x > 0$, $y \geq 0$. With z_C and β, find the optical path length l_C with the equation (E.2).

b. Calculate the vector normal to the two surfaces at each point P_i. The unitary normal vectors N_L and N_R associated with each surface are given by the unique solution of the system of equations that is derived from applying Snell's laws to the four incidences (two per ray) in P_i. Calculate also the vector T_C tangent to C_1 at P_i, which is parallel to $N_L \wedge N_R$ ($\wedge$ denotes vector product).

c. For each point P_i the point $L_{i0} = P_i + \varepsilon U_i$ is calculated, U_i being the unitary vector perpendicular to T_C and N_L in P_i pointing toward the interior of the lens (U_i is parallel to $T_C \wedge N_L$). The point L_{i0} is "approximately" a point of the surface D_L if the constant ε is sufficiently small. Consider that the normal in L_{i0} coincides with the normal in P_i.

d. Calculate the point R_{i1} of the surface D_R with the condition that the ray $e(-)$ refracted in L_{i0} is refracted again in R_{i1} toward A^*. This calculation can be done because we know L_{i0} from step c and the optical path length l_C from Ω^* to A^* from step a. The normal in R_{i1} can be found by the (inverse) application of Snell's law.

e. Calculate the point L_{i2} of the surface D_L with the condition that the ray that passes through R_{i1} and A, before refracting in R_{i1}, comes from the ray $e(+)$ refracted in L_{i2}. Again, this calculation is based on the fact that R_{i1} is known and that the optical path length from Ω to A is l_C. Again, the normal in L_{i2} is calculated verifying Snell's law.

f. Repeat steps d and e to calculate, for each point P_i of the edge, successive points $L_{i,2k}$ and $R_{i,2k+1}$ of the surfaces D_L and D_R that belong to the quadrant $y \geq 0$, $x > 0$. The (finite) sets of points $\{L_{i,2k}\}$ and $\{R_{i,2k+1}\}$ are isolated points of the surfaces. In fact, they constitute two (finite) sets of curves C_{2k}^L and C_{2k+1}^R contained in the surfaces D_L and D_R that are obtained by increasing arbitrarily the number N of points P_i-selected in step a. This curves "touch" on the curve C_1, as shown in Figure E.5. It remains to calculate, therefore, the intermediate points, situated between the curves C_{2k}^L and C_{2k+1}^R.

g. For each point P_i, interpolate a smooth curve between the consecutive points L_{i0} and L_{i2} that is perpendicular to the normal vectors at these points. Since L_{i0} and L_{i2} are points that are very close to one another and with very similar normals, the interpolated curve is practically a straight line.

h. Choose a finite number of points L_{i0}^j ($j = 1, 2, \ldots, M$) of the curve interpolated between L_{i0} and L_{i2} and repeat M times steps d and e, taking as initial point of the process L_{i0}^j. In this way the sets of points $\{L_{i,2k}^j\}$ and $\{R_{i,2k+1}^j\}$ of the surfaces are obtained.
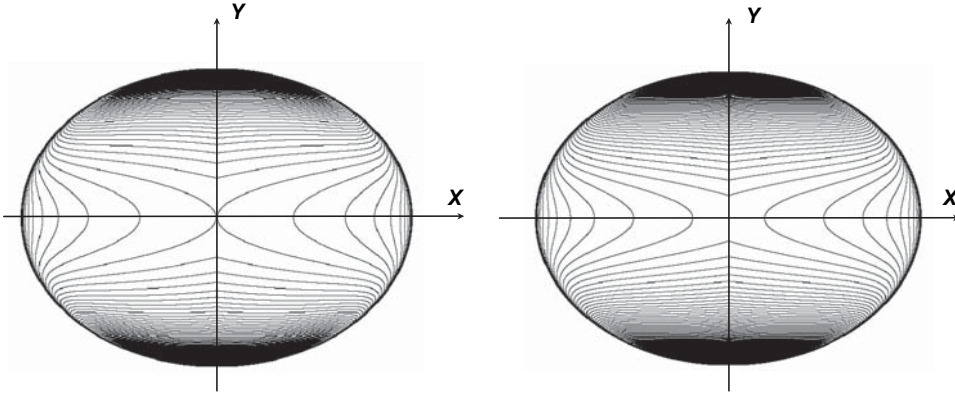
Figure E.5   When $N \rightarrow \infty$, the calculation process that starts with curve C generates curves contained in the surfaces D_L (left) and D_R (right). Design parameters: $n = 1.5$, $\beta = \pm 40°$, $z_C = 0.8801$.

The design of the RR(3D) concentrator is finished, since by choosing sufficiently large M and N the surfaces can be defined with any specific precision within the quadrant $x > 0$, $y \geq 0$. In fact, in step h points of curves (which result from taking very high N) have been calculated that are intermediate between those shown in Figure E.5a. Specifically, between each pair of successive curves, C^L_{2k} and C^L_{2k+2} on the surface D_L, and C^R_{2k+1} and C^R_{2k+3} on D_R, are defined M curves $C^{L,j}_{2k}$ y $C^{R,j}_{2k+1}$ to which the calculated points $\{L^j_{i,2k}\}$ and $\{R^j_{i,2k+1}\}$ belong. The lens is completed by symmetry with respect to the planes $x = 0$ and $y = 0$.

   The procedure just described is essentially valid for the case of edge C_2. The input parameters in this case are (a) the refractive index n, (b) the angle β that defines the plane wave fronts, and (c) the Z-coordinate of the ellipse center z_{ec} (which can be found in (E.4)). The only qualitative difference between both designs is that the lens is symmetric with respect to the planes $x = 0$ and $y = 0$, and as C_2 is placed in $x = 0$, we get

 i.   In step f, the successive points of the surfaces are calculated up to the point that they converge to the opposite side of curve C_2 (a condition of convergence must be stated). Depending on the initial parameters of the design, such convergence may not occur.
 ii. The symmetry of the problem will make that surface D_B is the symmetric of D_F with respect to the plane $x = 0$.

E.3.3   Results

Following the procedure described in Section E.3.2, a concentrator RR(3D) has been calculated with edge curve C_1, with input parameters $n = 1.5$, $\beta = \pm 40°$, and $z_C = 0.8801$. Figure E.6 shows a plan view of the surfaces D_L and D_R. Other data of the lens are $a = 1.0489$ and $b = 0.8035$. Figure E.7 shows the cross sections of the lens at the planes $x = 0$ and $y = 0$.

   In order to analyze the performance of the lens as an off-axis concentrator, a receptor disk of radius $r = a/36.83$ was positioned in the plane $z = 0$ with its center
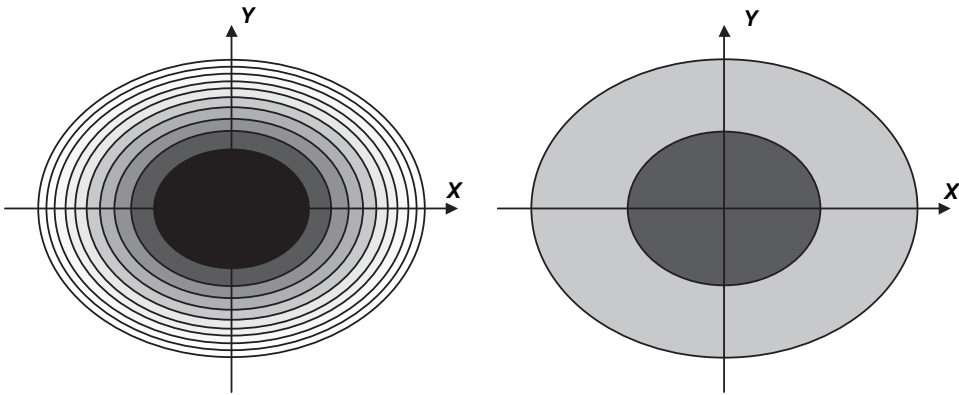
Figure E.6   Plan view of the surfaces D_L (left) and D_R (right). The changes in grays occur at intervals of $z_C/20$, with z_C being the height of the lens edge, $z_C = 0.8801$.
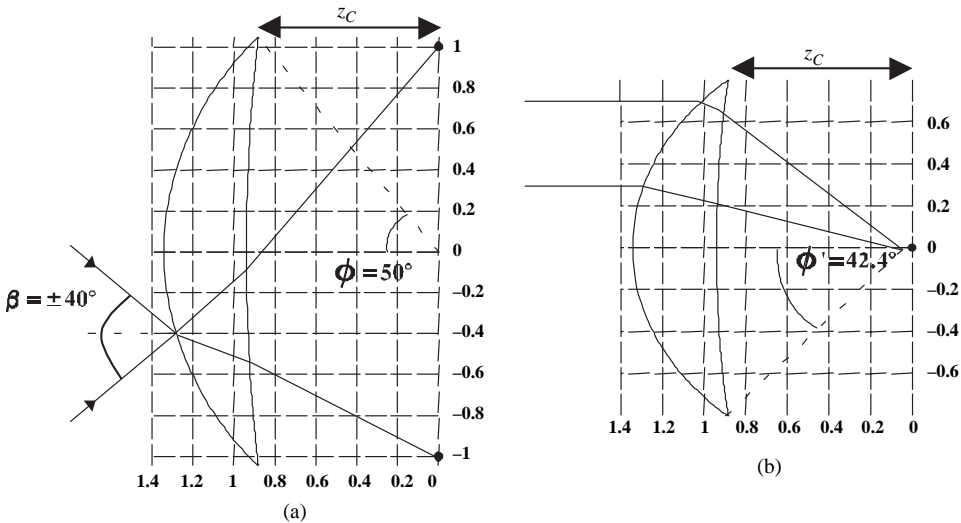
Figure E.7   Intersections of the designed RR(3D) with the planes (a) $y = 0$ and (b) $x = 0$.

at point A. The geometrical concentration is, therefore, $C_g = (\pi ab)/(\pi r^2) = 1{,}243\times$. By means of three-dimensional ray tracing the angular transmission of the system was calculated for the rays with $p = \sin 40°$ (sagital section) and the rays with $q = 0$ (meridian section). The results are shown in Figure E.8. The transmission for the direction of the design ($p = \sin 40°$, $q = 0$) is 100% and stays above 50% for angles of incidence that differ up to $0.75°$ from this direction.

It is interesting to point out that the bundles of the design of the RR(3D) cannot be considered as paraxial. This means that a classical lens with symmetry of revolution designed as a system for forming a paraxial image will present very high aberrations for these angles. Nevertheless, it is possible to design a rotational lens for nonparaxial functioning, which forms a stigmatic image of the meridian
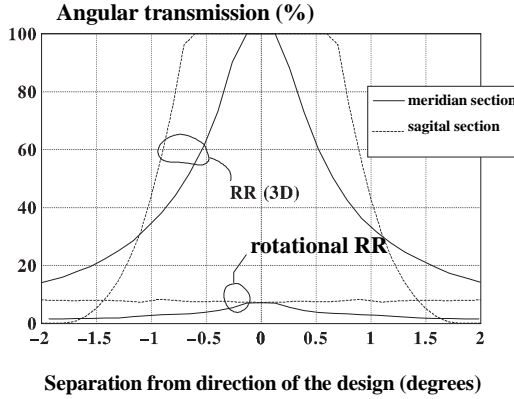
Figure E.8   Angular transmission for the rays with $q = 0$ (meridian section) and the rays with $p = \sin 40°$ (sagital section) of the lens RR(3D) and of a rotational lens RR coincident with RR(3D) in $y = 0$. The receptor is a disk centered on A of radius $a/36.83$.

rays that impinge with $\beta = \pm 40°$ on points A and A', and has a on-axis rim angle $\varphi = \pm 50$ (see Figure E.7). Such a lens is in fact an RR concentrator designed in 2D geometry—RR(2D)—whose meridian section coincides with the section $y = 0$ of the lens RR(3D), shown in Figure E.7a. If for this lens we calculate the angular transmission with the same receiver as before, we obtain a transmission for the desired direction ($p = \sin 40°$, $q = 0$) below 10%. This happens because, despite the fact that the design guarantees the transmission of all the meridian rays, the majority of the "skew" rays (those that are not contained in a meridian plane) hit outside of the receptor. In other words, the design criterion of the rotational RR is not optimal, given that the transmission of all the rays should be maximized, not only that of the meridian rays. From the point of view of image formation, the "spot" is clearly astigmatic, with the plane of the receptor situated at the tangential focus instead of in the plane of minimal confusion.

The contour lines of the surfaces of the RR(3D), shown in Figure E.6, approximate to ellipses. This suggests that if from the concentrator RR(2D) we generate a three-dimensional concentrator with "elliptical" symmetry instead of symmetry of revolution, we could expect better results. Figure E.9 shows the results of the ray tracing on the lens RR(3D) and on the elliptical lens, where it can be seen that while the results from this lens are inferior to those of the RR(3D), they are quite superior to those of the rotational lens. From a practical point of view, the elliptical lens may be easier to manufacture than the RR(3D).

The ray tracing has been done on two-variable polynomial approximations of the surfaces of the form

$$z = \sum_{i,j} a_{ij} x^{2i} y^{2j} \qquad (E.5)$$

The selected approximation criterion is the least-square fitting of the angle formed by the normal vectors at 2,500 uniformly distributed points of the surfaces. This seems to be the most appropriate criterion, because the accuracy in the performance of an optical device is mainly sensitive to slope errors (as shown in Chapter 15). In our case, polynomial approximations of 14th degree show a excellent accu-
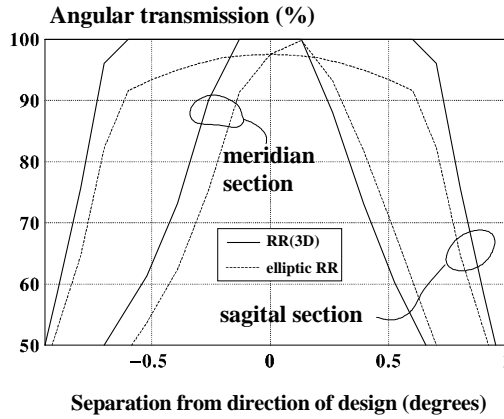
Figure E.9   As Figure E.7 for the concentrators RR(3D) and the elliptical RR. Note that the scale of the axis of ordinates is different from that of Figure E.8.

(a)   (b)
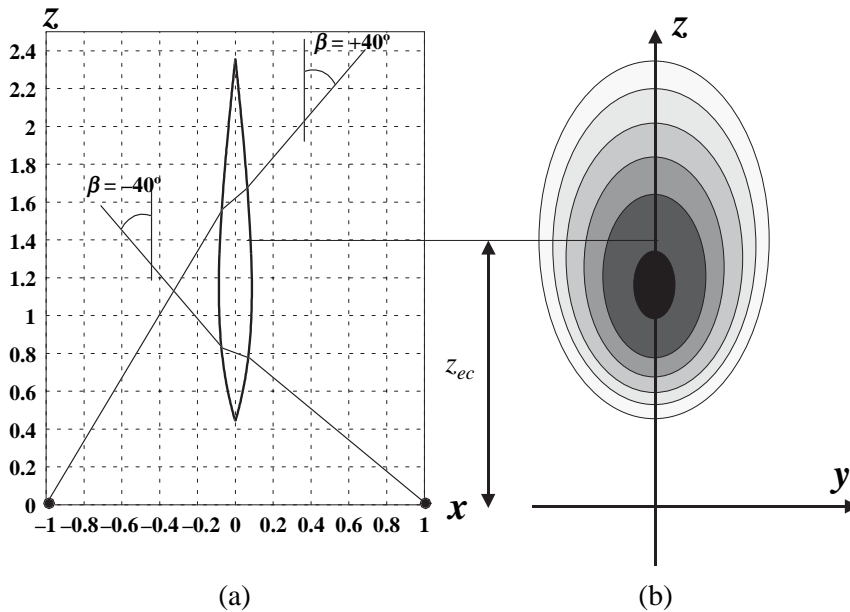
Figure E.10   Views of the RR(3D) design with edge C_2: (a) section $y = 0$, and (b) plan contour plot. The changes in grays occur at intervals $\Delta x = 0.0166$. Design parameters: $n = 1.5$, $\beta = \pm 40°$, $z_{ec} = 1.4$.

racy: for the RR(3D) in Figure E.7 they lead to standard deviations—in degrees—from the theoretical normals of $\sigma_{\text{Sup}} = 7.813 \; 10^{-4}$ and $\sigma_{\text{Inf}} = 0.91 \; 10^{-4}$ for the superior and inferior surfaces, respectively.

   The quality of these polynomial approximation can be also checked by doing a ray tracing for the direction of design and calculating the meridian and sagittal spot radius. The meridian (also tangential) spot radius (σ_T) can be defined in our case as the standard deviation of the Y-coordinate of the points of incidence of the rays on the plane $z = 0$. The sagittal spot radius (σ_S) is defined in a similar way,
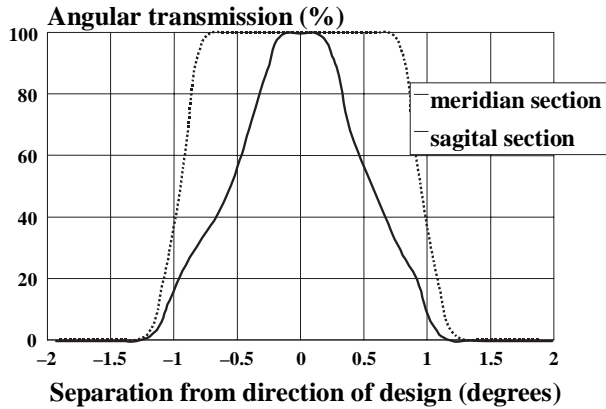
Figure E.11 Angular transmission for the rays with $q = 0$ (meridian section) and the rays with $p = \sin 40°$ (sagittal section) of the RR(3D) with edge C_2. The receptor is a disk in the plane $z = 0$ and centered on A of radius 0.02848.

but the coordinate X applies. In the case of the RR(3D) of Figure E.7, with 10^5 rays we get $\sigma_T = 0.00573\,\mu$m and $\sigma_S = 0.0198\,\mu$m. These values are very far from the diffraction limit of the device for the visible spectrum.

Figure E.10 shows the cross section and the contour plot of an RR(3D) with edge C_2. The parameters of the lens of Figure E.10 are $n = 1.5$, $\beta = \pm40°$, and $z_{ec} = 1.4$. Notice that the two refracting profiles are symmetric and its maximum thickness does not lie at the ellipse center.

The ray-tracing analysis on this lens has also been done using a similar polynomial approximation approach. In this case only one surface needs to be fitted (the other one is symmetric) with the expression

$$x = \sum_{i,j} b_{ij} y^{2i} z^j \tag{E.6}$$

The performance of the lens as an off-axis concentrator, using the same receptor as in the former cases, can be seen in Figure E.11.

REFERENCES

Benítez, P., Mohedano, R., and Miñano, J. C. (1997). Manufacturing tolerances for nonimaging concentrators. In *Nonimaging Optics: Maximum Efficiency Light Transfer IV* (Winston, R., ed.). Proceedings of SPIE, 98–109.

González, J. C. (1989–1990). *Diseño de lente y concentrador anidólico.* Proyecto Fin de Carrera, ETSI Telecomunicación, UPM, Madrid.

Miñano, J. C., and González, J. C. (1992). New method of design of nonimaging concentrators. *Appl. Opt.* **31**, 3051–3060.

Stavroudis, O. N. (1972). *The Optics of Rays, Caustics, and Wavefronts.* Academic Press, New York.

Welford. W. T., and Winston, R. (1979). On the problem of ideal flux concentrators: Addendum. *J. Opt. Soc. Am. A.* **69**, 367.

Williams, C. S., and Becklund, O. A. (1989). *Introduction to the Optical Transfer Function.* Wiley, New York, pp. 51, 77.

THE LUNEBURG LENS

The Luneburg lens, discussed in Chapter 3, is not, of course, of any use as a practical concentrator. It is, however, the simplest example of an ideal concentrator with maximum theoretical concentration for collecting angles up to $\pi/2$, and therefore we develop its theory and general properties in this appendix. The main reference is to Luneburg's own work (Luneburg, 1964). Some of the general geometrical optics background can also be found in Born and Wolf (1975), and extensions are given by Morgan (1958) and Cornbleet (1976).

Our starting point is the differential equation of the light rays. Let s be distance along a ray measured from some fixed origin, let r be the position vector of a point on a ray, and let $n(r)$ be the refractive index, varying continuously as a function of r. Then it can easily be shown [e.g., Born and Wolf (1975, Section 3.2.1)] that the rays are given by solutions of

$$\frac{d}{ds}\left(n(\mathbf{r})\frac{dr}{ds}\right) = grad\ n(\mathbf{r}) \tag{F.1}$$

Let the refractive index distribution have spherical symmetry. Then we can write $n(r) \equiv n(r)$, where r is a radial coordinate from the origin. Also, on account of the spherical symmetry, any light ray lies wholly in a plane through the origin, and we can describe its path by the polar coordinates (r, θ).

In order to do this we note that dr/ds is a unit vector along the tangent to the ray—say, s. Now we have

$$\frac{d}{ds}(\mathbf{r} \times n\mathbf{s}) = \mathbf{s} \times n\mathbf{s} + \mathbf{r} \times \frac{d(n\mathbf{s})}{ds} = \mathbf{r} \times grad\ n$$

from Eq. (F.1). But by spherical symmetry grad n is parallel to r and so we have

$$\frac{d}{ds}(r \times ns) = 0$$

and a first integral of Eq. (F.1) is

$$r \times ns = \text{const.}$$
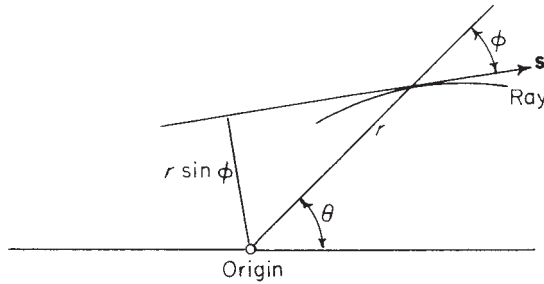
Figure F.1   Notation for ray paths in a medium with spherical symmetry.
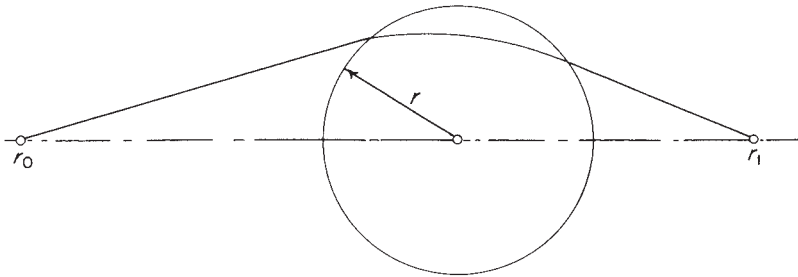
Figure F.2   Luneburg's general problem of an aberrationless unit sphere.

This is a vectorial equation that is equivalent to establishes the conservation of the skew relative to two orthogonal axes.

Since the rays are plane curves as just noted, this gives, say,

$$nr\sin\varphi = const. = h \qquad (F.2)$$

where ϕ is as in Figure F.1. Since

$$\tan\phi = r\frac{d\theta}{dr}$$

Eq. (F.2) yields

$$\frac{d\theta}{dr} = \frac{h}{r\sqrt{n^2r^2 - h^2}} \qquad (F.3)$$

and thus if $n(r)$ is specified, the ray paths are obtained by quadrature

$$\theta_1 - \theta_0 = h\int_{r_0}^{r_1} \frac{dr}{r\sqrt{n^2r^2 - h^2}} \qquad (F.4)$$

If we wish to find an index distribution $n(r)$ that will act as an aberration-free medium, we have to regard Eq. (F.4) as an integral equation for the function $n(r)$. Luneburg did this and considered the use of an index distribution inside the unit sphere and with index unity at the surface of the sphere. He formulated the conditions for focusing from any point r_0 to another r_1 as in Figure F.2, and he obtained an explicit solution for the case of present interest, $r_0 = \infty$, $r_1 = 1$.
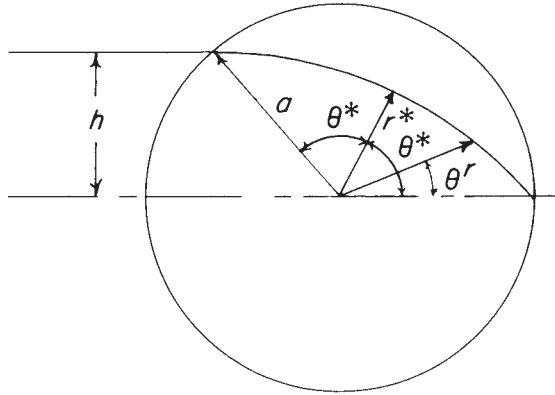
Figure F.3   Luneburg's problem for $r_0 = \infty$, $r_1 = 1$.

Luneburg's derivation is very complex and closely argued. Rather than reproduce his work we shall simply verify the solution he gave.

Let the lens have radius a and let the index vary in such a way that r has only one minimum value for the ray path in the lens—say, r^*—as in Figure F.3. Then from Eq. (F.2) we have, using values at the beginning of the trajectory in the lens,

$$nr\sin\phi \equiv a\cdot\frac{h}{a} = h$$

and thus

$$n(r^*)r^* = h \tag{F.5}$$

Since the two halves of the trajectory on either side of the radius r^* must be mirror images we have

$$\theta^* = \frac{\pi}{2} - \frac{1}{2}\arcsin\left(\frac{h}{a}\right) \tag{F.6}$$

and thus Eq. (F.4) takes the form

$$\frac{\pi}{2} - \frac{1}{2}\arcsin\left(\frac{h}{a}\right) = h\int_{r^*}^{a}\frac{dr}{r\sqrt{n^2r^2 - h^2}} \tag{F.7}$$

to be satisfied by a function $n(r)$ for any $h < a$.

Luneburg gave the distribution

$$n(r) = \sqrt{2 - \frac{r^2}{a^2}} \tag{F.8}$$

as the solution of this integral equation. If we substitute for $n(r)$, we have to evaluate

$$h\int_{r^*}^{a}\frac{dr}{r\sqrt{2r^2 - \dfrac{r^4}{a^2} - h^2}} \tag{F.9}$$

to verify the solution. From Eq. (F.5) we find

$$r^{*2} = a^2\left[1 - \sqrt{1 - \frac{h^2}{a^2}}\,\right] \tag{F.10}$$

and if we make the change of variable

$$r^2 = a^2\left[1 - \sin\beta\sqrt{1 - \frac{h^2}{a^2}}\,\right] \tag{F.11}$$

in the integral, it becomes

$$h\int_0^\pi \frac{d\beta}{2a\left(1 - \left[1 - \sin\beta\sqrt{1 - \dfrac{h^2}{a^2}}\,\right]\right)} \tag{F.12}$$

The indefinite integral has the value

$$\tan^{-1}\left\{\frac{\tan\dfrac{\beta}{2} - \sqrt{1 - \dfrac{h^2}{a^2}}}{\dfrac{h}{a}}\right\} \tag{F.13}$$

and on substituting the limits of integration and recalling that $h/a = \sin 2\theta^*$, we find that the value is θ^*, as required by Eq. (F.7). We have also obtained the actual ray paths, since these are specified by θ as a function of r for given h by

$$\theta_1 - \theta_0 = \tan^{-1}\left\{\frac{\tan\dfrac{\beta}{2} - \sqrt{1 - \dfrac{h^2}{a^2}}}{\dfrac{h}{a}}\right\}\Bigg|_{r_0}^{r_1} \tag{F.14}$$

with $\sin\beta = [1 - (r^2/a^2)](1 - h^2/a^2)^{-1/2}$. Figure F.4 shows several rays plotted to scale.

In order to use the Luneburg lens as an ideal concentrator we choose an input angle θ_i less than $\pi/2$ and use the opposite spherical cap as the absorbing surface,
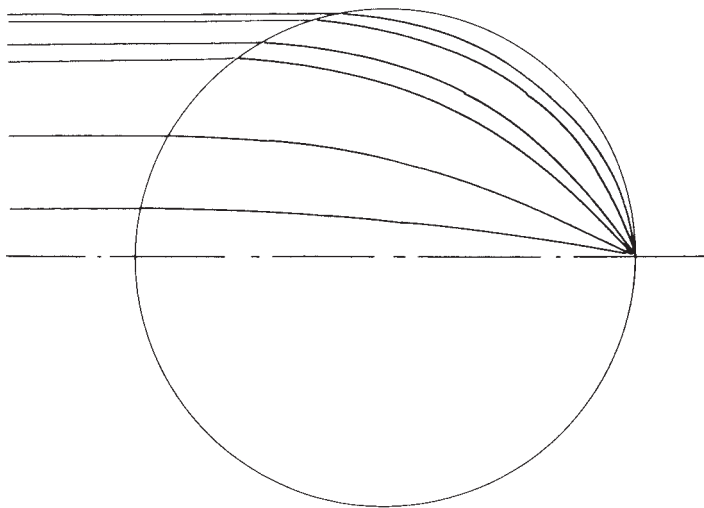
Figure F.4   Rays in the Luneburg lens.
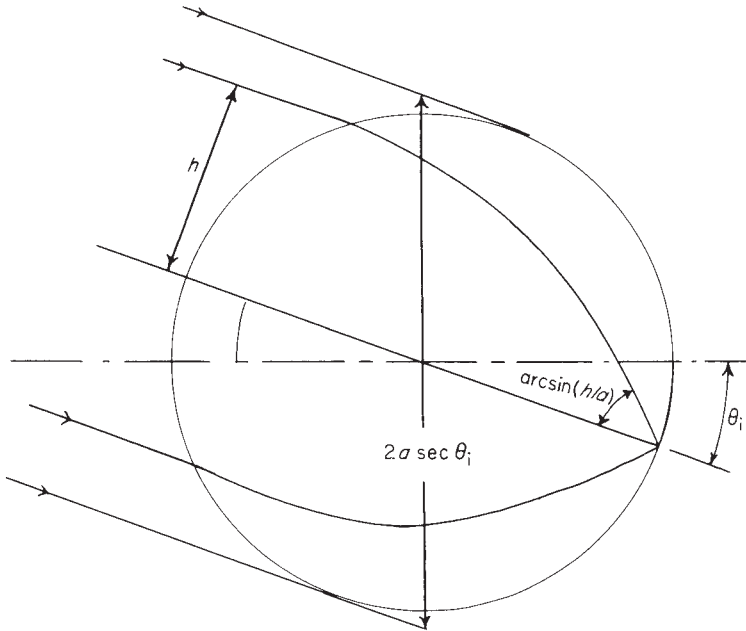
Figure F.5   The Luneburg lens as a concentrator of maximum theoretical concentration ratio for any entry angle θ_i less than $\pi/4$.

as in Figure F.5. A ray incident at height h from the center emerges at an angle to the surface arcsin (h/a) so that at any point on the exit surface rays emerge in all directions up to $\pi/2$ from the normal and the sines of the angles are distributed uniformly. Thus, the exit surface is filled with rays at all possible angles, and this must be in some sense a system with maximum theoretical concentration ratio. We can make this more explicit by calculating the entering étendue. This is, from Figure F.5, the area $\pi a^2 \sec\theta$ integrated over the direction cosines of the incoming beam—that is, it is

$$\int_0^{\theta_1}\int_0^{2\pi} \pi a^2 \sec\theta \, d(\sin\theta\cos\phi)\, d(\sin\theta\sin\phi)$$

where φ is an azimuthal angle. The value of this is $2\pi^2 a^2(1 - \cos\theta_i)$. Also, the integral over the direction cosines alone is

$$\int_0^{\theta_1}\int_0^{2\pi} d(\sin\theta\cos\phi)\, d(\sin\theta\sin\phi)$$

and this is $\pi\sin^2\theta_i$. Thus the effective entry area must be $\pi a^2/\cos^2(1/2)\theta_i$.

The area of the spherical cap that forms the exit surface is $4\pi a^2 \sin^2(1/2)\theta_i$ so that the concentration ratio is $1/\sin^2\theta_i$, the theoretical value for a collecting angle θ_i. Since all entering rays emerge, we have a 3D concentrator with maximum theoretical concentration ratio.

We had to proceed in this slightly oblique way because the collecting aperture of the lens shifts with angle and because the exit surface is not planar. Thus, the correspondence with systems with well-defined and plane entry and exit apertures is not perfect.

Appendix G

THE GEOMETRY OF THE BASIC COMPOUND PARABOLIC CONCENTRATOR

It is probably simples to obtain the basic properties of the CPC from the equation of the parabola in polar coordinates. Figure G.1 shows the coordinate system. The focal length f is the distance AF from the vertex to the focus. The equation of the parabola is, then,

$$r = \frac{2f}{1-\cos\phi} = \frac{f}{\sin^2\dfrac{\phi}{2}} \qquad\qquad \text{(G.1)}$$

This result is given in elementary texts on coordinate geometry, and it may be verified by transforming to polars the more familiar cartesian form, $z = y^2/4f$ with axes as indicated in the figure.

We apply this to the design of the CPC as in Figure G.2. We first draw the entrance and exit apertures PP' and QQ' with the desired ratio in aperture between them, and we choose the distance between them so that an extreme ray PQ' (or $P'Q$) makes the maximum collecting angle θ_i with the concentrator axis. Then according to Section 4.3 the profile of the CPC between P' and Q' is a parabola with axis parallel to PQ' and with focus at Q, and this parabola can be expressed in terms of the polar coordinates (r, φ) as on the diagram.

For the exit aperture we have, using Eq. (G.1)

$$QQ' = \frac{2f}{1-\cos(\pi/2+\theta_i)}$$

so that, if $QQ' = 2a'$, in our usual notation

$$f = a'(1+\sin\theta_i) \qquad\qquad \text{(G.2)}$$

Next we find

$$QP' = \frac{2f}{1-\cos 2\theta_i}$$

or

Figure G.1   The parabola in polar coordinates with origin at the focus, $r = 2f/(1 - \cos\phi) = f/\sin^2(\tfrac{1}{2})\,\phi$.

Figure G.2   Design of the CPC.

$$QP' = \frac{a'(1+\sin\theta_i)}{\sin^2\theta_i} \tag{G.3}$$

Thus,

$$a + a' = QP'\sin\theta_i = a'\frac{(1+\sin\theta_i)}{\sin\theta_i}$$

so that $a = a'/\sin\theta_i$, as required. Finally, for the length of the concentrator

$$L = QP' \cos\theta_i = a' \frac{(1+\sin\theta_i)}{\tan\theta_i \sin\theta_i} = \frac{a+a'}{\tan\theta_i} \tag{G.4}$$

This, of course, is again as required by the way we set up the design of the CPC.

It is obvious from the geometry of the initial requirements for the CPC that the profile at P and P' should have its tangent parallel to the axis. This also can be verified from the preceding formulation. For the tangent of the angle between the curve and the radius vector is, in polars,

$$\tan\psi = r\frac{d\phi}{dr} = \frac{2f}{r\sin\phi}$$

for the parabola, and if we put $r = QP'$ and $\phi = 2\theta_i$, we find

$$\tan\psi = \tan\theta_i$$

as expected.

The parametric representation of the CPC profile given in Section 4.3 is easily obtained from Figure G.2 if we take an origin for cartesian coordinates at the center of the exit aperture and z-axis along the concentrator axis. We have from the figure

$$y = r\sin(\phi-\theta_i) - a' = \frac{2f\sin(\phi-\theta_i)}{1-\cos\phi} - a'$$

$$= \frac{2a'(1+\sin\theta_i)\sin(\phi-\theta_i)}{1-\cos\phi} - a'$$

$$z = r\cos(\phi-\theta_i) = \frac{2a'(1+\sin\theta_i)\cos(\phi-\theta_i)}{1-\cos\phi}$$

Appendix H

THE θ_i/θ_o CONCENTRATOR

Figure H.1 shows the concentrator with appropriate notation. From the triangle $QQ'S$ we have

$$QS = 2a' \frac{\cos \dfrac{\theta_o - \theta_i}{2}}{\sin \dfrac{\theta_o + \theta_i}{2}} \tag{H.1}$$

Also, since the angle $R\hat{Q}S$ is $\pi - \theta_o - \theta_i$, the triangle RQS is isosceles, and so $QR = QS$.

From the polar equation of the parabola,

$$QR = \frac{2f}{1 - \cos(\theta_o + \theta_i)} \tag{H.2}$$

so from Eq. (H.1) we find for the focal length of the parabola

$$f = 2a' \cos \frac{\theta_o - \theta_i}{2} \sin \frac{\theta_o + \theta_i}{2} = a'(\sin \theta_i + \sin \theta_o) \tag{H.3}$$

Again using the polar equation of the parabola,

$$QP' = \frac{2f}{1 - \cos 2\theta_i} = a' \frac{\sin \theta_i + \sin \theta_o}{\sin^2 \theta_i}$$

and

$$a + a' = QP' \sin \theta_i = \frac{a'(\sin \theta_i + \sin \theta_o)}{\sin \theta_i} \tag{H.4}$$

From Eq. (H.4) we find immediately

$$\frac{a}{a'} = \frac{\sin \theta_o}{\sin \theta_i} \tag{H.5}$$

so that the θ_i/θ_o concentrator is actually ideal in two dimensions.
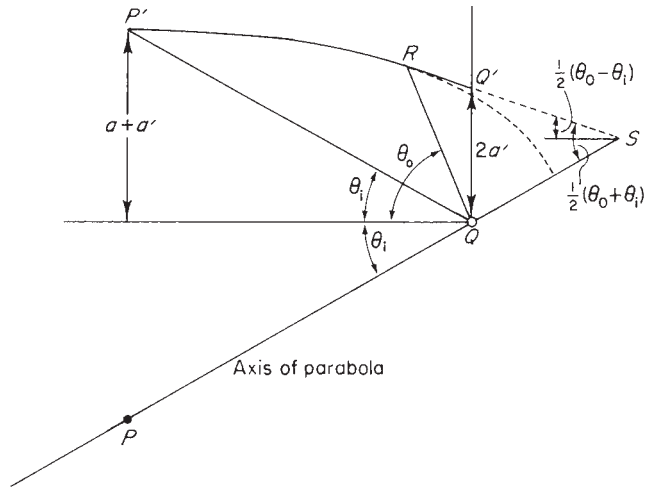
Figure H.1   Design of the θ_i/θ_o concentrator.

The overall length of the concentrator is, from the figure,

$$L = (a + a')\cot\theta_i \qquad\qquad\qquad (\text{H.6})$$

just as for the basic CPC ((Eq. (G.4)).

Appendix I

THE TRUNCATED COMPOUND PARABOLIC CONCENTRATOR

In this appendix we give the derivations of the formulas in Section 5.3. We show in Figure I.1 the parabolic section in polar coordinates. For these coordinates we recall (see Appendix G) that the equation of the parabola is

$$r = \frac{2f}{1 - \cos\phi} = \frac{f}{\sin^2 \dfrac{\phi}{2}}$$

$$f = a'(1 + \sin\theta_i)$$

(I.1)

If s is the arc length of the parabola, then

$$ds^2 = dr^2 + r^2 d\phi^2$$

so that

$$\frac{ds}{d\phi} = \sqrt{r^2 + (dr/d\phi)^2}$$

or

$$\frac{ds}{d\phi} = \frac{f}{\sin^3 \dfrac{\phi}{2}}$$

(I.2)

from Eq. (I.1). We have to integrate this to find the arc length, which is proportional to the reflector area in a 2D system. We have to evaluate the indefinite integral

$$s = f \int \frac{d\phi}{\sin^3 \dfrac{\phi}{2}}$$

On making the change of variable

473
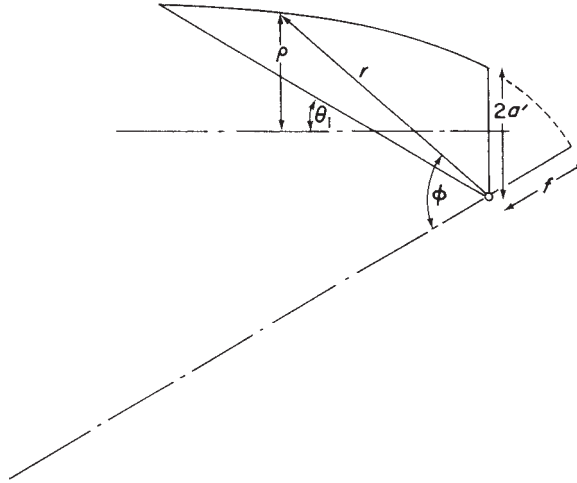
Figure I.1   Construction for the truncated CPC.

$$\cosh u = \frac{1}{\sin\dfrac{\phi}{2}}$$

we obtain

$$s_T = -f\left\{\frac{\cos\dfrac{\phi}{2}}{\sin^2\dfrac{\phi}{2}} - \ln\left(\tan\dfrac{\phi}{4}\right)\right\}\Bigg|_{\phi_T}^{\theta_i+\pi/2} \tag{I.3}$$

The radius of the entrance aperture is simply

$$a_T = r\sin(\phi - \theta_i) - a'$$

which gives Eq. (5.18) immediately. Thus, Eq. (5.22) for s_T/a_T is obtained. Likewise the length L_T of the truncated CPC is

$$L_T = r\cos(\phi - \theta_i)$$

from which Eq. (5.19) follows immediately, and thence Eq. (5.21).

For the 3D CPC the element of area of reflector is

$$dA = 2\pi\rho ds$$

where ρ is the radius of the CPC at the current point. Since ρ is $r\sin(\phi - \theta_i) - a'$ we have

$$dA = 2\pi\left\{\frac{f\sin(\phi-\theta_i)}{\sin^2\dfrac{\phi}{2}} - a'\right\}\frac{f}{\sin^3\dfrac{\phi}{2}}\,d\phi$$

This can be integrated by making the same change of variable, cosh $u = \mathrm{cosec}(\phi/2)$, and the result for the ratio of area of collection surface divided by area of collecting aperture is

$$
\frac{2f}{a_T^2}\left[f\sin\theta_i\left\{ \frac{3}{4}\ln\cot\frac{\phi}{4} + \frac{3}{4}\frac{\cos\frac{\phi}{2}}{\sin^2\frac{\phi}{2}} + \frac{1}{2}\frac{\cos\frac{\phi}{2}}{\sin^4\frac{\phi}{2}} \right\} \right.
$$
$$
\left. -(2f\sin\theta_i - a')\left\{ \ln\cot\frac{\phi}{4} + \frac{\cos\frac{\phi}{2}}{\sin^2\frac{\phi}{2}} \right\} - \frac{4f\cos\theta_i}{3\sin^3\frac{\phi}{2}} \right]_{\phi_T}^{\theta_i+\pi/2}
$$

(I.4)

Appendix J

THE DIFFERENTIAL EQUATION FOR THE 2D CONCENTRATOR PROFILE WITH NONPLANE ABSORBER

It was shown in Chapter 5 that the concentrator profile that gives the theoretical maximum concentration is obtained in two sections. The first section, that shadowed from the direct rays at angles less than θ_i, the maximum input angle, is an involute of the absorber cross section. The second section is such that rays at θ_i are tangent to the absorber after one reflection at the concentrator surface.

We suppose the absorber surface to be specified by its polar coordinates r, θ, as in Figure J.1. The current point P on the profile is at a distance t from the corresponding tangent point on the absorber. The condition for reflection is

$$\theta_i + \theta + \phi + 2\psi = 2\pi \tag{J.1}$$

The coordinates of the point P are

$$y = r\sin\theta - t\sin(\theta + \phi)$$
$$z = -r\cos\theta + t\cos(\theta + \phi) \tag{J.2}$$

and, as usual,

$$\frac{dy}{dz} = \tan\psi = -\tan\frac{\theta_i + \theta + \phi}{2} \tag{J.3}$$

Thus, on differentiating Eq. (J.2) and using Eq. (J.3) we obtain

$$\frac{\dot{r}\sin\theta + r\cos\theta - \dot{t}\sin(\theta + \phi) - t\cos(\theta + \phi)(1 + \phi)}{-\dot{r}\cos\theta + r\sin\theta + \dot{t}\cos(\theta + \phi) - t\sin(\theta + \phi)(1 + \phi)} + \tan\frac{\theta_i + \theta + \phi}{2} = 0 \tag{J.4}$$

In this equation t is the dependent variable and θ the independent variable; r and ϕ are both known functions of θ. Equation (J.4) is therefore a first-order linear differential equation for the profile that could be solved by well-known methods.

For the involute portion of the profile we have the classical result
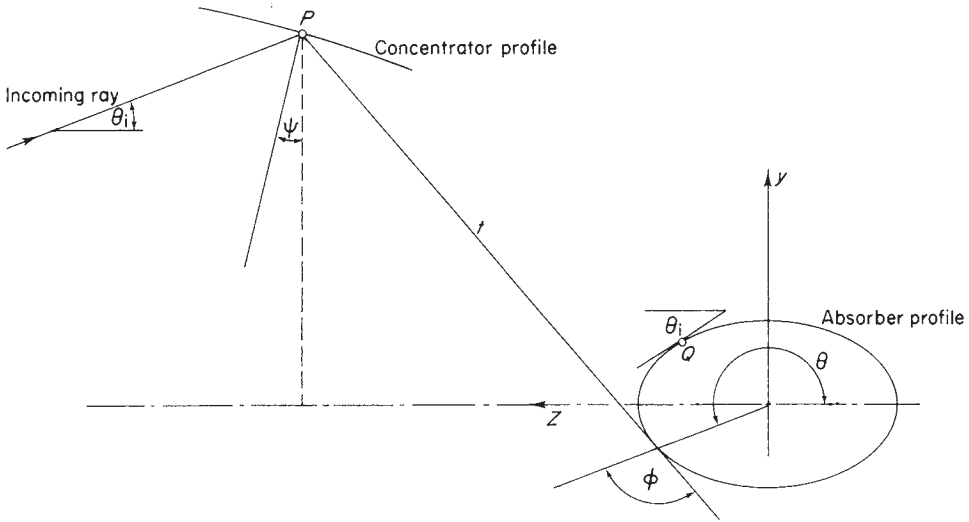
Figure J.1   The differential equation for a 2D concentrator for a nonplane absorber.

$$\frac{dt}{d\theta} = \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}$$
(J.5)

for the current point.

We can, however, write down the solution to Eq. (J.4) in closed form using the "string method" described in Chapter 5. Thus, let s be the perimeter of the absorber profile measured from the point corresponding to $\theta = 0$; θ_0, is the polar angle of the point Q that separates the shadowed from the directly viewed portion of the absorber. Then, by simply imposing the condition of constant string length on the geometry of Figure J.1, we obtain for the portion $\theta > \theta_0$

$$t = \frac{[s + s(\theta_0)] - r\cos(\theta - \theta_i) + r(\theta_0)\cos(\theta_0 - \theta_i)}{[1 - \cos(\theta + \phi - \theta_i)]}$$
(J.6)

For the involute portion $(\theta < \theta_0)$ we have

$$t = s$$
(J.7)

Rabl (1976a) set up these differential equations for the special case of great practical importance when the absorber cross section is circular. In this case $\dot{r} = 0$, $\phi = \pi/2$, and Eq. (J.4) reduces to the form

$$\frac{dt}{d\theta} - t\tan\left[\frac{\theta - \theta_i - \pi/2}{2}\right] - r = 0$$

or, setting $\frac{1}{2}(\theta - \theta_i - \pi/2) = \beta$

$$(dt/d\beta) - 2t\tan\beta = 2r$$
(J.8)

As Rabl showed, this can be integrated by making the substitution

$$t = u/\cos^2\beta$$

to give

$$t = \frac{r[(\theta - \theta_i - \pi/2) - \cos(\theta - \theta_i) + const]}{1 + \sin(\theta - \theta_i)} \qquad \text{(J.9)}$$

To determine the constant, we note that for $\theta = \theta_i + \pi/2$, we desire $t = r(\theta_i + \pi/2)$. This fixes the value of the constant in Eq. (J.9) and gives

$$t = \frac{r[(\theta + \theta_i + \pi/2) - \cos(\theta - \theta_i)]}{1 + \sin(\theta - \theta_i)} \qquad \text{(J.10)}$$

for $\pi/2 + \theta_i < \theta < 3\pi/2 - \theta_i$. For $\theta < \theta_i + \pi/2$,

$$t = r\theta \, (involute) \qquad \text{(J.11)}$$

Again, we could have skipped over the differential equation and obtained the solution directly by the string method, noticing that for the circle $s = r\theta$, $\theta_0 = \theta_i + \pi/2$.

REFERENCE

Rabl, A. (1976a). Solar concentrators with maximal concentration for cylindrical absorbers. *Appl. Opt.* **15**, 1871–1873.

Appendix K

SKEW RAYS IN HYPERBOLOIDAL CONCENTRATOR

We shall prove that any ray aimed at the rim of the entry aperture is reflected to some point on this rim. Let FF' in Figure K.1 be the diameter of the entry aperture, so that F and F' are the foci of the hyperbolas in the plane of the diagram, and let P be a point on a hyperbola. The bisector PR of angle FPF' is tangent to the hyperbola, and thus the tangent plane at P passes through PR. Any skew ray to the rim incident on the hyperbola at P passes through some point on the circle of which FF' is the diameter and that is in a plane perpendicular to the diagram, and all such rays generate an oblique circular cone. We now draw F_1F_1' through R to make the same angle with PR, as shown. A plane through F_1F_1' perpendicular to the diagram cuts the cone in a certain curve; the cone is a (degenerate) quadric surface, and therefore any plane intersects it in a conic section; but the plane through FF' intersects the cone in a circle by construction, and since the section through F_1F_1' has the same dimensions in and perpendicular to the plane of the diagram, it too must be a circle. An incident skew ray aimed at Q on the circle through FF' is reflected to some point Q' on the plane through FF', and by our construction this point is the mirror image in the tangent plane to the hyperboloid of the point in which the ray cuts the plane through F_1F_1'; thus Q' is on the circle through FF'.
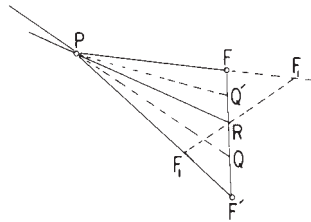
Figure K.1   Geometry of hyperboloid concentrator for skew rays.

Appendix L

SINE RELATION FOR HYPERBOLOIDAL/LENS CONCENTRATOR

With reference to Figure L.1 the lens images the input beams inclined at angles $\pm\theta$ to the axis onto the edges FF' of the virtual source. Therefore, by Fermat's principle $[QA'] + [A'F'] = [AF']$, where the brackets denote optical path length, and similarly for the other beam. From the geometry of the hyperbola, $[AF''] - [A'F''] = 2c$. Also $[QA'] = 2a \sin \theta$, since QA is normal to the beam. The desired result follows: $c = a \sin \theta$.
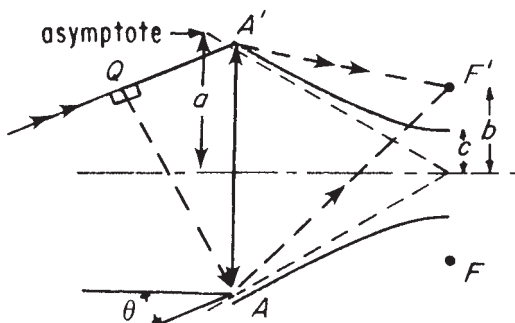
Figure L.1 The thin lens at AA' images the input beam inclined at angle θ to the axis onto F'.

Appendix M

THE CONCENTRATOR DESIGN FOR SKEW RAYS

M.1 THE DIFFERENTIAL EQUATION

The profile of the modified concentrator described in Section 5.6 can be described by means of the coordinate system in Fig. M.1. The origin is taken at the center of the exit aperture and this latter has radius a' as usual. We trace a ray in reverse from some point Q on the exit aperture to the point $R(0, y, z)$ on the desired profile and then reflect it. This ray must be constrained to leave Q with the chosen value h of the skew invariant and after reflection it must have the inclination θ_i to the axis. These constraints will determine the profile uniquely.

If Q is $(x_0, y_0, 0)$, the direction cosines of the ray QR are

$$(\alpha, \beta, \gamma) = (-x_0, y - y_0, z) \big/ \left(x_0^2 + (y - y_0)^2 + z^2 \right)^{\frac{1}{2}} \tag{M.1}$$

with the condition

$$x_0 y \big/ \left(x_0^2 + (y - y_0)^2 + z^2 \right)^{\frac{1}{2}} = h \tag{M.2}$$

Let the inclination of the normal at R be ψ, as indicated in the figure. Then the law of reflection [Eq. (2.1)] gives for the direction cosines of the reflected ray RP

$$(\alpha', \beta', \gamma') = (\alpha, -\beta \cos 2\psi + \gamma \sin 2\psi, \beta \sin 2\psi + \gamma \cos 2\psi) \tag{M.3}$$

Our second constraint requires that $\gamma' = \cos \theta_i$, i.e.,

$$\gamma \cos 2\psi + \beta \sin 2\psi = \cos \theta_i \tag{M.4}$$

Now $\tan \psi = dy/dz$, the slope of the required profile, so Eq. (M.4) takes the form

$$\gamma \left[1 - (dy/dz)^2 \right] + 2\beta(dy/dz) = \cos \theta_i \left[1 + (dy/dz)^2 \right] \tag{M.5}$$

This can be solved for dy/dz since β and γ are functions of y and z through Eqs. (M.1) and (M.2), and the result is a first-order nonlinear differential equation for the profile.

The actual solution must, of course, be done numerically.
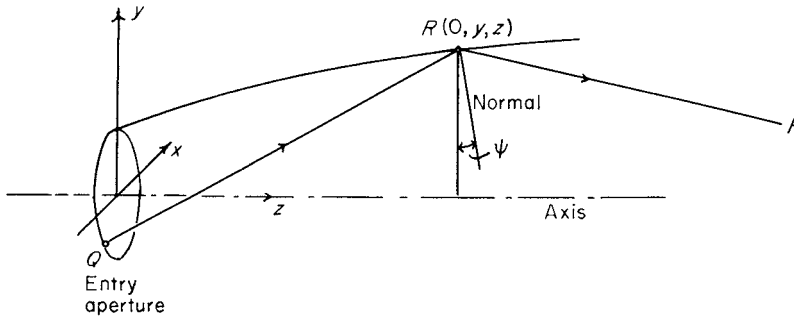
Figure M.1   Constructing the differential equation for the concentrator designed for skew rays.

M.2 THE RATIO OF INPUT TO OUTPUT AREAS FOR THE CONCENTRATOR

The process explained in Section M.1 determines a unique profile for given h and θ_i but nothing was put into the solution to ensure that the ratio of entry to exit areas has the correct value, $(1/\sin\theta_i)^2$. In other words, the length of the concentrator would be determined by where a ray starting in the plane of the exit aperture with skew invariant h meets the surface, and there is no reason why the diameter so found should equal $2a\sin\theta_i$. In fact, it turns out that this is always so, so that this concentrator could, if all rays were transmitted, have maximum theoretical concentration. We prove this result in this section.

We use in the proof a method of treating skew rays in an axisym-metric system due to Luneburg (1964). In this method a skew ray in a part of the system with uniform index of refraction is treated as a meridian ray in the same part of the system but with a radially varying index. We first derive this formalism. If $(\rho,\ \theta,\ z)$ are cylindrical coordinates, then the optical path length along a ray between planes z_0 and z_1 is, for any distribution of index $n(\rho, z)$,

$$V = \int_{z_0}^{z_1} n(\rho, z)\left(1+\dot\rho^2+\rho^2\dot\theta^2\right)^{\frac{1}{2}} dz \qquad (\text{M.6})$$

where the dots indicate differentiation with respect to z. In terms of these quantities it is easily seen that the skew invariant is

$$h = n\rho^2\dot\theta \Big/ \left(1+\dot\rho^2+\rho^2\dot\theta^2\right)^{\frac{1}{2}} \qquad (\text{M.7})$$

From Eq. (M.6)

$$
\begin{aligned}
V &= \int_{z_0}^{z_1} \frac{n\left(1+\dot\rho^2+\rho^2\dot\theta^2\right)}{\left(1+\dot\rho^2+\rho^2\dot\theta^2\right)^{\frac{1}{2}}} dz \\
&= \int_{z_0}^{z_1} \frac{h\left(1+\dot\rho^2+\rho^2\dot\theta^2\right)}{\rho^2\dot\theta} dz \\
&= \int_{z_0}^{z_1} \frac{h\left(1+\dot\rho^2\right)}{\rho^2\dot\theta} dz + h(\theta_1-\theta_0) \qquad (\text{M.8})
\end{aligned}
$$

Now, we can solve Eq. (M.7) for $\dot{\theta}$, giving

$$\dot{\theta} = \frac{h(1+\dot{\rho}^2)^{1/2}}{\rho(n^2\rho^2 - h^2)^{1/2}}$$

and if this is substituted in the integral in Eq. (M.8), we obtain

$$V = \int_{z_0}^{z_1} \left(n^2 - \frac{h^2}{\rho^2}\right)^{1/2} (1+\dot{\rho}^2)^{\frac{1}{2}} dz + h(\theta_1 - \theta_0) \tag{M.9}$$

This result, due to Luneburg, has the significance that the optical path length along a skew ray between planes z_0 and z_1 can be computed as that along a meridian ray between $z_0 z_1$ in a medium of "refractive index"

$$m(\rho, z) = [n^2(\rho, z) - h^2/\rho^2]^{1/2} \tag{M.10}$$

together with an azimuthal contribution $h(\theta_1 - \theta_0)$. Let the longitudinal and azimuthal components of the right-hand side of Eq. (M.9) be denoted by V_z and V_θ. Then the result shows that in the ρ, z plane the "rays" obey Fermat's principle for V_z with the fictitious index m given by Eq. (M.10), so that there is true image formation in this plane.

Some rays in the concentrator designed for nonzero h appear as in Fig. M.2 in the Luneburg representation. We show rays entering with the collecting angle θ_i and parallel to each other. Ray 1 strikes the entry aperture at A and is reflected to the rim of the exit aperture at B, while ray 2 grazes the entry aperture at D and is reflected at the exit aperture at E. Ray 2 has, from the design, the skew invariant h, so that it travels perpendicular to the axis and meets the exit aperture again at C. Points B and C are in fact coincident, but we have not yet shown this. We wish to calculate the optical path lengths V_{z_1} and V_{z_2} from the source point at infinity along rays 1 and 2 to points B and C on the exit rim. However, since rays 1 and 2 are parallel before they enter the concentrator, it suffices to compute
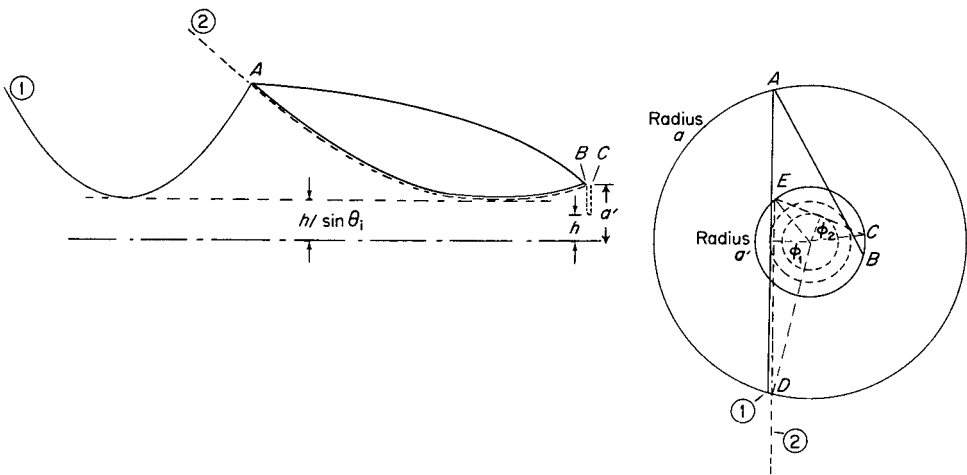
Figure M.2   Rays in the Luneburg representation.

the paths from a plane through D perpendicular to both rays. We have for these paths

$$V_1 = 2a \sin \varphi_1 \sin \theta_i + V_1(AB)$$
$$V_2 = V_2(DE) + 2a' \sin \varphi_2$$

(M.11)

and $V_1(AB) = V_2(DE)$, since these two paths are geometrically identical.

Now, in the Luneburg representation the longitudinal path lengths from infinity to the exit rim are equal, since our solution of the differential equation in Section M.1 has ensured this. Likewise, the longitudinal path lengths from A to B and from D to E are equal, so that

$$V_{z_1} = V_{z_2}$$

or

$$V_1 - h\theta_1 = V_2 - h\theta_2$$

(M.12)

where θ_1 and θ_2 are the appropriate azimuthal angles for the two paths. Combining Eqs. (M.12) and (M.12), we obtain

$$V_1 - V_2 = 2a \sin \varphi_1 \sin \theta_i - 2a' \sin \varphi_2 = h(\theta_1 - \theta_2)$$

(M.13)

If we compute the angles θ_1 and θ_2 from infinity to B and C, respectively, and remember that AB and DE subtend the same angles, we find

$$\theta_1 - \theta_2 = 2(\varphi_1 - \varphi_2)$$

(M.14)

and also from the skew invariant we find

$$\cos \varphi_1 = h/(a \sin \theta_i), \quad \cos \varphi_2 = h/a'$$

(M.15)

We use Eqs. (M.14) and (M.15) to eliminate a, a', θ_i, and θ_2 from Eq. (M.13) and obtain

$$\tan \varphi_1 - \tan \varphi_2 = \varphi_1 - \varphi_2$$

(M.16)

Now $\tan \varphi - \varphi$ is monotonic for φ less than $\pi/2$ and the angles in Eq. (M.16) are less than $\pi/2$, so we conclude that $\varphi_1 = \varphi_2$ and thus, from Eq. (M.15),

$$a'/a = \sin \theta_i$$

(M.17)

Thus the concentrator designed for the nonzero skew invariant has the ratio of apertures required to give the maximum theoretical concentration ratio.

We can also calculate the length L of the concentrator:

$$L = \left[(a^2 - h^2/\sin^2 \theta_i)^{\frac{1}{2}} + (a'^2 - h^2/\sin^2 \theta_i)^{\frac{1}{2}} \right] \cot \theta_i$$

(M.18)

M.3 PROOF THAT EXTREME RAYS INTERSECT AT THE EXIT APERTURE RIM

We stated in Section 5.6 in connection with Fig. 5.8 that rays in a chordal plane at θ_i and touching either side of the entry rim (points A and D in Fig. 5.8) will emerge at the same point C of the exit aperture rim. We can show this result very

Figure M.3   Coincidence of the points C and F at which two rays emerge.

simply, as in Fig. M.3; this is the same as Fig. 5.8 but with some construction lines added.

We assume that the ray reflected from B meets the exit rim at C and that reflected from D meets it at F, and we shall show that F and C coincide. First, by construction the triangles AOB and DOF are congruent so that $\angle GOB = \angle DOF$. Thus the external angle from G to F is given by

$$\angle GOF = \angle 2GOB + \angle BOD \qquad\qquad (\text{M}.19)$$

Next, the triangles AOD and BOC are similar, as may be seen from the radii of the pairs of circles involved in the construction; the ratio of sides is $\sin\theta_i$. Thus for the external angle from G to C we have

$$\angle GOC = \angle GOB + \angle BOC = \angle GOB + \angle GOD$$
$$= \angle 2GOB + \angle BOD \qquad\qquad (\text{M}.20)$$

Thus comparing Eqs. (M.19) and (M.20) we have proved our result.

M.4   ANOTHER PROOF OF THE SINE RELATION FOR SKEW RAYS

We may also obtain the sine relation for the concentrator designed for skew rays by formally evaluating the étendue in the p, ρ plane using the fictitious index m

(bearing in mind that in these variables the concentration has ideal properties). We need to calculate

$$H = \iint dp\, d\rho \tag{M.21}$$

where p is the optical direction cosine in the ρ direction,

$$p = m(\rho,z)\dot{\rho}/(1+\dot{\rho}^2)^{\frac{1}{2}} \tag{M.22}$$

and $m(\rho, z)$ is given by Eq. (M.10).

To find the limits on p, the following identity is useful:

$$m/(1+\dot{\rho}^2)^{\frac{1}{2}} = n\cos\theta \tag{M.23}$$

where θ is the inclination of the ray to the z axis. Then the étendue over the entrance aperture is

$$H = 2\int_{h/\sin\theta_i}^{a} \left(\sin^2\theta_i - h^2/\rho^2\right)^{\frac{1}{2}} d\rho \tag{M.24}$$

while the étendue over the exit aperture is

$$H' = 2\int_{h}^{a'} \left(1 - h^2/\rho^2\right)^{\frac{1}{2}} d\rho \tag{M.25}$$

Performing the integrations and setting $H = H'$, we obtain

$$2h[\tan\varphi_1 - \varphi_1] = 2h[\tan\varphi_2 - \varphi_2] \tag{M.26}$$

the same relation as Eq. (M.16). Therefore, the sine relation is proved.

M.5 THE FREQUENCY DISTRIBUTION OF h

Having found a solution for fixed h, we may ask for the frequency of occurrence of h in an ideal 3D system with rotational symmetry. This is readily found by writing the étendue in cylindrical coordinates (ρ, θ, z):

$$H = \int dh\, dp\, d\rho\, d\theta \tag{M.27}$$

Here p is the generalized momentum conjugate to ρ obtained from differentiation the integrand of Eq. (M.6) with respect to $\dot{\rho}$. Then h is distributed according to

$$dH/dh = 2\pi \int dp\, d\rho \tag{M.28}$$

We can just as well work in the Luneburg representation and obtain p from Eq. (M.9) (the two are of course identical). The desired integration is therefore already performed in the previous section and we may take over the result:

$$dH/dh = 4\pi h[\tan\varphi_2 - \varphi_2] \tag{M.29}$$

For purposes of plotting, this expression is more conveniently written

$$dH/dh = 4\pi a'\left[(1-x^2)^{\frac{1}{2}} - x\cos^{-1}x\right] \tag{M.30}$$
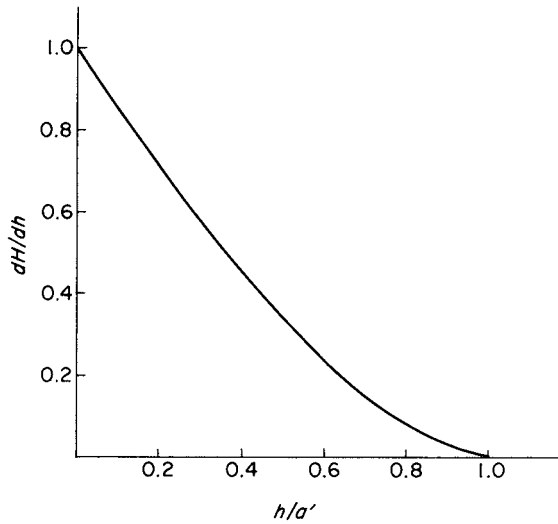
Figure M.4   Frequency distribution in h for an ideal 3D system. The ordinate scale is chosen to have maximum value = 1.

where $x = h/a'$ lies in the interval $(-1, 1)$. This function is plotted in Fig. M.4. We observe that the frequency distribution strongly peaks toward $h = 0$. We may check that the distribution is properly normalized by integrating Eq. (M.30) over all h values:

$$H = \int_{-a'}^{a'} (dH/dh)dh = \pi^2 a'^2 \qquad (M.31)$$

We recognize this to be the étendue of a diffusely illuminated disk of radius a'.

INDEX