Geometrical optics

Geometrical optics calls from physics just Snell law. All is dealt with in terms of light rays, although wavefronts are mentioned as surfaces normal to a ray bundle. The systematic application of Snell law allows tracing a ray, that is, to follow it along its path in an optical system, and so evaluate and improve its performance.

Ray tracing through a centered optical system

Almost all optical systems are formed by coaxial surfaces of revolution. We call these surfaces "diopters", and a mirror is also a diopter.

In such cases ray tracing is relatively simple.

In Fig. 4.1 it is seen a typical centered optical system. It is characterized by a sequence of four parameters pertaining to each optical surface *i*

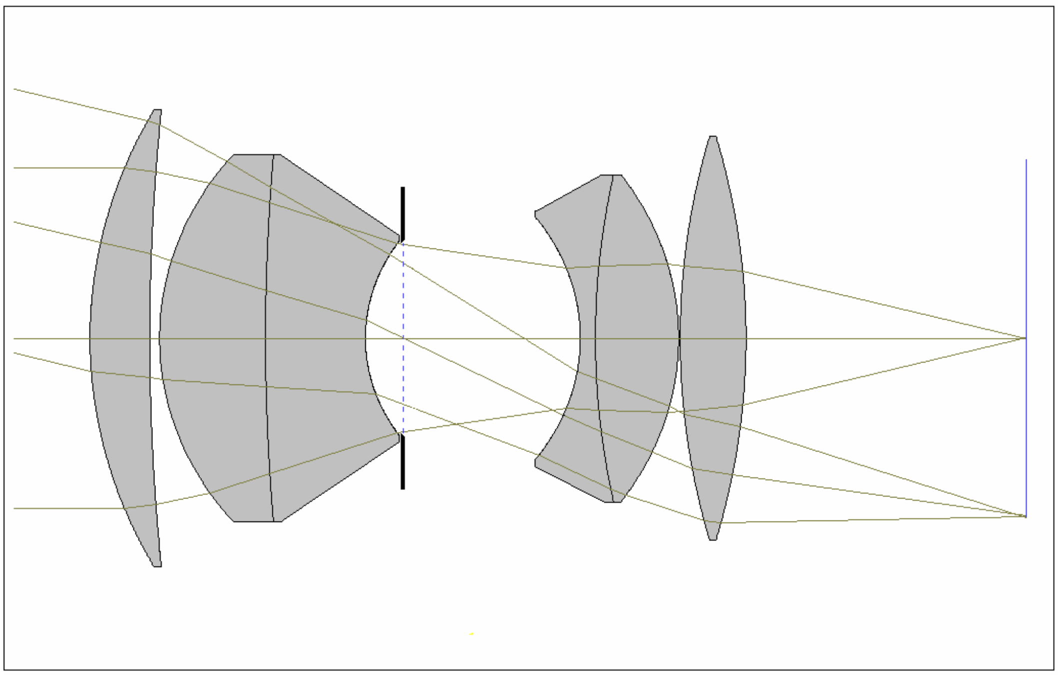


Fig. 4.1

Convention for prescriptions.

In order to define a c.o.s. it is needed a listing of its parameters. At this point the literature is irregular. We shall follow a logical way.

The parameters are listed in columns according the format

i	Radius curv.	Clear radius	Distance	Glass	Comment

i is the number of the surface

In the "Glass" column, it may appear the index following the surface or the glass name to calculate its index in several λ. If "*1*" or " –1" appears, that space is air. (See below).The index before the first surface must be separately specified.

The last column is used to make any comment about that surface, e.g, it is aperture stop, aspheric, etc.

For this case, for example, the prescription is

i	Radius curv.	Clear radius	Distance	Glass	Comment
1	*72.62*	*34*	*7.9*	*BASF51*	
2	*379.36*	*34*	*1.3*	*1*	
3	*43.22*	*28*	*14*	*LAKN7*	
4	*355.75*	*28*	*13.2*	*SF1*	
5	*25.6*	*14.5*	*5*	*1*	
6	*0*	*14*	*23.4*	*1*	*diaphragm*
7	*-30.27*	*18*	*2*	*BASF2*	
8	*124.81*	*25.5*	*11*	*LAF2*	
9	*-42.63*	*25.5*	*0.2*	*1*	
10	*118.05*	*30*	*8.7*	*LAF2*	
11	*-115.61*	*30*	*-*	*1*	

In the formulae there appears the curvatures $c = 1/R$ because they are always bounded, but for an assessment by the maker is better the radius of curvature. For a plane surface, it may be stated for example, "∞", or (symbolically) "0"

The prescription is complemented with such data as the units of measure, the object position, the working wavelength range, the index of the first medium if it is not 1, etc.

In some programs the object and image are defined as surfaces, but the classical notation is followed here.

The sign of the radius of curvature obeys a convention to be explained later.

The c.o.s. establish a correspondence between the object and its image.

If all rays travel near the axis, they are called paraxial, and the correspondence is perfect, that is point-to-point. The practical systems must accept rays far from paraxial conditions, and these don't met at a point, but in a region of maximum concentration with finite extent.

The diagnosis and partial cure of this failure (aberration is the devoted name) is made with ray tracing.

The procedure calls for a great mass of numerical computation only accessible with modern computers, but there are analytical methods patiently developed in the past, that has the value of giving a more clear idea about the behavior of the aberration, decomposing it in various types, analyzing its variations, and pointing the contribution of each surface.

Meridional ray tracing

Rays travelling within a plane containing the axis are called meridional. The image has more others, but these are simpler to trace, and are enough for many cases.

In Fig. 4.2 y 4.3 there are illustrated the formulas.

A ray incident on the first surface is defined by a segment Q, from the vertex of the surface to the ray, intersecting it normally, and the angle U respect to the axis. At the surface it forms an angle I of incidence and one I' of refraction, and takes a new slope U'.

Sign conventions.

To follow the literature, it is stated

1- Light incides from left.

2- Angles U and U' are positive if they go clockwise from axis to ray.

3- Angles I, I' are positive if they go counter-clockwise from normal to ray.

4- Curvatures are positive if the concavity faces to right.

5- Coordinates has their usual signs of geometry, with the origin at the intersection of the axis with the surface.

The convention on the sign of curvature is consistent with analytical geometry. If a surface is developed near the origin is $x = y^2 / 2\overline{R}$, where $\overline{R}$ is the radius of curvature at the vertex, and the sign of *x* is the same as of $\overline{R}$

The convention on U is not consistent with the sign of the ray slope, but it will be justified later when dealing with paraxial rays.

From an examination of the figures

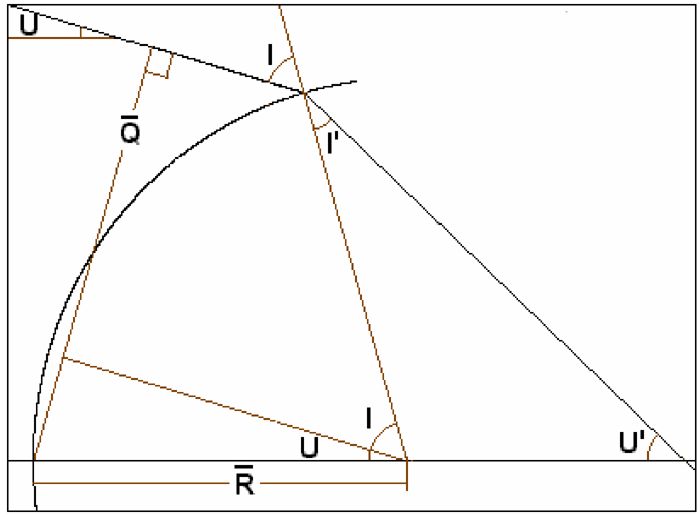


Fig. 4.2

Refraction

(4.1) $\overline{Q} = \overline{R}\,\sin I + \overline{R}\,\sin U$

Then

(4.2) $\sin I = \overline{Q}\,c - \sin U$

Also

(4.3) $\sin I' = \dfrac{n}{n'}\sin I$

Applying successive rotations

(4.4) $U' = U + I - I'$

Putting primes to (4.1)

(4.5) $\overline{Q'} = \dfrac{\sin I' + \sin U'}{c}$

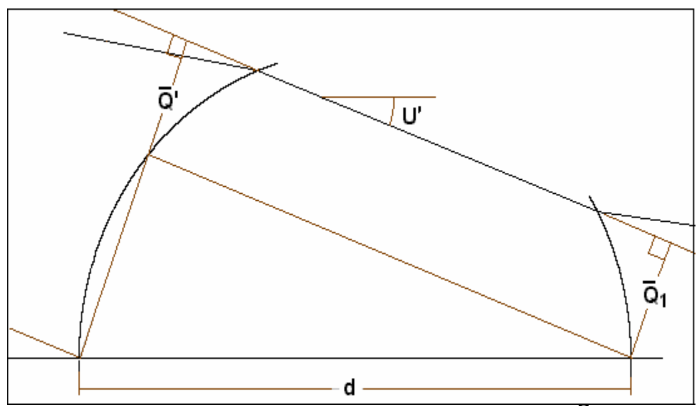


Fig. 4.3

Transport

(4.6) $\overline{Q}_1 = \overline{Q'} - d\,\sin U'$

Starting with $\overline{Q}, U$ reaches to $\overline{Q}_1, U_1 = U'$ to repeat the process with the following surface.

The prime indicates a quantity after refraction, and sub-index *1* means a quantity associated to the following surface. It should be *i+1*, but so is for short.

If it is wanted to programming the formulae, it must be modified (4.5) because it diverges when $c = 0$. But if about programming is dealt, is better to use the general tracing formulas to be described later. Meridional formulae where developed for mechanical calculators, and here they are of value only as an introduction to the paraxial trace.

Paraxial rays

If the rays are kept at an infinitesimal distance to the axis they are called paraxial. They must obey two conditions

1- To incide near the axis, that is, $\overline{Q} \to 0$
2- To incide with a small angle, that is $U \to 0$, for otherwise, on propagation they would not fulfill the former condition. Also it implies $I \to 0$.

Results $\overline{Q} \to \overline{Q'} \to y =$ height or distance to axis.

Denoting with lower case the corresponding quantities, formulae (4.2) - (4.6) become

(4.7) $i = y\,c - u$

(4.8) $i' = \dfrac{n}{n'}\,i$

(4.9) $u' = u + i - i'$

(4.10) $\mathbf{y_1 = y - d\,u'}$

Eq. (4.5) is redundant. Eliminating i, i' yields

(4.11) $$\mathbf{n'u' = n\,u + (n' - n)\,y\,c}$$

This formula, together with that of transport (4.10), are of immense practical importance because they permits a rapid approximate calculation of the image *position* .

On using the paraxial formulae with numerical data, they ceases to be paraxial. If a ray is traced with the exact formulae, the difference of the results is the geometrical aberration of the ray.

That is, paraxial rays inform us about the position of the image but not about its structure.

To appreciate paraxial rays is useful to imagine tracing exact rays very near the axis on a drawing of the system made on a rubber sheet. Therefore they will not be seen because they merged with the axis. If the sheet is stretched vertically, they will appear with their slopes and heights depending on the curvatures, separations and indices. But now the curvatures has been stretched to become straight lines.

Analytical application of paraxial tracing.

Due to their simplicity, paraxial formulae are suited to be put in algebraic expressions, and they are source of countless mathematical relations. It's necessary to master the method because otherwise any discussion becomes incomprehensible, as may be seen looking at the special literature. Formulae (4.10), (4.11), (4.22) are essential, and is indispensable all derived from them.

*Some examples follow***.**

Elementary diopter (Fig. 4.4).

Replacing $u = \frac{y}{s}$ < *0* and $u' = \frac{y}{s'}$ in (4.11) is

(4.12) $$\frac{n'}{s'} = \frac{n}{s} + (n' - n)\,c$$

Here we find a justification for the sign convention for *U, u*:

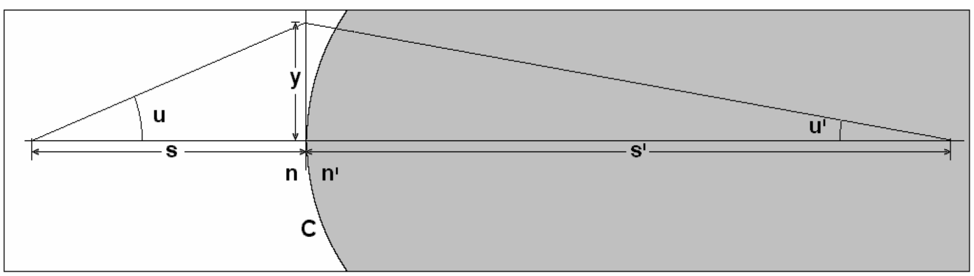


Fig 4.4

They are set by the signs of *y, s*.

In this drawing and the following there appears coarse errors. Here, for instance, the refraction does not occurs at the surface but in the tangent plane. This happens by using paraxial formulae on a drawing which is not, remember that rubber sheet.

Thin lens in air (Fig. 4.5)

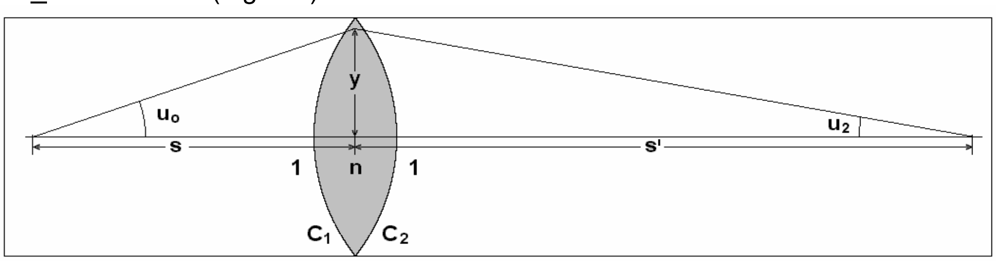


Fig. 4.5

Applying in succession eqs (4.11) y (4.10)

$$n\,u_1 = 1\,u_0 + (n-1)\,y\,c_1$$

$$y_2 \approx y_1 = y \text{ because } d \approx 0$$

$$1\,u_2 = n\,u_1 + (1-n\,)\,y\,c_2$$

$$= u_0 + (n-1)\,y\,c_1 + (1-n)\,y\,c_2\,, \qquad \text{results}$$

(4.13) $$\frac{1}{s'} = \frac{1}{s} + (n-1)(c_1 - c_2)$$

This uses to be called the Gauss and lens maker formula.

If $s \to \infty, \quad s' \to f'$ and if $s' \to \infty, \quad s \to f$

If the lens is on air is

(4.14) $$-\frac{1}{f} = \frac{1}{f'} = (n-1)(c_1 - c_2)$$

From now on we will call *f* to *f '* by convenience of writing.

A lens is thin when its thickness is negligible compared to their radii of curvature and not to the radius of the lens. In this drawing, the coarseness lies in assuming that lens as thin, and the two refractions taking place in the middle.

Graphical constructions (Fig. 4.6)

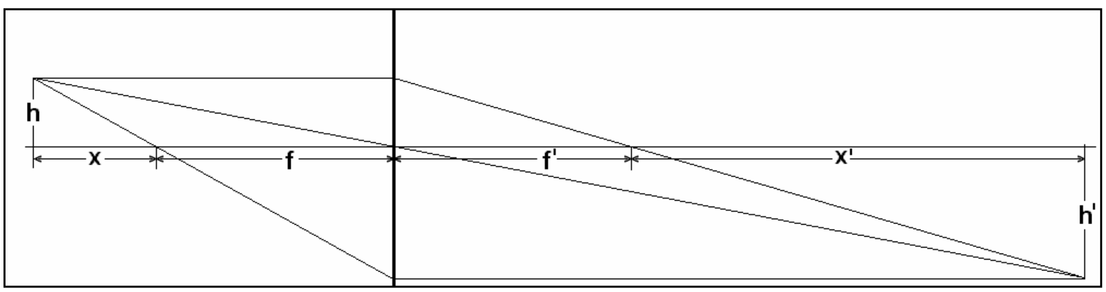


Fig. 4.6

This diagram is in all elementary books. All of us have learned it in the school, but nobody told they where paraxial formulae.

By similarity of triangles is

(4.15) $$m = \frac{h'}{h} = \frac{-f}{x} = \frac{x'}{f}$$

And the Newton formula

(4.16) $$x\,x' = -f\,f'$$

With this diagram there is discussed the basic workings of the loupe, telescope and microscope.

Exercise

At what distance from a screen must be placed a lens of focal distance *f* if the separation between object and image is *D?*

$$\frac{1}{s'} = \frac{1}{s} + \frac{1}{f}\,, \qquad s' - s = D$$

On substitution, the quadratic equation has roots

(4.17) $$s' = \tfrac{1}{2}\left(D \pm \sqrt{D^2 - 4Df}\right)$$

There are two positions, that reduces to one if $D = 4\,f$. If $D < 4\,f$ there is none.

Thick lens and principal planes.
(Fig. 4.7)

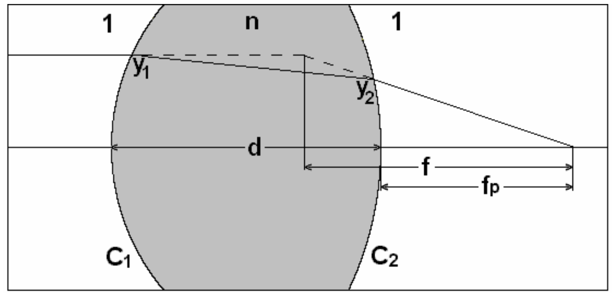


Fig. 4.7

Let $u_0 = 0$ (ray parallel to axis).

As by custom,

$$n u_1 = 0 + (n-1) y_1 c_1$$

$$y_2 = y_1 - d u_1$$

$$= y_1 \left(1 - \frac{n-1}{n} c_1 d \right)$$

$$u_2 = n u_1 + (1-n) y_2 c_2$$

$$= y_1 (n-1) c_1 + y_1 (1-n) \left(1 - \frac{n-1}{n} \right) d c_1 c_2$$

(4.18) $$\frac{u_2}{y_1} = \frac{1}{f} = (n-1)\left(c_1 - c_2 + \frac{n-1}{n} d c_1 c_2 \right)$$

If $d \to 0$ return (4.14). Note that f is the focal distance of an equivalent thin lens placed at the intersection of the prolongation of the rays. It is not the distance between the last surface and the focus. The latter is called f_p, back focal length, and may be calculated by the same routine.

f is called the effective focal length, and defines the second principal plane. The first is obtained in the same way, turning the lens. The relations between object and image are the same, if we get rid of the distance between principal planes.

Mirrors

The formulae are valid for mirrors if all distances d and indices n following a reflection change sign. For example, if a glass has its back surface silvered, is $n = 1.5$ (say) and $n' = -1.5$

The index of a mirror in air is -1.

Reflection

Mirrors reminds our conscience. This is why many people hate them. (J. .L. Borges said). But well used, they discover new worlds. (Astronomers .reply)

Example

Cassegrain telescope (Fig. 4.8).

We call Cassegrain a configuration with concave primary mirror and convex secondary, irrespective of the detailed shape of the mirrors.

Let be $n_0 = 1, \quad n_1 = -1, \quad n_2 = 1, \quad u_0 = 0$

$$-u_1 = 0 + (-1-1) y_1 c_1$$

$$u_1 = 2 y_1 c_1$$

$$y_2 = y_1 - d u_1$$

$$= y_1 - 2 d y_1 c_1$$

$$u_2 = -2 y_1 c_1 + 2 c_2 (y_1 - 2 d y_1 c_1)$$

(4.19) $$\frac{u_2}{y_1} = \frac{1}{f} = 2 (c_2 - c_1 - 2 d c_1 c_2)$$

For a numerical calculation remember that $d < 0$

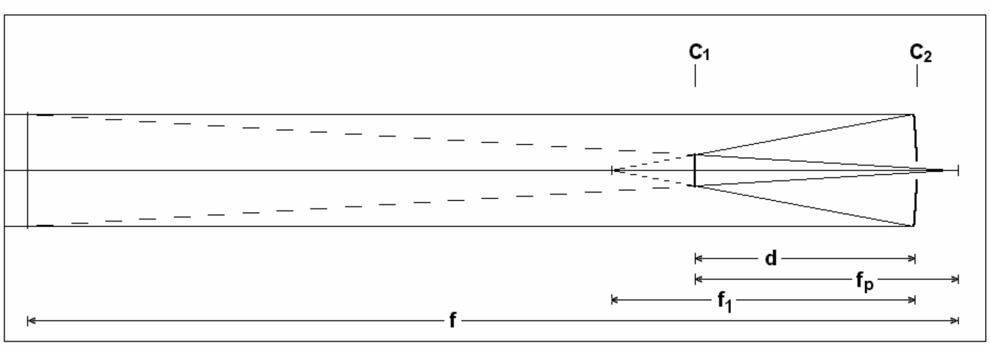


Fig. 4.8

Focussing the Cassegrain

A small shift of the secondary produces a larger one of the image, by which in Cassegrain telescopes focussing is often done moving the secondary.

Is worth then to find the back focal distance, between the secondary and the image.

This is

$$f_p = \frac{y_2}{u_2} = \frac{1}{2}\ \frac{1 - 2\,d\,c_1}{c_2 - c_1 - 2\,d\,c_1\,c_2} \tag{4.20}$$

To find the amplification factor, we derive respect to d, and the expression turns to be

$$\frac{df_p}{dd} = \frac{-4\,c_1\,(c_2 - c_1 - 2\,d\,c_1\,c_2) + (1 - 2\,d\,c_1)\,4\,c_1\,c_2}{4\,(c_2 - c_1 - 2\,d\,c_1\,c_2)^2}$$

$$= \frac{-4\,c_1\,c_2 + 4\,c_1^2 + 8\,c_1^2\,d\,c_2 + 4\,c_1\,c_2 - 8\,c_1^2\,d\,c_2}{\frac{1}{f^2}}$$

$$= (2\,c_1\,f)^2 = \left(\frac{f}{f_1}\right)^2 \tag{4.21}$$

The ratio $f\,/\,f_1$ uses to be called “telephoto factor", which permits a large focal length within a small physical length. Fig 4.8 is truly proportional to the CAsLeo telescope and it is appreciated this feature, much needed in large telescopes.

Example of the example

CAsLeo telescope

$R_1 = -11.176\ m \qquad R_2 = -4.432\ m \qquad d = -4.051\ m$

Results $\qquad f = 18.237\ m \qquad f_p = 5.016\ m \qquad f_1 = 5.588\ m$

$$\left(\frac{f}{f_1}\right)^2 = 10.65$$

That is, if the secondary moves away from the primary by *1 mm*, is *dd* = *–1mm*. (*d* goes from *–4.051 m* to *–4.052 m*). The positive quantity f_p varies by *–10.65 mm*. (The focus shifts towards the secondary).

But this is referred to the secondary. On an instrument mounted in the back of the telescope the shift is *–11.65 mm*

More about this telescope.

How much must be shifted ***the eyepiece*** to focus on a turbulent air current at a height of 5000 or 10000 m?

As this distance is >> *f*, it can be used Newton formula (4.16), with *x* = 5000 or 10000 m.

Results *x'* = 66 mm or 33 mm, respectively.

Yes, *the eyepiece*, because this is best seen with the eye than with any other detector.

Systems of thin lenses

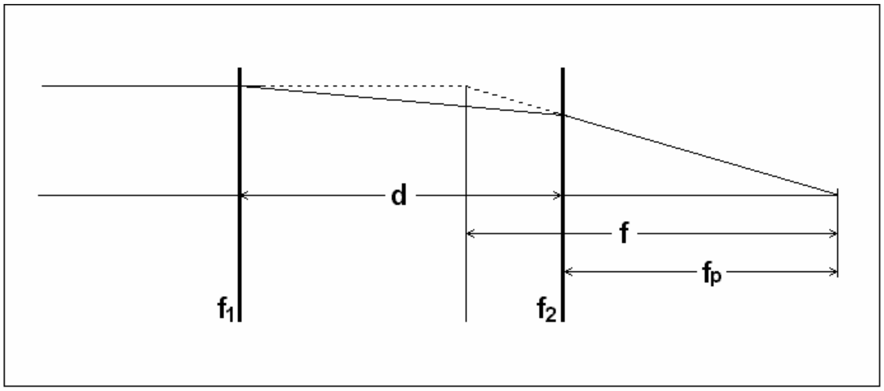
In a thin lens is

$$\frac{1}{s'} = \frac{1}{s} + \frac{1}{f}$$

Being $\qquad s = \frac{y}{u} \quad ; \quad s' = \frac{y}{u'}\ ; \qquad$ results

$$\mathbf{u' = u + \frac{y}{f}} \tag{4.22}$$

That replaces (4.11) in paraxial tracing. The thin lens is considered as a whole. For a combination of two thin lenses separated by a distance *d* (Fig. 4.9)



$$u_1 = 0 + \frac{y_1}{f_1}$$

Fig. 4.9

$$y_2 = y_1 - d\,u_1$$

$$u_2 = u_1 + \frac{y_2}{f_2} = \frac{y_1}{f_1} + \frac{y_1 - d\,\frac{y_1}{f_1}}{f_2}$$

(4.23) $$\frac{u_2}{y_1} = \frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1\,f_2}$$

Thin lenses are shown as straight segments remembering that rubber sheet.

Lagrange invariant

Tracing two arbitrary paraxial rays, one of them labeled with sub-index *p*, for any surface holds

(4.24) $$n'\,u' = n\,u + (n' - n)\,y\,c$$

(4.25) $$n'\,u'_p = n\,u_p + (n' - n)\,y_p\,c$$

That is

(4.26) $$(n' - n)\,c = \frac{n'\,u' - n\,u}{y} = \frac{n'\,u'_p - n\,u_p}{y_p}$$

(4.27) $$n\,u\,y_p - n\,u_p\,y = n'\,u'\,y_p - n'\,u'_p\,y$$

Hence the quantity

4.28) $$I = n\,u\,y_p - n\,u_p\,y$$

Is invariant through a surface. Its easy to see that is also invariant from one surface to another. It is used in analytical traces to save calculations. Applied to the object and image planes, we can take the *p*-ray as coming from the tip of the object, and the one coming from the base without sub-index. At the object and image is $y = 0$, and y_p its height *h.*

(4.29) $$(h\,n\,u)_{object} = (h\,n\,u)_{image}$$

Paraxial aperture, pupil and field

As it has been shown, any image forming optical system is equivalent to a diopter. When it is really a diopter, the light gathering is proportional to its area, or its clear radius squared. We call it semiaperture. Each point from an extended object sends a cone of rays defined by the rim of the aperture. The axis of the cone from the tip of the object at height *h* is called principal ray, and pass through the aperture with slope u_p. If the object is at infinity, (astronomical objects) there are only angles. In both cases we call semifield angle that formed by the principal ray and the axis.

In the general case of a compound system, it must be defined the aperture. For example, let be the photographic objective in Fig. 4.1. In general, it has a lens group ahead and other behind a diaphragm of adjustable radius (the dummy surface $N^o\ 6$), that sets a limit on the incident cone of rays. If we look from ahead, its is seen the diaphragm through the first lens group, that is, the virtual image of the diaphragm given by that group.

In the same way, when we look at someone's eyes, we see the image of the iris diaphragm through the cornea, and call it the pupil.

Pardon is begged if geometrical optics kills the magic of looking at the eyes.

The radius of the pupil defines the aperture. In the photographic objective the complete name is "entrance pupil", because looking from behind, there is also an exit pupil.

An interesting case is a visual telescope. The exit pupil is the circle seen on looking to the eyepiece with the telescope aimed to the day sky.

If it is wanted to observe dim objects, as nebulae, the diameter of that circle cannot exceed that of the eye pupil, to avoid loss of light. As the magnification of the telescope is given by the ratio of the entrance and exit pupils, this condition defines a magnification, the equipupilar magnification. On the other hand, if it is desired the greatest resolution on bright objects, the Airy disk produced by the objective must subtend on the retina an angle about 1', the resolution of a normal eye alone (resolvent magnification).

The equipupilar magnification is less than the resolvent one because the pupil dilates under dim light.

(Remembrances from an age when astronomers look through their telescopes)

The black part within the limit of the pupil is the retina, at the back of the eye. If it is illuminated with an intense beam, a luminous point is formed in it and the reflected light exits again in reverse direction. This is clearly seen in the cats and dogs illuminated at night with a flashlight; their eyes shine with an intense green, that is the color of the retina. In humans is feeble and purple, and is seen in photographs taken with flash. Walking at night in a forest with a flashlight near the eyes one is able to see shinning eyes of all animals looking at the beam. *(¡ Even the spiders !).*

The diaphragm may not exist as a separated physical object, but there is always an element whose rim sets a limit on the beam. The pupils and diaphragms are conjugated, that is, they are image one of the other. In these planes all principal rays pass through the center and in other parts they have different heights.

The procedure to find the pupils once the diaphragm is identified, is obvious: a paraxial ray is traced from its center through all lenses in front and rear of it, and the intersections (or their prolongations if it is a virtual image) with the axis marks the position of the pupils. Their radii are defined tracing rays from the border of the diaphragm, and they are the heights of intersection with the planes of the pupils. Or, if one wants to save algebra, using the Lagrange invariant.

The ray without sub-index from the base of the object to the border of the entrance pupil is called marginal. Both rays give all information about the working aperture and field of the system.

In Fig. 4.1 there has been plotted the beams defined by marginal and principal rays

If the system departs too much from paraxial regime, all said ceases to apply. The pupils are determined with exact rays and behave in a more complex way. This is very striking in the gran-angular objectives; the lens appears "to follow one with its gaze".

Exercise

To find the exit pupil of CasLeo telescope.

This is a very simple case because the aperture stop (diaphragm) is the primary mirror.

As it is the first surface, it is also the entrance pupil. The exit pupil is the image of the primary mirror given by the secondary. It suffices to use the formula of the diopter (4.12), and to follow the ritual that light comes from left, we turn the mirror and put the object at the center of the primary. Hence, reviewing former data:

$$n = 1\,,\quad n' = -1\,,\quad R = 4432\,,\quad d = 4051$$

$$\frac{1}{s'} = \frac{1}{d} + \frac{2}{R} \quad s' = \frac{R\,d}{R + 2\,d} = 1432.4$$

The clear radius of the primary is $h = 1075$, then

$$h' = \frac{s'}{s}\,h$$

$$h' = 380.1$$

The exit pupil is at *1432.4* mm behind the secondary (on the concave side) and has radius *380.1* mm

Field lens

The important function of this lens is better understood with a very clear example, a submarine periscope. It is intended to transport an image through a long narrow tube, keeping constant magnification. In a former exercise it was shown that if the object is at twice the focal length, the image too, and has unit magnification.

Placing a sequence of lenses of the same focal length f separated by a distance $4\,f$, the image can be transported with unit magnification, but looking at Fig. 4.10 (a), it is seen that in the following image the rays incide very excentric, that requires a larger lens partially illuminated. If it is tried to form a sequence, this leads to an optical catastrophe.

In Fig. 4.10 (b), it was inserted an equal additional lens in the image plane.

In both figures there are drawn the principal ray *p* and the *s, i* (superior and inferior rays). The ray *p* is marginal for the inserted lens, what indicates that this lens forms the image of L_1 on L_2. Now, at paraxial level, the sequence can proceed indefinitely.

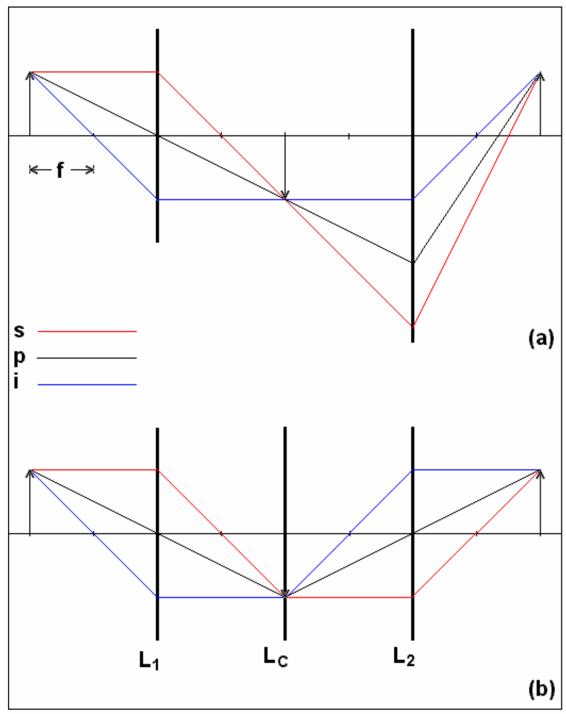


Fig. 4.10

The "field lens" is the inserted, L_c.

Generally it is used when an intermediate image is formed in a compound system, is placed on this image, and forms an image of the exit pupil of the former sub-system on the entrance pupil of the following one.

By technical reasons, the field lens is placed slightly offset in order to its surface defects do not superpose with the image.

In Astronomy it is used, for example, in focal reducers. These systems increase the convergence of the direct beam to match the image to a CCD. The first sub-system is the telescope, and behind its image there is an objective of photographic type. In the intermediate image goes the field lens

Exercise (Fig. 4.11)

This is a standard numerical problem applying formulae (4.11) and (4.10) in sequence.

The prescription is

i	Radius curv.	Clear radius	Distance	Glass	Comment
1	*30*	-	*1*	*BK7*	*n = 1.517*
2	*-20*	-	*-10*	*1*	
3	*-20*	-	*-1*		*mirror*

The clear radii doesn't matter in this case.

It is intended to find the position and size of the image.
The object coordinates are $x_o = -10, \quad y_o = 1, \quad (n_o = 1)$

Arbitrarily it is set an incidence height at the first surface $y_1 = 1$

To find the position of the image we trace a ray from the bottom of the object.
The quantities obtained by the sequential application of the formulae are

i	y	u
1	1	- 0.0545594
2	1.054559	- 0.05550631
3	1.609623	- 0.105456

The position of the image is $x_i = y_3 / u_3 = -15.26346$ from the last surface?

No ! There is a problem, the light pass again through the lens.
The true system has two more surfaces. To the former prescription it must be added

4	*-20*	-	*-1*	- *BK7*	*n = -1.517*
5	*30*	-	-	*-1*	

And distance *3* is not unknown, is *–10*. Results

4	*0.555063*	*- 0.07897452*
5	*0.4760885*	*- 0.1280089*

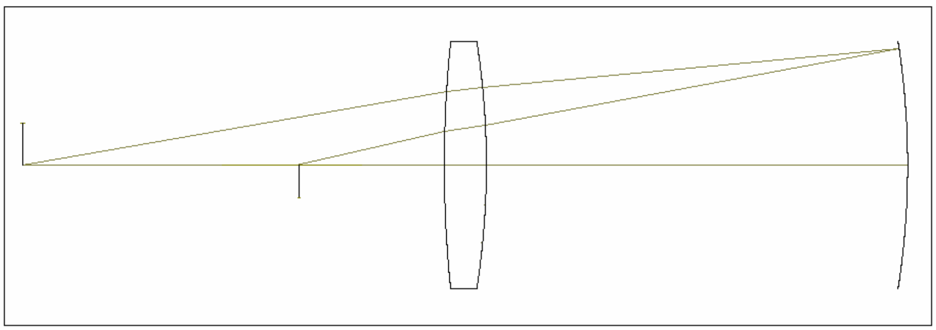


Fig. 4.11

The position of the image is $x_i = -3.719182$, its height follows from the Lagrange invariant (4.29) and is $y_i = -0.7811956$

In some programs the ray is traced in non sequential form and this problem is dealt with automatically.

The figure represents a real system because the errors of the paraxial calculation can be indulged in this case.

Chromatic aberration

The refractive index of glass depends on λ. Hence, at paraxial level, the position of the image varies with color. This is called chromatic aberration or chromatism. The aberration may be characterized considering the positions *x, y* of the image for two colors *r ,b* at the extremes of the spectrum (red and blue, for example).

There is a transverse, or lateral aberration $y_r - y_b$ and a longitudinal, or axial $x_r - x_b$.

In what follows it will be used sub-index *F* to designate blue, *D* yellow-green and *C* red.

The letters identify standard spectral lines for index measurement.

At paraxial level, $y = u_p x$, where u_p is the slope of the principal ray and *x* is the distance from the second principal plane to the image. (If the object is at infinity it is *f*, the effective focal distance). If this distance depends on color, it appears lateral chromatism.

Axial chromatism is found when the distance from the last surface to the image varies with color. (If the object is at infinity, it is f_p, the back focal distance).

Achromatization

We shall see simple cases where some of these aberrations can be eliminated. For a single lens is

(4.30) $$\frac{1}{f} = (n-1)(c_1 - c_2) = (n-1)\,k\,,$$

where k does not depends on λ.

In a compound system of two lenses there will be, putting for brevity,

(4.31) $$k_1 = c_{1,1} - c_{2,1} \quad ; \quad k_2 = c_{1,2} - c_{2,2}$$

(4.32) $$\frac{1}{f} = k_1\,(n_1 - 1) + k_2\,(n_2 - 1) - d\,k_1\,k_2\,(n_1 - 1)(n_2 - 1)$$

If it is wanted achromatization in a certain spectral range, the differential of this expression respect to n_1 y n_2 must vanish.

Results

(4.33) $$d = \frac{1}{k_1\,k_2}\;\frac{k_1\,\delta n_1 + k_2\,\delta n_2}{(n_2 - 1)\,\delta n_1 + (n_1 - 1)\,\delta n_2}$$

Where δn_1 is approximately $n_{1F} - n_{1C}$, and similarly δn_2

If the lenses are of the same glass,

(4.34) $$d = \frac{k_1 + k_2}{2\,(n-1)\,k_1 k_2} = \frac{f_1 + f_2}{2}$$

This arrangement, discovered by Huygens, is used in simple eyepieces and corrects the lateral, or magnification chromatism.

Achromatization of thin lenses in contact (telescope objectives)

In this case the problem is to correct the longitudinal chromatism. The lateral one is zero because the principal planes coincide with the plane of the thin lens and $f = f_p$

There must be used different glasses and this lead us to the study of the optical properties of glass.

By (4.30) is

(4.35) $$(n_F - n_C)\,k = \frac{1}{f_F} - \frac{1}{f_C} = \frac{f_C - f_F}{f_C\,f_F} \approx \frac{f_C - f_F}{f_D^2}$$

Then

(4.36) $$\frac{f_C - f_F}{f_D} = (n_F - n_C)\,k\,f_D = \frac{n_F - n_C}{(n_D - 1)} = \frac{1}{v_D}$$

Where v_D is called the Abbe number, (referred to D line) indicating the chromatic dispersion for that glass. The larger is v, the lesser the dispersion. It may be put, for a single lens

(4.37) $$\left(\frac{\Delta f}{f}\right)_{primary} \approx \frac{1}{v_D}$$

To achromatize, we use a doublet (two glued lenses)

(4.38) $$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} = (n_1 - 1)\,k_1 + (n_2 - 1)\,k_2$$

If it is achromatic both focus F and C coincide

(4.39) $$\frac{1}{f_F} - \frac{1}{f_C} = (n_{F1} - n_{C1})\,k_1 + (n_{F2} - n_{C2})\,k_2 = 0$$

In terms of v's, (f_D can be called f without confusion)

(4.40) $$k_1\,\frac{n_{D1} - 1}{v_1} + k_2\,\frac{n_{D2} - 1}{v_2} = \frac{1}{v_1\,f_1} + \frac{1}{v_2\,f_2} = 0$$

That is

(4.41) $$\frac{f_1}{f_2} = -\frac{v_2}{v_1}$$

The design of the achromatic doublet consist of fulfilling equations (4.38) and (4.41), giving

(4.42) $$f_1 = \frac{v_1}{f\,(v_1 - v_2)} \qquad \text{and} \qquad f_2 = \frac{v_2}{f\,(v_2 - v_1)}$$

The spectrum of a single lens is displayed over a segment of length f/v .

In an achromatic doublet it is folded so that the extreme colors *F* and *C* coincide, but *D* remains at a distance called secondary spectrum.

Estimation of the secondary spectrum

As *F* and *C* coincide, in the formulae there may be used any, let be *F*.

(4.43) $$\frac{1}{f_F} - \frac{1}{f_D} = (n_{F1} - n_{D1})\,k_1 + (n_{F2} - n_{D2})\,k_2 \approx \frac{f_F - f_D}{f_D^2} = \frac{\Delta f}{f^2}$$

(4.44) $$= \frac{\Delta f}{f}\left((n_{D1} - 1)\,k_1 + (n_{D2} - 1)\,k_2\right)$$

Here Δf is the secondary spectrum

Then

(4.45) $$\left(\frac{\Delta f}{f}\right) = \frac{(n_{F1} - n_{D1})\,k_1 + (n_{F2} - n_{D2})\,k_2}{(n_{D1} - 1)\,k_1 + (n_{D2} - 1)\,k_2}$$

Because it is an achromatic doublet, is

(4.46) $$\frac{k_1}{k_2} = -\frac{n_{F2} - n_{C2}}{n_{F1} - n_{C1}}$$

On substitution in the former results

(4.47) $$\frac{\Delta f}{f} = \frac{-\dfrac{n_{F1} - n_{D1}}{n_{F1} - n_{C1}} + \dfrac{n_{F2} - n_{D2}}{n_{F2} - n_{C2}}}{-\dfrac{n_{D1} - 1}{n_{F1} - n_{C1}} + \dfrac{n_{D2} - 1}{n_{F2} - n_{C2}}}$$

This complex expression is dealt with in the same way as that of the primary spectrum, defining a typical quantity similar to the Abbe number, called partial dispersion.

(4.48) $$P_{FD} = \frac{n_F - n_D}{n_F - n_C}$$

Results

(4.49) $$\left(\frac{\Delta f}{f}\right)_{secondary} = \frac{P_2 - P_1}{v_2 - v_1}$$

To sum up, knowing the refractive indices of both glasses in λ_{short} , λ_{medium} , λ_{long} , it may be designed an achromatic lens and judged its secondary spectrum.

The primary spectrum of a single lens may lie between $1/30$ and $1/70$ of its focal distance. The secondary spectrum of a compound one is near $1/1600$ a $1/2000$ with common glass combinations. With special combinations it can reach $1/25000$ for the so called apochromats.

Use of catalogs

The optical glass factories edits large catalogs that are references to study all properties of glass, not just optical. The most ancient and complete is that of Schott. There is introduced an interpolation formula allowing the calculation of the index to 5 decimal places. The Schott formula is

(4.50) $$n^2 = A_0 + A_1\lambda^2 + \frac{A_2}{\lambda^2} + \frac{A_3}{\lambda^4} + \frac{A_4}{\lambda^6} + \frac{A_5}{\lambda^8}$$

Where λ must be in microns. In each catalog sheet there is a list of the 6 coefficients measured over a set of 18 values of λ, besides tables of experimental indices.

On designing a lens it may be important to know the resistance to atmospheric agents, the hardness, the bubble content, the dilatation coefficient and the Young modulus (On cementing two large lenses there appear stress if temperature changes).

The Schott standard nomenclature to identify an optical glass is, for instance

KzFS 1 − 613 443 , BK7− 517 642 etc.

First there is a code that relates to chemical composition.

Then, the first group of three figures is *1000 (n − 1),* e.g, *1.517 → 517*

The second group is *10 v*, e.g*, 64.2 → 642*.

In the catalogs there are data about partial dispersions and they may be searched for apochromatic combinations with formula (4.49)

General ray tracing

It is not indispensable for all would be astronomer to know about this, but always there is some (is hoped) astronomer interested in optics and she or he may find it useful. The ray tracing formulas can be more general than what follows, and they are the *ultimate truth* in geometrical optics, computer willing. *It is important to rescue them before they sink in the inner workings of canned programs.*

Meridional rays are a subset of the rays forming the image.

The general trace is more complex, and the formulae are useful only if they are intended to be programmed.

The equations will be developed for coaxial surfaces of revolution of any shape, not only spheres.

Each surface is linked to a local coordinate system with its x–axis as the one of revolution.

The y,z plane containing the pole of the surface is called polar tangent plane.

A ray is defined by 6 parameters, (although one is redundant). The first three are the components of a vector from the origin to the point, and the others are the components of a unit vector along the ray.

The trace can be decomposed in three operations

-Intersection. With the ray in the polar tangent plane, to carry it to the surface

-Refraction. To find the new direction of the ray.

-Transport. To carry the ray to the next polar tangent plane.

Intersection

The initial data of the ray are its position y_0, z_0 in the polar tangent plane and its direction cosines k, l, m.

The mathematical procedure differs depending on the surface being a conic of revolution (quadric) or a higher order aspheric. In the first case the intersection is done by an exact closed method, and in the second it must be resorted to an iteration.

Almost all surfaces in optics are quadrics, with the sphere as leading case.

Due to the simplicity of the formulae, quadrics are separately dealt with to speed the calculation. The final data of intersection are the coordinates *x, y, z* of the intersection point, but also there must be found the direction cosines α, β, γ of the normal in that point, to be used later to find the direction of the refracted ray.

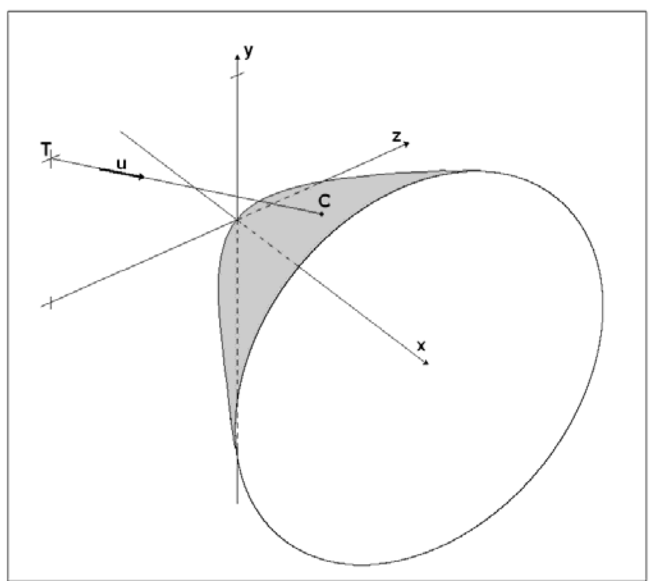


Fig. 4.12

Intersection with quadrics.

A quadric has an equation

(4.51) $$f(x,y,z) = x - \tfrac{1}{2} c \left((1+b)\,x^2 + H\right) = 0 \text{, where}$$

$$H = y^2 + z^2$$

$c = 1/R$ is the vertex curvature and has single value only at this place. If $c > 0$, the concavity lies on positive x.

R is the radius of curvature·

b is a parameter defining the type of quadric and we will call it citric constant.

$b > 0$	Flat ellipsoid	(mandarin)
$b = 0$	Sphere	(orange)
$0 > b > -1$	Long ellipsoid	(lemon)
$b = -1$	Paraboloid	
$b < -1$	Hyperboloid	

For closed quadrics the formula represents only the first half.

The segment of ray $\overline{TC}$ (Fig 4.12) has a vector equation

(4.52) $$\mathbf{S} = \mathbf{T} + \lambda \mathbf{u}$$

$\mathbf{u}$ is the unit vector along the ray.

In components

(4.53) $$x = x_0 + \lambda\, k$$

(4.54) $$y = y_0 + \lambda\, l$$

(4.55) $$z = z_0 + \lambda\, m$$

On the polar tangent plane is $x_0 \equiv 0$, but the symbol is retained because in other cases it is not. λ is a distance along the ray, to be found.

As x, y, z satisfy simultaneously the ray and quadric equation, it follows that

(4.56) $$\lambda\, k - \frac{c}{2}\left((1+b)(\lambda\, k)^2 + (y_0 + \lambda\, l)^2 + (z_0 + \lambda\, m)^2 \right) = 0$$

(Here is explicitly $x_0 \equiv 0$)

Decomposing the expression and taking account that

(4.57) $$k^2 + l^2 + m^2 = 1$$

Results

(4.58) $$A\,\lambda^2 - 2\,B\,\lambda + D = 0$$

Where

(4.59) $$A = c\left(1 + b\,k^2\right)$$

(4.60) $$B = k - c\left(l\,y_0 + m\,z_0\right)$$

(4.61) $$D = c\left(y_0^2 + z_0^2\right)$$

The roots of the equation are

----(4.62) $$\lambda = \frac{B \pm \sqrt{B^2 - A\,D}}{A}$$

It must be taken the root belonging to the near sheet. It is identified by considering the case $c = 0$, in which the surface reduces to the plane y,z; and $\lambda = 0$

The $+$ sign gives $\lambda = \infty$ and the $-$ sign an indetermination type $0/0$. It is removed transforming the expression so that the root in the numerator disappears.

Results

(4.63) $$\lambda = \frac{D}{B + \sqrt{B^2 - A\,D}}$$

The intersection follows from (4. 59), (4. 60), (4. 61), (4. 63), (4. 53),(4. 54), (4. 55)

If $B^2 \prec A\,D$, λ is imaginary and the ray does not intersects the surface.

Normal line

Let be a displacement $d\mathbf{s}$ over a surface given by $f(x,y,z)=0$

It will be

(4.64) $$df = f_x dx + f_y dy + f_z dz = 0 = \nabla f \cdot d\mathbf{s}$$

Where, following the usual notation is $f_x = \partial f / \partial x$, etc.

The gradient ∇f is then perpendicular to any displacement over the surface.

Then, ∇f is on the normal. The components of the unit normal vector are the partial derivatives normalized to *1*.

Calling them α, β, γ , is

(4.65) $$\alpha = f_x / N$$

(4.66) $$\beta = f_y / N$$

(4.67) $$\gamma = f_z / N$$

Where the norm is

(4.68) $$N = \sqrt{f_x^2 + f_y^2 + f_z^2}$$

From (4.51) results

(4.69) $$f_x = 1 - c(1+b)x$$

(4.70) $$f_y = -c\,y$$

(4.71) $$f_z = -c\,z$$

The normal is given by equations (4.69), (4.70), (4.71), (4.68), (4.65), (4.66),(4.67)

Intersection with higher order aspheric. (up to order 8).

The iterative method is illustrated in Fig 4.13, where to simplify it was used a meridional ray.

In the drawing, the curve was much exaggerated to show the successive approximations and avoid merging the points.

The point P of intersection is search by the sequence of points *0* , $\underline{0}$, *1* , $\underline{1}$, *2* , $\underline{2}$, etc.

Points *0* , *1* , *2* are on the ray but not on the aspheric, and points $\underline{0}$, $\underline{1}$, $\underline{2}$ are on the aspheric but not on the ray.

Due to the fast convergence, point *0* is arbitrary. (Here it is not on the polar tangent plane also by clarity of the drawing).

From point $0\,(0, y_0, z_0)$ it goes to $\underline{0}\,(\underline{x_0}, y_0, z_0)$ replacing $\underline{x_0}$ found by introducing y_0, z_0 in equation (4.73), explicit of the aspheric. After it is traced the tangent plane (4.74) through $\underline{0}$ and it is found the intersection of the ray with it. This gives point $1\,(x_1, y_1, z_1)$

It goes to $\underline{1}$ in the same way as to $\underline{0}$ and to *2* as to *1.*

The process is repeated *N* times until $x_N - \underline{x_N}$ be negligible.

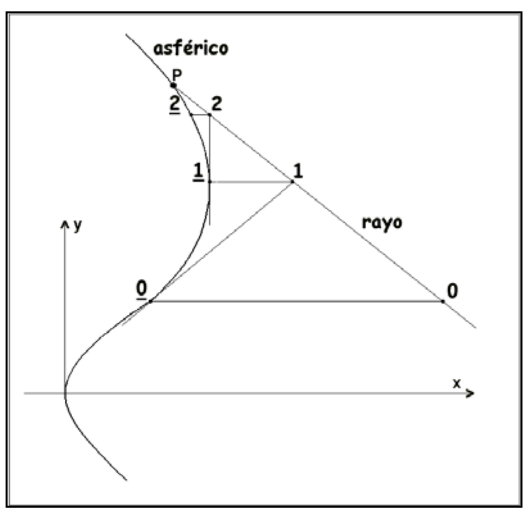


Fig. 4.13

Explicit equation of the aspheric.

It will be so that its first term belongs to a quadric and the following ones to its higher order deformations.

As the surface is of revolution about the *x* axis, its development has only terms of even degree.

Solving for *x* in (4.51), the first term is

(4.72) $$x = \frac{1-\sqrt{1-(1+b)\,c^2 H}}{c\,(1+b)}$$

This expression may be put in an alternative form in the same way it goes from (4.62) to (4.63), to which is added later higher order terms

(4.73) $$x = \frac{c\,H}{1+\sqrt{1-c^2(1+b)\,H}} + d\,H^2 + e\,H^3 + f\,H^4$$

Tangent plane

If in equation (4.64) there are replaced differentials by finite increments and the derivatives are normalized, it follows the equation of the tangent plane.

(4.74) $$(x - \underline{x_0})\,\alpha + (y - y_0)\,\beta + (z - z_0)\,\gamma = 0$$

This is the equation of the tangent plane passing by point $\underline{0}\,(\underline{x_0}, y_0, z_0)$ and having a normal with direction cosines α, β, γ . The derivatives are found with the explicit equation (4.73) in the form

(4.75) $$x - \frac{c\,H}{1+\sqrt{1-c^2(1+b)\,H}} - d\,H^2 - e\,H^3 - f\,H^4 = 0$$

Results

(4.76) $$f_x = 1$$

(4.77) $$f_y = -G\,y$$

(4.78) $$f_z = -G\,z$$

Where

(4.79) $$G = \frac{c}{\sqrt{1-(1+b)\,c^2 H}} + 4\,d\,H + 6\,e\,H^2 + 8\,f\,H^3$$

The norm is

(4.80) $$N = \sqrt{1+G^2 H}$$

These are the values to be replaced in equations (4.65), (4.66), (4.67)

Although it seems simpler, it cannot be differentiated the implicit function (4.51) with the addition of the higher order terms because the series development of the quadric also has higher order terms, that should be added to the others.

The higher order terms are defined as departures from the quadric.

The intersection of the tangent plane with the ray is found in a similar way as (4.56), replacing in (4.74) the equations of the ray passing through $(0, y_0, z_0)$, which are

(4.81) $$x_1 = x_0 + \lambda\,k \qquad (\,x_0 = 0\,)$$

(4.82) $$y_1 = y_0 + \lambda\,l$$

(4.83) $$z_1 = z_0 + \lambda\,m$$

Results

(4.84) $$(x_0 + \lambda\,k - \underline{x_0})\,\alpha + \lambda\,l\,\beta + \lambda\,m\,\gamma = 0$$

The value of λ here is

(4.85) $$\lambda = \frac{(\underline{x_0} - x_0)\,\alpha}{\alpha\,k + \beta\,l + \gamma\,m}$$

The quantities with sub-index *1* replace those with sub-index *0* and the process is repeated, but no longer will be $x_0 = 0$

The iterated intersection is given by equations (4.76), (4.79), (4.77), (4.78), (4.80), (4.65), (4.66), (4.67), (4.73), (4.85), (4.81), (4.82), (4.83).

The lack of intersection is pointed out by the root in (4.73).

Refraction

Vector form of Snell law

Let be **N** the unit vector normal to the surface, **u**,**u**′ unit vectors along the incident and refracted rays, and *n,n′* the respective indices of refraction.

The coplanarity of **N**,**u**,**u**′ is established by saying that there exists three parameters to be found, such that

(4.86) $$\lambda_1 \mathbf{u} + \lambda_2 \mathbf{u}' + \lambda_3 \mathbf{N} = 0$$

Defining I = incidence angle, I' = refraction angle

Results, on scalar multiplication by **N**

(4.87) $$\lambda_1 \cos I + \lambda_2 \cos I' + \lambda_3 = 0$$

And on vector multiplication by **N**

(4.88) $$\lambda_1 \,sen\, I + \lambda_2 \,sen\, I' = 0$$

Using Snell law

(4..89) $$n \,sen\, I - n' \,sen\, I' = 0$$

It can be found the values of $\lambda_1, \lambda_2, \lambda_3$

(4.90) $$\lambda_1 = n$$

(4.91) $$\lambda_2 = -n'$$

(4.92) $$\lambda_3 = n' \cos I' - n \cos I$$

Equation (4.86) permits to solve for the versor **u**′

(4.93) $$\mathbf{u}' = \frac{n}{n'}\mathbf{u} + \left(\cos I' - \frac{n}{n'}\cos I\right)\mathbf{N}$$

The right member may be written in terms of known quantities

(4.94) $$\cos I = \alpha k + \beta l + \gamma m$$

(4.95) $$\cos I' = \sqrt{1 - \frac{n^2}{n'^2}\left(1 - \cos^2 I\right)}$$

Calling

(4.96) $$\Omega = \cos I' - \frac{n}{n'}\cos I$$

Results the components of (4.93)

(4.97) $$k' = \frac{n}{n'}k + \Omega\alpha$$

(4.98) $$l' = \frac{n}{n'}l + \Omega\beta$$

(4.99) $$m' = \frac{n}{n'}m + \Omega\gamma$$

Refraction is given by equations (4.94), (4.95), (4.96), (4.97), (4.98), (4.99)

If $\cos I'$ is imaginary in (4.95) the ray suffers total reflection.

Transport

Let be x, y, z the coordinates in the actual surface. The following polar tangent plane is at a distance D_x from the actual. Then, if y'_0, z'_0 are the ray coordinates on the following

(4.100) $$D_x = x + \lambda k'$$

(4.101) $$y'_0 = y + \lambda l'$$

4.102) $$z'_0 = z + \lambda m'$$

The value of λ is obtained from (4.100)

(4.103) $$\lambda = \frac{D_x - x}{k'}$$

The coordinates of transport to the following polar tangent plane are given by equations (4.103), (4.101), (4.102)

Next, all primes are dropped and it proceeds with the next surface. The reason of passing through this plane are that here it can be applied rotations or translations.

Lost rays

It was seen under what conditions some rays may be lost on intersection or refraction, but this seldom happens. The most common way for a ray to be lost is when it fall outside the rim defined by the clear radius of some surface, when the beam enters the system at large obliquity.

This results in a decrease of luminosity in the image towards the border of the field, and is called "vignetting". The lost rays has large aberrations and hence the image quality improves at expense of luminosity.

Spot diagram

A plot of y, z in the image plane is called spot diagram. It is the most accurate representation of the geometrical image. If the quantity of rays is large and their distribution is random, the diagram resembles an actual image taken with a grainy film. But this is an end result, and no clue remains about what parts of the system originated the aberrations.

Neither are taken account the diffraction effects, and by this reason is less accurate as the aberrations vanishes. The diffraction figure of aberrations can be modeled too, but a practical criterion valid in *almost* all cases is to consider a "perfect" system when the aberrations are of the order of the Airy disk. The exception are the systems with large central obscuration, see equation (3.39) and what follows.

Focussing

In a well corrected system, the image changes very much near the focus, and this data is not known in advance to place there the detector. The paraxial plane is often inadequate. It is possible to find the average back focal distance by a statistical treatment of the rays emerging from the last surface, using the criterion that the best geometrical image in presence of aberrations is such where the quadratic dispersion of the transverse coordinates y, z is least.

Any program plotting spot diagrams must have focussing to be useful.

To follow we must define some nomenclature.

Given a set of quantities $A_1, A_2 \ldots A_N$, the mean value is denoted by the symbol $\langle\ \rangle$, so that

(4.104) $$\langle A \rangle = \frac{\Sigma A_i}{N} \qquad (i = 1,2,\ldots N)$$

And the mean quadratic dispersion by

(4.105) $$\sigma_{AA} = \langle (A - \langle A \rangle)^2 \rangle$$

From the definition

(4.106) $$\sigma_{AA} = \langle \left(A^2 - 2\,A\,\langle A \rangle + \langle A \rangle^2\right) \rangle = \langle A^2 \rangle - \langle A \rangle^2$$

By analogy it may be defined for two sets A, B

(4.107) $$\sigma_{AB} = \langle A\,B \rangle - \langle A \rangle \langle B \rangle$$

Application of the former to a bundle of rays.

Let be y_0, z_0 the coordinates of the ray in the last polar tangent plane. On propagation over a distance x, it will be

(4.108) $$y = y_0 + x\,\frac{l}{k} = y_0 + x\,L$$

(4.109) $$z = z_0 + x\,\frac{m}{k} = z_0 + x\,M$$

Where it was defined $L = \frac{l}{k}$ and $M = \frac{m}{k}$ to keep the formulas simple.

The focus is the coordinate x_f on the plane where $\sigma_{yy} + \sigma_{zz}$ = minimum.

(4.110) $$\langle y\rangle = \langle y_0\rangle + x\,\langle L\rangle$$

(4.111) $$\langle y\rangle^2 = \langle y_0\rangle^2 + 2\,x\,\langle y_0\rangle\,\langle L\rangle + x^2 L^2$$

Also

(4.112) $$y^2 = y_0^2 + 2\,x\,y_0 L + x^2 L^2$$

(4.113) $$\langle y^2\rangle = \langle y_0^2\rangle + 2\,x\,\langle y_0 L\rangle + x^2 \langle L^2\rangle$$

Hence

(4.114) $$\sigma_{yy} = \langle y^2\rangle - \langle y\rangle^2 = \sigma_{y_0 y_0} + 2\,x\,\sigma_{y_0 L} + x^2 \sigma_{LL}$$

There is a similar expression for σ_{zz}

Hence

(4.115) $$\sigma_{yy} + \sigma_{zz} = \sigma_{y_0 y_0} + \sigma_{z_0 z_0} + 2\,x\left(\sigma_{y_0 L} + \sigma_{z_0 M}\right) + x^2\left(\sigma_{LL} + \sigma_{MM}\right)$$

The value of x_f is obtained by setting to zero the derivative respect to x

(4.116) $$x_f = -\frac{\sigma_{y_0 L} + \sigma_{z_0 M}}{\sigma_{LL} + \sigma_{MM}}$$

For a numerical calculation there are formed the sums

(4.117) $$S_{y_0} = \sum_N y_0 \quad ; \quad S_{y_0 L} = \sum_N y_0 L \quad ; \quad S_{LL} = \sum_N L^2$$; etc., similar

Results

(4.118) $$x_f = \frac{S_{y_0} S_L + S_{z_0} S_M - N\left(S_{y_0 L} + S_{z_0 M}\right)}{S_L^2 + S_M^2 - N\left(S_{LL} + S_{MM}\right)}$$

A continuation of this treatment may include:

1) Transverse displacements and tilts among the surfaces. It is useful for discussion of mounting tolerances and some special systems.

2) Toroidal surfaces. For very special systems.

3) Diffracted rays. Although it may seem paradoxical, they exist, and may be defined from the Fermat principle, that is more general than Snell law.–Remind that Snell law can be derived from Fermat principle--.They are very useful for a study of spectroscopes with non-planar gratings, and can be generalized to holograms (a diffraction grating is a particular case of hologram)

4) Diffraction pattern instead of spot diagram. It is used ray tracing to approximate the wavefront in the exit pupil and then applying the ideas discussed in sections Fraunhofer diffraction and aberrations, using FFT (Fast Fourier Transform) in the numerical calculation. This treatment is necessary for evaluation of a system placed in space. (The Hubble Space Telescope could be fixed thanks to an accurate model of its diffraction pattern transmitted to Earth).

5) Anisotropic or inhomogeneous materials. For study such things as Lyot and Solc filters, used in solar observation.

6) Vector diffraction. It is left the scalar Kirchhoff approximation, and may be studied anomalous behaviors of diffraction gratings.

7) The above is optical system evaluation. All intervening parameters may be put within an optimization procedure, where they are varied so as to maximize a given quality criterion.

To achieve all this is a very complex but very beautiful task.

Third order aberrations

The development of optical systems much preceded computers, and this made room for the invention of subtle theories to get maximum information with minimum numeric calculation.

The subtlety lies principally in the management of approximations.

A condition for the approximations be possible is that all slope angles U be small enough so

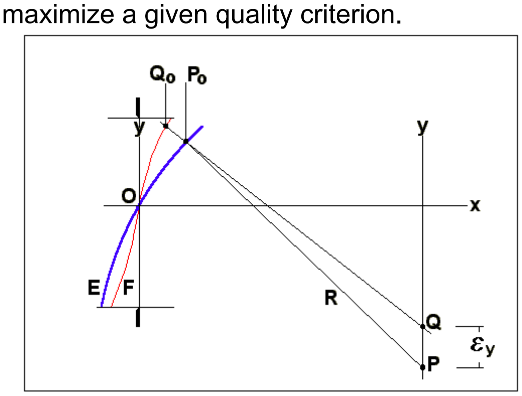


Fig. 4.14

4.125) $\sin U \approx U - \frac{U^3}{3}$; $\cos U \approx 1 - \frac{U^2}{2}$

If only the first term is retained, it follows paraxial formulae.

This approximation may therefore be called extended paraxial.

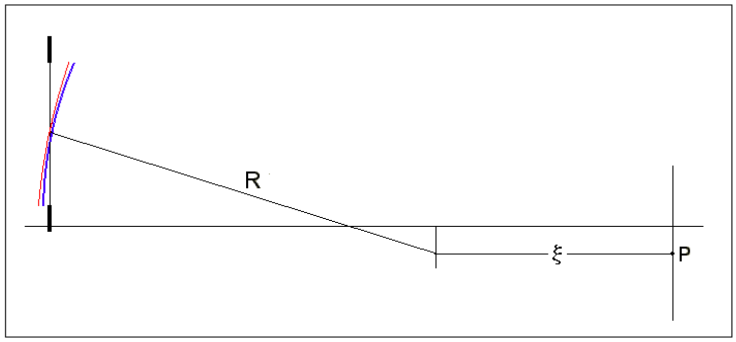


Fig. 4.15

Ray aberration and wavefront aberration.

Making more precise a former definition, it is called ray aberration to the segment located at the paraxial image plane going from the paraxial ray intersection to the exact intersection.

Its components are $\varepsilon_y, \varepsilon_z$.

There is also an alternative definition in terms of wavefronts.

In Fig. (4.14), F is an aberrated wavefront emerging from the exit pupil.

The normal to F intersects the paraxial plane at Q, and may have a z component.

E is a sphere with center at the paraxial image point P, and has a radius R such that at the center of the pupil there intersects the optical axis, the sphere and the aberrated wavefront.

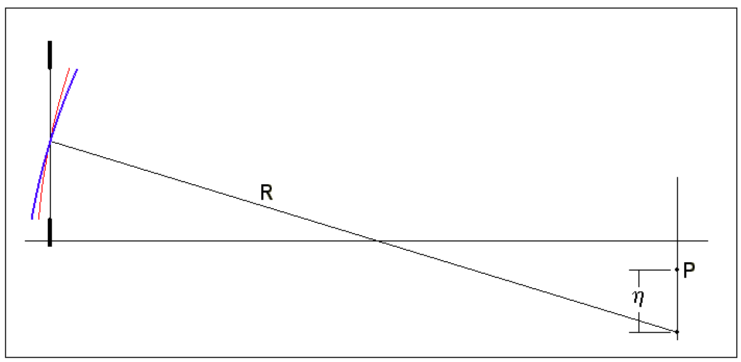


Fig. 4.16

E is called the reference sphere. A radius of E intersects it at P_0 and F at Q_0.

The segment $\overline{P_0 Q_0}$ is the wavefront aberration. W is a function of the coordinates over E, and approximately the same function over the coordinates y,z in the plane of the pupil.

If the image is perfect, F is also a sphere and $W \equiv 0$ by definition.

But still in this case, there may appear an aberration if the center of the reference sphere is displaced.

If it is displaced a quantity ξ along the x axis, the aberration is a quadratic function over the pupil: $W = A_1 \left(y^2 + z^2\right)$(Fig.4.15)

Of course, it is a valid approximation if one accepts that R changes in magnitude but not in direction, and the spheres are paraboloids.

If the center is displaced a quantity η along the y axis, the aberration is a linear function of y: $W = A_2\, y$ (Fig. 4.16). Here we accept that R changes direction but not magnitude.

In the drawings there is needed to make a compromise between accuracy and visibility, for the very small quantities involved are important if compared to a wavelength, also itself very small.

By now let the center be at the paraxial point.

The normal to the aberrated wavefront is an exact ray, and the normal to the sphere is a paraxial ray.

The angle between both is approximately $|\nabla W|$.

On reaching the paraxial image plane, their coordinates differ by a segment $\overline{PQ}$. Its components $\varepsilon_y, \varepsilon_z$ are the ray aberration. Hence , approximately,

(4.126) $$\varepsilon_y = R\frac{\partial W}{\partial y}$$

(4.127) $$\varepsilon_z = R\frac{\partial W}{\partial z}$$

Note that the coordinate systems x,y are identical over the pupil and over the paraxial plane.

For each object point at a different distance to the axis, or subtending a different angle with it if the object is at infinity; the form of $W(x,y)$ is different. Then, W depends on two parameters h,k, that are the coordinates of the object point in its plane, or the direction tangents if it is at infinity.

Hence

(4.128) $$W = W(y,z,h,k)$$

Derivation of the aberration polynomial by symmetry considerations

The expression may be developed in series, and to find a most simple form it is convenient to make symmetry considerations. As the optical system has cylindrical symmetry, a block rotation of the coordinate systems y,z and h,k about the x axis by an angle θ does not alter W.

To better imagine this, think about of a rotation of the optical system letting fixed the coordinates.

The invariance of W implies that the variables y,z,h,k enter only as combinations of type

(4.129) $$y^2+z^2\ ,\ h^2+k^2\ ,\ y\ h+z\ k$$

This may be checked with the formulae of transformation by rotation (Fig. 4.17).

(4.130) $Y = y\ cos\ \theta - z\ sin\ \theta \quad H = h\ cos\ \theta - k\ sin\ \theta$

(4.131) $Z = y\ sin\ \theta + z\ cos\ \theta \quad K = h\ sin\ \theta + k\ cos\ \theta$

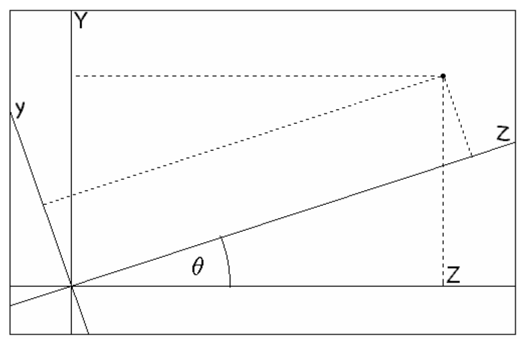


Fig. 4.17

For example

$Z\,K+Y\,H =$

$= (y\,sin\,\theta + z\,cos\,\theta)(h\,sin\,\theta + k\,cos\,\theta) + (y\,cos\,\theta - z\,sin\,\theta)(h\,cos\,\theta - k\,sin\,\theta)$

$= y\,h\,sin^2\,\theta + y\,k\,sin\,\theta\,cos\,\theta + z\,h\,cos\,\theta\,sin\,\theta + z\,k\,cos^2\,\theta + y\,h\,cos^2\,\theta$

$- y\,k\,cos\,\theta\,sin\,\theta - z\,h\,sin\,\theta\,cos\,\theta + z\,k\,sin^2\,\theta$

$= (z\,k + y\,h)sin^2\,\theta + (z\,k + y\,h)cos^2\,\theta$

$= z\,k + y\,h$

The symmetry also allows us to make identically zero the k coordinate because the system may be rotated until it be. That is, no generality is lost if the object lies in the meridional plane.

Hence

(4.132) $W = W\left(y^2+z^2\quad,\quad y\,h\quad,\quad h^2\right) =$

$= C + A_1\left(y^2+z^2\right) + A_2\ y\,h + A_3\ h^2 +$

$+ B_1\left(y^2+z^2\right)^2 + B_2\left(y^2+z^2\right)y\,h + B_3\,y^2h^2 + B_4\left(y^2+z^2\right)h^2 + B_5\,y\,h^3 +$

$+ B_6\ h^4 +$ higher order terms.

The A_i terms are linear in the new variables, and the B_i terms are quadratic.

As *E* and *F* intersects at the center of the pupil, $W(0,0,h) = 0$. Hence, terms without y or z vanishes. That is, $C = A_3 = B_6 = 0$.

Terms A_1 and A_2 vanishes because the reference sphere is centered at the paraxial point.

The survivors are terms $B_1 \cdots B_5$, and they do not vanishes in principle.

They form a polynomial in two variables producing a map of the pupil on the paraxial image plane.

This map is the image structure, or the image of a point object; is formed from *W* with equations (4.126), (4.127) and depends on the relative magnitude of the coefficients.

Equations (4.126), (4.127) are of degree *3* in *y, z, h* and this is why this theory is called "third order approximation".

<u>S</u>tudy of each term

We shall see the form of an image when W has only one term.

Sometimes there will be used polar coordinates for convenience, when it is easier to see the mapping and the influence of a diaphragm. To simplify we take $R = 1$

<u>S</u>pherical aberration

(4.133) $$W = B_1\left(y^2 + z^2\right)^2$$

(4.134) $$\varepsilon_y = 4\,B_1 y\left(y^2 + z^2\right) = 4\,B_1 \rho^3 \cos\theta$$

(4.135) $$\varepsilon_z = 4\,B_1 z\left(y^2 + z^2\right) = 4\,B_1 \rho^3 \sin\theta$$

As *W* does not depends on *h*, this is the only aberration present at the axis, and has the same form everywhere on the image.

Each circle $\rho = const$ in the pupil becomes a circle of radius $4\,B_1\rho^3$ in the image.

In Fig. 4.18 there are shown traces of these circles in the meridional plane. They may be considered rays, but they are neither paraxial nor exact, are of third order. In the pupil they are equidistant, and in the paraxial image plane their distances to the axis are proportional to ρ^3.

They form a typical fan called caustic. If $B_1 > 0$ the tip is to the right of the paraxial plane.

Shifting the receiving plane, a zone is reached where the rays are contained within a circle of least radius, called disk of least confusion. It is placed at *3 /4* of the road from the paraxial plane to the intersection of the marginal ray with the axis, as can be checked from the figure and also analytically.

Not necessarily is the disk of least confusion the best focus. The image has generally a nucleus and a halo, and it may be preferable a smaller disk with a dim halo. That is, the best focus may lie between this disk and the paraxial plane.

The caustic is the well known figure appearing in a cup of coffee.

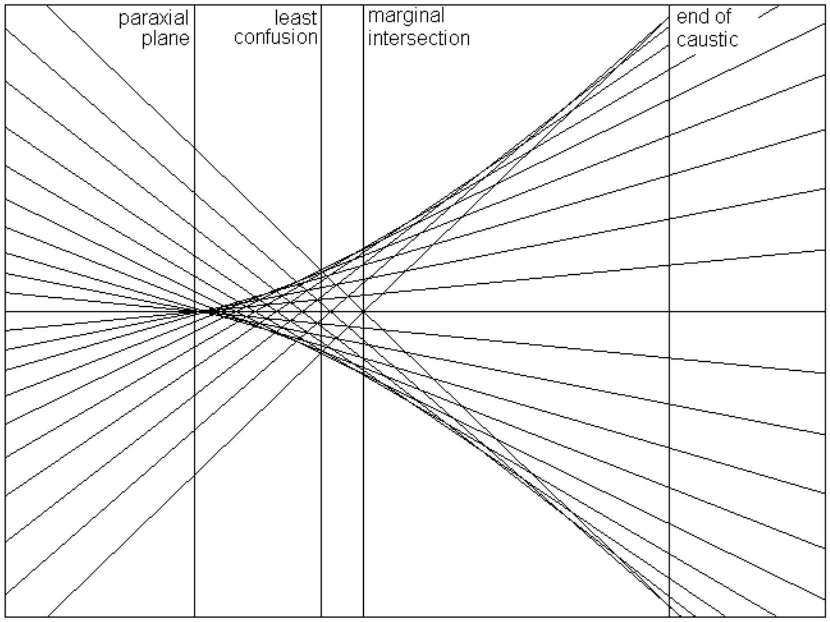


Fig. 4.18

Exercise.

Write a program to generate these figures (Program "CAUSTIC"). The versions .BAS and .EXE are in the page of contents

Coma

(4.136) $$W = B_2\left(y^2 + z^2\right) y\, h$$

(4.137) $$\varepsilon_y = B_2 h\left(3\, y^2 + z^2\right) = B_2 h\, \rho^2 \left(2 + \cos 2\theta\right)$$

(4.138) $$\varepsilon_z = 2\, B_2 h\; y\; z = B_2 h\, \rho^2\, \sin 2\theta$$

In (4.137) the parenthesis evolves in this way

$$3\cos^2\theta + \sin^2\theta = 1 + 2\cos^2\theta = 2 + \cos^2\theta - 1 + \cos^2\theta = 2 + \cos^2\theta - \sin^2\theta = 2 + \cos 2\theta$$

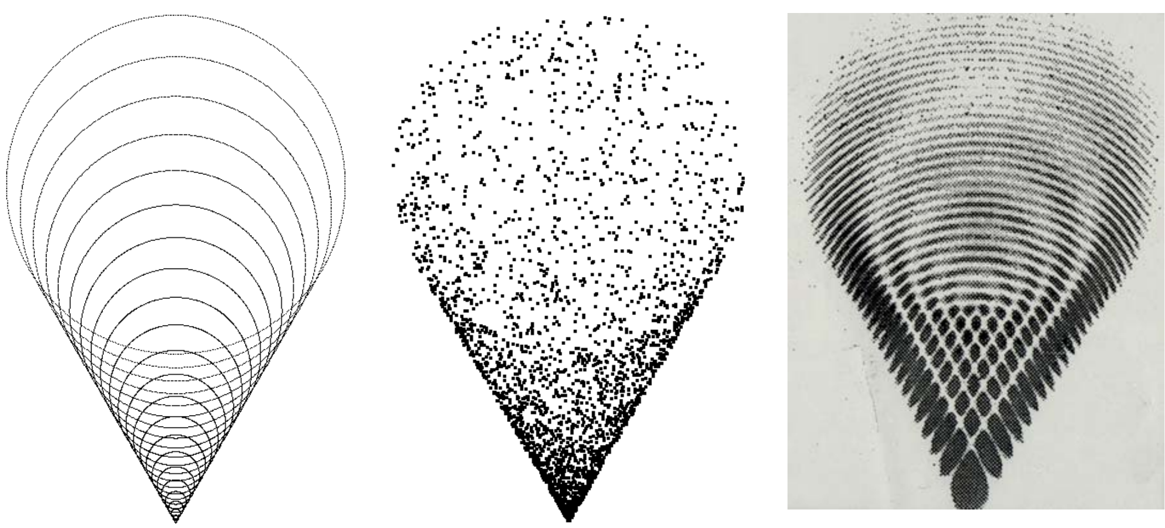


Fig 4.19 (a) (b) (c) (Born and Wolf)

When a circle is described in the pupil, a circle is described twice in the image, with radius $B_2\rho^2 h$ and center in $\varepsilon_y = 2B_2\rho^2 h$. Fig. 4.19(a) is simply a graph of equations (4.137) y (4.138)

IF $B_2 > 0$ the vertex points to the axis. The scale of the figure increases in a proportional way to h.

Fig. 4.19(b) is the coma of a paraboloidal mirror obtained by ray tracing. The image is more realistic because rays are randomly distributed over the pupil, and so it may be seen in a common telescope. The luminous intensity is proportional to the ray density.

La Fig. 4.19(c) is a photograph of an actual coma obtained in a laboratory. The fine structure is due to diffraction and is not seen in an earth-based telescope.

Coma owe its name to the appearance of a comet.

Field curvature

(4.139) $$W = B_4 h^2 \left(y^2 + z^2\right)$$

(4.140) $$\varepsilon_y = 2\,B_4 h^2 y$$

(4.141) $$\varepsilon_z = 2\,B_4 h^2 z$$

Field curvature and astigmatism are more clearly discussed if there is introduced an axial shift ξ of the center of the reference sphere. As it was shown before, this may be considered as an aberration, it is of the first order and will be called defocus.

In Fig. 4.20 the light cone converges to the point P without aberration.

Now will be taken a reference sphere displaced by ξ, of radius $R-\xi$, with $\xi << R$. The wavefront aberration is

(4.142) $$W_{defocus} = \left(\frac{1}{2\,(R-\xi)} - \frac{1}{2\,R}\right)\left(y^2 + z^2\right) \approx \frac{\xi\,\left(y^2 + z^2\right)}{2\,R^2}$$

And as has been taken $R = 1$, is

(4.143) $$\varepsilon_y = \xi\,y$$

(4.144) $$\varepsilon_z = \xi\,z$$

The same result may be reached by a shortcut considering similar triangles

$$\frac{\varepsilon_y}{\xi} = \frac{y}{R}$$

Including a defocus, the aberration is

(4.145) $$\varepsilon_y = 2\,B_4 h^2 y + \xi\,y$$

(4.146) $$\varepsilon_z = 2\,B_4 h^2 z + \xi\,z$$

The image is pointlike if $\xi = -2\,B_4 h^2$. The shift ξ is a quadratic function of *h,* and also of the height of the image, because at paraxial level it is proportional to the object's.

Hence, the image is pointlike on a quadratic surface ($\approx$ sphere). If $B_4 > 0$, the concavity points to the left. This aberration has no effect on the definition provided the receiving surface is adequately curved.

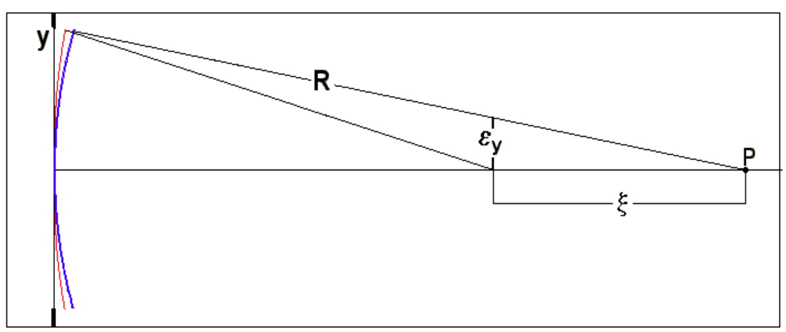


Fig. 4.20

Astigmatism

(4.147) $$W = B_3 y^2 h^2$$

(4.148) $$\varepsilon_y = 2\,B_3 h^2 y$$

(4.149) $$\varepsilon_z = 0$$

This term produces a wavefront having different curvature in the direction *y*, that is the meridional plane, while in the sagittal plane the curvature is only that of the reference sphere.

Including a defocus, the aberration is

(4.150) $$\varepsilon_y = 2\,B_3 h^2 y + \xi\, y$$

(4.151) $$\varepsilon_z = \xi\, z$$

We see that if $\xi = 0$, the image is a segment of length $2\,B_3 h^2 y$ on the meridional plane.

If $\xi = -2\,B_3 h^2$, it is a segment with the same length normal to the former, and in an intermediate point both components has a magnitude equal to half these segments. Here it is also formed a circle of least confusion analogous to the one of spherical aberration.

Fig. 4.21 shows the structure of an astigmatic beam.

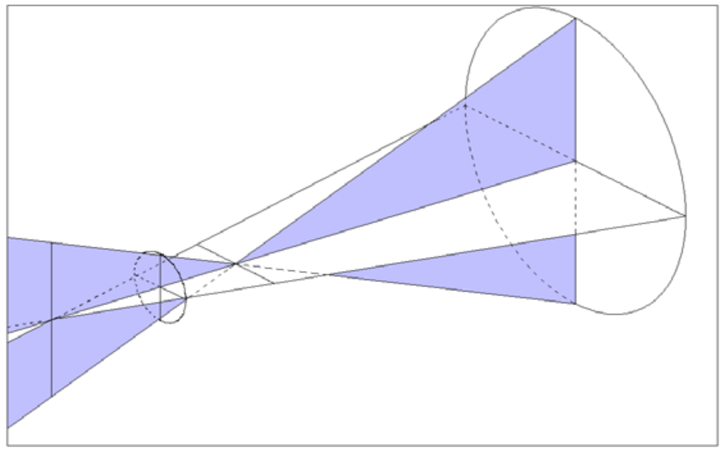


Fig. 4.21

By the factor h^2, the places where the aberration has this form lie on two curved surfaces, and their curvatures depends on the joint contribution of the coefficients B_3 and B_4.

To say somewhat a bit more intuitive, the image of a wagon wheel (if its axis coincides with the optical axis), focused on the paraxial plane has its rays sharp and the rim blurred; shifting the focus a place is reached where the rim is sharp and the rays blurred, and in a middle point there is a compromise of sharpness.

<u>D</u>istortion

(4.152) $$W = B_5 y\, h^3$$

(4.153) $$\varepsilon_y = B_5 h^3$$

(4.154) $$\varepsilon_z = 0$$

This aberration does not depends on the pupil coordinates, that is, the image is sharp but laterally displaced, and its height is no longer proportional to the object.

In Figs. 4.22 (a) and (b) there appears the image of a reticle in the paraxial plane.

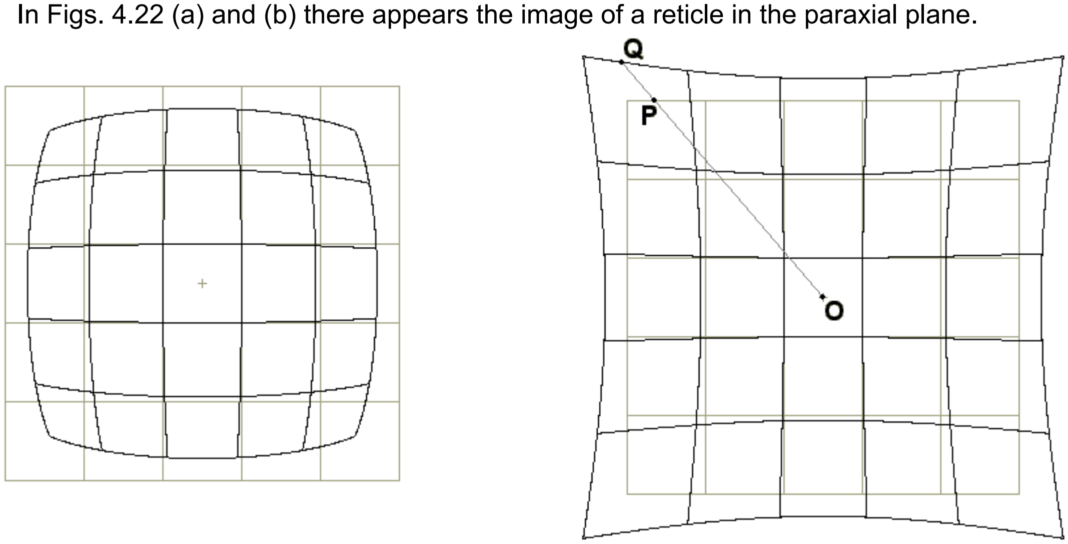


Fig. 4.22 (a) (b)

It is an extended image chosen for explaining the aberration and not an aberration figure.

To a more clear understanding of how the image of the reticle is formed, remember that it is always possible to place the object in the meridional plane.

The segment $\overline{OP}$ is h, and $\overline{OQ}$ goes from the axis to the aberrated intersection.

It is

(4.155) $$\overline{PQ} = B_5 h^3$$

Or

(4.156) $$\overline{OQ} = h\left(1 + B_5 h^2\right)$$

Figure(a) is for $B_5 < 0$ and (b) for the same value positive.

The correct image can be found by restitution, knowing the value of B_5.

Exercise

Write a program to generate these figures. (Program "DISTORTI") The versions .BAS and .EXE are in the page of contents

Aberrations "á la carte"

The structure of the image given by an optical system operating in the third order regime will be a composite of the various pure forms shown above. In Fig 4.23 there are several examples. In each image there is an arbitrary proportion of spherical aberration, coma, astigmatism and defocus, and it is defined by the condition that if presented alone, a value of 100 fills the square frame. The figures are generated from the aberration polynomial, not being linked to any optical system . The size in all cases, and the position in case of coma, are manually adjusted a posteriori, so that the figures are not to the same scale. The curious figures seen in an actual optical system are variations on this theme with some dissonances due to higher order terms.

Exercise

Write a program to generate these figures. (Program ABACART). The versions .BAS and .EXE are in the page of contents.

The former formulae are valid for the extended paraxial region, in the same sense that a series development with two terms applies in a larger interval than with a single term.

After the third order comes the fifth, the seventh, etc.

All optical systems have the same type of aberrations; what differs from one to other is the magnitude of the coefficients. *Moreover, in the derivation of the aberration polynomial it was not mentioned Snell law, only his symmetry properties. Then, the above is valid also , for example , to an electron microscope*.

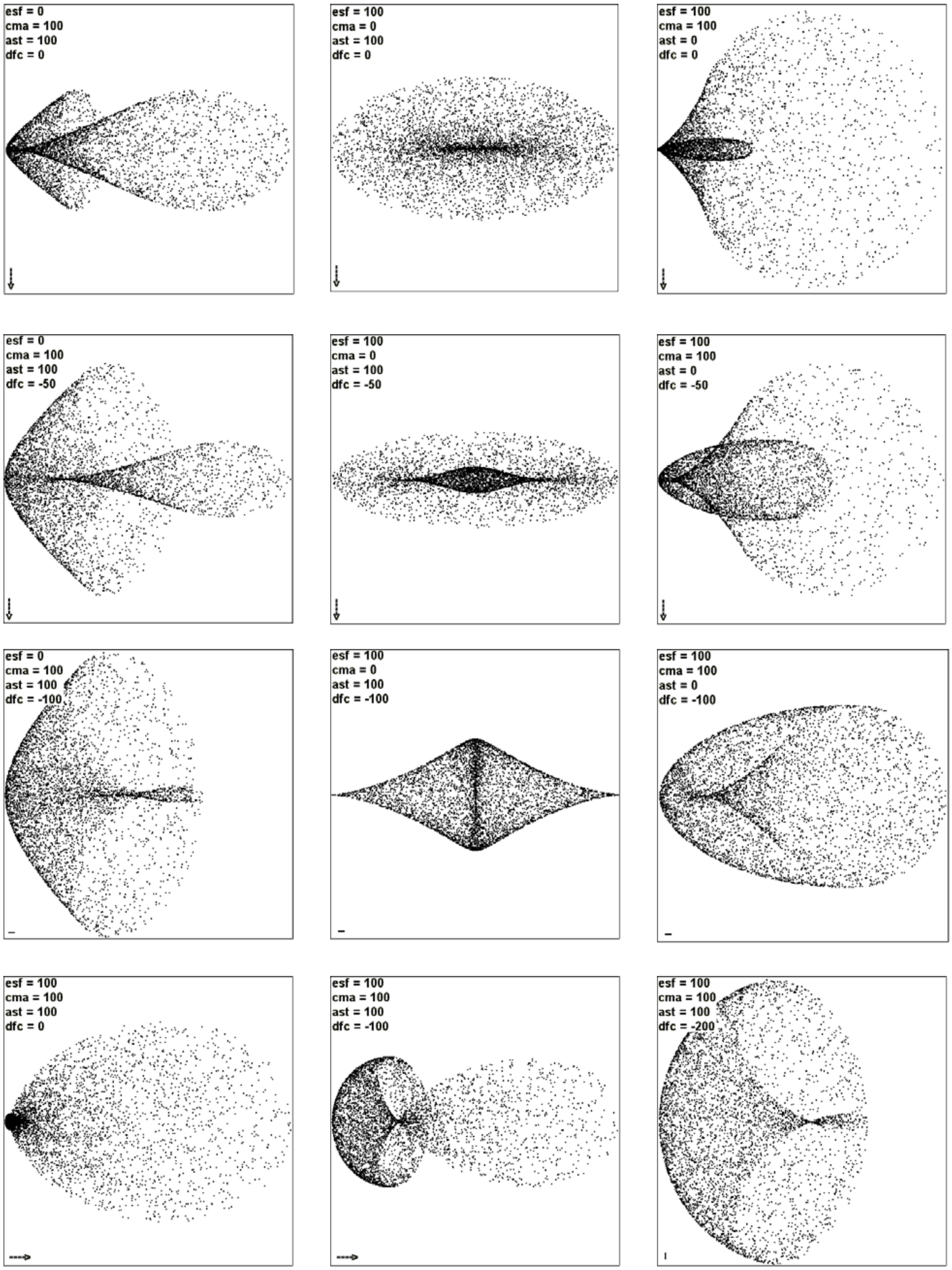
The most important feature of third order coefficients is that they can be computed in terms of the system parameters, and that they are the result of a sum over the surfaces, so that can be pointed the surfaces with heavier contribution to the aberration.

The formulae are an approximation to ray tracing, and it may be generated synthetic spot diagrams, to the third order. In the same measure as the field and aperture decreases, they must merge with the diagrams generated by ray tracing.

The generation of these synthetic diagrams imply a calculation of the numeric values of the coefficients in terms of construction parameters and object data,

....and will be left for another opportunity...

Figure 4.23


esf = 0
cma = 100
ast = 100
dfc = 0
esf = 100
cma = 0
ast = 100
dfc = 0
esf = 100
cma = 100
ast = 0
dfc = 0
esf = 0
cma = 100
ast = 100
dfc = -50
esf = 100
cma = 0
ast = 100
dfc = -50
esf = 100
cma = 100
ast = 0
dfc = -50
esf = 0
cma = 100
ast = 100
dfc = -100
esf = 100
cma = 0
ast = 100
dfc = -100
esf = 100
cma = 100
ast = 0
dfc = -100
esf = 100
cma = 100
ast = 100
dfc = 0
esf = 100
cma = 100
ast = 100
dfc = -100
esf = 100
cma = 100
ast = 100
dfc = -200