Grazing Incidence Focusing

Hongjie Lu

2018-02

Contents

1 Geometrical optics

1.1 Fermat's principle

Fermat's principle states that "light travels between two points along the path that requires the least time, as compared to other nearby paths." Thus it is mostly called "Fermats principle of least time".

A more accurate statement of Fermat's principle: Any hypothetical small change in the actual path of a light ray would only result in a second order change in the optical path length. The first order change in the optical path length would be zero.

1.1.1 Reflection

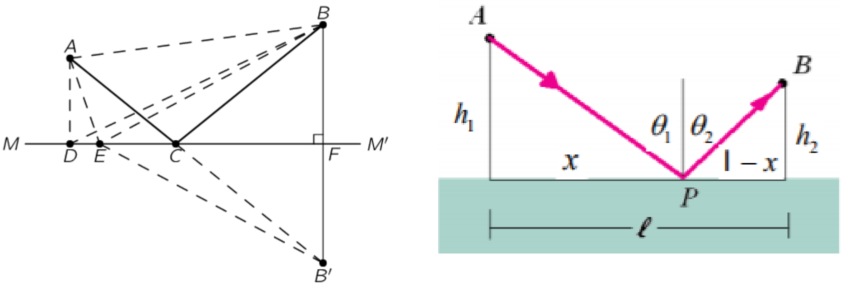


Figure 1: Illustration of Fermats principle for reflection

We can get the optical path length from the reflection case that:

$$L = \sqrt{x^2 + h_1^2} + \sqrt{(l-x)^2 + h_2^2} \tag{1}$$

To minimize the optical path length or travel time we set the dericative of the equation with respect to x equal to zero.

$$\frac{dL}{dx} = \frac{x}{\sqrt{x^2 + h^2}} + \frac{-(l-x)}{\sqrt{(l-x)^2 + h_2^2}} = 0 \tag{2}$$

$$\frac{x}{\sqrt{x^2 + h^2}} = \frac{(l-x)}{\sqrt{(l-x)^2 + h_2^2}} \tag{3}$$

$$sin\theta_1 = sin\theta_2 \tag{4}$$

$$\theta_1 = \theta_2 \tag{5}$$

1.1.2 Refraction

Now we consider a light ray traveling from point A to point B in media with different indices of refraction. The optical path length, which is also proportional to the travel time, between the two points is the product of the geometric length of the path and the index of refraction of the medium.

$$L = n_1\sqrt{x^2 + h_1^2} + n_2\sqrt{(l-x)^2 + h_2^2} \tag{6}$$

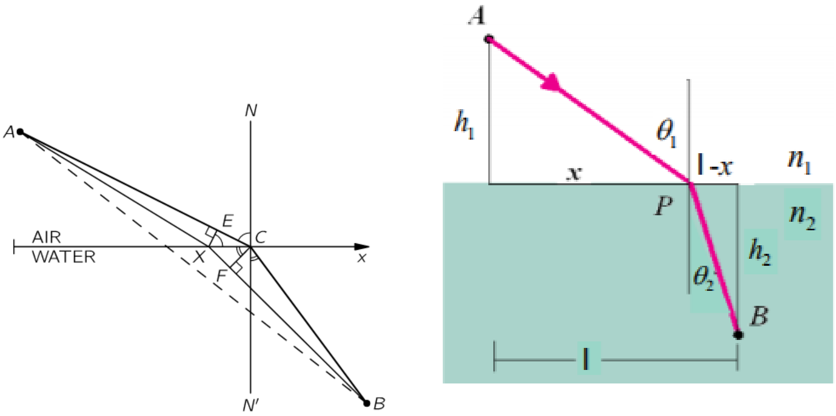


Figure 2: Illustration of Fermats principle for refraction

To minimize the optical path length we set the dericative of the equation with respect to x equal to zero.

$$\frac{dL}{dx} = n_1 \frac{x}{\sqrt{x^2 + h^2}} + n_2 \frac{-(l-x)}{\sqrt{(l-x)^2 + h_2^2}} = 0 \tag{7}$$

$$n_1 \frac{x}{\sqrt{x^2 + h^2}} = n_2 \frac{(l-x)}{\sqrt{(l-x)^2 + h_2^2}} \tag{8}$$

$$n_1 sin\theta_1 = n_2 sin\theta_2 \tag{9}$$

which is Snell's law

1.2 Paraxial optics

1.2.1 Single spherical surface

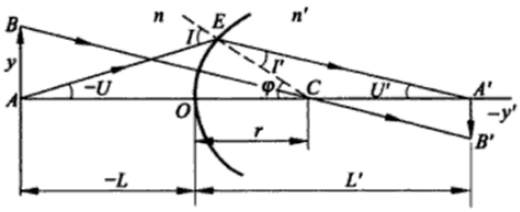


Figure 3: Single spherical surface

From the figure, we can have:

$$\frac{sin(-U)}{r} = \frac{sin(I)}{r-L} \tag{10}$$

$$\Rightarrow sinI = \frac{L-r}{r} sinU \tag{11}$$

$$n' sinI' = nsinI \tag{12}$$

$$\varphi = I + U = I' + U' \tag{13}$$

$$\Rightarrow U' = I + U - I' \tag{14}$$

$$\frac{sinU'}{r} = \frac{sinI'}{L'-r} \tag{15}$$

$$\Rightarrow L' = r + r\frac{sinI'}{sinU'} \tag{16}$$

U and L are the function of U' and L', so we can get U' and L' when provided U and L. The specific rays from object point with fixed U and L will get the image point with the same U' and L', but one object point normally can have different angles of U, which means the image through the spherical refractive surface will not focus at one point, as shown in the figure.

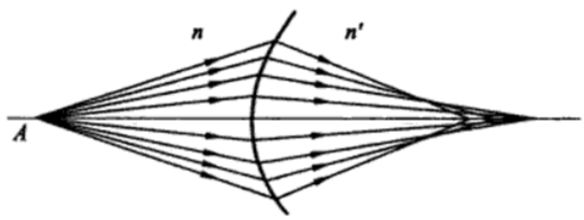


Figure 4: Non-perfect imaging of single spherical surface

1.2.2 Paraxial approximation

If we confine the rays in a small angle to axis, which means the value of angle U, I, U', and I' is small, here replaced by u, i, u', and i' respectively. And L and L' are replaced by l and l'. So we can get following equations when we approcimate the angle sin value with its arc value:

$$\frac{-u}{r} = \frac{i}{r-l} \tag{17}$$

$$\Rightarrow i = \frac{L-r}{r} u \tag{18}$$

$$n'i' = ni \tag{19}$$

$$\varphi = i + u = i' + u' \tag{20}$$

$$\Rightarrow u' = i + u - i' \tag{21}$$

$$\frac{u'}{r} = \frac{i'}{l'-r} \tag{22}$$

$$\Rightarrow l' = r + r\frac{i'}{u'} \tag{23}$$

We can derive that $l' = \frac{n'l}{n'l-n(l-r)}$, l' is independent with the angle u, which means the paraxial rays of one object point can perfectly form a image point. Another form of the last equation is:

$$\frac{n'}{l'} - \frac{n}{l} = \frac{n'-n}{r} = \varphi \tag{24}$$

We call φ the focal power of the specific medium and surface, which indicates the degree of optical system converging or diverging light. The SI unit for optical power is the inverse metre(m^{-1}). or in form of:

$$n'\left(\frac{1}{r} - \frac{1}{l'}\right) = n\left(\frac{1}{r} - \frac{1}{l}\right) = Q \tag{25}$$

Q is called the Abbe Invariant.

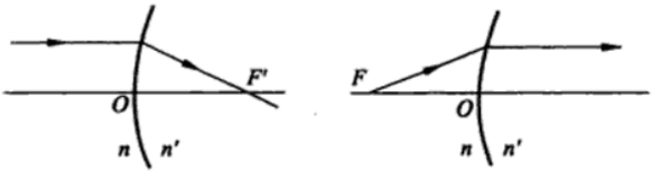


Figure 5: Object/Image at infinite distance

If the object is at infinite distance, $l \to -\infty$

$$l'_{l=-\infty} = f' = \frac{n'}{n'-n}r \tag{26}$$

Similiarly, if the image is at infinite distance, $l' \to \infty$

$$l_{l'=-\infty} = f = -\frac{n}{n'-n}r \tag{27}$$

$$f + f' = r \tag{28}$$

$$\varphi = \frac{n'}{f'} = -\frac{n}{f} \tag{29}$$

$$\frac{f'}{f} = -\frac{n'}{n} \tag{30}$$

1.2.3 Transverse magnification

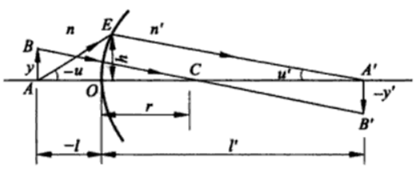


Figure 6: Transverse magnification

Transerve magnification is the height of the image divided by that of object.

$$\beta = \frac{y'}{y} \tag{31}$$

$$\frac{-y'}{y} = \frac{l'-r}{-l+r} \tag{32}$$

$$\Rightarrow \frac{y'}{y} = \frac{l'-r}{l-r} \tag{33}$$

$$\beta = \frac{y'}{y} = \frac{nl'}{n'l}(based\ on\ equation\ before) \tag{34}$$

1.2.4 Longitude magnification

Longitude magnification is the relation between displacements of a pair of conjugate points. Longitude magnification α is the object axial displacement divided by the image axial displacement.

$$\alpha = \frac{dl'}{dl} \tag{35}$$

from former equation $\frac{n'}{l'} - \frac{n}{l} = \frac{n'-n}{r}$, we take the derivative and get,

$$-\frac{n'dl'}{l'^2} + \frac{ndl}{l^2} = 0 \tag{36}$$

$$\alpha = \frac{dl'}{dl} = \frac{nl'^2}{n'l^2} = \frac{n'}{n}\beta^2 \tag{37}$$

We can see the longitude magnification is always positive, which means the image moves in the same direction as object moves. This equation is only applible for the adjacence of the point. Similarly, we can derive the longitude magnification for two axial object point,

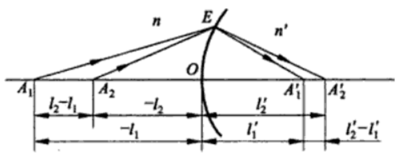


Figure 7: Longitude magnification

$$\alpha = \frac{l'_2 - l'_1}{l_2 - l_1} = \frac{n'}{n}\beta_1\beta_2 \tag{38}$$

1.2.5 Angular magnification

Angular magnification is defined as the ratio of angles between conjugate paraxial rays and optical axis.

$$\gamma = \frac{u'}{u} \tag{39}$$

$$for\ paraxial\ case,\ lu = l'u',\ then, \tag{40}$$

$$\gamma = \frac{l}{l'} = \frac{n}{n'}\frac{1}{\beta} \tag{41}$$

1.2.6 Relation among magnifications

From above, the relation among three magnifications shows $\alpha\gamma = \beta$

1.2.7 Lagrange-Helmoltz invariant

From $\beta = \frac{y'}{y} = \frac{nl'}{n'l}$ and $lu = l'u'$, we can get the Lagrange-Helmoltz invariant $J = nuy = n'u'y'$. It is the production of object height, aperture angle and medium index.

1.2.8 Reflective spherical mirror

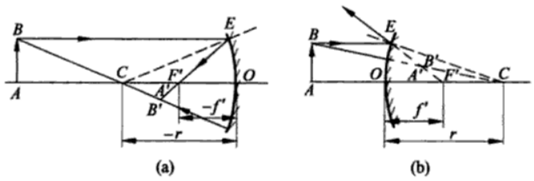


Figure 8: Reflective spherical mirror

Similarly, for reflective spherical lens, we can make $n' = -n$, and will get corresponding equations.

$$\frac{1}{l'} + \frac{1}{l} = \frac{2}{r} \tag{42}$$

for infinite image or object

$$f' = f = \frac{r}{2} \tag{43}$$

and for the magnifications,

$$\beta = \frac{y'}{y} = \frac{nl'}{n'l} = -\frac{l'}{l} \tag{44}$$

$$\alpha = \frac{n'}{n}\beta^2 = -\beta^2 \tag{45}$$

$$\gamma = \frac{n}{n'}\frac{1}{\beta} = -\frac{1}{\beta} \tag{46}$$

1.2.9 Thin lens

A thin lens is a lens with a thickness(distance along the optical axis between the two surfaces of the lens) that is negligible compared to the radii of curvature of the lens surfaces. The thin lens approximation ignores optical rffects due to the thickness of lenses and simplifies ray tracing calculations. It is often combined with the paraxial approximation in techiques such as ray transfer matrix analysis. The assumptions for thin lens are $n_1 = n_2' = 1$(air), $n_1' = n_2 = n$(lens index)

$$n\left(\frac{1}{r_1}-\frac{1}{l_1'}\right)=1*\left(\frac{1}{r_1}-\frac{1}{l_1}\right) \tag{47}$$
$$n\left(\frac{1}{r_2}-\frac{1}{l_2'}\right)=1*\left(\frac{1}{r_2}-\frac{1}{l_2}\right) \tag{48}$$
$$l_2=l_1' \tag{49}$$
$$\Rightarrow (n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)=\frac{1}{l'}-\frac{1}{l} \tag{50}$$

For infinite object or image, we can get the focus length,

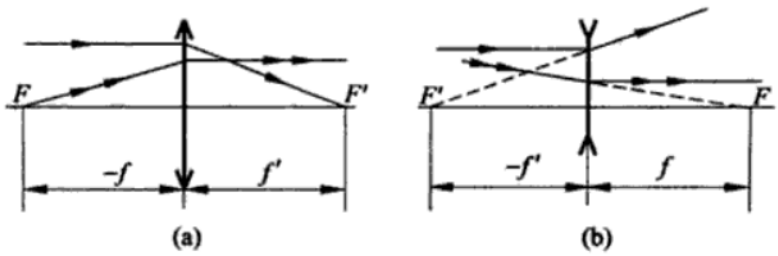


Figure 9: Convex and concave

$$f'=-f=\frac{1}{(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)} \tag{51}$$
$$\varphi=\varphi_1+\varphi_2=\frac{1}{f'}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right) \tag{52}$$
$$\frac{1}{l'}-\frac{1}{l}=\frac{1}{f'} \tag{53}$$

1.2.10 Cardinal points

Gaussian optics(paraxial ray-tracing), or coaxial ideal optical system, normally extends the results of the paraxial approximation.
In Gaussian optics, the cardinal points consist of three pairs of points located on the optical axis of a rotationally symmetric, focal, optical system. These are the focal points, the principal points, and the nodal points. For ideal systems, the basic imaging properties such as image size, location, and orientation are completely determined by the locations of the cardinal points; in fact only four points are necessary: the focal points and either the principal or nodal points. The only ideal system that has been achieved in practice is the plane mirror, however the cardinal points are widely used to approximate the behavior of real optical systems. Cardinal points provide a way to analytically simplify a system with many components, allowing the imaging characteristics of the system to be approximately determined with simple calculations.
By definition, the image focus F' of an optical system is the image of the infinite point on the axis. The beam issued out of this point is made of parallel rays to the axis.
These rays focalise in F' after crossing the system. The location of the intersection points of

each incident ray with its corresponding image ray is, in paraxial approximation, a plane which shall be called image principal plane of the optical system.
This plane cuts the axis in H', $f' = H'F'$ is the focal image distance of the optical system. H' is the image principal point. We proceed in the same way as for the object focus F, the object principal plane, the object focal distance $f = HF$.

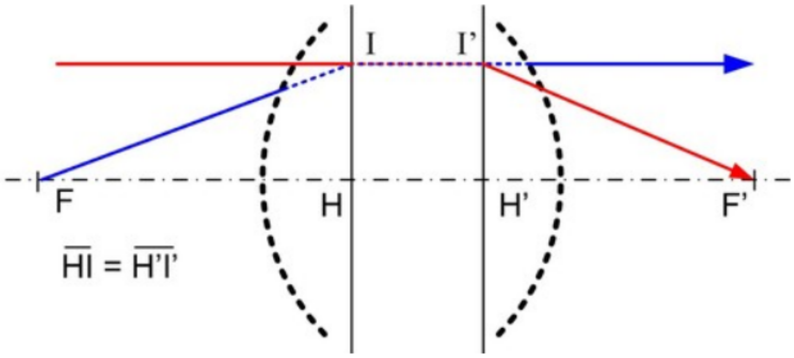


Figure 10: Cardinal points

Any luminous ray issued from F cust the object principal plane in I and comes out parallel to the axis, it cuts the image principal plane in I'. An incident ray parallel to the axis groing through I, also goes through I' the converges in F'. These two rays cross each other in I in the object space then in I' in the image space I and I' are therefore conjugated.
The principal planes are conjugated with an associated transversal magnification equal to 1.
Nodal points

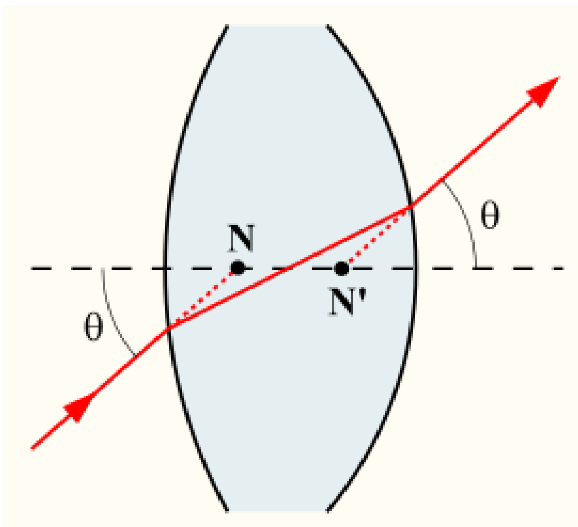


Figure 11: Nodal points

The front and rear nodal points have the property that a ray aimed at one of them will be refracted by the lens such that it appears to have come from the other, and with the same angle

with respect to the optical axis.
If the medium on both sides of the optical system is the same (e.g., air), then the front and rear nodal points coincide with the front and rear principal points, respectively.

1.2.11 Marginal and Chief Rays

the ray that passes from the center of the object, at the maximum aperture of the lens, is normally known as the marginal ray. It therefore passes through the edge of the aperture stop. Conventionally, this ray is in the y-z plane, usually called the meridian plane.

The chief ray is defined to be the ray from an off-axis point in the object passing through the center of the aperture stop; although there can be an infinite number of such rays, we can usually assume, at least for centered systems, that the chief ray is also restricted to the meridian plane.

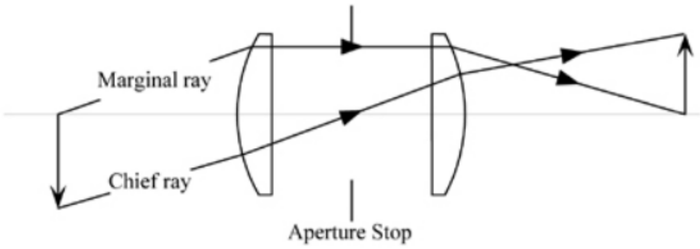


Figure 12: Simple lens system and aperture stop.

1.2.12 Reference

`https://books.google.de/books?id=vI7-WDw7yxkC&printsec=frontcover#v=onepage&q&f=false`

1.3 Aberration

The optic design based on paraxial approximation owns small numerical aperture and small field angle. The need for an aberration theory valid for more complicated optical systems with larger values of numerical aperture and field angle was felt when photography emerged.
Aberration can be defined as a departure of the performance of an optical system from the predictions of paraxial optics.
The Five Seidel Aberrations
The five basic types of aberration which are due to the geometry of lenses or mirrors, and which are applicable to systems dealing with monochromatic light, are known as Seidel aberrations, from an 1857 paper by Ludwig von Seidel. These are the aberrations that become evident in third-order optics, also known as Seidel optics. As we know,

$$sinx = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^11}{11!} + \frac{x^13}{13!} - ... \tag{54}$$

$$cosx = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^10}{10!} + \frac{x^12}{12!} - ... \tag{55}$$

When we neglect the later terms in the series, so that we behave as if $sin(x) = x$, and $cos(x) = 1$, we obtain first-order optics, in which all lenses are perfect. When we include the x squared and x cubed terms, then we have proceeded to third-order optics, in which the aberrations resulting from the nature of real lenses, exclusive of chromatic aberration, become evident. The five Seidel aberrations are:

1.3.1 Spherical Aberration

this is the aberration affecting rays from a point on the optical axis; because rays from this point going out in different directions pass through different parts of the lens, then, if the lens is spherical, or otherwise not the exact shape needed to bring them all to a focus, then these rays will not all be focused at the same point on the other side of the lens.

1.3.2 Coma

this aberration affects rays from points off the optical axis. If spherical aberration is eliminated, different parts of the lens bring rays from the axis to the same focus. But the place where the image of an off-axis point is formed may still change when different parts of the lens are considered.

1.3.3 Astigmatism

this is another aberration affecting rays from a point off the optical axis. These rays, as they head through the lens to the point in the image where they will be focused, pass through a lens that is, from their perspective, tilted. Even if neither spherical aberration nor coma prevents them from coming to a sharp focus, if we consider the rays of light that are in the plane of the tilt, and the rays of light that are in the plane perpendicular to that, these rays pass through a part of the lens with a different profile. So they may not be focused at the same distance from the lens, even if they do come to a focus in each case.

1.3.4 Curvature of Field

even when light from every point in the object is brought to a sharp focus, the points at which they are brought into focus might lie on a curved surface instead of a flat plane.

1.3.5 Distortion

even when all the previous aberrations have been corrected, the light from points in the object might be brought together on the image plane at the wrong distance from the optical axis, instead of being linearly proportional to the distance from the optical axis in the object. If distance increases faster than in the object, one has pincushion distortion, if more slowly, barrel distortion.

1.3.6 Reference

`http://www.quadibloc.com/science/opt0505.htm`

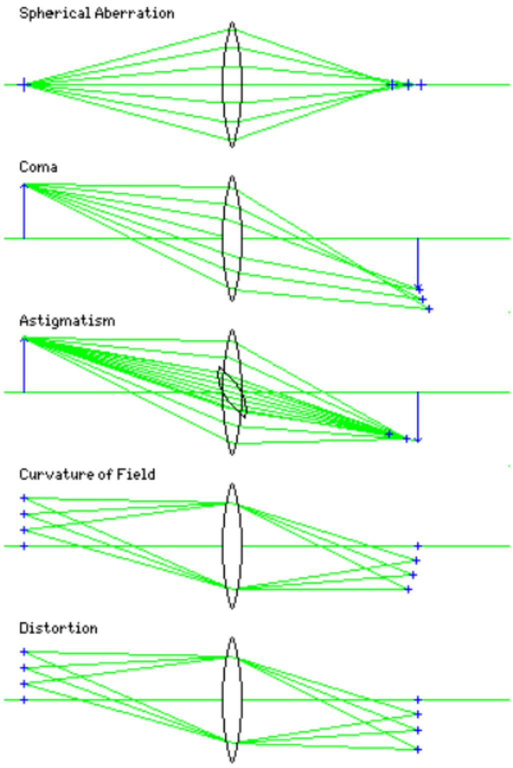


Figure 13: Seidel Aberrations

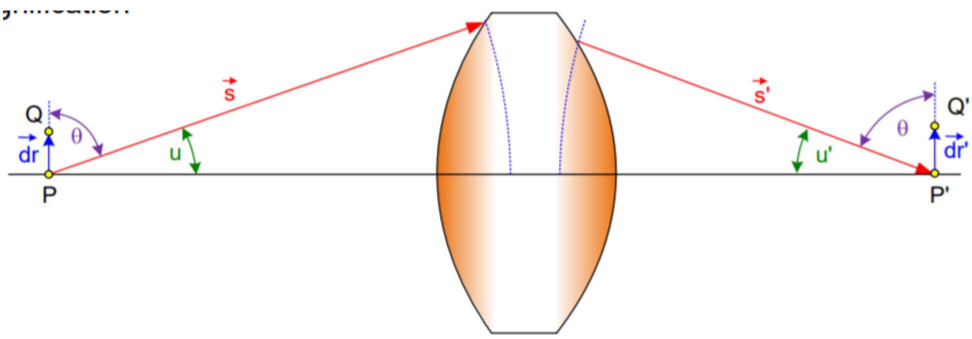


Figure 14: eikonal equation

1.3.7 Abbe sine condition

From equations 11 12 15 above we can also get the abbe sine condition, using $\frac{AB}{AC} = \frac{A'B'}{A'C'}$, then $nysinU = n'y'sinU'$
We can also derive the Abbe sine condition by eikonal equation.

$$\delta L = n'\vec{s'}d\vec{r'} - n\vec{s}d\vec{r} = 0 \tag{56}$$

$$n'\vec{s'}d\vec{r'} = n\vec{s}d\vec{r} \tag{57}$$

$$n'd\vec{r'}cos\theta' = nd\vec{r}cos\theta \tag{58}$$

$$nsin\varphi = n'\beta sin\varphi' \tag{59}$$

1.4 Index

Taylor series: $sinx = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ...$
$\sqrt{1-x^2} = 1 - \frac{x^2}{2} - \frac{x^4}{8} - ...$
numerical aperture v.s. f-number, written f/ or N
$NA = nsin\theta$
$N = \frac{f}{D}$

2 Aspheric surface

2.1 Spheric

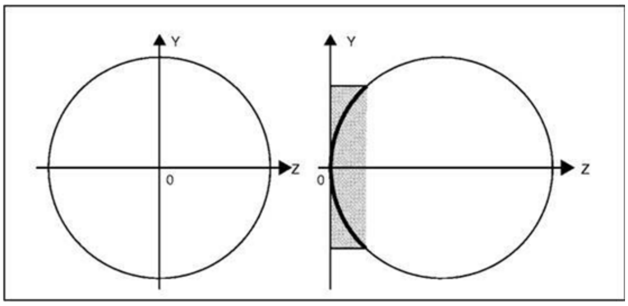


Figure 15: Coordinate translation

Figure 15 shows the coordinate translation to the left vertex of the surface. The math equations before and after the coordinate translation are as follows:

$$z^2 + y^2 = R^2 \tag{60}$$

$$z^2 - 2zR + y^2 = 0 \tag{61}$$

$$Pz^2 - 2zR + y^2 = 0 \tag{62}$$

where $P = 1 + K$, and $K = -e^2$. e is the eccentricity with $e^2 = \frac{a^2 - b^2}{a^2}$, a is the major axis for the conic section and b is the minor axis. For the left half part we are interested in,

$$z = \frac{R - \sqrt{R^2 - Py^2}}{P} \tag{63}$$

$$z = \left(\frac{R}{P}\right)\left[1 - \sqrt{1 - P\left(\frac{y}{R}\right)^2}\right] \tag{64}$$

$$z = \frac{cy^2}{1 + \sqrt{1 - (1 + K)c^2y^2}} \tag{65}$$

We call z the sag of the surface – the displacement from the vertex, K is the conic constant, c is the surface curvature and y is the radial height on the surface. We can get different conic section surface types when we choose different K values.

Conic mirrors give perfect geometric imagery when an axial point object is located at one conic focus and the axial point image is located at the other conic foucs.

2.2 Aspheric

General aspheres are surfaces with fourth- and higher-order surface deformation on top of a flat or curved surface. The surface deformation of a rotationally sy,,etric general asphere is given by the relation

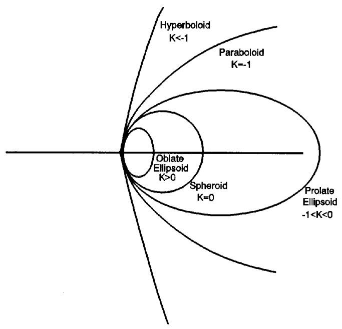
$$z = \frac{cy^2}{1 + \sqrt{1 - (1 + K)c^2y^2}} + Ay^4 + By^6 + Cy^8 + Dy^10 \tag{66}$$

where A,B,C and D are 4th-, 6th-, 8th-, and 10th=order coeffcients that determine the sign and magnitude of the deformation produced by that order. Although general aspheres allow

Table 3.1
Conic constant associated with different surface types.

Surface Type	Conic constant (K)	$P = 1 + K$
Circle	0	1
Parabola	–1	0
Hyperbola	< –1	< 0
Prolate Ellipse	–1 < K < 0	0 < P < 1
Oblate Ellipse	> 0	> 1

(a) Conic constant



(b) Relative shapes of conic surfaces in two dimensions

Figure 16: Conic surfaces

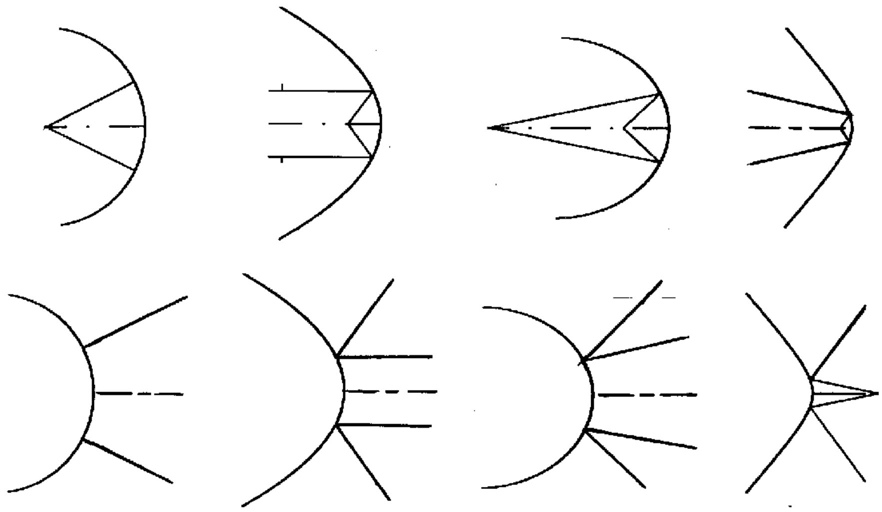


Figure 17: Ray paths for perfect axial imagery

correction of third- and higher-order aberrations and may reduce the number of elements in an optical system, general aspheres are more expensive than spheres or conics. If aspheric deformation is required, conic surfaces should be tried first, especially since a conic offers higher-order correction.

2.3 Reference:

https://zhuanlan.zhihu.com/p/29930116
Handbook of Optics, Chapter 18

3 Optical components

3.1 Refraction

3.1.1 Planar interface

- Snells law
- external reflection ($n_1 < n_2$): ray refracted away from the interface
- internal reflection ($n_1 > n_2$): ray refracted towards the interface
- total internal reflection (TIR) for: $\theta_2 = \frac{\pi}{2} \Rightarrow sin\theta_1 = sin\theta_{TIR} = \frac{n_2}{n_1}$ Normally, the definitions of "internal" and "external" are with respect to the medium the light is emerged from. Total internal reflection is more common in application, however, total external reflection normally refers to X-rays whose indices of refraction for all materials are all slightly below 1. This entails that total reflection of X-rays only can occur when they travel through vacuum and impinge on a surface (at a small glancing angle). Since this kind of total reflection takes place outside of the material it is termed total external reflection.

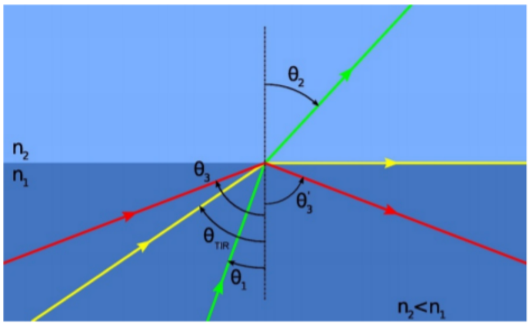


Figure 18: Planar interface

3.1.2 Spherical interface (paraxial)

as shown in "Geometrical optics" part

3.1.3 Spherical thin lens (paraxial)

as shown in "Geometrical optics" part

3.2 Reflection(Mirror)

3.2.1 Planar mirror

- Rays originating from P1 are reflected and seem to originate from P2.

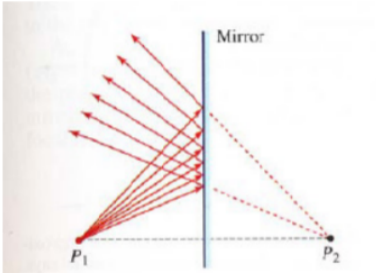


Figure 19: Planar mirror

3.2.2 Paraboloidal mirror

- parabolic or paraboloid or paraboloidal mirror
- Parallel rays converge in the focal point (focal length f).
- Applications: Telescope, collimator

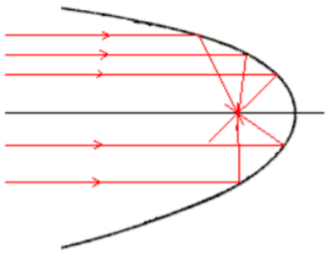


Figure 20: paraboloidal mirror

3.2.3 Ellipsoidal mirror

- Rays originating from focal point P_1 converge in the second focal point P_2

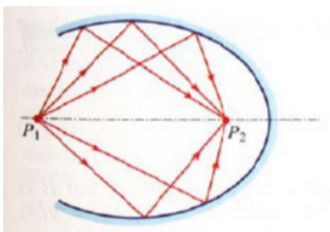


Figure 21: Ellipsoidal mirror

3.2.4 Cylindrical mirror

3.2.5 Spherical mirror

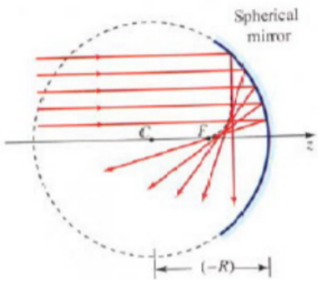


Figure 22: Spherical mirror

- Neither imaging like elliptical mirror nor focusing like parabolic mirror
- parallel rays cross the optical axis at different points
- connecting line of intersections of rays → caustic
- parallel, paraxial rays converge to the focal point $f = (-R)/2$
- convention: $R < 0$ - concave mirror; $R > 0$ - convex mirror
- for paraxial rays the spherical mirror acts as a focusing as well as an imaging optical element. paraxial rays emitted in point P_1 are reflected and converge in point P_2. $\frac{1}{z_1} + \frac{1}{z_2} \approx \frac{2}{-R}$
- paraxial imaging: imaging formula and magnification. $m = -\frac{z_2}{z_1}$

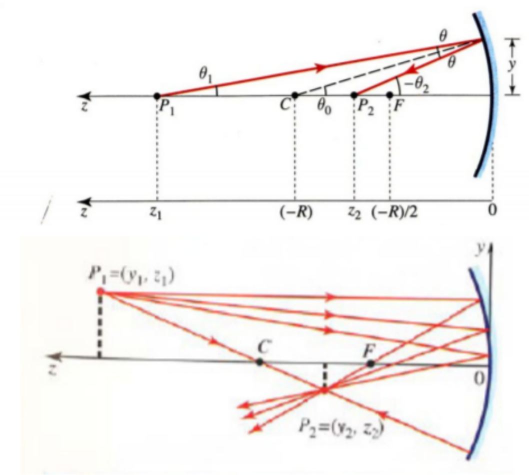


Figure 23: Spherical mirror paraxial

3.2.6 Toroidal mirror

Toroidal mirrors are focusing devices having two different radii whose axes are oriented perpendicularly. They are utilized in instances where a beam must be focused and folded. Rather than using both a spherical mirror and a plane mirror for this purpose, both functions may be combined in one element. Toroidal mirrors also correct for the astigmatism that result when a spherical mirror is used off axis.
Toroidal mirrors are aspherical mirrors where each curvature of orthogonal two axes (horizontal and vertical one) are different. Barrel or Tire shaped with a rotation axis are available.

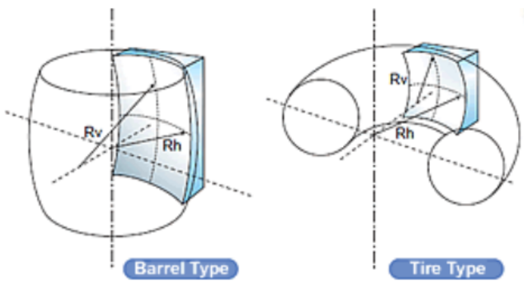


Figure 24: Toroidal mirror

3.3 reference

Lecture slide "Fundamentals of Modern Optics" of Jena

3.4 Index

The Polar coordinates form of conic sections

4 Conic surface optical properties

5 Ray tracing

5.1 Ray tracing

5.1.1 Introduction

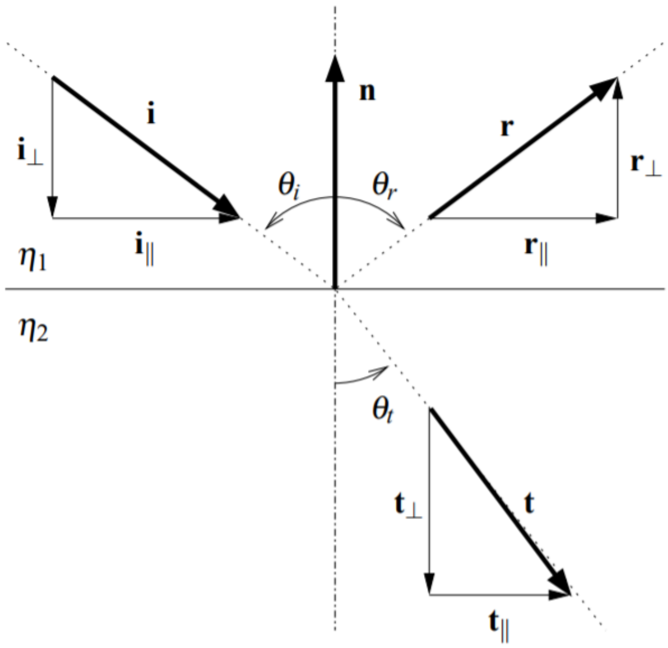


Figure 25: Scheme

In the figure we have a interface between two materials with different refractive indices n_1 and n_2. The direction vector of the incident ray is $\mathbf{i}$ and we assume this vector is normalized. The normalized direction vectors of the reflected and transmitted rays are $\mathbf{r}$ and $\mathbf{t}$ and will be calculated. The normal vector $\mathbf{n}$, orthogonal to the interface and pointing towards the first material n_1.

$$|\mathbf{i}| = |\mathbf{r}| = |\mathbf{t}| = |\mathbf{n}| \tag{67}$$

The direction vectors of the rays can be split in components orthogonal and parallel to the interface. Taking the generic vector $\mathbf{v}$, we have

$$\mathbf{v}_\perp = \frac{\mathbf{v}\cdot\mathbf{n}}{|\mathbf{n}|^2}\mathbf{n} = (\mathbf{v}\cdot\mathbf{n})\mathbf{n} \tag{68}$$

$$\mathbf{v}_\parallel = \mathbf{v} - \mathbf{v}_\perp \tag{69}$$

$$|\mathbf{v}|^2 = \left|\mathbf{v}_\parallel\right|^2 + |\mathbf{v}_\perp|^2 \tag{70}$$

$$\cos\theta_v = \frac{|\mathbf{v}_\perp|}{|\mathbf{v}|} = |\mathbf{v}_\perp| \tag{71}$$

$$\sin\theta_v = \frac{\left|\mathbf{v}_\parallel\right|}{|\mathbf{v}|} = \left|\mathbf{v}_\parallel\right| \tag{72}$$

5.1.2 Reflection

According to the law of reflection $\theta_r = \theta_i$, and 117 72, thus

$$|\mathbf{r}_\perp| = \cos\theta_r = \cos\theta_i = |\mathbf{i}_\perp| \tag{73}$$

$$\left|\mathbf{r}_\parallel\right| = \sin\theta_r = \sin\theta_i = \left|\mathbf{i}_\parallel\right| \tag{74}$$

$$\mathbf{r} = \mathbf{r}_\parallel + \mathbf{r}_\perp = \mathbf{i}_\parallel - \mathbf{i}_\perp = \mathbf{i} - 2\mathbf{i}_\perp = \mathbf{i} - 2(\mathbf{i}\cdot\mathbf{n})\mathbf{n} \tag{75}$$

And we can find that the reflection vector $\mathbf{r}$ is also nomarlized.

5.1.3 Refraction

According to Snell's Law for the refraction case,

$$n_1 \sin\theta_i = n_2 \sin\theta_t \tag{76}$$

We leave out the total internal reflection case first here. With equations 117 72, we have

$$n_1\left|\mathbf{i}_\parallel\right| = n_2\left|\mathbf{t}_\parallel\right| \tag{77}$$

$$\mathbf{t}_\parallel = \frac{n_1}{n_2}\mathbf{i}_\parallel = \frac{n_1}{n_2}(\mathbf{i} + \cos\theta_i\mathbf{n}) \tag{78}$$

$$\mathbf{t}_\perp = -\sqrt{1 - \left|\mathbf{t}_\parallel\right|^2}\,\mathbf{n} \tag{79}$$

$$\mathbf{t} = \mathbf{t}_\parallel + \mathbf{t}_\perp = \frac{n_1}{n_2}\mathbf{i} + \left(\frac{n_1}{n_2}\cos\theta_i - \sqrt{1 - \left|\mathbf{t}_\parallel\right|^2}\right)\mathbf{n} = \frac{n_1}{n_2}\mathbf{i} + \left(\frac{n_1}{n_2}\cos\theta_i - \sqrt{1 - \sin^2\theta_t}\right)\mathbf{n} \tag{80}$$

With 76, the refraction vector $\mathbf{t}$ can be conducted.

5.1.4 Reference

`https://graphics.stanford.edu/courses/cs148-10-summer/docs/2006--degreve--reflection_refraction.pdf`

5.2 Real ray tracing

Here the term real ray tracing refers to the 'case 5' part of the ZEMAX DLL file.
The real ray trace portion of the DLL, type code 5, allows definition of the surface transmittance. The surface transmittance must be a number between 0.0 and 1.0, and indicates the relative fraction of intensity that the ray transmits through the surface. In this context, transmit means

```
int Refract(double thisn, double nextn, double *l, double *m,
double *n, double ln, double mn, double nn)
{
double nr, cosi, cosi2, rad, cosr, gamma;
if (thisn != nextn)
      {
      nr = thisn / nextn;
      cosi = fabs((*l) * ln + (*m) * mn + (*n) * nn);
      cosi2 = cosi * cosi;
      if (cosi2 > 1) cosi2 = 1;
      rad = 1 - ((1 - cosi2) * (nr * nr));
      if (rad < 0) return(-1);
      cosr = sqrt(rad);
      gamma = nr * cosi - cosr;
      (*l) = (nr * (*l)) + (gamma * ln);
      (*m) = (nr * (*m)) + (gamma * mn);
      (*n) = (nr * (*n)) + (gamma * nn);
      }
return 0;
}
```

Figure 26: Refraction code in ZEMAX dll file

"continues on", so for reflective surfaces, the transmitted portion is that which normally reflects to the next surface.
In standard dll file, real ray tracing calculates the intersection point (x, y, z) of the sag surface and the ray with the direction vector (l, m, n), and the normal vector (ln,mn,nn) at this point.

5.2.1 Intersection

With the ray starting point $\mathbf{O} = (x_0, y_0, z_0)$ and the normalized ray direction vector $\mathbf{D} = (l, m, n)$, we can express the ray $\mathbf{P} = (x, y, z)$ in parametric form

$$\mathbf{P} = \mathbf{O} + t\mathbf{D} \tag{81}$$

$$\begin{cases} x = x_0 + tl \\ y = y_0 + tm \\ z = z_0 + tn \end{cases} \tag{82}$$

We can first transform the surface sag equation into another form:

$$z = \frac{c\rho^2}{1 + \sqrt{1 - (1 + K)c^2\rho^2}} \tag{83}$$

$$\rho^2 + (1 + K)z^2 - \frac{2z}{c} = 0 \tag{84}$$

$$where\ \rho^2 = x^2 + y^2 \tag{85}$$

To find the intersection between this surface and an arbitrary ray, substitute the ray equation in the surface equation:

$$(x_0 + tl)^2 + (y_0 + tm)^2 + (1 + K)(z_0 + tn)^2 = \frac{2(z_0 + tn)}{c} \tag{86}$$

$$(l^2 + m^2 + n^2(1 + K))t^2 + 2\left(x_0 l + y_0 m + (1 + K)z_0 n - \frac{n}{c}\right)t + (x_0^2 + y_0^2 + z_0^2(1 + K) - \frac{2z_0}{c}) = 0 \tag{87}$$

Because the ray direction vector is normalized, so we have $l^2 + m^2 + n^2 = 1$. And if we assume the ray start point satifies $z_0 = 0$, the equation can be simplifed as

$$(1 + n^2K)t^2 + 2\left(x_0 l + y_0 m - \frac{n}{c}\right)t + (x_0^2 + y_0^2) = 0 \tag{88}$$

$$choose \begin{cases} a_1 = a = (1 + n^2K) \\ b_1 = -\frac{b}{2} = \frac{n}{c} - x_0 l - y_0 m \\ c_1 = c = x_0^2 + y_0^2 \end{cases} \tag{89}$$

Here we choose a_1, b_1 and c_1 which ar econsistent with that in ZEMAX DLL file 'case 5' part. Therefore the solution is

$$t = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{b_1 \pm \sqrt{b_1^2 - a_1c_1}}{a_1} = \frac{c_1}{b_1 \mp \sqrt{b_1^2 - a_1c_1}} \tag{90}$$

We then can calculate the intersection using equation 82

5.2.2 Normal vector

First, we will derive the normal vector form of an arbitrary point on an arbitrary surface. We consider the surface $U(x, y, z) = k$, it is said to be smooth if the gradient $\nabla U = \langle U_x, U_y, U_z\rangle$ is continuous and non-zero at any point on the surface. Equivalently, we ofter write:

$$\nabla U = U_x\mathbf{e}_x + U_y\mathbf{e}_y + U_z\mathbf{e}_z \tag{91}$$

$$where\ \mathbf{e}_x = \langle 1, 0, 0\rangle, \mathbf{e}_y = \langle 0, 1, 0\rangle, \mathbf{e}_z = \langle 0, 0, 1\rangle \tag{92}$$

Suppose that $\mathbf{r}(t) = \langle x(t), y(t), z(t)\rangle$ lies on a smooth surface $U(x, y, z) = k$. Applying the derivative with respect to t to both sides of the equation of the surface yields

$$\frac{dU}{dt} = \frac{d}{dt}k = 0 \tag{93}$$

$$\nabla U \cdot \mathbf{v} = 0 \tag{94}$$

$$where\mathbf{v} = \langle x'(t), y'(t), z'(t)\rangle \tag{95}$$

$\mathbf{v}$ is tangent to the surface because it is tangent to a curve on the surface, which implies that ∇U is orthogonal to each tangent vector $\mathbf{v}$ at a given point on the surface. We thus say that the gradient ∇U is normal to the surface $U(x, y, z) = k$ at each point on the surface. To get the normalized normal vector, we can simply divide the gradient ∇U with its length.
Have known the way to calculate the normal vector, we come back to our case, where we want to calculate the normal vector at the intersection point that we have already got in the former

part.

$$
\begin{aligned}
U(x,y,z) &= \rho^2 + (1+K)z^2 - \frac{2z}{c} = 0 && (96)\\
\frac{dU}{d\rho} &= 2\rho && (97)\\
\Rightarrow \frac{dU}{dx} &= 2x, \frac{dU}{dy} = 2y && (98)\\
\frac{dU}{dz} &= 2(1+K)z - \frac{2}{c} && (99)\\
|U| &= \sqrt{\left(\frac{dU}{dx}\right)^2 + \left(\frac{dU}{dy}\right)^2 + \left(\frac{dU}{dz}\right)^2} && (100)\\
&= \sqrt{4\rho^2 + 4\left((1+K)z - \frac{1}{c}\right)^2} && (101)\\
&= 2\sqrt{-(1+K)z^2 + \frac{2z}{c} + \left((1+K)z - \frac{1}{c}\right)^2} && (102)\\
&= 2\sqrt{K(1+K)z^2 - \frac{2k}{c}z + \frac{1}{c^2}} && (103)\\
&= \frac{2}{c}\sqrt{K(1+K)c^2z^2 - 2kcz + 1} && (104)
\end{aligned}
$$

And we can get the normalized normal vector by diveding the $\nabla U = \langle U_x, U_y, U_z\rangle$ with its length $|U|$, which is consistent with the code in DLL file.

```
case 5:
  /* ZEMAX wants a real ray trace to this surface */
   if (FD->cv == 0.0)
          {
           UD->ln =  0.0;
        UD->mn =  0.0;
     UD->nn = -1.0;
                     if (Refract(FD->n1, FD->n2, &UD->l, &UD->m, &UD->n, UD->ln, UD->mn, UD->nn)) return(-FD->surf);
      return(0);
      }
   /* okay, not a plane. */
                  a = (UD->n) * (UD->n) * FD->k + 1;
                  b = ((UD->n)/FD->cv) - (UD->x) * (UD->l) - (UD->y) * (UD->m);
                  c = (UD->x) * (UD->x) + (UD->y) * (UD->y);
                  rad = b * b - a * c;
                  if (rad < 0) return(FD->surf);  /* ray missed this surface */
                  if (FD->cv > 0) t = c / (b + sqrt(rad));
                  else           t = c / (b - sqrt(rad));
                  (UD->x) = (UD->l) * t + (UD->x);
                  (UD->y) = (UD->m) * t + (UD->y);
                  (UD->z) = (UD->n) * t + (UD->z);
                  UD->path = t;
                  zc = (UD->z) * FD->cv;
                  rad = zc * FD->k * (zc * (FD->k + 1) - 2) + 1;
                  casp = FD->cv / sqrt(rad);
                  UD->ln = (UD->x) * casp;
                  UD->mn = (UD->y) * casp;
                  UD->nn = ((UD->z) - ((1/FD->cv) - (UD->z) * FD->k)) * casp;
   if (Refract(FD->n1, FD->n2, &UD->l, &UD->m, &UD->n, UD->ln, UD->mn, UD->nn)) return(-FD->surf);
   break;
```

Figure 27: Real ray tracing code in ZEMAX dll file 'case 5'

5.2.3 Reference

ZEMAX manual-Chapter11(SurfaceType)-UserDefined
https://www.cl.cam.ac.uk/teaching/1999/AGraphHCI/SMAG/node2.html
https://www.scratchapixel.com/lessons/3d-basic-rendering/minimal-ray-tracer-rendering-simple-shapes/ray-sphere-intersection
http://math.etsu.edu/multicalc/prealpha/chap3/chap3-6/printversion.pdf

6 Aberration derivation

7 Ray transfer matrix analysis

8 Abbe sine condition

8.1 Derivation of Abbe and Herschel condition

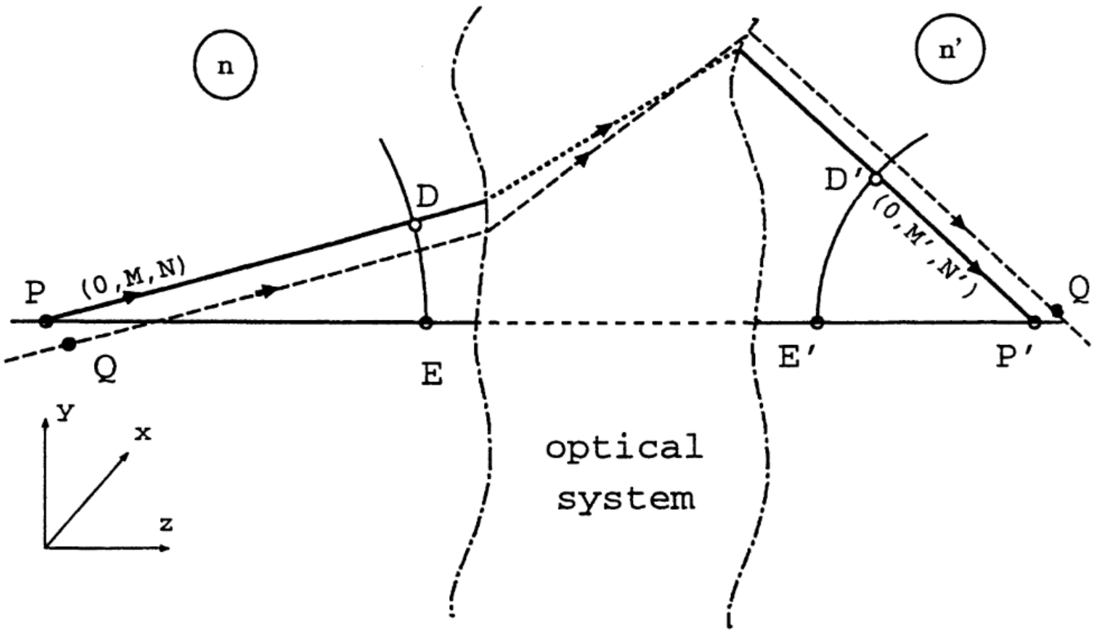


Figure 28: Abbe sine condition

In figure 28 we schematically show an optical system with its entrance and exit pupil located at E and E'. From a pencil of rays leaving the object point P only a certain ray PDD'P' is shown with direction cosines$(0, M, N)$ and $(0, M', N')$ in respectively the object and image space (D and D' are located on the pupil reference spheres). The plane of the drawing is the yz-plane with x=0. We suppose that the pencil of rays is exactly focused at P' (stigmatic imaging). We now want to know how a pencil of rays leacing a neighbouring object Q is focused in image space. According to paraxial optics, the position of an image point Q' derived from the object point Q, which has been subjected to infinitesimal shifts δy and δz with respect to P, is found by applying the correct (paraxial) magnification factors. The lateral magnification factor is denoted by

$$\beta' = \delta y' / \delta y \tag{105}$$

and the axial magnification factor is

$$\delta' = \delta z' / \delta z \tag{106}$$

It can be shown that the relationship between the lateral and axial magnification is given by

$$\delta' = \frac{n'}{n} \beta'^2 \tag{107}$$

It now has to be seen under what condition the image quality of Q' can be equal to the (perfect) quality of P'. To this goal, we take, from the pencil of rays through Q, a particular ray (dashed in the figure) which is parallel to the ray through P shown in the drawing. Physically spoken, we consider this ray to represent a small part of a plane wave with the direction cosines(0,M,N). If the reference point for measuring path length is shifted from P towards Q, the change ΔW in path of the propagating wave disturbance along the ray from P to P' is given by the scalar product of the ray vector $\vec{s} = (0, M, N)$ and the displacement vector $\delta\vec{r} = (0, \delta y, \delta z)$ multiplied with the refractive index of the object space:

$$\Delta W = -n\delta\vec{r}\vec{s} \tag{108}$$

The minus sign is needed here because a shift of Q in the positive z-direction leads to shorter optical path for the ray along the optical axial.
In the image space, given the image displacement vector $\delta\vec{r'}$ and the ray vector $\vec{s'}$, we observe a path difference $\Delta W'$ according to

$$\Delta W' = -n'\delta\vec{r'}\vec{s'} \tag{109}$$

Equal imaging quality in Q and P(isoplanatism) is obtained when the residual δW of the path difference in object and image space is zero for arbitary values of the ray vectors of all rays belonging to the object and image space pencils,

$$\delta W = \Delta W + \Delta W' = 0 \tag{110}$$

We discern two particular cases:

8.1.1 Abbe sine condition

$\delta\vec{r} = (0, \delta y, 0)$
The isoplanatic condition now becomes:

$$\delta W_C = n'\delta y' M' - n\delta y M = 0 \tag{111}$$

$$\delta W_C = \left[M' - \frac{nM}{n'\beta'}\right](n'\delta y') \tag{112}$$

Here we use the paraxial magnification β'. This condition, which guarantees the absence of aberration if an infinitesimal lateral excursion off-axis is applied, is generally known as Abbe's sine condition. We have used index C for the aberration because this aberration is called coma.

8.1.2 Herschel condition

$\delta\vec{r} = (0, 0, \delta z)$
The isoplanatic condition now becomes:

$$\delta W_S = n'\delta z' N' - n\delta z N = 0 \tag{113}$$

$$\delta W_S = \left[N' - \frac{n^2 N}{n'^2\beta'^2}\right](n'\delta z') \tag{114}$$

This condition, which guarantees an extended axial range over which the object point can be shifted, is known as Herschel's condition. The subscript S has been used because the aberration which could appear is circularly symmetric spherical aberration.
In general, the constant pathlength difference $n'\delta z' - n\delta z$ encountered for the axial ray (N=N'=1) is subtracted from the expression above and we obtain:

$$\delta W_S = \left[(N' - 1) - \frac{n^2(N-1)}{n'^2\beta'^2}\right](n'\delta z') = 0 \tag{115}$$

8.1.3 Comparsion between Abbe and Herschel

Abbe's sine condition is a prerequisite for an opticla system which needs to image an extended flat object, e.g., for a photographic camera objective, a reduction objective but also for a microscope objective or an astronomical telescope.
Herschel's condition is required when a system needs to operate at different magnifications, e.g., a narrow field telescope for both remote and close observation.
Unfortunately, both conditions seriously conflict when the numerical aperture becomes high. A particular case arises when $\beta' = \pm\frac{n}{n'}$, a case which reduces to unit magnification when $n = n'$; here, both conditions can be satisfied simultaneously. In most applications, it is the sine condition which will prevail because otherwise the useful lateral or angualr image field would become unacceptably small.

8.1.4 Reference

- Braat, J.: The Abbe sine condition and related imaging conditions in geometrical optics. In Proc. SPIE, checked on 11/7/2016.

8.2 Modified sine relationship for spherical mirror at grazing incidence

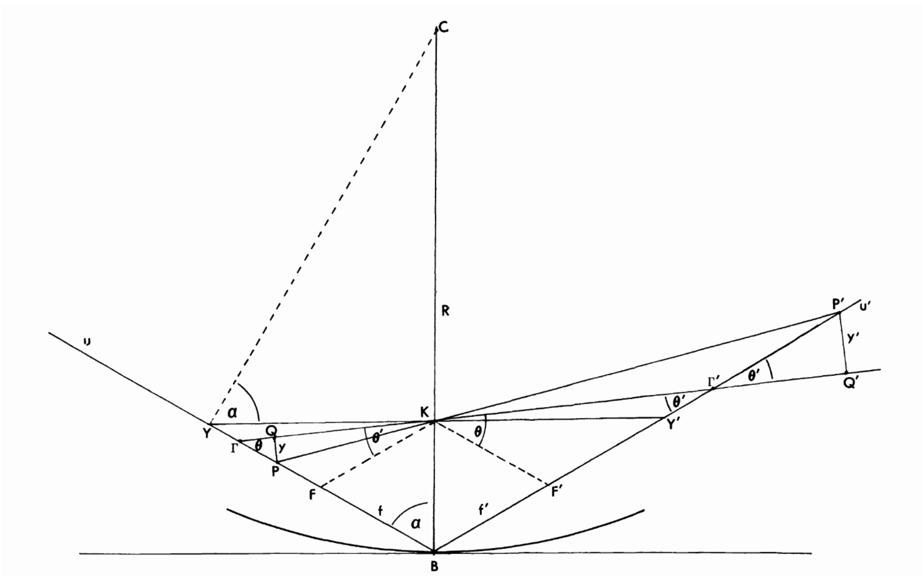


Figure 29: Ray construction for reflection at grazing incidence leading to a modified sine condition.

In figure29 a concave reflector with its apex at B and of radius R has points Y and Y' which make CY and CY' perpendicular to the incoming and outgoing ray direction μ respectively. The center of perspective is at K. $CK = Rsin^2\alpha$ and consequently $KB = R(1 - sin^2\alpha$ which is very small for the case of grazing incidence where α approaches 90°.
In the figure, the distance $PB = p$ the object distance and $BP' = q$ the image distance then

since the triangles PFK and KF'P' are similiar, the ratio of corresponding sides results in

$$\frac{p-f}{f'} = \frac{f}{q-f'} \tag{116}$$
$$(p-f)(q-f') = ff' \tag{117}$$
$$\tag{118}$$

which is recognized as Newton's Equation in optics. By substituting $f' = f = \frac{Rsini}{2}$, the well known focal length in the meridian plane for grazing incidence, where $i = 90° - \alpha$ and R is the radius of curvature of the reflecting surface, the equation reduces to

$$\frac{1}{p} + \frac{1}{q} = \frac{2}{Rsini} \tag{119}$$

the equation commonly used to determining image formation at grazing angles of incidence in the meridian plane.
For point P in figure the distance $(p - f)$ and $(q - f)$is given by

$$(p-f) = PF = \Gamma F - \Gamma P \tag{120}$$
$$(q-f) = F'P' = F'\Gamma' + \Gamma' P' \tag{121}$$

By construction

$$PQ = y = \Gamma P sin\theta \tag{122}$$
$$P'Q' = y' = \Gamma' P' sin\theta' \tag{123}$$
$$\frac{\Gamma F}{sin\theta'} = \frac{f}{sin\theta} \tag{124}$$
$$\frac{\Gamma' F'}{sin\theta} = \frac{f}{sin\theta'} \tag{125}$$

Substitution in Eq. 117 yields

$$f^2 = (\Gamma F - \Gamma P)(\Gamma' F' - \Gamma' P') \tag{126}$$
$$f^2 = \left(f\frac{sin\theta'}{sin\theta} - \frac{y}{sin\theta}\right)\left(f\frac{sin\theta}{sin\theta'} + \frac{y'}{sin\theta'}\right) \tag{127}$$

and get the modified sine condition for grazing incidence optics

$$y' sin\theta' - y sin\theta = \frac{yy'}{f} \tag{128}$$

8.2.1 Reference

- Mcgee, J. F.; Keski-Kuha, Ritva A. M. (1986): A Modified Sine Condition for Grazing Incidence Optics. In Proc. SPIE, p.640, checked on 7/19/2016.

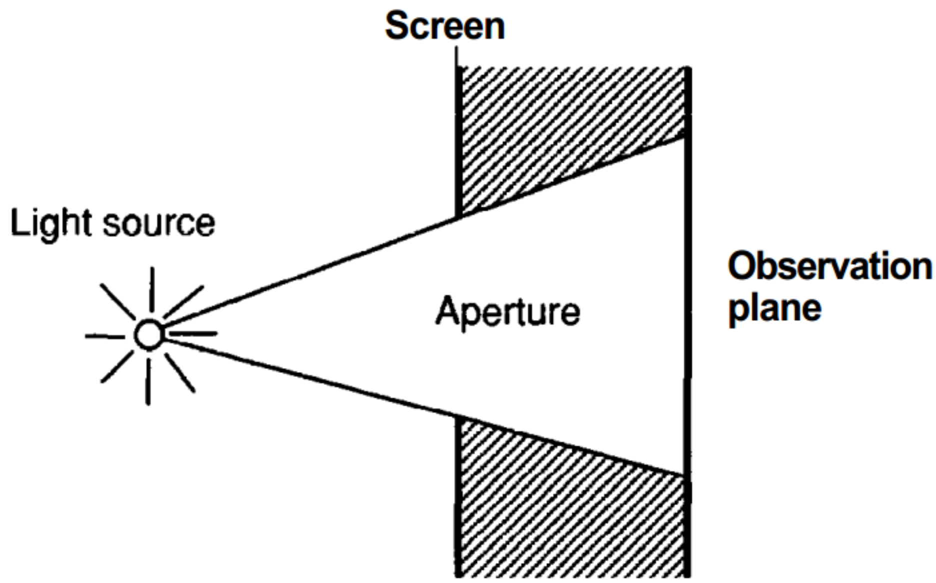


Figure 30: Grimaldi's observations

8.3 generalized sine condition for Wolter type II and WS telescope

9 GI focusing mirror deisgn

10 Wave optics

10.1 History of Diffraction theory

1665, Grimaldi's observations showed the shadow behind the opaque screen with a aperture was not with sharp borders but with gradual transition. This phenomenon can not be explained by the corpuscular theory of light propagation.30

1678, Huygens expressed the intuitive conviction that if each point on the wavefront of a disturbance were considered to be a new source of a "secondary" spherical disturbance, then the wavefront at a later instant could be found by constructing the "envelope" of the secondary wavelets.31

1804, Thomas Young strengthen the wave theory of light by introducing the critical concept of interference. The idea was a radical one at the time, for it stated that under proper conditions, light could be added to light and produce darkness.

1818, Augustion Jean Fresnel brought together the ideas of Huygens and Young. By making some rather arbitary assumptions about the amplitudes and phases of Huygens' secondary sources, and by allowing the various wavelets to mutually interfere, Fresnel was able to calculate the distribution of light in diffraction patterns with excellent accuracy.

At Fresnel's presentation of his paper to a prize committee of the French Academy of Sciences, his theory was strongly disputed by the great French mathematician S. Poisson, a member of the committee. He demonstrated the absurdity of the theory by showing that it predicted the existence of a bright spot at the center of the shadow of an opaque disk. F. Arago, who chaired

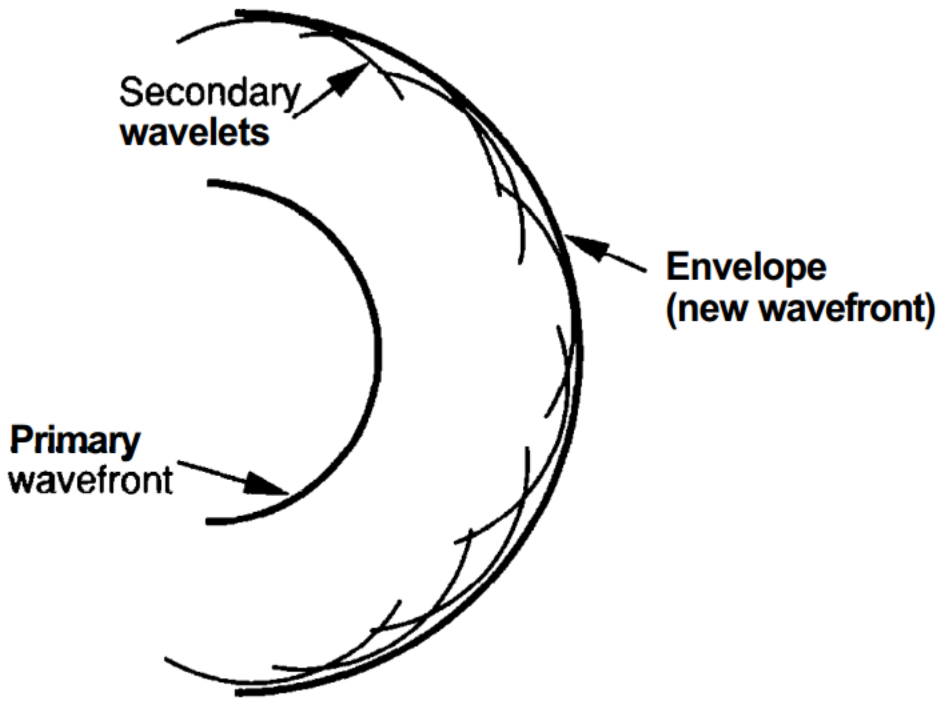


Figure 31: Huygens' envelope construction

the prize committee, performed such an experiment and found the predicted spot. Fresnel won the prize, and since then the effect has been known as "Poisson's spot".

1860, Maxwell identified light as an electronmagnetic wave, a step of enormous importance.

1882, Gustav Kirchhoff put the ideas of Huygens and Fresnel on a firmer mathematical foundation with two assumptions about the boundary values of the light incident on the surface of an obstacle placed in the way of propagation of light. These assumptions were later proved to be inconsistent with each other, by PoincarC in 1892 and by Sommerfeld in 1894.' As a consequence of these criticisms, Kirchhoff's formulation of the so-called Huygens-Fresnel principle must be regarded as a first approximation, although under most conditions it yields results that agree amazingly well with experiment.

Sommerfeld modified the Kichhoff theory by eliminating one of the assumptions concerning the light amplitude at the boundary by making use of the theory of Green's functions. This is called Rayleigh-Sommerfeld diffraction theory.

The Kirchhoff and Rayleigh-Sommerfeld theories share certain major simplifications and approximations thta light is treated as scalar phenomenon, neglecting the fundamentally vectorial nature of the electromagnetic fields. The scalar theory yields very accurate results if two conditions are met: (1) the diffracting aperture must be large compared with a wavelength, and (2) the diffracting fields must not be observed too close to the aperture.

10.2 From a vector to a scalar theory

Maxwell's equations:

$$\bigtriangledown \times \vec{\varepsilon} = -\mu \frac{\delta \vec{H}}{\delta t} \tag{129}$$

$$\bigtriangledown \times \vec{H} = \epsilon \frac{\delta \vec{\varepsilon}}{\delta t} \tag{130}$$

$$\bigtriangledown \cdot \epsilon \vec{\varepsilon} = 0 \tag{131}$$

$$\bigtriangledown \cdot \mu \vec{H} = 0 \tag{132}$$

Here $\vec{\varepsilon}$ is the electric field, with rectilinear components $(\vec{\varepsilon}_X, \vec{\varepsilon}_Y, \vec{\varepsilon}_Z)$, and $\vec{H}$ is the magnetic field, with components $(\vec{H}_X, \vec{H}_Y, \vec{H}_Z)$. $\vec{\varepsilon}$ and $\vec{H}$ are functions of both position P and time t. The symbols $\times$ and $\cdot$ represent a vector cross product and a vector dot product respectively, while $\bigtriangledown = \frac{\delta \vec{i}}{\delta x} + \frac{\delta \vec{j}}{\delta y} + \frac{\delta \vec{k}}{\delta z}$, where $\vec{i}, \vec{j}, \vec{k}$ are unit vector in the x, y z directions respectively.

Assumption:

- The wave is propagating in a dielectric medium.
- The medium is linear if it satisfies the linearity properties.
- The medium is isotropic if its properties are independent of the direction of polarization of the wave.
- The medium is homogeneous if the permittivity is constant throughout the region of propagation.
- The medium is nondispersive if the permittivity is independent of wavelength over the wavelength region occupied by the propagating wave.
- The medium is nonmagnetic, the magnetic permeability is always equal to po, the vacuum permeability.

Applying the $\bigtriangledown \times$ operation to the left and right sides of the first equation for $\vec{\varepsilon}$, and make use of Vector calculus identities, we can get

$$\bigtriangledown \times (\bigtriangledown \times \vec{\varepsilon}) = \bigtriangledown(\bigtriangledown \cdot \vec{\varepsilon}) - \bigtriangledown^2 \vec{\varepsilon} \tag{133}$$

If the propagation medium is linear, isotropic, homogeneous, and nondispersive, we can get

$$\bigtriangledown^2 \vec{\varepsilon} - \frac{n^2}{c^2} \frac{\delta^2 \vec{\varepsilon}}{\delta t^2} = 0 \tag{134}$$

where n is the refractive index of the medium and c is the velocity of propagation in vacuum.

The same for magnetic field:

$$\bigtriangledown^2 \vec{H} - \frac{n^2}{c^2} \frac{\delta^2 \vec{H}}{\delta t^2} = 0 \tag{135}$$

Since the vector wave equation is obeyed by both $\vec{\varepsilon}$ and $\vec{H}$, an identical scalar wave equation is obeyed by all components of those vectors, for example:

$$\bigtriangledown^2 \vec{\varepsilon}_X - \frac{n^2}{c^2} \frac{\delta^2 \vec{\varepsilon}_X}{\delta t^2} = 0 \tag{136}$$

To summarize, the behavior of all components of $\vec{\varepsilon}$ and $\vec{H}$ through a single scalar wave equation:

$$\bigtriangledown^2 u(P,t) - \frac{n^2}{c^2}\frac{\delta^2 u(P,t)}{\delta t^2} = 0 \tag{137}$$

where $u(P,t)$ represents any of the scalar field components dependent on position P in space and time t.

10.3 Kirchhoff diffraction

10.3.1 The Helmholtz Equation

Only monochromatic wave considered here, the scalar field for monochromatic wave is

$$u(P,t) = A(P)\cos\left[2\pi vt + \phi(P)\right] \tag{138}$$

where A(P) and $\phi(P)$ are the amplitude and phase of the wave at position P, while v is the optical frequency. The equation can be formed in complex notation:

$$u(P,t) = Re\{U(P)exp(-j2\pi vt)\} \tag{139}$$

where $Re\{\}$ signifies "real part of", and U(P) is a complex function of position (sometimes called a phasor)

$$U(P) = A(P)exp[-j\phi(P)] \tag{140}$$

according to equation 137, we can get

$$(\bigtriangledown^2 + k^2)U = 0 \tag{141}$$

here k is termed the wave number and is given by

$$k = 2\pi n\frac{v}{c} = \frac{2\pi}{\lambda} \tag{142}$$

The equation 146 is known as the Helmholtz equation

10.3.2 Green's theorem

Let U(P) and G(P) be any two complex-valued functions of position, and let S be a closed surface surrounding a colume V. If U, G, and their first and second partial derivatives are single-valued and continuous within and on S, then we have

$$\iiint_V (U \bigtriangledown^2 G - G \bigtriangledown^2 U)dv = \iint_S (U\frac{\delta G}{\delta n} - G\frac{\delta U}{\delta n})ds \tag{143}$$

10.3.3 The integral theorem of Helmholtz and Kirchhoff

Let the point of observation be denoted P_0, and let S denote an arbitary closed surface surrounding P_0. The problem is to express the optical disturbance at P_0 in terms of its values on the surface S. To solve this problem, we follow Korchhoff in applying Green's theorem and in choosing as an auxiliary function a unit-amplitude spherical wave expanding about the point P_0 (the so-called free space Green's function). Thus the value of Kirchhoff's G at an arbitary point P_1 is given by

$$G(P_1) = \frac{exp(jkr_{01})}{r_{01}} \tag{144}$$

where we adopt the notation that r_{01} is the length of the vector $\vec{r_{01}}$ pointing from P_0 to P_1

...

To be legitimately used in Green's theorem, the function G must be continuous within the enclosed volume V. Therefore to exclude the discontinuity at P_0, a small spherical surface S_ϵ of radius ϵ is inserted about the point P_0. Green's theorem is then applied, the volume of integration V' being that volume lying between S and S_ϵ, and the surface of integration being the composite surface

$$S' = S + S_\epsilon \tag{145}$$

as indicated in fig 32. Note that the outward normal normal to the composite surface points outward in the conventional sense on S, but inward(towards P_0) on S_ϵ

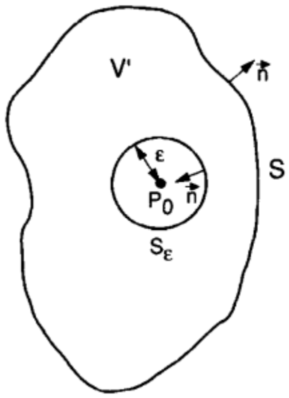


Figure 32: Surface of integration.

Within the volume V', the disturbance G being simply an expanding spherical wave, satisfies the Helmholtz equation

$$(\bigtriangledown^2 + k^2)G = 0 \tag{146}$$

Substituting the equation in the left-hand side of Green's theorem, we find

$$\iiint_{V'} (U \bigtriangledown^2 G - G \bigtriangledown^2 U)dv = -\iiint_{V'} (UGk^2 - GUk^2)dv \equiv 0 \tag{147}$$

Thus the theorem reduces to

$$\iint_{S'} (U\frac{\delta G}{\delta n} - G\frac{\delta U}{\delta n})ds = 0 \tag{148}$$

$$-\iint_{S_\epsilon} (U\frac{\delta G}{\delta n} - G\frac{\delta U}{\delta n})ds = \iint_{S} (U\frac{\delta G}{\delta n} - G\frac{\delta U}{\delta n})ds \tag{149}$$

Note for a general point P_1 on S', we have

$$G(P_1) = \frac{exp(jkr_{01})}{r_{01}} \tag{150}$$

$$\frac{\delta G(P_1)}{\delta n} = \cos(\vec{n}, \vec{r}_{01})(jk - \frac{1}{r_{01}})\frac{exp(jkr_{01})}{r_{01}} \tag{151}$$

where $\cos(\vec{n}, \vec{r_{01}})$ represents the cosine of the angle between the outward normal $\vec{n}$ and the vector $\vec{r_{01}}$ joining P_0 to P_1. For the particular case of P_1 on S_ϵ, $\cos(\vec{n}, \vec{r_{01}}) = -1$, and the equations become

$$G(P_1) = \frac{exp(jk\epsilon)}{\epsilon} \tag{152}$$

$$\frac{\delta G(P_1)}{\delta n} = \frac{exp(jk\epsilon)}{\epsilon}(\frac{1}{\epsilon} - jk) \tag{153}$$

Letting ϵ become arbitarily small, the continuity of U and its derivatives at P_0 allows us to write

$$\lim_{\epsilon\to 0} \iint_{S_\epsilon} \left(U\frac{\delta G}{\delta n} - G\frac{\delta U}{\delta n} \right) ds = \lim_{\epsilon\to 0} 4\pi\epsilon^2 \left[U(P_0)\frac{exp(jk\epsilon)}{\epsilon}\left(\frac{1}{\epsilon} - jk\right) - \frac{\delta U(P_0)}{\delta n}\frac{exp(jk\epsilon)}{\epsilon} \right] = 4\pi U(P_0) \tag{154}$$

$$U(P_0) = \frac{1}{4\pi} \iint_S \left\{ \frac{\delta U}{\delta n}\left[\frac{exp(jkr_{01})}{r_{01}}\right] - U\frac{\delta}{\delta n}\left[\frac{exp(jkr_{01})}{r_{01}}\right] \right\} ds \tag{155}$$

This result is known as the integral theorem of Helmholtz and Kirchhoff. It plays an important role in the development of the scalar theory of diffraction, for it allows the field at any point P_0 to be expressed in terms of the boundary values of the wave on any closed surface surrounding that point.

10.3.4 Kirchhoff diffraction by a planar screen

Consider now the problem of diffraction of light by an aperture in an infinite opaque screen. As illustrated in Fig 33, a wave disturbance is assumed to impinge on the screen and the aperture from the left, and the field at the point Pobehind the aperture is to be calculated. Again the field is assumed to be monochromatic.

Following Kirchhoff, the closed surface S is chosen to consist of two parts, as shown in Fig 33. Let a plane surface S_1, lying directly behind the diffracting screen, be joined and closed by a large spherical cap, S_2, of radius R and centered at the observation point P_0. The total closed surface S is simply the sum of S_1 and S_2. Thus,

$$U(P_0) = \frac{1}{4\pi} \iint_{S_1+S_2} \left(G\frac{\delta U}{\delta n} - U\frac{\delta G}{\delta n} \right) ds \tag{156}$$

... It can be proved that the integral over S_2 will yield a contribution pf zero so we can have

$$U(P_0) = \frac{1}{4\pi} \iint_{S_1} \left(G\frac{\delta U}{\delta n} - U\frac{\delta G}{\delta n} \right) ds \tag{157}$$

For the opaque screen with open aperture denoted as Σ, it therefore seems intuitively reasonable that the major contribution to the integral arises from the points of S_1 located within the aperture Σ where we would expect the integrand to be largest. Kirchhoff accordingly adopted the following assumptions:

- Across the surface Σ the field distribution U and its derivative $\frac{\delta U}{\delta n}$ are exactly the same as they would be in the absence of the screen.

- Over the portion of S_1 that lies in the geometrical shadow of the screen, the field distribution U and its derivative $\frac{\delta U}{\delta n}$ are identically zero

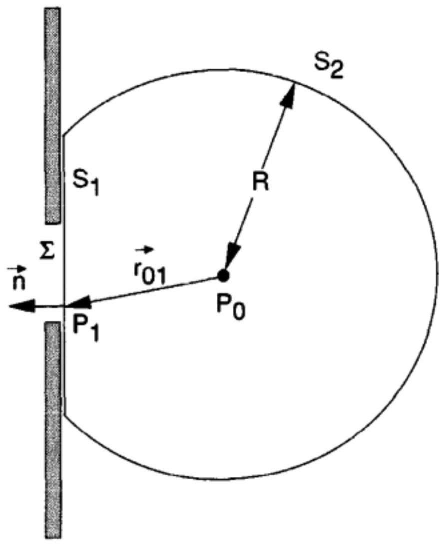


Figure 33: Kirchhoff formulation of diffraction by a plane screen

These conditions are commonly known as the Kirchhoff boundary conditions. The first allows us to specify the disturbance incident on the aperture by neglecting the presence of the screen. The second allows us to neglect all of the surface of integration except that portion lying directly within the aperture itself.

And the equation can be reduced to

$$U(P_0) = \frac{1}{4\pi} \iint_{\Sigma} \left(G\frac{\delta U}{\delta n} - U\frac{\delta G}{\delta n} \right) ds \tag{158}$$

While the Kirchhoff boundary conditions simplify the results considerably, it is important to realize that neither can be exactly true. The presence of the screen will inevitably perturb the field on Σ to some degree, for along the rim of the aperture certain boundary condition must be met that would not be required in the absence of the screen. In addition, the shadow behind the screen is never perfect, for fields will inevitably extend behind the screen for a distance of several wavelengths. However, if the dimensions of the aperture are large compared with a wavelength, these fringing effects can be safely neglected, and the two boundary conditions can be used to yield results that agree very well with experiment.

10.3.5 The Fresnel-Kirchhoff diffraction formula

A further simplification of the expression for $U(P_0)$ is obtained by noting that the distance r_{01} from the aperture to the observation point is usually many optical wavelengths; and if we made the same assumption for the point source P_2 and the distance r_{21} from P_1 we have

$$k \gg \frac{1}{r_{01}} \tag{159}$$

$$k \gg \frac{1}{r_{02}} \tag{160}$$

$$\frac{\delta G(P_1)}{\delta n} = \cos(\vec{n}, \vec{r}_{01})(jk - \frac{1}{r_{01}})\frac{exp(jkr_{01})}{r_{01}} \approx jk\cos(\vec{n}, \vec{r}_{01})\frac{exp(jkr_{01})}{r_{01}} \tag{161}$$

$$U(P_1) = \frac{Aexp(jkr_{21})}{r_{21}} \tag{162}$$

$$\frac{\delta U(P_1)}{\delta n} = A\cos(\vec{n}, \vec{r}_{21})(jk - \frac{1}{r_{21}})\frac{exp(jkr_{21})}{r_{21}} \approx Ajk\cos(\vec{n}, \vec{r}_{21})\frac{exp(jkr_{21})}{r_{21}} \tag{163}$$

$$\begin{aligned} U(P_0) &= \frac{1}{4\pi}\iint_\Sigma \left\{ \frac{\delta U}{\delta n}\left[\frac{exp(jkr_{01})}{r_{01}}\right] - U\frac{\delta}{\delta n}\left[\frac{exp(jkr_{01})}{r_{01}}\right]\right\} ds \\ &= \frac{1}{4\pi}\iint_\Sigma \frac{exp(jkr_{01})}{r_{01}}\left[\frac{\delta U}{\delta n} - jkU\cos(\vec{n}, \vec{r_{01}})\right] ds \\ &= \frac{A}{j\lambda}\iint_\Sigma \frac{exp(jk(r_{01}+r_{02})}{r_{01}r_{02}}\left[\frac{\cos(\vec{n}, \vec{r_{01}}) - \cos(\vec{n}, \vec{r_{21}})}{2}\right] ds \end{aligned} \tag{164}$$

This result is known as the Fresnel-Kirchhoff diffraction formula. This equation can be written in another form:

$$U(P_0) = \iint_\Sigma U'(P_1)\frac{exp(jkr_{01})}{r_{01}} ds \tag{165}$$

where

$$U'(P_1) = \frac{1}{j\lambda}\left[\frac{Aexp(jkr_{21}}{r_{21}}\right]\left[\frac{\cos(\vec{n}, \vec{r_{01}}) - \cos(\vec{n}, \vec{r_{21}})}{2}\right] \tag{166}$$

This equaiton may be interpreted as implying that the field at P_0 arises from an infinity of fictitious secondary point sources located within the aperture itself.

Note that the above derivation has been restricted to the case of an aperture illumination consisting of a single expanding spherical wave. However, such a limitation can be removed by the Rayleigh-Sommerfeld theory.

10.3.6 Rayleigh-Sommerfeld theory

First Rayleigh-Sommerfeld solution:

$$G_-(P_1) = \frac{exp(jkr_{01})}{r_{01}} - \frac{exp(jk\tilde{r}_{01})}{\tilde{r}_{01}} \tag{167}$$

$$U_I(P_0) = \frac{-1}{4\pi}\iint_\Sigma U\frac{\delta G_-}{\delta n} ds = \frac{-1}{2\pi}\iint_\Sigma U\frac{\delta G}{\delta n} ds = \frac{1}{j\lambda}\iint_\Sigma U(P_1)\frac{exp(jkr_{01})}{r_{01}}\cos(\vec{n}, \vec{r}_{01}) ds \tag{168}$$

Second Rayleigh-Sommerfeld solution:

$$G_+(P_1) = \frac{exp(jkr_{01})}{r_{01}} + \frac{exp(jk\tilde{r}_{01})}{\tilde{r}_{01}} \tag{169}$$

$$U_{II}(P_0) = \frac{1}{4\pi}\iint_\Sigma \frac{\delta U}{\delta n} G_+ ds = \frac{1}{2\pi}\iint_\Sigma \frac{\delta U}{\delta n} G ds = \frac{1}{2\pi}\iint_\Sigma \frac{\delta U(P_1)}{\delta n}\frac{exp(jkr_{01})}{r_{01}} ds \tag{170}$$

10.4 Fresnel and Fraunhofer Diffraction

10.4.1 The intensity of a wave field

10.4.2 The Huygens-Fresnel Principle in Rectangular Coordinates

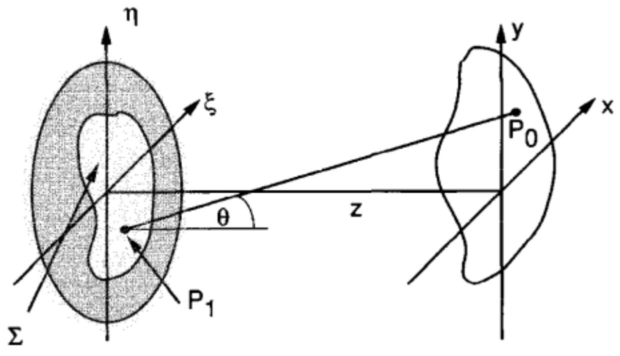


Figure 34: Diffraction geometry

First state the principle in more explicit form for the case of rectangular coordinates. As shown in fig 34, the diffraction aperture is assumed to lie in the left plane and is illuminated in the positive z direction. We will calculate the wavefield across the (x,y) plane, which is parallel to the aperture plane.

According to First Rayleigh-Sommerfeld solution equation 168, the Huygens-Fresnel principle can be stated as

$$U(P_0) = \frac{1}{j\lambda}\iint_\Sigma U(P_1)\frac{exp(jkr_{01})}{r_{01}}\cos(\theta)ds \tag{171}$$

where θ is the angle between the outward normal and the vector pointing from P_0 to P_1. The term $cos\theta$ is given exactly by

$$\cos\theta = \frac{z}{r_{01}} \tag{172}$$

Therefore we can rewrite the equation:

$$U(x,y) = \frac{z}{j\lambda}\iint_\Sigma U(\xi,\eta)\frac{exp(jkr_{01})}{r_{01}^2}d\xi d\eta \tag{173}$$

where the distance r_{01} is given by

$$r_{01} = \sqrt{z^2+(x-\xi)^2+(y-\eta)^2} \tag{174}$$

There have been only two approximations inreaching this expression. One is the approximation inherent in the scalar theory. The second is the assumption that the observation distance is many wavelangths from the aperture, $r_{01} \gg \lambda$.

10.4.3 Fresnel Diffraction

To reduce the Huygens-Fresnel principle to a more simple and usable expression, we introduce approximations for the distance r_{01} between P_1 and P_0.

$$\sqrt{1+b} = 1+\frac{1}{2}b-\frac{1}{8}b^2+... \tag{175}$$

where the number of terms needed for a given accuracy depends on the magnitude of b. Apply this binomial expansion to the r_{01}, we can get

$$r_{01} = z\sqrt{1 + \left(\frac{x-\xi}{z}\right)^2 + \left(\frac{y-\eta}{z}\right)^2} \approx z\left[1 + \frac{1}{2}\left(\frac{x-\xi}{z}\right)^2 + \frac{1}{2}\left(\frac{y-\eta}{z}\right)^2\right] \tag{176}$$

The question now arises as to whether we need to retain all the terms in the approximation, or whether only the first term might suffice. The answer to this question depends on which of the several occurrences of r_{01} is being approximated.

- for the r_{01}^2 in the denominator of equation 173, the error introduced by dropping all terms but z is generally acceptably small.
- for r_{01} appearing in the exponent, errors are much more critical. First, they are multiplied by a very large number k. Second, phase changes of as little as a fraction of a radian can change the value of the exponential significantly. For this reason we retain both terms of the binomial approximation in the exponent.

The resulting expression for the field at (x,y) of equation 173 becomes:

$$U(x,y) = \frac{exp(jkz)}{j\lambda z}\iint_{-\infty}^{+\infty} U(\xi,\eta) exp\left\{j\frac{k}{2z}\left[(x-\xi)^2 + (y-\eta)^2\right]\right\} d\xi d\eta \tag{177}$$

where we have incorporated the finite limits of the aperture in the definition of $U(\xi,\eta)$, in accord with the usual assumed boundary conditions.

This equation 177 can be seen be to a convolution, expressible in the form

$$U(x,y) = \iint_{-\infty}^{+\infty} U(\xi,\eta) h(x-\xi, y-\eta) d\xi d\eta \tag{178}$$

$$h(x,y) = \frac{exp(jkz)}{j\lambda z} exp\left[\frac{jk}{2z}(x^2+y^2)\right] \tag{179}$$

Another form the the equation 177 is Fourier transform of the product of the complex field just to the right of the aperture and a quadratic phase exponential.

$$U(x,y) = \frac{exp(jkz)}{j\lambda z} exp\left[\frac{jk}{2z}(x^2+y^2)\right] \iint_{-\infty}^{+\infty} \left\{U(\xi,\eta)\left[\frac{jk}{2z}(\xi^2+\eta^2)\right]\right\} exp\left[-j\frac{2\pi}{\lambda z}(x\xi + y\eta)\right] d\xi d\eta \tag{180}$$

We refer to both forms of the results as the Fresnel diffraction integral. When this approximation is valid, the observer is said to be in the region of Fresnel diffraction, or equivalently in the near field of the aperture.

10.4.4 Positive or negative phases

It is common practice when using the Fresnel approximation to replace expressions for spherical waves by quadratic-phase exponentials. The question often arises as to whether the sign of the phase should be positive or negative in a given expression.

The critical fact to keep in mind is that we have chosen our phasors to rotate in the clockwise direction, i.e. their time dependence is of the form $exp(-j2\pi vt)$. For this reason, if we move in space in such a way as to intercept portions of a wavefield that were emitted later in time, the

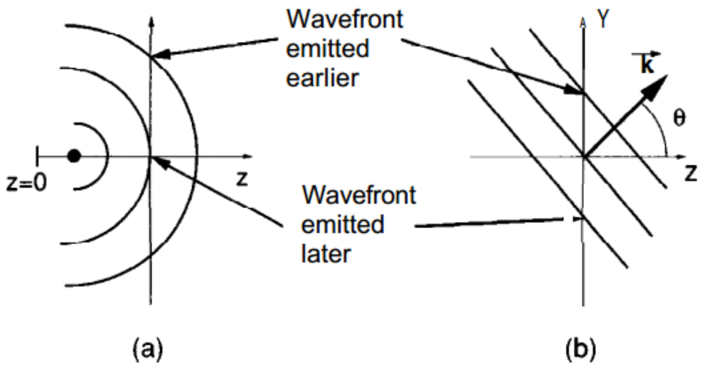


Figure 35: Determining the sign of the phases of exponential representations of spherical/plane waves

phasor will have advanced in the clockwise irection, and therefore the phase must become more negative.

If we imagine observing a spherical wave that is diverging from a point on the z axis, the observation being in an (x, y) plane that is normal to that axis, then movement away from the origin always results in observation of portions of the wavefront that were emitted earlier in time than that at the origin, since the wave has had to propagate further to reach those points. For that reason the phase must increase in a positive sense as we move away from the origin. Therefore the expressions $exp(-jkr_{01})$and $exp[-j\frac{k}{2z}(x^2+y^2)$(for positive Z)represent a diverging spherical wave and a quadraticphase approximation to such a wave, respectively.

10.4.5 Accuracy of the Fresnel approximation

It can be seen from the equation that the spherical secondary wavelets of the Huygens-Fresnel principle have been replaced by wavelets with parabolic wavefronts. A sufficient condition for accuracy would be that the maximum phase change induced by dropping the $b^2/8$ term be much less than 1 radian, which means,

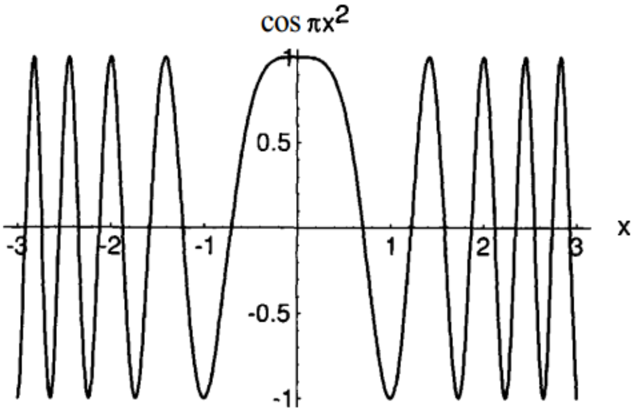
$$z^3 \gg \frac{pi}{4\lambda}[(x-\xi)^2+(y-\eta)^2]^2_{max} \tag{181}$$

But this sufficient condition can be overly stringent. It is not necessary that the higher-order terms of the expansion be small, only that they not change the value of the Fresnel diffraction integral significantly. Considering the concolution form of the result, equation 178, if the major contribution to the integral comes from points (ξ, η) for which $\xi \approx x$ and $\eta \approx y$, then the particular values of the higher-order terms of the expansion are umimportant.

We expand the quadratic-phase exponential of equation 179 into its real and imaginary parts,

$$\frac{1}{j\lambda z}exp\left[\frac{j\pi}{\lambda z}(x^2+y^2)\right] = \frac{1}{j\lambda z}\left\{\cos\left[\frac{\pi}{\lambda z}(x^2+y^2)\right] + j\sin\left[\frac{\pi}{\lambda z}(x^2+y^2)\right]\right\} \tag{182}$$

where we have dropped the unit magnitude phasor $exp(jkz)$ simply by redifing the phase reference. The volume under this function can readily be shown to be unity.



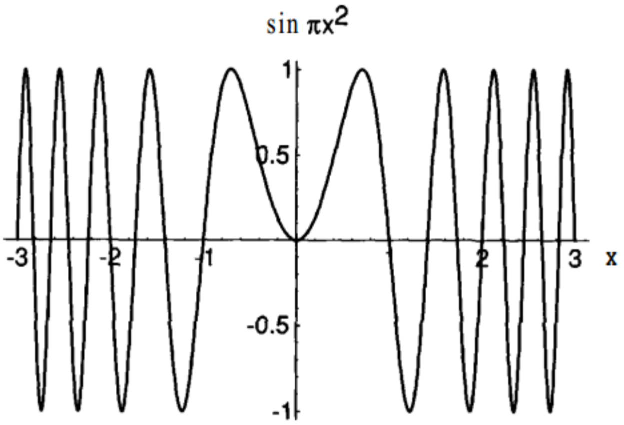


Figure 36: Quadratic phase cosine and sine functions

Figure 36 shows plots of one-dimensional quadratic-phase cosine and sine functions $\cos(\pi x^2)$ and $\sin(\pi x^2)$, each of them has the area of $1/\sqrt{2}$. Using this fact it can be shown that all of the unit area under the two-dimensional quadratic-phase exponential is contributed by the two-dimensional sinusoidal term.

Figure 37 shows the magnitude of the integral of a quadratic-phase exponential function

$$\left| \int_{-X}^{X} exp(j\pi x^2) dx \right| = \tag{183}$$

10.4.6 Fraunhofer Diffraction

$$\beta' = \delta y' / \delta y \tag{184}$$

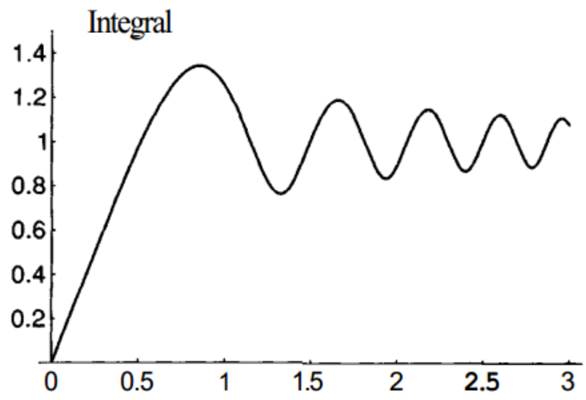


Figure 37: Magnitude of the integral of the quadratic-phase exponential function

11 Q&A

11.1 Question

- How ZEMAX deals with paraxial ray tracing and real ray tracing? Some parameters are based on paraxial approximation, such as EFL, F number and magnification.
- Is it acceptable to model the light source from the slit as a object point?
- Does Abbe sine condition need to be complied with in focusing case?
- Four schemes:
  Kirkpatrick-Baez mirror using two cylindrical mirrors;
  Toroidal mirror;
  Ellipsoidal mirror;
  Conic mirror pair, such as Wolter mirror
- Hints on off-axis sag in user-defined surface.
- Literature download in rwth. OSA, SPIE.