DOING PHYSICS WITH MATLAB
COMPUTATIONAL OPTICS
FOUNDATIONS OF SCALAR DIFFRACTION THEORY

Ian Cooper
School of Physics, University of Sydney
ian.cooper@sydney.edu.au

DOWNLOAD DIRECTORY FOR MATLAB SCRIPTS

View document: Numerical evaluation of the Rayleigh-Sommerfeld diffraction integral of the first kind

SCALAR DIFFRACTION INTEGRALS

The integral theorem of Helmholtz and Kirchhoff

Kirchhoff boundary conditions

on aperture S_A $\quad E(\vec{r}) = E_A(\vec{r})$ and $\nabla E(\vec{r}) = \nabla E_A(\vec{r})$

$$E(\mathrm{P}) = \frac{1}{4\pi} \iint_{S_A} E_A(\vec{r}) \left(\frac{\mathrm{e}^{jkr}}{r^2} \right) \left((jkr - 1)\hat{\vec{r}} - r\nabla E_A(\vec{r}) \right) \cdot \vec{n}\, \mathrm{d}S$$

Rayleigh-Sommerfeld diffraction integral of the second kind

$$E(\mathrm{P}) = \frac{-1}{2\pi} \iint_{S_A} \frac{e^{jkr}}{r} \nabla E(\vec{r}) \cdot \vec{n}\, \mathrm{d}S$$

Rayleigh-Sommerfeld diffraction integral of the first kind

$$E(\mathrm{P}) = \frac{1}{2\pi} \iint_{S_A} E(\vec{r}) \frac{e^{jkr}}{r^3} z_p\, (jkr - 1)\, \mathrm{d}S$$

The **Rayleigh-Sommerfeld region** includes the entire space to the right of the aperture. It is assumed that the Rayleigh-Sommerfeld diffraction integral of the first kind is valid throughout this space, right down to the aperture. There are no limitations on the maximum size of either the aperture or observation region, relative to the observation distance, because **no approximations have been made**.

THE INTEGRAL THEOREM OF HELMHOLTZ AND KIRCHHOFF

The basic idea of the **Huygens-Fresnel theory** is that the light disturbance at a point P arises from the superposition of secondary waves that proceed from a surface situated between this point P and the light source. Kirchhoff formulated a sound mathematical basis for this principle based upon solutions of the homogeneous wave equation, where the value at an arbitrary point P in the field can be expressed in terms of the values of the solution and its first derivatives at all points on an arbitrary closed surface surrounding the point P.

We consider strictly monochromatic scalar waves that are solutions of the wave equation

$$\nabla^2 E(\vec{r},t) - \frac{1}{c^2} \frac{\partial^2 E(\vec{r},t)}{\partial t^2} = 0 \tag{1}$$

where $E(\vec{r},t)$ is electric field (or monochromatic optical disturbance) and can be written as

$$E(\vec{r},t) = E(\vec{r})\, e^{-j\omega t} = E(\vec{r})\, e^{-jkct} \tag{2}$$

Solutions of the wave equation for ***plane waves*** traveling the direction $+\vec{r}$ and $-\vec{r}$ are

$+\vec{r}$ direction: $\qquad E(\vec{r},t) = E_o\, e^{j(\vec{k}\bullet\vec{r}-\omega t)}$

$-\vec{r}$ direction: $\qquad E(\vec{r},t) = E_o\, e^{j(-\vec{k}\bullet\vec{r}-\omega t)} \tag{3}$

Solutions of the wave equation for ***spherical waves*** traveling the direction $+\vec{r}$ and $-\vec{r}$ are

$+\vec{r}$ direction: $\qquad E(\vec{r},t) = \dfrac{E_o}{r} e^{j(\vec{k}\bullet\vec{r}-\omega t)}$

$-\vec{r}$ direction: $\qquad E(\vec{r},t) = \dfrac{E_o}{r} e^{j(-\vec{k}\bullet\vec{r}-\omega t)}$

In vacuum, the space-dependent part then satisfies the ***Helmholtz equation***

$$\nabla^2 E(\vec{r}) + k^2\, E(\vec{r}) = 0 \tag{4}$$

In problems involving boundaries it is often convenient to study the properties of the difference between two solutions of an equation rather than one of the solution alone, since the boundary conditions become simpler to handle.

Take the origin at the observation point P, at which the electric field has the value $E(0)$. The two wave fields, $E(\vec{r})$ the unknown field to be calculated and $E_t(\vec{r})$ a trial solution of the Helmholtz equation satisfy the equation

$$E(\vec{r})\nabla^2 E_t(\vec{r}) - E_t(\vec{r})\nabla^2 E(\vec{r}) = -E(\vec{r})k^2 E_t(\vec{r}) + E_t(\vec{r})k^2 E(\vec{r}) = 0 \quad (5)$$

at all points except at the origin $\vec{r} = 0$. Now, integrate (5) throughout a volume V bounded by a surface S and by *Green's theorem*, the volume integral can be changed to a surface integral:

$$\iiint_V \left(E(\vec{r})\nabla^2 E_t(\vec{r}) - E_t(\vec{r})\nabla^2 E(\vec{r}) \right) dV = \oiint_S \left(E(\vec{r})\nabla E_t(\vec{r}) - E_t(\vec{r})\nabla E(\vec{r}) \right) \bullet \vec{n}\, dS = 0$$

$$\oiint_S \left(E(\vec{r})\nabla E_t(\vec{r}) - E_t(\vec{r})\nabla E(\vec{r}) \right) \bullet \vec{n}\, dS = 0 \quad\quad (6)$$

$\vec{n}$ being the inward pointing normal to the surface S. We assume that both $E(\vec{r})$ and $E_t(\vec{r})$ possess continuous first- and second-order partial derivatives within and on this surface. Because the integrand is zero in (5), the integrals in (6) are also zero. Furthermore, the volume containing P must be free of any sources of optical radiation (including secondary sources of reradiated or scattered light).

Note: terms such as $\nabla\psi \bullet \vec{n}$ give the normal component of the gradient at the surface of integration.

Consider a trial solution to be

$$E_t(\vec{r}) = \frac{E_{ot}}{r}e^{jkr} \quad\quad\quad\quad (7)$$

which is a spherical wave radiating from the origin at P. This spherical wave $E_t(\vec{r})$ plays the role of a mathematical auxiliary function – a "probe" which we use for investigating the optical field $E(\vec{r})$. This is a virtual spherical wave and has nothing to do with any real spherical waves that act as the source functions.

This trial function has a singularity at the point $r = 0$ and since $E_t(\vec{r})$ was assumed to be continuous and differentiable, the point P must be excluded from the domain of integration in equation (6). This is done by surrounding P by a small sphere of radius ε and taking the surface S to be divided into two parts, an arbitrary outer surface S_o and the small inner spherical surface S_i of radius ε.
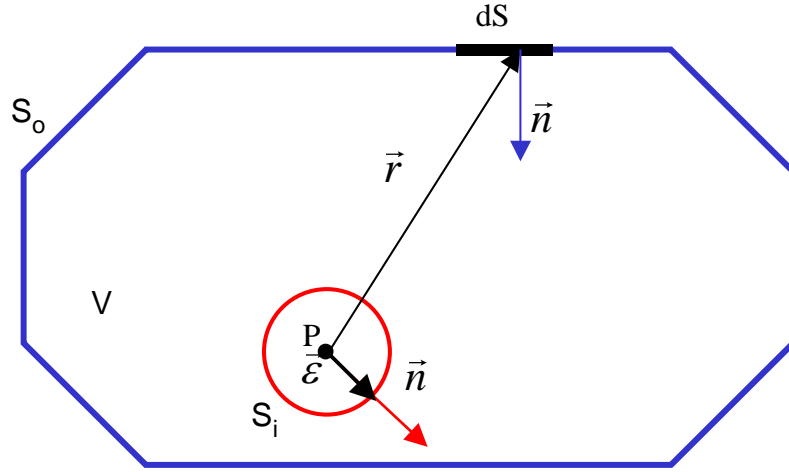
Therefore, the gradient of the trial function is

$$\nabla E_t(\vec{r}) = \nabla\left(\frac{E_{ot}e^{jkr}}{r}\right) = \hat{r}\,\frac{\partial}{\partial r}\left(\frac{E_{ot}e^{jkr}}{r}\right) = \frac{1}{r^2}\left(jkr-1\right)E_{ot}e^{jkr}\,\hat{r} \qquad (8)$$

However, after we make the substitutions into of equations (7) and (8) into (6) we can take these waves to describe spherical waves which are radiated by elements of the surface S and which arrive at the point P at a distance r from dS.

Therefore, (6) become

$$\oiint_{S_o}\left(E(\vec{r})\nabla E_t(\vec{r})-E_t(\vec{r})\nabla E(\vec{r})\right)\bullet\vec{n}\,dS + \oiint_{S_i}\left(E(\vec{r})\nabla E_t(\vec{r})-E_t(\vec{r})\nabla E(\vec{r})\right)\bullet\vec{n}\,dS = 0$$

$$(9)$$

The contribution from S_i can be evaluated directly, since over the small sphere of radius ε we consider $E(\varepsilon)$ to be constant and equal to $E(0)$. The unit vector $\vec{n}$ is parallel to $\vec{\varepsilon}$, so we have $\vec{\varepsilon}\bullet\vec{n}=1$ and we can substitute $\varepsilon^2 d\Omega$ for dS where dΩ is an element of the solid angle. Therefore, the integral over the inner surface can be written as

$$I_{\mathrm{i}} = \oiint_{S_{\mathrm{i}}} \left(E_{\mathrm{ot}}\, e^{jk\varepsilon} \right) \left[E(0)\,(jk\varepsilon - 1)\frac{\hat{\vec{\varepsilon}}}{\varepsilon^2} - \frac{1}{\varepsilon}\nabla E(0) \right] \bullet \vec{n}\,\varepsilon^2\, d\Omega$$

$$I_{\mathrm{i}} = \left(E_{\mathrm{ot}}\, e^{jk\varepsilon} \right) \left\{ \oiint_{S_{\mathrm{i}}} E(0)\,(jk\varepsilon)\,d\Omega - \oiint_{S_{\mathrm{i}}} E(0)\,d\Omega - \oiint_{S_{\mathrm{i}}} \varepsilon\,\nabla E(0)\bullet\vec{n}\,d\Omega \right\}$$

$$\text{limit } \varepsilon \to 0$$

$$I_{\mathrm{i}} = -4\pi\, E_{\mathrm{ot}}\, E(0)$$

$$(10)$$

The value of the electric field at the observation point P is

$$\oiint_{S_{\mathrm{o}}} \left(E(\vec{r})\frac{\hat{\vec{r}}}{r^2}\,(jkr-1)\,E_{\mathrm{ot}}\, e^{jkr} - \frac{E_{\mathrm{o}}\, e^{jkr}}{r}\nabla E(\vec{r}) \right) \bullet \vec{n}\, dS - 4\pi\, E_{\mathrm{ot}}\, E(0) = 0$$

$$E(0) = \frac{1}{4\pi} \oiint_{S_{\mathrm{o}}} E(\vec{r}) \left(\frac{e^{jkr}}{r^2} \right) \left((jkr-1)\,\hat{\vec{r}} - r\,\nabla E(\vec{r}) \right) \bullet \vec{n}\, dS \qquad (10)$$

Thus $E(0)$ can be found if the values for $E(\vec{r})$ and $\nabla E(\vec{r})$ are known at all points on the outer surface surrounding the point P. This is one form of the ***integral theorem of Helmholtz and Kirchhoff***.

This theorem expresses $E(0)$ in terms of both $E(\vec{r})$ and $\nabla E(\vec{r})$. It may, however be shown that from the theory of Green's functions, that either $E(\vec{r})$ or $\nabla E(\vec{r})$ on S_{o} are sufficient to specify $E(0)$ at every point P with S_{o}.

In applying the integral theorem of Helmholtz and Kirchhoff to diffraction problems, the surface S_{o} is divided into three: 1) the diffraction aperture (opening) S_{A}, 2) the screen blocking the radiation S_{B}, and 3) since the surface must be closed, we must include a mathematical surface S_{∞} of radius R centre about the point P, so that when $R \to \infty$ the integral over the integral in (10) will approach zero or R is chosen so large there is insufficient time for a contribution from this surface to have reached the point P.

$S_o = S_A + S_B + S_\infty$
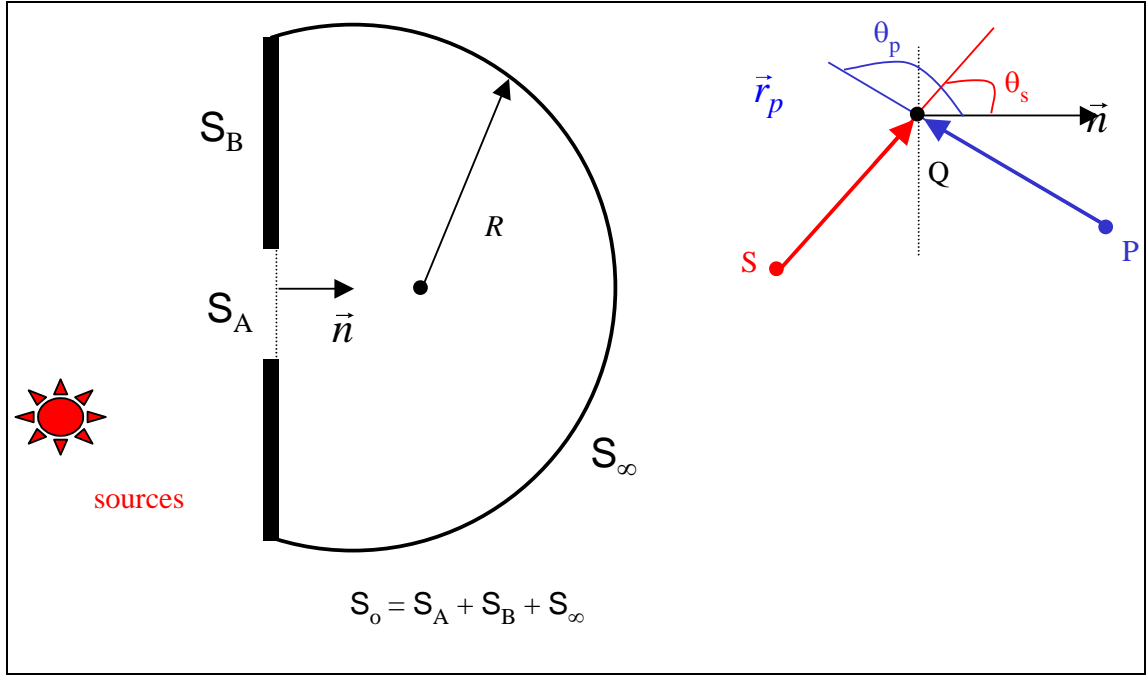
It remains to integrate (10) over $S_A + S_B$. However, the actual field and its gradient over the surface are almost impossible to determine except in the simplest cases because for real physical apertures will partially absorb, reflect and scatter the incident radiation.

The first approximation to make in integrating (10) is to ignore the real boundary values and apply the so-called *Kirchhoff boundary conditions* and are the basis of *Kirchhoff's diffraction theory*. These replace the actual field in the presence of the aperture by the incident field for the open aperture $E_A(\vec{r})$ and by zero elsewhere:

Kirchhoff boundary conditions:

on S_B $\quad E(\vec{r}) = 0 \quad$ and $\quad \nabla E(\vec{r}) = 0$ $\qquad\qquad$ (11a)

on S_A $\quad E(\vec{r}) = E_A(\vec{r}) \quad$ and $\quad \nabla E(\vec{r}) = \nabla E_A(\vec{r})$ $\qquad$ (11b)

Then (10) becomes

$$E(0) = \frac{1}{4\pi} \iint_{S_A} E_A(\vec{r}) \left(\frac{e^{jkr}}{r^2} \right) \left((jkr - 1)\hat{\vec{r}} - r\nabla E_A(\vec{r}) \right) \cdot \vec{n} \, dS \qquad (12)$$

The Kirchhoff boundary conditions are an approximation, since at the edge of the opening, the amplitude of the wave does not suddenly jump from a finite value to zero. The approximation is better when the aperture is large compared to the wavelength.

Strictly speaking the boundary conditions (11) are mathematically inadmissible. A theorem in Riemann's theory of functions implies that if the value of a function and its first derivative vanish along a segment then the function vanishes everywhere, in fact the assumed boundary conditions even contradict each other and if we would calculate the value at P on the surface S, the boundary conditions would not be reproduced.

In principle, a scalar approach can be performed for each component of the vector wave, but in practice this is rarely necessary. We can assume that the radiation from the opening is much larger than from the blocking screen. Also, well into the opening, all possible directions of polarization are possible for the radiation. Often, edge effects are only predominant in a region within less than a wavelength from the edges. Therefore, in most practical applications the use of a scalar approach is justified. However, in the scalar approach, all information about the polarization of the wave is lost.

GREEN'S FUNCTION
A SIMPLIFIED FORMULATION OF HUYGENS'S PRINCIPLE

Rayleigh-Sommerfeld diffraction integral of the first kind

Rayleigh-Sommerfeld diffraction integral of the second kind

The mathematical contradiction posed by the Kirchhoff boundary conditions can be avoided by substituting for the trial spherical wave function $E_t(\vec{r})$ given by (7) by the *Green's function* G belonging to the surface surrounding the point P. This function is defined by the following conditions:

within the volume V $\qquad\qquad \nabla^2 G + k^2 G = 0$ $\qquad\qquad\qquad$ (13a)

on S_B $\qquad\qquad\qquad\qquad G = 0$ $\qquad\qquad\qquad\qquad\qquad$ (13b)

as $r \to 0$ $\qquad\qquad\qquad\qquad G \to E_t(\vec{r})$ $\qquad\qquad\qquad\qquad$ (13c)

as $r \to \infty$ $\qquad\qquad\qquad r(\nabla G \bullet \vec{n} - j k G) \to 0$ $\qquad\qquad$ (13d)

as before, r is the distance from the point P and (13d) is called the radiation condition. G has a singularity only at the point P ($r = 0$) and is continuous everywhere else within the volume V. G differs from $E_t(\vec{r})$ because of the additional condition (13b). As a result of this condition the term containing the term $\nabla E(\vec{r})$ in equation (9) vanishes and therefore (9) simplifies to

$$E(P) = \frac{1}{4\pi} \oiint_{S_o} E(\vec{r}) \ \nabla G(\vec{r}) \bullet \vec{n}\, dS \qquad\qquad (14)$$

Now, we need to specify only the boundary values of $E(\vec{r})$ and not its gradient. This approach has the practical advantage of leading to the simpler form of the integral as given in (14) as compared to (12).

This approach again is only an approximation and is valid for sufficiently small wavelengths. The real field does not vanish completely behind the blocking screen, nor is the field entirely unaffected by the presence of the screen, at least not within distances of the order of magnitude of a wavelength from the edge of the screen.

However, the applicability of the Green's function method is restricted to the special case of a *plane screen*, for this is the only case for which the Green's function can be conveniently expressed.

Note: The wavelength λ is really only defined for plane monochromatic waves and can lose its significance in wave types encountered in diffraction. However, we can always interpret λ as the length $2\pi c/\omega$ which is defined for all monochromatic radiation processes and which for plane waves is identical with the actual wavelength.

The method of images is used to determine the appropriate Green's function.

Construction of the Green's function for a plane screen
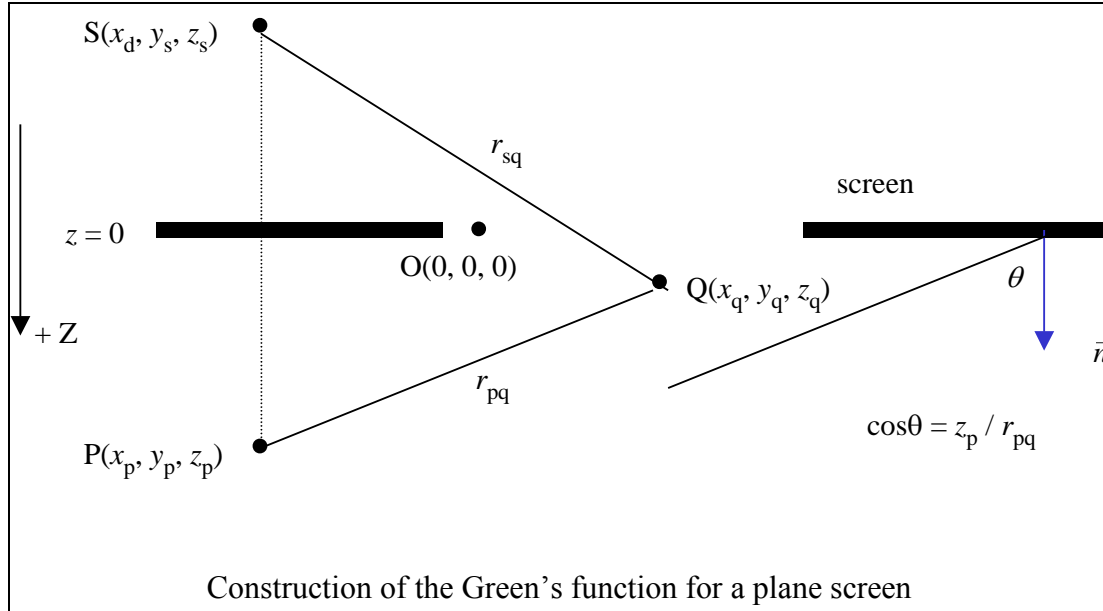
We construct the mirror image S of the point P with respect to the plane of the screen $z = 0$. For an arbitrary point $Q(x_q, y_q, z_q)$ where $z_q > 0$, we form the Green's function

$$G = \frac{e^{jk\,r_{pq}}}{r_{pq}} - \frac{e^{jk\,r_{sq}}}{r_{sq}} \tag{15}$$

where

$$r_{pq}{}^2 = (x_p - x_q)^2 + (y_p - y_q)^2 + (z_p - z_q)^2$$

$$r_{sq}{}^2 = (x_s - x_q)^2 + (y_s - y_q)^2 + (z_s - z_q)^2$$

$$x_p = x_s \qquad y_p = y_s \qquad z_p = -z_s \tag{16}$$

We have to determine ∇G to evaluate (14). Firstly,

$$\frac{\partial G}{\partial z_q} = \frac{d}{dr_{pq}}\left(\frac{e^{jk\,r_{pq}}}{r_{pq}}\right)\frac{\partial r_{pq}}{\partial z_q} - \frac{d}{dr_{sq}}\left(\frac{e^{jk\,r_{sq}}}{r_{sq}}\right)\frac{\partial r_{sq}}{\partial z_q} \tag{17}$$

If we now place Q on the screen, we have

$$r_{pq} = r_{sp} = r$$

$$\frac{\partial r_{pq}}{\partial z_q} = -\frac{\partial r_{sq}}{\partial z_q} = \cos\theta = \frac{z_p}{r_{pq}} = \frac{z_p}{r} \tag{18}$$

$$|\nabla G| = \frac{\partial G}{\partial n} = \frac{\partial G}{\partial z_q} = 2\frac{\partial}{\partial r}\left(\frac{e^{jkr}}{r}\right)\cos\theta \tag{19}$$

$$\frac{\partial}{\partial r}\left(\frac{e^{jkr}}{r}\right) = \frac{1}{r^2}(jkr-1)e^{jkr} \tag{20}$$

Substituting (19) and (20) into (14) we obtain and we can integrate over the area S_A.

$$E(\mathrm{P}) = \frac{1}{2\pi}\iint_{S_A} E(\vec{r})\frac{e^{jkr}}{r^3}z_p(jkr-1)\,\mathrm{d}S \tag{21}$$

(21) is known as the ***Rayleigh-Sommerfeld diffraction integral of the first kind***.

If we make the approximation that $k\,r \gg 1$ then $\dfrac{1}{r^2}(jk\,r-1) \approx \dfrac{2\pi}{\lambda\,r}j$

then (21) can be written as

$$E(\mathrm{P}) = \frac{j}{\lambda}\iint_{S_A} E(\vec{r})\,\frac{e^{jkr}}{r}\cos\theta\,\mathrm{d}S \tag{22}$$

This expression (22) is equivalent to Huygens's principle – *a light wave falling on an aperture propagates as if every element* d*S emitted a spherical wave, the amplitude and phase of which are given by that of the incident wav*e. The factor $\cos\theta$ relates to Lambert's law of surface brightness.

Instead of the Green's function (15) which satisfies the boundary conditions (13) in which $G = 0$ at $z = 0$, we now form

$$G = \frac{e^{jk\,r_{pq}}}{r_{pq}} + \frac{e^{jk\,r_{sq}}}{r_{sq}} \tag{23}$$

This is a function which satisfies the boundary condition $\partial G/\partial z = 0$ at $z = 0$.

As a result of this condition the term containing the term $E(\vec{r})$ in equation (9) vanishes and therefore (9) simplifies to

$$E(\mathrm{P}) = \frac{-1}{4\pi} \oiint_{S_o} G(\vec{r}) \, \nabla E(\vec{r}) \bullet \vec{n} \, \mathrm{d}S \quad (24)$$

Now, we need to specify only the boundary values of $\nabla E(\vec{r})$ and not the value of $E(\vec{r})$.

If we substitute the values of the Green's function on the plane $z = 0$ we obtain

$$E(\mathrm{P}) = \frac{-1}{2\pi} \iint_{S_{\mathrm{A}}} \frac{e^{jkr}}{r} \, \nabla E(\vec{r}) \bullet \vec{n} \, \mathrm{d}S \qquad (25)$$

(25) is known as the ***Rayleigh-Sommerfeld diffraction integral of the second kind***.

The ***Rayleigh-Sommerfeld region*** – the entire space to the right of the aperture. It is assumed that the Rayleigh-Sommerfeld diffraction integrals are valid throughout this space, right down to the aperture. There are no limitations on the maximum size of either the aperture or observation region, relative to the observation distance, because no approximations have been made.

The ***Fresnel region*** (**near field**) – that portion of the Rayleigh-Sommerfeld region within which the Fresnel conditions are satisfied. Both the input and output signals are restricted to regions lying near the z axis, i.e., the lateral dimensions of which are much smaller than the separation between the input and output planes.

$$\left| z_{pq} \right| >> L_1 + L_2 \qquad (26)$$

where L_1 and L_2 are the maximum lateral extents of the aperture and observation planes respective. Also it is assumed that

$$\left| z_{pq} \right|^3 >> \frac{\pi \left(L_1 + L_2 \right)^4}{4\lambda} \qquad (27)$$

Conditions (26) and (27) are sometimes known as the **Fresnel conditions**.

The *Fraunhofer region* (**far field**) – that portion of the Fresnel region within which the Fraunhofer condition is satisfied

$$\left| z_{pq} \right| \gg \frac{\pi L_1^2}{\lambda} \qquad\qquad (30)$$

In the Fraunhofer region the size of the diffraction pattern increases with increasing distance but its shape is invariant.

The Rayleigh distance d$_{RL}$

The **Rayleigh distance** in optics is the axial distance from a radiating aperture to a point an observation point P at which the path difference between the axial ray and an edge ray is $\lambda / 4$. A good approximation of the Rayleigh Distance d_{RL} is

$$d_{RL} = \frac{4 a^2}{\lambda}$$

where *a* is the radius of the aperture. Rayleigh distance is also a distance beyond which the distribution of the diffracted light energy no longer changes according to the distance z_P from the aperture.

$$z_P \; < \; d_{RL} \qquad \text{Fresnel diffraction}$$

$$z_P \; > \; d_{RL} \qquad \text{Fraunhofer diffraction.}$$

If we consider a circular aperture of radius *a*, then much of the energy passing through the aperture is diffracted through an angle of the order $\theta \sim \lambda / a$ from its original propagation direction. When we have travelled a distance $\sim d_{RL}$ from the aperture, about half of the energy passing through the opening will have left the cylinder made by the geometric shadow if $a / d_{RL} \sim \theta$. Putting these formulae together, we see that the majority of the propagating energy in the "far field region" at a distance greater than the Rayleigh distance $d_{RL} = 4 a^2 / \lambda$ will be diffracted energy. In this region then, the polar radiation pattern consists of diffracted energy only, and the angular distribution of propagating energy will then no longer depend on the distance from the aperture.

References

Born 417   Hect 501,648   Klein 360   Lipson 156   Slater 168   Sommerfeld 197
Gaskill p385