2006

From Physics to Illumination Models of Subsurface Scattering

Huiying Xu
Yi Ru
Yinlong Sun

Report Number:
06-002
FROM PHYSICS TO ILLUMINATION MODELS OF SUBSURFACE SCATTERING

Huiying Xu
Yi Ru
Yinlong Sun

Department of Computer Sciences
Purdue University
West Lafayette, IN 47907

CSD TR #06-002
January 2006
From Physics to Illumination Models of Subsurface Scattering

Huiying Xu, Yi Ru, Yinlong Sun
Department of Computer Sciences, Purdue University, W. Lafayette, IN, 47907-1398, USA

Abstract

The transport of electromagnetic waves is essential in many scientific and engineering fields. In recent years, substantial researches have been conducted in computer graphics to model and simulate the optical phenomena of subsurface scattering. The ultimate goal is to develop accurate scattering models for complex surfaces and media, as well as efficient algorithms for simulating the related phenomena. This report reviews on the subsurface scattering problem from the fundamental physics principles to the practical methods developed recently in computer graphics.

Keywords: wave scattering, subsurface scattering, translucent media, multiple layers, illumination models, Monte Carlo method
1 Introduction

The transport of electromagnetic waves are an essential problem in many scientific and engineering fields. Traditional applications include optical engineering, radiophysics, remote sensing, radar imaging, and metrology. In recent years, a tremendous amount of research on subsurface scattering has been conducted in computer graphics to synthesize realistic images of optical scattering though translucent media. The ultimate goal is to develop accurate and efficient scattering models for complex media and surfaces, such as for objects of translucent material or with surfaces consisting of multiple layers. Based on this research development, this report attempts to make a review on subsurface scattering with a focus on the linkage between the fundamental physics and the practical illumination models and methods.

Although the applications of subsurface scattering involved in this report are entirely optical, we believe that it is beneficial to discuss the physical aspect of the scattering in terms of general electromagnetic waves. The advantage is that all the principles and equations are meant to be applicable to all kinds of electromagnetic waves, with optical as a special case. Therefore, we started with review from the Maxwell’s equations. When using generic equations to optical scattering, we just need to specify that the waves are light waves.

In computer graphics, the interactions between light and objects (or surfaces) may be classified into local and global illumination. Global illumination handles light transport amongst objects, and local illumination addresses the scattering behavior from a small surface area. Because the area considered for local illumination is sufficiently small with respect to the eye or to the typical size of the interested objects, the scattering details in the area can be encapsulated into a function which serves as a building block for global illumination. This strategy can greatly reduce the computational complexity in analysis and simulation of optical processes.

The task of deriving local illumination models is extremely challenging. One reason is that the problem depends on multiple factors including the incident and outgoing directions, surface roughness, surface anisotropy, surface composition, material properties, wavelength, and polarization. Besides, surface reflection may involve not only single scattering, but also multiple scattering. Since a local illumination area may have a complicated profile, an incident wave may be actually reflected once, twice or more times by the surface. If the surface material is transparent, a scattering path may contain both reflection and transmission events. Moreover, natural surfaces not only are often rough, but also differ statistically. Finally, surface reflection may involve complex scattering mechanisms and considerable volume absorption. For example, light reflection at a surface of a translucent medium involves surface scattering at the medium’s boundary as well as volume scattering by small particles within the medium. One particularly challenging case is a surface of some translucent material or consisting of multiple layers. In this case, surface reflection may significantly involve scattering within the translucent medium or the multiple layers, which is often called subsurface scattering.

The following sections are organized as follows. Section 2 reviews the electromagnetic theory. Section 2 reviews the scattering theory of single particles. Section 4 discusses the transport theory of light in media containing particles. Section 5 reviews Monte Carlo simulation of light transport. Section 6 reviews the illumination models of subsurface scattering of light.
2 Electromagnetic Wave Theory

In this section, we first present the fundamental equations — Maxwell equations. Then, we will describe how electromagnetic waves (including damped waves and plane waves) propagate in space. Some fundamental concepts like the Poynting vector and energy density will be defined. Then we will describe polarizations of electromagnetic waves and Stokes parameters. Finally, we will discuss reflection and refraction of electromagnetic waves at an interface between two media. Further information may be found in literature on electromagnetic theory 

2.1. Maxwell Equations

Classical electromagnetic phenomena are governed by Maxwell equations:

\[ \nabla \cdot \mathbf{D} = \rho \]
\[ \nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \]
\[ \nabla \cdot \mathbf{B} = 0 \]
\[ \nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \]

where \( \rho \) is the charge density, \( \mathbf{D} \) the displacement current, \( \mathbf{H} \) the magnetic field intensity, \( \mathbf{J} \) the electric current density, \( \mathbf{B} \) the magnetic induction, and \( \mathbf{E} \) the electric field intensity. In a uniform isotropic linear medium, the following relations apply

\[ \mathbf{D} = \varepsilon \mathbf{E} \]
\[ \mathbf{B} = \mu \mathbf{H} \]
\[ \mathbf{J} = g \mathbf{E} \]

where \( \varepsilon \) is the permittivity, \( \mu \) the permeability, and \( g \) the conductivity.

Maxwell equations are the basis of classical electrodynamics. Eq. (2.1b) shows that an electric field varying with time will give rise to a magnetic field, and Eq. (2.1d) shows that a magnetic field varying with time will give rise to an electric field. These equations indicate the co-existence of an electric and a magnetic field. Maxwell equations predict that the light is an electromagnetic wave. They also show that electromagnetic waves may be generated at any frequencies.

2.2. Isotropic and Homogeneous Media

We consider an electromagnetic field in an isotropic and homogeneous medium. Substituting Eq. (2.2b) into Eq. (2.1d) and performing the curl operation to it, we obtain

\[ \nabla \times \nabla \times \mathbf{E} = -\mu \frac{\partial}{\partial t} (\nabla \times \mathbf{H}) = -\mu g \frac{\partial \mathbf{E}}{\partial t} - \mu \varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} \]
Since \( \nabla \times \nabla \times \mathbf{E} = -\nabla^2 \mathbf{E} + \nabla (\nabla \cdot \mathbf{E}) = -\nabla^2 \mathbf{E} + \frac{\nabla \rho}{\varepsilon} \), we obtain

\[
\mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} + \mu g \frac{\partial \mathbf{E}}{\partial t} - \nabla^2 \mathbf{E} = -\frac{\nabla \rho}{\varepsilon}.
\]  

(2.4)

Similarly, substituting Eqs. (2.2a) and (2.2c) into Eq. (2.1b) and performing curl,

\[
\mu \epsilon \frac{\partial^2 \mathbf{H}}{\partial t^2} + \mu g \frac{\partial \mathbf{H}}{\partial t} - \nabla^2 \mathbf{H} = 0.
\]  

(2.5)

Both Eqs. (2.4) and (2.5) describe the propagation of an electromagnetic wave. If we assume that the space is charge-free \((\rho = 0)\) and the wave is monochromatic (with a single frequency), the solution of the electromagnetic wave can be written as

\[
\begin{pmatrix}
\mathbf{E} \\
\mathbf{H}
\end{pmatrix} = \begin{pmatrix}
\mathbf{E}_0(x) \\
\mathbf{H}_0(x)
\end{pmatrix} e^{-i\omega t}.
\]  

(2.6)

We adopt the convention that the electromagnetic waves are described with complex numbers (the actual fields are only their real parts).

Substituting Eq. (2.6) into Eqs. (2.4) and (2.5), we obtain

\[
\begin{pmatrix}
\mu \epsilon \omega^2 + i\mu g \omega + \nabla^2 \\
\mu \mu \epsilon \omega^2 + i\mu g \omega + \nabla^2
\end{pmatrix} \mathbf{F} = \mathbf{0},
\]

where \( \mathbf{F} = \begin{pmatrix} \mathbf{E}_0(x) \\ \mathbf{H}_0(x) \end{pmatrix} \).

In the case \( g \neq 0 \), Eq. (2.7) describes a damped wave. To understand this, consider the one-dimensional case and let \( g/(\varepsilon \omega) \gg 1 \); Eq. (2.7) can be simplified into

\[
\begin{pmatrix}
i\mu g \omega + \frac{d^2}{dx^2}
\end{pmatrix} \mathbf{F} = \mathbf{0}.
\]  

(2.8)

The solution can be written in the form of

\[
\mathbf{F}(x) = \mathbf{F}_e e^{\pm i(\omega t + k x)},
\]  

(2.9)

where the signs \( \pm \) stand for two waves propagating in opposite directions, and \( \delta = \sqrt{8\pi/(\mu g \omega)} \) is called the skin depth, the distance that the wave is damped to \( 1/e \) of its original amplitude.

In the case \( g = 0 \), Eq. (2.7) can be simplified into

\[
\begin{pmatrix}
\mu \epsilon \omega^2 + \nabla^2
\end{pmatrix} \mathbf{F} = \mathbf{0}.
\]  

(2.10)

Eq. (2.10) describes a plane electromagnetic wave. Solving Eq. (2.10), we obtain

\[
\mathbf{F}(x) = \mathbf{F}_e e^{i k x},
\]  

(2.11)

where \( k \) is called the propagation vector, \( k = \sqrt{\mu \epsilon \omega} \), and its direction \( n \) is the wave propagation direction. The phase velocity of the wave is

\[
v = \frac{\omega}{k} = \frac{1}{\sqrt{\mu \epsilon}}.
\]  

(2.12)
Finally, $E(x,t)$ and $H(x,t)$ are given as

$$
\begin{pmatrix}
  E(x,t) \\
  H(x,t)
\end{pmatrix} =
\begin{pmatrix}
  E_0 e^{-i(\omega t + k \cdot x)} \\
  H_0 e^{-i(\omega t + k \cdot x)}
\end{pmatrix}
$$

(2.13)

The wave described by Eq. (2.13) is a transverse wave propagating in direction $n \cdot n$ (the propagation direction), $e_1$ (the direction of $E$), and $e_2$ (the direction of $H$) are perpendicular to each other and they satisfy the right-hand rule (see Figure 2.1).

Figure 2.1: Directions of $E$, $H$, and $k$.

Substituting Eq. (2.13) into Eq. (2.1d), we obtain

$$
H_0 = \frac{\varepsilon}{\mu} n \times E_0.
$$

(2.14)

The flow of the time-averaged electromagnetic energy (energy per unit area per unit time) is given by the real part of the complex Poynting vector:

$$
S = \frac{1}{2} E \times H^* = \frac{1}{2} \frac{\varepsilon}{\mu} |E_0|^2 n.
$$

(2.15)

The time-averaged energy density $u$ is given as

$$
u = \frac{1}{4} \left( \varepsilon E \cdot E^* + \frac{1}{\mu} B \cdot B^* \right) = \frac{\varepsilon}{2} |E_0|^2 = \frac{\mu}{2} |H_0|^2.
$$

(2.16)

The ratio of the magnitude of Eq. (2.15) to Eq. (2.16) shows that the speed of energy flow is $v = 1/\sqrt{\mu \varepsilon}$, as expected from Eq. (2.12).

### 2.3. Polarizations and Stokes Parameters

The plane wave described by Eq. (2.13) is a wave with its electric field vector always in direction $e_1$. Such a wave is called a linearly polarized wave (with polarization vector $e_1$). Consider another wave polarized in $e_2$ and is linearly independent of the first wave; the two waves are
The combination gives a general homogeneous plane wave propagating in \( \mathbf{k} = kn \),
\[
\mathbf{E}(x, t) = (e_1 E'_0 + e_2 E''_0) e^{i(k \cdot x - \omega t)}.
\] (2.18)

The amplitudes \( E'_0 \) and \( E''_0 \) are complex numbers, so it is possible that there is a phase difference between the two waves with different polarizations.

If \( E'_0 \) and \( E''_0 \) have the same phase, Eq. (2.18) describes a linearly polarized wave whose polarization vector makes an angle \( \theta = \tan^{-1}(E''_0/E'_0) \) with \( e_1 \). The total field magnitude is 
\[
E_0 = \sqrt{E'_0^2 + E''_0^2}
\]

Figure 2.2: Electric field of a linearly polarized wave.

In general, when \( E'_0 \) and \( E''_0 \) have different phases, Eq. (2.18) describes an elliptically polarized wave. Assume that \( E'_0 \) has phase \(-\alpha'\) and \( E''_0 \) has phase \(-\alpha^*\) (\( \alpha = \alpha^* - \alpha' \)), the components of the actual electric field, obtained by taking the real part of Eq. (2.18), are given as
\[
\begin{align*}
E_x &= |E'_0| \cos (k \cdot z - \omega t - \alpha') \\
E_y &= |E''_0| \cos (k \cdot z - \omega t - \alpha^*) \\
E_z &= 0
\end{align*}
\] (2.19)

Thus
\[
\frac{E_y}{|E''_0|} = \frac{E_x}{|E'_0|} \cos \alpha + \sin(k \cdot z - \omega t - \alpha') \sin \alpha.
\] (2.20)

From Eq. (2.20), we obtain
Eq. (2.21) is an elliptic equation, shown in Figure 2.3. If $\sin \alpha > 0$, the polarization is left handed; if $\sin \alpha < 0$, it is right handed. The polarization is defined as right-handed when the electric field rotates as a right-handed screw advancing in the direction of propagation.

![Figure 2.3: Elliptic polarization (right-handed).](image)

When $|E'_0| = |E''_0|$ and $\alpha = \pm \pi/2$, the elliptically polarized wave becomes a circularly polarized wave. The left-handed or right-handed polarizations can still be judged by the signs of the phase difference $\alpha$.

To describe polarized waves, Stokes introduced some parameters, which are called the Stokes parameters today. They are

\begin{align*}
I &= |E'_0|^2 + |E''_0|^2 \\
U &= 2|E'_0||E''_0|\cos \alpha \\
Q &= |E'_0|^2 - |E''_0|^2 \\
V &= 2|E'_0||E''_0|\sin \alpha
\end{align*}

with the relation

\[ I^2 = Q^2 + U^2 + V^2. \]

As an example, for a linearly polarized electromagnetic wave ($\alpha = 0$), $U = 2|E'_0||E''_0|$ and $V = 0$. But for a right-handed circularly polarized wave ($|E'_0| = |E''_0|$ and $\alpha = -\pi/2$), $Q = 0$, $U = 0$, and $V = -2|E'_0|^2 = -2|E''_0|^2$.

In radiative transfer theory, it is more common to use the modified Stokes parameters
In the more generous case of a polychromatic wave with a certain bandwidth $\Delta \omega$, the amplitude and the phase difference undergo continuous variations at a rate within $\Delta \omega$. Therefore, the Stokes parameters should be expressed by the average. Denoting the time average by the bracket $\langle \rangle$, we have

\begin{align}
I_1 &= \langle |E_0'|^2 \rangle \\
I_2 &= \langle |E_0''|^2 \rangle \\
U &= 2 \langle |E_0'| |E_0''| \cos \alpha \rangle \\
V &= 2 \langle |E_0'| |E_0''| \sin \alpha \rangle
\end{align}

with the relation

$$I^2 \geq Q^2 + U^2 + V^2.$$ (2.26)

For the modified Stokes parameters $(I_1', I_2', U', V')$, we have $I_1' = \langle |E_0'|^2 \rangle$ and $I_2' = \langle |E_0''|^2 \rangle$.

Natural light is characterized by the facts that the intensity is the same in any direction perpendicular to the propagation direction, and that there is no correlation between rectangular components of the field. Therefore, the necessary and sufficient conditions for light to be natural are

\begin{align}
I &= 2 \langle |E_0'|^2 \rangle = 2 \langle |E_0''|^2 \rangle \\
Q &= U = V = 0
\end{align}

In general, a wave may be partially polarized. The degree of polarization $m$ is defined by the ratio

$$m = (Q^2 + U^2 + V^2)^{1/2} / I,$$ (2.28)

and $m=1$ for elliptic polarization, $0 < m < 1$ for partial polarization, and $m=0$ for an unpolarized wave (the case of natural light).

### 2.4. Reflection and Refraction of Plane Waves

This subsection mainly discusses reflection and refraction of a plane electromagnetic wave at a smooth interface between two media. Consider the interface shown in Figure 2.4, the incident wave, refracted wave and reflected wave are given as:

\begin{align}
\text{INCIDENT} & \quad \begin{cases}
    E = E_0 e^{i(kx - \omega t)} \\
    B = \sqrt{\mu \varepsilon} \frac{k \times E}{k}
\end{cases} \\
\text{REFRACTED} & \quad \begin{cases}
    E' = E_0' e^{i(k'x - \omega t)} \\
    B' = \sqrt{\mu' \varepsilon'} \frac{k' \times E'}{k'}
\end{cases}
\end{align} (2.29)
The existence of the boundary conditions at \( z = 0 \), which must be satisfied at all points on the plane at any time, implies that the spatial (and time) variation of all fields must be the same at \( z = 0 \). Consequently, we must have the phase factors all equal at \( z = 0 \),

\[
(k \cdot x)_{z=0} = (k' \cdot x)_{z=0} = (k'' \cdot x)_{z=0} ,
\]

which gives rise to the Snell's law

\[
\sin \theta = \frac{k'}{k} = \frac{n'}{n} = \frac{\mu' \varepsilon'}{\mu \varepsilon} ,
\]

where \( n' \) and \( n \) are the refraction indexes of the two media, given as

\[
n = \frac{c}{v} = c \sqrt{\mu \varepsilon}
\]

\[
n' = \frac{c}{v'} = c \sqrt{\mu' \varepsilon'}
\]

Here \( c \) is the speed of electromagnetic wave in vacuum.

At the interface between two media, the following conditions are satisfied:

\[
(D - D') \cdot n = K \quad (E - E') \times n = 0
\]

\[
(B - B') \cdot n = 0 \quad (H - H') \times n = L
\]

(2.36)
where $K$ is the surface charge density, and $L$ is the surface current density. For the interface between two dielectric media, we have $K = 0$ and $L = 0$. Therefore, in terms of electric fields in Eqs. (2.29, 2.30, 2.31), the boundary conditions become

$$\left[ \varepsilon (E_0 + E'_0) - \varepsilon' E'_0 \right] \cdot n = 0$$

$$\left[ k \times E_0 + k'' \times E'_0 - k' \times E'_0 \right] \cdot n = 0$$

$$\left[ E_0 + E''_0 - E'_0 \right] \times n = 0$$

(2.37)

In applying these boundary conditions, it is convenient to consider two separate situations: one is that the incident light is linearly polarized with its polarization vector perpendicular to the plane of incidence, and another is that the polarization vector of the linearly polarized incident light is parallel to the plane of incidence. These two situations are shown in Figure 2.5. The reflection and refraction of an elliptically polarized incident light can be obtained by a linear combination of the two fields.

Figure 2.5: Reflection and refraction with electric field (a) perpendicular, and (b) parallel to the plane of incidence.

In the first situation, as shown in Figure 2.5(a), the boundary conditions in Eq. (2.37) give

$$E_0 + E''_0 - E'_0 = 0$$

$$\sqrt{\frac{\varepsilon}{\mu}} (E_0 - E''_0) \cos \theta_i - \sqrt{\frac{\varepsilon'}{\mu'}} E'_0 \cos \theta_r = 0$$

(2.38)

Therefore, we obtain Fresnel equations:
In the second situation, as shown in Figure 2.5(b), the boundary conditions in Eq. (2.37) give
\[(E_0 - E') \cos \theta_i - E' \cos \theta_r = 0\]
\[(\sqrt{\frac{\varepsilon}{\mu}} (E_0 + E') - \sqrt{\frac{\varepsilon'}{\mu'}} E'_0) = 0\]
Therefore, we can obtain the Fresnel equations:
\[t_\parallel = \frac{E'_0}{E_0} = \frac{2}{\sqrt{\frac{\varepsilon}{\mu} \cos \theta_i + \frac{\varepsilon'}{\mu'} \cos \theta_r}} = \frac{2 \cos \theta_i}{\mu' n' \cos \theta_i + n \cos \theta_r},\]
\[r_\parallel = \frac{E''_0}{E_0} = \frac{2}{\sqrt{\frac{\varepsilon}{\mu} \cos \theta_i - \frac{\varepsilon'}{\mu'} \cos \theta_r}} = \frac{\mu' n' \cos \theta_i - n \cos \theta_r}{\mu' n' \cos \theta_i + n \cos \theta_r},\]

For most dielectric media, \(\mu/\mu' \approx 1\). Thus, we can remove the term \(\mu/\mu'\) in Eqs. (2.39) and (2.41)
\[r_\perp = -\frac{\sin(\theta_i - \theta_r)}{\sin(\theta_i + \theta_r)} t_\perp = \frac{2 \cos \theta_i \sin \theta_r}{\sin(\theta_i + \theta_r)}\]
\[r_\parallel = \frac{\tan(\theta_i - \theta_r)}{\tan(\theta_i + \theta_r)} t_\parallel = \frac{2 \cos \theta_i \sin \theta_r}{\sin(\theta_i + \theta_r) \cos(\theta_i - \theta_r)}\]

The reflection coefficients \((R_\perp, R_\parallel)\) and transmission coefficients \((T_\perp, T_\parallel)\) are defined by the ratios of the time-average normal components of the Poynting vectors of the reflected and transmitted waves to the time average of the normal component of the incident wave. They are given as
\[R_\perp = \frac{\langle \mathbf{n} \cdot \mathbf{S}' \rangle}{\langle \mathbf{n} \cdot \mathbf{S} \rangle} = r_\perp^2 T_\perp = \frac{\langle \mathbf{n} \cdot \mathbf{S}' \rangle}{\langle \mathbf{n} \cdot \mathbf{S} \rangle} = \frac{n' \cos \theta'}{n \cos \theta} t_\perp^2\]
\[R_\parallel = \frac{\langle \mathbf{n} \cdot \mathbf{S}' \rangle}{\langle \mathbf{n} \cdot \mathbf{S} \rangle} = r_\parallel^2 T_\parallel = \frac{\langle \mathbf{n} \cdot \mathbf{S}' \rangle}{\langle \mathbf{n} \cdot \mathbf{S} \rangle} = \frac{n' \cos \theta'}{n \cos \theta} t_\parallel^2\]
3 Scattering by Single Particles

An essential problem is how light is absorbed and scattered when it travels inside a translucent medium (which contains randomly distributed particles). To analyze this problem, we follow two steps. First, we consider a single particle and examine its scattering and absorption characteristics. Second, we consider the effects of many particles. In this section, we focus on a single particle, and also we only consider the cases that the scattered light has the same wavelength as the incident light.

Light scattering by a single particle depends on the relative length scales involved — the wavelength of light and the size of the particle. When the wavelength is significantly larger than the dimension of the particle, a simple description in terms of the lowest order induced multipoles is sufficient. When the wavelength and particle size are comparable, a more mathematical treatment with multipole fields is required. In the limit of very small wavelength compared to the particle size, semi-geometrical methods can be used to describe the scattering.

We will first present the concepts of cross-sections, and then the electromagnetic theory about vector and scalar potentials, multipole fields and radiation. Finally, we will present the theories of Rayleigh scattering and MIE scattering. Further information may be found in literature on electromagnetic theory.

3.1. Cross-Sections

Let us consider the case that the particle size is much greater than the wavelength first (see Figure 3.1). When a wave encounters a single particle, a portion of the incident power is scattered out and another portion is absorbed by the particle.

![Figure 3.1: Total cross-section and geometrical cross-section.](image)

If the incident wave has a power flux density $S_i$, the total flux $S_i \sigma_g$ through area $\sigma_g$ is either scattered out or absorbed by the particle. Behind the particle, there should be a region where practically no wave exists. In this shadow region, the scattered wave from the particle is exactly equal to the incident wave but 180° out of phase, and this scattered flux is equal to $S_i \sigma_g$ in
magnitude. Therefore, the total scattered and absorbed flux approaches $S_j \sigma_g + S_j \sigma_g$ and the total cross-section $\sigma_t$ approaches

$$\sigma_t \to 2 S_j \sigma_g / S_j = 2 \sigma_g .$$

(3.1)

We also see that the total absorbed power, when the particle is very large, cannot be greater than $S_j \sigma_g$, and thus the absorption cross-section $\sigma_a$ approaches a constant somewhat less than the geometric cross-section:

$$\sigma_a \to \sigma_g .$$

(3.2)

If the size of a particle is comparable or much smaller than the wavelength, the cross-sections are more complicated. Consider a linearly polarized plane wave propagating in a medium with dielectric constant $\varepsilon_0$ and permeability $\mu_0$. Along the incident direction $n_0$, the incident polarization vector is $e_{inc} (i=1,2)$ and the incident fields are

$$E_{inc} = e_{inc} E_0 e^{i k n_0 \cdot x}$$

$$H_{sc} = \frac{\varepsilon_0}{\mu_0} n_0 \times E_{inc}$$

(3.3)

where $k = \omega / c$ and a time-dependence $e^{i \omega t}$ is omitted.

As shown in Figure 3.2, the scatter is a particle with the permittivity $\varepsilon$ and the permeability $\mu$, and the scattered electric field is $E_{sc}$ with distance $r$ far away from the particle and the direction $n$. The differential scattering cross-section (area per unit solid angle) is defined as the ratio of the power flux radiated in the direction $n$ per unit solid angle to the incident flux density:

$$\frac{d\sigma_s}{d\Omega} = \frac{1}{2} \left[ \frac{\varepsilon_0}{\mu_0} |e_{inc} \cdot E_{inc}|^2 \right]$$

(3.4)
where \( \Omega \) is the solid angle and \( e_{sc} \) is the polarization vector of the scattered electric field.

The differential scattering-cross has the relation with the total cross-section \( \sigma \) as

\[
\frac{d\sigma}{d\Omega} = \frac{\sigma_s}{4\pi} p(n, n_o) .
\]

(3.5)

The dimensionless quantity \( p(n, n_o) \) is called phase function. The scattering cross-section is given by

\[
\sigma_s = \frac{\sigma_s}{4\pi} \int_0^{2\pi} p(n, n_o) d\Omega .
\]

(3.6)

The ratio of the scattering cross-section \( \sigma_s \) to the total cross-section \( \sigma \), is called the albedo of a single particle:

\[
W_0 = \frac{\sigma_s}{\sigma} = \frac{1}{4\pi} \int_0^{2\pi} p(n, n_o) d\Omega .
\]

(3.7)

According to the conservation of energy, the absorption cross-section \( \sigma_a \) satisfies

\[
\sigma_a = \sigma - \sigma_s .
\]

(3.8)

### 3.2. Vector and Scalar Potentials

Maxwell equations consist of a set of coupled first-order partial differential equations. These equations can be solved in simple situations. In general cases, it is often helpful to introduce potentials to obtain a smaller number of second-order equations while satisfying some of Maxwell equations identically. Now we introduce a scalar potential \( \Phi \) and the magnetic potential \( A \). Since \( \nabla \cdot B = 0 \), we define

\[
B = \nabla \times A .
\]

(3.9)

Therefore, Eq. (2.1d) can be written as

\[
\nabla \times \left( E + \frac{\partial A}{\partial t} \right) = 0 .
\]

(3.10)

So we can introduce a scalar potential \( \Phi \),

\[
E + \frac{\partial A}{\partial t} = -\nabla \Phi
\]

or

\[
E = -\frac{\partial A}{\partial t} - \nabla \Phi .
\]

(3.11)

Substituting Eqs. (3.11) and (3.9) into Eq. (2.1), we obtain

\[
\nabla^2 \Phi + \frac{\partial}{\partial t} (\nabla \cdot A) = -\rho / \varepsilon
\]

(3.12)

\[
\nabla^2 A - \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} - \nabla \left[ \nabla \cdot A + \frac{1}{c^2} \frac{\partial \Phi}{\partial t} \right] = -\mu J
\]

(3.13)

These two equations are still coupled. Since \( A \) is introduced by Eq. (3.9), \( B \) is unchanged if we make the transformation...
\[ \mathbf{A} \rightarrow \mathbf{A}' = \mathbf{A} + \Delta \mathbf{A} \, . \]  
(3.14)

For \( \mathbf{E} \) to remain the same, the scalar potential must be simultaneously transformed
\[ \Phi \rightarrow \Phi' = \Phi - \frac{\partial \Delta}{\partial t} \, . \]  
(3.15)

Consider the freedom presented by Eqs. (3.14) and (3.15), we may choose a set of potentials \((\mathbf{A}, \Phi)\) such that
\[ \nabla \cdot \mathbf{A} + \frac{1}{c^2} \frac{\partial \Phi}{\partial t} = 0 \, . \]  
(3.16)

Eq. (3.16) will uncouple the pair of Eqs. (3.12) and (3.13), and lead to two inhomogeneous wave equations,
\[ \nabla^2 \Phi - \frac{1}{c^2} \frac{\partial^2 \Phi}{\partial t^2} = -\rho/\varepsilon \]  
(3.17)
\[ \nabla^2 \mathbf{A} - \frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} = -\mu \mathbf{J} \]  
(3.18)

The transformation of Eqs. (3.14) and (3.15) is called \textit{gauge transformation}, and the invariance of the fields under such transformation is called \textit{gauge invariance}. The relation between \( \mathbf{A} \) and \( \phi \) in Eq. (3.16) is called \textit{Lorentz condition}.

The wave equations in Eqs. (3.17) and (3.18) have the following pattern:
\[ \nabla^2 \psi - \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} = -4\pi f(x, t) \, , \]  
(3.19)
where \( f(x, t) \) is a known source distribution. To solve Eq. (3.19), it is useful to find a Green function. The corresponding time-dependent Green function satisfies
\[ \left( \nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) G(x, t; x', t') = -4\pi \delta(x - x') \delta(t - t') \, . \]  
(3.20)

The solutions of Eq. (3.20) are given as
\[ G^\pm(x, t; x', t') = \delta \left( t' - \left[ t' \mp \frac{|x - x'|}{c} \right] \right) \, , \]  
(3.21)
where the Green function \( G^+ \) is called the \textit{retarded Green function}, and \( G^- \) is called the \textit{advanced Green function}.

Using Green functions, we can give the solution to Eq. (3.19),
\[ \psi^\pm(x, t) = \int \int G^\pm(x, t; x', t') f(x', t') d^3 x' dt' \, . \]  
(3.22)

Consider a source distribution \( f(x, t) \) that is localized in time and space, and is different from zero only for a finite interval of time around \( t' = 0 \). Two limiting situations are envisioned. In the first it is assumed that at time \( t \rightarrow -\infty \), there exists a wave \( \psi_{in}(x, t) \) that satisfies the homogeneous wave equation. The complete solution for this situation at all times is given as
\[ \psi(x, t) = \psi_{in}(x, t) + \int \int G^+(x, t; x', t') f(x', t) d^3 x' dt' \, . \]  
(3.23)
The second case is that the wave is given as \( \psi_{\text{out}}(x,t) \) at remotely late times \( t \to +\infty \). The complete solution is given as
\[
\psi(x,t) = \psi_{\text{out}}(x,t) + \int \int G^-(x,t;x',t')f(x',t)d^3x'dt'.
\] (3.24)

Following Eq. (3.23), we can obtain the vector potential \( A(x,t) \) in the Lorentz gauge as
\[
A(x,t) = \frac{\mu}{4\pi} \int d^3x' \int dt' \frac{J(x',t')}{|x-x'|} \delta \left( t' + \frac{|x-x'|}{c} - t \right)
\] (3.25)
provided no boundary surfaces are present. Similarly, the scalar potential \( \Phi(x,t) \) is
\[
\Phi(x,t) = \int d^3x' \int dt' \frac{\rho(x',t')}{|x-x'|} \delta \left( t' + \frac{|x-x'|}{c} - t \right).
\] (3.26)

### 3.3. Multipole Fields and Radiation

For a system of charges and currents varying in time, we can perform a Fourier analysis of the time dependence and handle each Fourier component separately. We therefore lose no generality by considering the potentials, fields, and radiation from a localized system of charges and currents that vary sinusoidally in time:
\[
\rho(x,t) = \rho(x)e^{-i\omega t},
\]
\[
J(x,t) = J(x)e^{-i\omega t}
\] (3.27)

As usual, the real part of such expressions is to be taken to obtain physical quantities. The electromagnetic potentials and fields are assumed to have the same time dependence.

With the sinusoidal time dependence, the solution for \( A \) becomes
\[
A(x) = \frac{\mu}{4\pi} \int J(x')e^{i|\mathbf{k}|x-x'}d^3x',
\] (3.28)
where a sinusoidal time dependence is not shown explicitly.

Given a current distribution \( J(x') \), if the source dimensions are of order \( d \) and the wavelength is \( \lambda = 2\pi c/\omega \), and if \( d \ll \lambda \), then generally there are three spatial regions of interest:
- The near (static) zone: \( d \ll r \ll \lambda \)
- The intermediate (induction) zone: \( d \ll r \sim \lambda \)
- The far (radiation) zone: \( d \ll \lambda \ll r \)

In this section, we are only interested in the scattered electromagnetic field far away from the particle. Therefore, we only discuss in the far zone. In this region, it is sufficient to approximate
\[
|x-x'| = r - \mathbf{n} \cdot \mathbf{x}'.
\] (3.29)

Therefore, the vector potential becomes
\[
\lim_{kr \to \infty} A(x) = \frac{\mu}{4\pi} \frac{e^{ikr}}{r} \int J(x')e^{-i\omega n \cdot x'}d^3x'.
\] (3.30)
This demonstrates that in the far zone the vector potential behaves as an outgoing spherical wave with an angular-dependent coefficient. If the source dimensions are small compared to the wavelength, it is helpful to expand the integral in Eq. (3.30) in powers of \( k \):

\[
\lim_{kr \to \infty} A(x) = \frac{\mu}{4\pi} \frac{e^{ikr}}{r} \sum_{n} \frac{(-ik)^n}{n!} \int J(x')(n \cdot x')^n d^3x'
\]

(3.31)

where the magnitude of the \( n \)th term is given by

\[
\frac{1}{n!} \int J(x')(kn \cdot x')^n d^3x'.
\]

(3.32)

Since the order of the magnitude of \( x' \) is \( d \) and \( kd \) is small compared to unity by assumption, the successive terms in the expansion of \( A \) evidently fall off rapidly with \( n \). Consider the first term \( (n = 0) \), the vector potential is

\[
A_0(x) = \frac{\mu}{4\pi} \frac{e^{ikr}}{r} \int J(x')d^3x'.
\]

(3.33)

After performing partial integration in Eq. (3.33), we can obtain

\[
\int Jd^3x' = -\int x'(\nabla \cdot J)d^3x'.
\]

(3.34)

Consider the continuity equation

\[ \nabla \cdot J = -\frac{\partial \rho}{\partial t} = i\omega \rho , \]

(3.35)

Eq. (3.33) becomes

\[
A_1(x) = -\frac{i\mu \omega}{4\pi} \frac{e^{ikr}}{r} \frac{p}{r},
\]

(3.36)

where

\[
p = \int x'\rho(x')d^3x'.
\]

(3.37)

is the electric dipole moment.

The electric dipole fields are given as

\[
H_1 = \frac{ck^2}{4\pi} (n \times p) \frac{e^{ikr}}{r} \left( \frac{1}{r} - \frac{1}{ikr} \right)
\]

(3.38)

\[
E_1 = \frac{1}{4\pi\varepsilon} \left\{ k^2 (n \times p) \times n \frac{e^{ikr}}{r} \left[ 3n(n \cdot p) - p \right] \left( \frac{1}{r^3} - \frac{ik}{r^2} \right) e^{ikr} \right\}
\]

In the radiation zone, the fields take the limit forms

\[
H_1 = \frac{ck^2}{4\pi} (n \times p) \frac{e^{ikr}}{r}
\]

(3.39)

\[
E_1 = \frac{\mu}{\varepsilon} H \times n
\]

Consider the second term \( (n = 1) \), the vector potential is

\[
A_2(x) = \frac{\mu}{4\pi} \frac{e^{ikr}}{r} \left( \frac{1}{r} - ik \right) \int J(x')(n \cdot x')d^3x'.
\]

(3.40)

Here the correct terms (refer to equation 9.12 in Jackson99) have been included so that the expression is valid everywhere outside the source. This vector potential can be written as the sum
of two terms: one gives a transverse magnetic induction and the other gives a transverse electric field. Thus
\[
(n \cdot x')J = \frac{1}{2}[(n \cdot x')J + (n \cdot J)x'] + \frac{1}{2}(x' \times J) \times n.
\] (3.41)
The second, anti-symmetrical part gives rise to a magnetic dipole moment:
\[
m = \frac{1}{2} \int (x \times J) d^3 x.
\] (3.42)
Therefore the vector potential caused from the magnetic dipole moment is given as
\[
A_2'(x) = \frac{ik\mu}{4\pi}(n \times m) \frac{e^{ikr}}{r} (1 - \frac{1}{ikr}).
\] (3.43)
Correspondently,
\[
H_2' = \frac{1}{4\pi} \left\{ k^2(n \times m) \times n \frac{e^{ikr}}{r} + [3n(n \cdot m) - m] \left( \frac{1}{r^3} - \frac{ik}{r^2} \right) e^{ikr} \right\}
\] (3.44)
\[
E_2' = -\frac{k^2}{4\pi} \sqrt{\frac{\mu}{\varepsilon}} (n \times m) \frac{e^{ikr}}{r} \left( 1 - \frac{1}{ikr} \right).
\]
The integral term of the symmetrical term in Eq. (3.41) can be transformed by an integration by parts and some rearrangement:
\[
\frac{1}{2} \int [(n \cdot x')J + (n \cdot J)x'] d^3 x' = -\frac{i\omega}{2} \int x'(n \cdot x') \rho(x') d^3 x'.
\] (3.45)
Then the vector potential is
\[
A_2''(x) = -\frac{\mu ck^2}{8\pi} \frac{e^{ikr}}{r} \left( 1 - \frac{1}{ikr} \right) \int x'(n \cdot x') \rho(x') d^3 x'.
\] (3.46)
Consequently the magnetic field is
\[
H_2'' = -\frac{i ck^3}{8\pi} \frac{e^{ikr}}{r} \int (n \times x')(n \cdot x') \rho(x') d^3 x'.
\] (3.47)
With the definition of quadrupole moment tensor,
\[
Q_{ab} = \int (3x_a x_b - r^2 \delta_{ab}) \rho(x) d^3 x,
\] (3.48)
the integral in Eq. (3.47) can be written as
\[
n \times \int x'(n \cdot x') \rho(x') d^3 x' = \frac{1}{2} n \times Q(n),
\] (3.49)
where the vector \(Q(n)\) is defined as having the components
\[
Q_a = \sum_b Q_{ab} n_\beta.
\] (3.50)
With these definitions we have the fields
\[
H_2'' = \frac{i ck^3}{24\pi} \frac{e^{ikr}}{r} n \times Q(n)
\] (3.51)
\[
E_2'' = \frac{i}{k} \sqrt{\frac{\mu}{\varepsilon}} \nabla \times H_2''
\]
3.4. Rayleigh Scattering

3.4.1 Scattering by Dipoles Induced in Small Scatters
In practice, a common case is that the scattering objects have dimensions smaller than the wavelength. In such interaction, it is convenient to view the incident (radiation) fields as inducing electric and magnetic multipoles. They oscillate in definite phase relationship with the incident wave and radiate energy in directions other than the direction of incidence. If the wavelength of the radiation is long compared to the size of the scatter, only the lowest multipoles, usually the electric and magnetic dipoles are important.

For the case shown in Figure 3.2, the scattered (radiated) fields, which are far away from the particle, can be found from Eq. (3.38) and Eq. (3.44),

\[ E_{sc} = \frac{1}{4\pi\varepsilon_0} k^2 e^{ikr} \left[ (n \times p) \times n - n \times m/c \right] \]

\[ H_{sc} = \frac{\varepsilon_0}{\mu_0} n_0 \times E_{inc} \]  

(3.52)

Substituting Eq. (3.52) into Eq. (3.4), we obtain

\[ \frac{d\sigma_s}{d\Omega} = \frac{k^4}{(4\pi\varepsilon_0 E_0)^2} \left| \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \right| e_{sc} \cdot p + (n \times e_{sc}) \cdot m/c \right|^2 . \]

(3.53)

The variation of the differential (and total) scattering cross-section with wave number as \( k^4 \) (or in wavelength as \( \lambda^{-4} \)) is an almost universal characteristics of the scattering of long-wavelength radiation by any finite system. This dependence on frequency is known as Rayleigh's law. Only if both static dipole moments vanish does the scattering fail to obey Rayleigh's law; the scattering then behaves as a quadrapole or higher multipoles (or frequency-dependent dipole moments), and the scattered energy varies as \( \omega^6 \) or even higher power of \( \omega \). Sometimes the dipole scattering is known as Rayleigh scattering, but this term is usually reserved for the incoherent scattering by a collection of dipole scatters.

3.4.2. Scattering by a Small Dielectric Sphere
A simple example of dipole scattering is a small dielectric sphere of radius \( a \) with \( \mu = \mu_0 \) and a uniform isotropic dielectric constant \( \varepsilon_r(\omega) \) \( (\varepsilon = \varepsilon_r \varepsilon_0) \). The electric dipole moment is given as

\[ p = 4\pi\varepsilon_0 \left( \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \right) a^3 E_{inc} . \]

(3.54)

There is no magnetic dipole moment. The differential scattering cross-section is

\[ \frac{d\sigma_s}{d\Omega} = k^4 a^6 \left| \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \right| e_{sc} \cdot e_{inc} \right|^2 . \]

(3.55)

The polarization dependence is typical of purely electric dipole scattering.
Typically the incident light is unpolarized. The scattering plane is defined by the vectors \( \mathbf{n}_0 \) and \( \mathbf{n} \). The differential scattering cross-sections for the polarizations of the incident and scattered lights parallel and perpendicular to the scattering plane are given as

\[
\frac{d\sigma_{s,\parallel}}{d\Omega} = \frac{k^4 a^6}{2} \left| \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \right|^2 \cos^2 \theta
\]

\[
\frac{d\sigma_{s,\perp}}{d\Omega} = \frac{k^4 a^6}{2} \left| \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \right|^2
\]

(3.56)

where the subscripts \( \parallel \) and \( \perp \) indicate the polarization parallel and perpendicular to the scattering plane, respectively, and \( \theta \) is the angle between \( \mathbf{n}_0 \) and \( \mathbf{n} \).

The polarization of the scattered radiation is defined by

\[
\Pi(\theta) = \frac{d\sigma_{s,\perp}/d\Omega - d\sigma_{s,\parallel}/d\Omega}{d\sigma_{s,\perp}/d\Omega + d\sigma_{s,\parallel}/d\Omega}
\]

(3.57)

From Eq. (3.56), we find the polarization for the (electric dipole) scattering by a small dielectric sphere as

\[
\Pi(\theta) = \frac{\sin^2 \theta}{1 + \cos^2 \theta}
\]

(3.58)

The differential cross-section, summed over all scattered polarization, is

\[
\frac{d\sigma}{d\Omega} = \frac{k^4 a^6}{2} \left| \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \right|^2 \frac{1}{2} (1 + \cos^2 \theta)
\]

(3.59)

and the total scattering cross-section is

\[
\sigma_s = \int d\Omega \frac{d\sigma}{d\Omega} = \frac{8\pi}{3} k^4 a^6 \left| \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \right|^2.
\]

(3.60)

From Eq. (3.56), we can see that \( d\sigma_{s,\parallel}/d\Omega = 0 \) and \( \Pi(\theta) = 1 \) when \( \theta = \pi/2 \). At this angle, the scattered radiation is 100% linearly polarized perpendicular to the scattering plane, and for an appreciable range of angles around \( \theta = \pi/2 \) is quite significantly polarized. The polarization characteristics of the blue sky are an example of this phenomenon. (In fact, the blue sky was the motivation that led Rayleigh to investigate the problem.) From Eq. (3.60), we can see that the scattering is strongest when \( \theta = 0, \pi \). Both the differential scattering cross-section and the polarization versus the angle \( \theta \) are shown in Figure 3.3.
3.4.3. Scattering by a Small Perfectly Conducting Sphere

An example with interesting aspects involving coherence between different multipoles is the scattering by a small perfectly conducting sphere of radius $a$. The electric dipole moment of such a sphere is

$$ p = 4\pi\varepsilon_0 a^3 E_{\text{inc}} \quad \text{(3.61)} $$

and the magnetic moment is

$$ m = -2\pi a^3 H_{\text{inc}} \quad \text{(3.62)} $$

For a linearly polarized incident wave, the two dipoles are at right angles to each other and to the incident direction.

Substituting Eqs. (3.61) and (3.62) into Eq. (3.53), we obtain

$$ \frac{d\sigma}{d\Omega} = k^4 a^6 \left| e_{\text{sc}} \cdot e_{\text{inc}} - \frac{1}{2} (n \times e_{\text{sc}}) \cdot (n_0 \times e_{\text{inc}}) \right|^2. \quad \text{(3.63)} $$

The differential cross-sections for polarization of the incident and scattered lights parallel and perpendicular to scattering plane are

$$ \frac{d\sigma_{\parallel}}{d\Omega} = \frac{k^4 a^6}{2} \left| \cos \theta - \frac{1}{2} \right|^2 \quad \text{(3.64)} $$

$$ \frac{d\sigma_{\perp}}{d\Omega} = \frac{k^4 a^6}{2} \left| -\frac{1}{2} \cos \theta \right|^2 $$

The polarization $\Pi(\theta)$ is given as

$$ \Pi(\theta) = \frac{3\sin^2 \theta}{5(1 + \cos^2 \theta) - 8\cos \theta}. \quad \text{(3.65)} $$

The differential cross-section summed over both states of scattered polarization can be written

$$ \frac{d\sigma}{d\Omega} = k^4 a^6 \left[ \frac{5}{8} (1 + \cos^2 \theta) - \cos \theta \right]. \quad \text{(3.66)} $$
Both the differential cross-section and the polarization versus the angle $\theta$ are shown in Figure 3.4.

![Figure 3.4: Differential scattering cross section and the polarization of scattered radiation for a small perfectly conducting sphere (electric and magnetic dipole approximation).](image)

From Eq. (3.64) and Figure 3.4, we can see that the scattered light is 100% linearly polarized perpendicular to the scattering plane when $\theta = \pi/3$, and is positive through the whole angular range. From Figure 3.4, we can find that the scattering has a strong backward peak caused by the electric dipole and magnetic dipole interference.

### 3.4.4. Born Approximation

If the medium is uniform in its properties, the wave propagates undisturbed and undeflected. If there are spatial (or temporal) variations in the medium, the wave is scattered. From Eq. (2.1), we can obtain a wave equation for $D$ as

$$\nabla^2 D - \mu_0 \epsilon_0 \frac{\partial^2 D}{\partial t^2} = -\nabla \times \nabla \times (D - \epsilon_0 E) + \epsilon_0 \frac{\partial}{\partial t} \nabla \times (B - \mu_0 H). \quad (3.67)$$

Following Eq. (3.23) and with a time dependence $e^{-i\omega t}$ left out, we obtain

$$D = D^{(0)} + \frac{1}{4\pi} \int d^3 x' \frac{e^{i|\mathbf{x} - \mathbf{x}'|}}{|\mathbf{x} - \mathbf{x}'|} \left\{ \nabla' \times \nabla' \times (D - \epsilon_0 E) + i\epsilon_0 \omega \nabla' \times (B - \mu_0 H) \right\}. \quad (3.68)$$

If $D^{(0)}$ describes a wave incident in some direction, the field far away from the scattering region can be written as

$$D \rightarrow D^{(0)} + A_{\infty} \frac{e^{i\theta}}{r}. \quad (3.69)$$

Then we obtain
Correspondently, the differential scattering cross-section can be obtained as

\[ \frac{d\sigma_s}{d\Omega} = \left| \mathbf{e}_{sc} \cdot \mathbf{A}_{sc} \right|^2, \]

where \( \mathbf{e}_{sc} \) is the polarization vector of the scattered radiation.

We now consider the scattering characteristics of a scatter whose permittivity \( \varepsilon \) close to \( \varepsilon_0 \), and permeability \( \mu \) close to \( \mu_0 \). Thus, we assume that the connections between \( \mathbf{D} \) and \( \mathbf{E} \), and \( \mathbf{B} \) and \( \mathbf{H} \) are given as

\[ \mathbf{D}(x) = [\varepsilon_0 + \delta \varepsilon(x)]\mathbf{E}(x), \]
\[ \mathbf{B}(x) = [\mu_0 + \delta \mu(x)]\mathbf{H}(x) \]

where \( \delta \varepsilon(x) \) and \( \delta \mu(x) \) are small in magnitude compared with \( \varepsilon_0 \) and \( \mu_0 \).

Consider the first-order approximation (Born approximation):

\[ \mathbf{D}(x) = \varepsilon_0 \mathbf{E}(x) + \delta \varepsilon(x) \mathbf{E}_0(x) = \varepsilon_0 \mathbf{E}(x) + \delta \varepsilon(x) \mathbf{D}^{(0)}(x) \]
\[ \mathbf{B}(x) = \mu_0 \mathbf{H}(x) + \delta \mu(x) \mathbf{H}_0(x) = \mu_0 \mathbf{H}(x) + \delta \mu(x) \mathbf{B}^{(0)}(x) \]

If the unperturbed fields are those of a plane wave propagating in a direction \( \mathbf{n}_0 \), then \( \mathbf{D}^{(0)} \) and \( \mathbf{B}^{(0)} \) are

\[ \mathbf{D}^{(0)}(x) = \mathbf{e}_{inc} \mathbf{D}_0 e^{i\mathbf{n}_0 \cdot \mathbf{x}} \]
\[ \mathbf{B}^{(0)}(x) = \mathbf{\mu}_0 \mathbf{n}_0 \times \mathbf{D}^{(0)}(x) \]

Following Eq. (3.70), we obtain

\[ \frac{\mathbf{e}_{sc} \cdot \mathbf{A}_{sc}}{D_0} = \frac{k^2}{4\pi} \int d^3x e^{i\mathbf{q} \cdot \mathbf{x}} \left\{ \mathbf{e}_{sc} \cdot \mathbf{e}_{inc} \frac{\delta \varepsilon(x)}{\varepsilon_0} + (\mathbf{n} \times \mathbf{e}_{sc}) \cdot (\mathbf{n}_0 \times \mathbf{e}_{inc}) \frac{\delta \mu(x)}{\mu_0} \right\}, \]

where \( \mathbf{q} = k(\mathbf{n}_0 - \mathbf{n}) \) is the difference between the incident and scattered wave vectors. The absolute square of Eq. (3.75) gives the differential scattering cross-section.

If the wavelength is large compared with the spatial extent of \( \delta \varepsilon \) and \( \delta \mu \), the exponential in Eq. (3.75) can be set equal to unity. Consider the scattering region is a uniform dielectric sphere of radius \( a \) in vacuum, the integral in Eq. (3.75) can be performed for arbitrary \( |\mathbf{q}| \), with the result

\[ \frac{\mathbf{e}_{sc} \cdot \mathbf{A}_{sc}}{D_0} = k^2 \left( \frac{\varepsilon_{sc} \cdot \varepsilon_{inc}}{\varepsilon_0} \right) \left[ \frac{\sin(qa) - qa \cos(qa)}{q^2} \right]. \]

In the limit \( a \rightarrow 0 \), the Born approximation to the differential cross-section for scattering by dielectric sphere of radius \( a \) is
Comparison with Eq. (3.55), we can see that the Born approximation and the exact frequency result have the expected relationship.

### 3.5. Mie Scattering

Section 3.4 mainly works for the condition that the wavelength of incident wave is much larger than the size of scatter. Now we will present an exact solution of the scattering of the plane electromagnetic wave by an isotropic, homogeneous sphere, which works for both the wavelength much larger than and comparable to the size of scatter. This solution was obtained by Mie in 1908 and today is commonly referred to as Mie scattering. The scattering of incident wave by a sphere is shown in Figure 3.5.

![Figure 3.5: Scattering of radiation by a sphere.](image)

**3.5.1 The Hertz Vectors**

In order to present the Mie theory, we first introduce the Hertz vector. Let us confine ourselves for the present to a region of an isotropic homogeneous medium, within which there are neither conduction currents nor free charges, i.e., \( J = 0 \) and \( \rho = 0 \). The vector potential \( \mathbf{A} \) is assumed to be proportional to the time derivative of the Hertz vector \( \mathbf{\pi} \),

\[
\mathbf{A} = \mu e \frac{\partial \mathbf{\pi}}{\partial t}.
\]

Consequently,

\[
\mathbf{B} = \mu e \nabla \times \frac{\partial \mathbf{\pi}}{\partial t},
\]

\[
\mathbf{E} = \nabla \Phi - \mu e \frac{\partial^2 \mathbf{\pi}}{\partial t^2}.
\]

Substituting Eq. (3.79) into (2.1b), we obtain

\[
\frac{\partial}{\partial t} \left( \nabla \times \nabla \times \mathbf{\pi} + \nabla \Phi + \mu e \frac{\partial^2 \mathbf{\pi}}{\partial t^2} \right) = 0.
\]
Since $\rho = 0$, the scalar function $\Phi$ is arbitrary as long as it satisfies Eq. (3.17). In the present instance, it will be enough such that

$$\Phi = -\nabla \cdot \pi.$$  \hfill (3.81)

Then upon integrating Eq. (3.80) with respect to the time, we obtain

$$\nabla \times \nabla \times \pi - \nabla \nabla \cdot \pi + \mu \varepsilon \frac{\partial^2 \pi}{\partial t^2} = \text{constant}.$$  \hfill (3.82)

The particular value of the constant does not affect the determination of the field and we are therefore free to place it equal to zero. Eq. (2.1a) is also satisfied since the divergence of the curl of any vector vanishes identically. Then, we may state that every solution of the vector equation

$$\nabla \times \nabla \times \pi - \nabla \nabla \cdot \pi + \mu \varepsilon \frac{\partial^2 \pi}{\partial t^2} = 0$$  \hfill (3.83)

determines an electromagnetic field through

$$\mathbf{B} = \mu \varepsilon \nabla \times \nabla \frac{\partial \pi}{\partial t}$$  \hfill (3.84)

$$\mathbf{E} = \nabla \nabla \cdot \pi - \mu \varepsilon \frac{\partial^2 \pi}{\partial t^2}.$$

Eq. (3.83) can be transformed into

$$\nabla^2 \pi - \mu \varepsilon \frac{\partial^2 \pi}{\partial t^2} = 0.$$  \hfill (3.85)

Since the vector $\mathbf{D}$ and $\mathbf{B}$ are solenoidal in a charge free region, an alternative solution can be constructed of the form

$$\mathbf{A}' = \mu \varepsilon \frac{\partial \pi'}{\partial t} \quad \Phi' = -\nabla \cdot \pi'$$  \hfill (3.86)

$$\mathbf{D}' = -\mu \varepsilon \nabla \times \nabla \frac{\partial \pi'}{\partial t} \quad \mathbf{H}' = \nabla \nabla \cdot \pi' - \mu \varepsilon \frac{\partial^2 \pi'}{\partial t^2}.$$  \hfill (3.87)

where $\pi'$ is any solution of Eq. (3.83) or (3.85).

From these results we conclude that the electromagnetic field within a region throughout which $\varepsilon$ and $\mu$ are constant, $\mathbf{J} = 0$ and $\rho = 0$, may be resolved into two partial fields, the one derived from the vector $\pi$, and another from the vector $\pi'$. To determine the physical significance of the Hertz vectors, it is necessary to relate them to the sources.

Let us express the vector $\mathbf{D}$ in terms of $\mathbf{E}$ and the electric polarization $\mathbf{P}$

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}.$$  \hfill (3.88)

Then Eqs. (2.1a) and (2.1b) for $\mathbf{J} = 0$ and $\rho = 0$ can be transformed into

$$\nabla \cdot \mathbf{E} = -\frac{1}{\varepsilon_0} \nabla \cdot \mathbf{P}$$  \hfill (3.89)

$$\nabla \times \mathbf{H} - \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \frac{\partial \mathbf{P}}{\partial t}.$$  \hfill (3.90)

We can find that Eq. (3.84) still satisfies Eq. (3.89) provided that $\varepsilon$ be replaced by $\varepsilon_0$ and
Similarly, if we express \( B \) as
\[
B = \mu_0 (H + M)
\]
then Eqs. (2.1c) and (2.1d) can be transformed into
\[
\nabla \cdot H = -\nabla \cdot M
\]
\[
\nabla \times \mathbf{E} + \mu_0 \frac{\partial \mathbf{H}}{\partial t} = -\mu_0 \frac{\partial \mathbf{M}}{\partial t}
\]
These equations can be satisfied by Eq. (3.87) provided that \( \mu \) be replaced by \( \mu_0 \) and
\[
\nabla^2 \pi' - \mu_0 \varepsilon \frac{\partial^2 \pi'}{\partial t^2} = -\mathbf{M}
\]
We can see that the Hertz vectors come from two sources; \( \pi \) is from the electric dipole, and \( \pi' \) from the magnetic dipole. The total fields should be the addition of the contributions of these two vectors. Therefore, we can obtain
\[
\mathbf{E}_{\text{total}} = \nabla \nabla \cdot \pi - \mu_0 \varepsilon \frac{\partial^2 \pi}{\partial t^2} - \mu \nabla \times \frac{\partial \pi'}{\partial t}
\]
\[
\mathbf{H}_{\text{total}} = \varepsilon \nabla \times \frac{\partial \pi}{\partial t} + \nabla \nabla \cdot \pi' - \mu_0 \varepsilon \frac{\partial^2 \pi'}{\partial t^2}
\]
### 3.5.2. Exact Solution of Scattering by a Sphere

The field which comes from \( \pi \), is called the electric wave or the transverse magnetic wave (TM). In the spherical coordinates, it is characterized by a zero value of the radial components of its magnetic field intensity \( H' = 0 \). For the wave which comes from \( \pi' \), it is called the magnetic wave or transverse electric wave (TE), and in the spherical coordinates, \( E' = 0 \). So the electromagnetic wave can be conceived arising from a distribution in space of oscillating electric dipoles (TM) and of oscillating magnetic dipoles (TE), or alternatively from the super-position of oscillating electric and magnetic multipoles located in the origin.

The Hertz vectors \( \pi \) and \( \pi' \) can be derived in a simple way from corresponding scalar potential functions known either as Hertz potential or as Debye potentials, depending precisely upon how the original vector quantity has been defined
\[
-\nabla \cdot \pi = \pi
\]
\[
-\nabla \cdot \pi' = \pi'
\]
The Hertz-Debye potentials \( \pi \) and \( \pi' \) are solutions of the scalar wave equation
\[
\nabla^2 u - \mu_0 \varepsilon \frac{\partial^2 u}{\partial t^2} = 0,
\]
where \( u = \pi, \pi' \).

For sinusoidal time dependence \( e^{i\omega t} \), the non-homogeneous scalar wave equation reduces to the homogeneous form
\[ \nabla^2 u' + k^2 u' = 0 \]  \hspace{1cm} (3.97)

where \( u = u' e^{i\omega t} \) and \( k = \omega \sqrt{\mu \varepsilon} \).

In spherical coordinates, Eq. (3.97) becomes
\[
\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial (ru')}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial (ru')}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u'}{\partial \phi^2} + k^2 u' = 0 . \hspace{1cm} (3.98)
\]

The solution of Eq. (3.98) is given as
\[
ru' = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \left\{ c_l \psi_l(\psi kr) + d_l \chi_l(\chi kr) \right\} \left\{ P_l^{(m)}(\cos \theta) \right\} \left\{ a_m \cos(m\phi) + b_m \sin(m\phi) \right\} , \hspace{1cm} (3.99)
\]
where \( l \) is integral and \( m = -l, \ldots, 0, \ldots, +l \); \( \psi_l(\psi kr) \) and \( \chi_l(\chi kr) \) are the Ricatti-Bessel functions defined as
\[
\psi_l(\psi kr) = (\pi kr/2)^{1/2} J_{l+1/2}(kr) \hspace{1cm} (3.100)
\]
\[
\chi_l(\chi kr) = -(-\pi kr/2)^{1/2} N_{l+1/2}(kr)
\]
where \( J_{l+1/2}(kr) \) and \( N_{l+1/2}(kr) \) are the half integral order Bessel and Neumann functions; \( P_l^{(m)}(\cos \theta) \) is Legendre polynomial; \( a_m \) and \( b_m \) are the coefficients.

The Bessel and Neumann functions are given as
\[
J_v(x) = \left( \frac{x}{2} \right)^v \sum_{j=0}^{\infty} \frac{(-1)^j}{j! \Gamma(j + 1)} \left( \frac{x}{2} \right)^{2j} \hspace{1cm} (3.101)
\]
\[
N_v(x) = J_v(x) \cos(v\pi) - J_{-v}(x) \sin(v\pi)
\]
and the Legendre polynomial is given as
\[
P_l^{(m)}(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_l(x) \hspace{1cm} (3.102)
\]
where
\[
P_l(x) = \frac{1}{2^l l!} \frac{d^l}{dx^l} (x^2 - 1)^l . \hspace{1cm} (3.103)
\]

We assume that the Hertz-Debye potentials for the incident wave are \( \pi_i \) and \( \pi'_i \), those for the wave inside the particle, \( \pi_r \) and \( \pi'_r \), and those for the scattered wave \( \pi_s \) and \( \pi'_s \). We will ignore the time dependence \( e^{i\omega t} \) in the following discussion. The wave number for the wave outside of the particle is defined as \( k \), and that inside of the particle \( k_r \). The relation between \( k \) and \( k_r \) can be obtained by Eq. (2.34). The incident plane wave can be expanded in the form of Eq. (3.99) as
\[
r \pi_i = \frac{1}{k^2} \sum_{l=1}^{\infty} i^{n-1} \frac{2l+1}{l(l+1)} \psi_l(\psi kr) P_l^{(i)}(\cos \theta) \cos \phi \hspace{1cm} (3.104)
\]
\[
r \pi'_i = \frac{1}{k^2} \sum_{l=1}^{\infty} i^{n-1} \frac{2l+1}{l(l+1)} \psi_l(\psi kr) P_l^{(i)}(\cos \theta) \sin \phi \]
The scattered wave can be expressed as

\[ r\pi_s = -\frac{1}{k^2} \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} a_l \zeta_l(kr) P_l^{(1)}(\cos \theta) \cos \phi \]

\[ r\pi'_s = -\frac{1}{k^2} \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} b_l \zeta_l(kr) P_l^{(1)}(\cos \theta) \sin \phi \]

where the function \( \zeta_l(kr) \) is given as

\[ \zeta_l(kr) = \psi_l(kr) + i \chi_l(kr) . \]

The wave inside the particle can be expressed as

\[ r\pi_r = -\frac{1}{k^2} \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} c_l \psi_l(kr) P_l^{(1)}(\cos \theta) \cos \phi \]

\[ r\pi'_r = -\frac{1}{k^2} \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} d_l \psi_l(kr) P_l^{(1)}(\cos \theta) \sin \phi \]

The boundary conditions are that the tangential components of \( E \) and \( H \) be continuous across the spherical surface \( r = a \). Following the boundary conditions, the coefficients \( a_l, b_l, c_l \) and \( d_l \) can be determined. Since we are interested in the scattered wave, \( a_l \) and \( b_l \) are given as

\[
\begin{align*}
a_l &= \frac{\psi_l(\alpha)\psi'_l(\beta) - m\psi'_l(\beta)\psi_l(\alpha)}{\zeta_l(\alpha)\psi'_l(\beta) - m\psi'_l(\beta)\zeta'_l(\alpha)} \\
b_l &= \frac{m\psi_l(\alpha)\psi'_l(\beta) - \psi_l(\beta)\psi'_l(\alpha)}{m\zeta_l(\alpha)\psi'_l(\beta) - \psi_l(\beta)\zeta'_l(\alpha)}
\end{align*}
\]

where

\[
\begin{align*}
\alpha &= ka \\
\beta &= kr, a
\end{align*}
\]

3.5.3. Far-Field Solution

We now consider a scattered field at distance sufficiently far away from the scatter so that \( kr \gg l \) where \( l \) is the order of the Ricatti-Bessel function. The scattered electric field intensity \((E_\phi, E_\theta)\) is given as

\[ E_\phi = \frac{-ie^{-i\beta r}}{kr} S_1(\theta) \sin \phi \]

\[ E_\theta = \frac{ie^{-i\beta r}}{kr} S_2(\theta) \cos \phi \]

where

\[ S_1(\theta) = \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} \{a_l \pi_l(\cos \theta) + b_l \tau_l(\cos \theta)\} \]

\[ S_2(\theta) = \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} \{a_l \pi_l(\cos \theta) + b_l \tau_l(\cos \theta)\} \]

Here the angular functions \( \pi_l(\cos \theta) \) and \( \tau_l(\cos \theta) \) are given as
\[ \pi_r(\cos \theta) = \frac{P_r^{(l)}(\cos \theta)}{\sin \theta} \]  
\[ \tau_r(\cos \theta) = \frac{d}{d\theta} P_r^{(l)}(\cos \theta) \]  

The scattering cross-section, absorption cross-section and total cross-section can be obtained as

\[ \sigma_s = \frac{\lambda^2}{2\pi} \sum_{l=1}^{\infty} (2l+1) \left\{ |a_l|^2 + |b_l|^2 \right\} \]  
\[ \sigma_a = \frac{\lambda^2}{2\pi} \sum_{l=1}^{\infty} (2l+1) \left\{ 2 - |a_l + 1|^2 - |b_l + 1|^2 \right\} \]  
\[ \sigma_t = -\frac{\lambda^2}{2\pi} \sum_{l=1}^{\infty} (2l+1) \{ \text{Re}(a_l + b_l) \} \]  

(3.113)
4 Transport Theory of Light in Random Particles

In a medium with randomly distributed particles, the entire scattering process is very complicated. The scattering process in principle may include multiple particle scattering, diffraction, and interference. In practice, it is unfeasible to obtain a formulation that includes all the effects. Various theories that yield useful solutions are all approximate, and each is only applicable in a specific range of parameters.

In this section, we will introduce an approximate theory, transport theory. The transport theory does not start from wave equations, but from the transport of energy through a medium containing particles. The development of this theory is heuristic and lacks the mathematical rigor of an analytic theory. Even though the diffraction and interference effects are included in the description of the scattering characteristics of a single particle, transport theory does not include the diffraction effects. It is assumed in transport theory that there is no correlation between the fields, and therefore, the addition of powers is equivalent to the addition of fields. The polarization effect can be included through the Stokes matrix. However, in most cases, polarization is neglected mainly for mathematical convenience. Finally, all sections will not include the polarization effect.

In the following, we will first present the concepts in radiometry, then the transport theory. Finally, we will present the approximation solutions for tenuous media—the first order multiple scattering approximation, and the approximation solutions for dense media—the diffusion approximation. Refer to details in the relevant literature.2,5,6

4.1. Radiometry

The term “radiometry” means the measurement of energy radiation. The most primitive beginning of radiometry may have been the observation of the different brightness of stars and the sensing of the warmth from the Sun and fires. To quantitatively measure energy radiation, four basic quantities are needed:

- Radiant flux, \( \Phi \), which has the unit [watts]
- Radiant flux density (irradiance and radiant exitance), \( E \) [watts m\(^{-2}\)]
- Radiant intensity, \( I \) [watts sr\(^{-1}\)]
- Radiance, \( L \) [watts m\(^{-2}\) sr\(^{-1}\)]

Here sr (steradians) is the unit for a solid angle. A solid angle is the two-dimensional equivalence of a linear angle. It is the projection of an area onto a unit sphere, as shown in Figure 4.1. An equivalent definition is the projection of an area on a sphere divided by the square of the radius of the sphere. A sphere subtends 4\( \pi \) steradians, while a hemisphere subtends 2\( \pi \) steradians. Similarly, the projected solid angle is defined as the angle corresponding to the projected area of any area onto the unit sphere.
Consider the energy flow of a wave passing through an elementary area $dA$ around a point $r$ in a random medium, as shown in Figure 4.2. The phase and amplitude of the wave undergo some random variations in time, and therefore the magnitude and direction of its power flux density vector vary continuously in time. Assume that the wave propagation has reached a stable status so that the average amount of the energy passing through $dA$ per unit time does not change with the time. Then the radiant flux, $\Phi$, is defined by this amount as

$$\Phi = \frac{dQ}{dt},$$

where $dQ$ is the energy passing through $dA$ within the time interval $dt$.

The radiant flux density $E$ is defined as the flux per unit area at point $r$ with surface normal $\mathbf{n}$. 

![Figure 4.1: Solid angle $\Omega$ and projected solid angle $\Omega^\perp$.](image)

![Figure 4.2: Energy flow through an elementary area $dA$ around point $r$.](image)
There are two cases of radiant flux density: irradiance and radiant exitance. The irradiance describes the incident illumination at a point on a surface, and radiant exitance works for the radiative flux leaving a point on a surface.

The radiant intensity describes the flux per solid angle passing through a point in space

\[ I = \frac{d\Phi}{d\Omega}. \]  
(4.3)

The intensity is most useful for describing the angular distribution of emission from an infinitesimal point source.

The radiance \( L \) is defined as the flux per solid angle per projected area as

\[ L = \frac{d^2\Phi}{d\Omega dA^\perp}, \]  
(4.4)

where \( dA^\perp \) is the projected area measure on a plane perpendicular to \( \Omega \). When we consider the radiance at a surface, we can equivalently write

\[ L = \frac{d^2\Phi}{d\Omega dA^\perp} = \frac{d^2\Phi}{d\Omega dA \left| \hat{\Omega} \cdot \mathbf{n} \right|} = \frac{d^2\Phi}{d\Omega^2 dA}. \]  
(4.5)

### 4.2. Radiance in Free Space and at Surfaces

We first prove the invariance of radiance along the ray path in the free space. Consider the specific radiances \( L_1 \) and \( L_2 \) at two points \( r_1 \) and \( r_2 \) separated by a distance \( r \) along the direction \( \mathbf{n} \) and two small areas \( dA_1 \) and \( dA_2 \) perpendicular to \( \mathbf{n} \), as shown in Figure 4.3.

![Figure 4.3: Proof of invariance of radiance in free space.](image)

We can express the power flux received by \( dA_2 \) in two different ways. In terms of \( L_1 \), it is \( L_1 d\Omega_1 dA_1 \); in terms of \( L_2 \), it is \( L_2 d\Omega_2 dA_2 \). These two terms should be equal. Since \( dA_1 = r^2 d\Omega_2 \),
and \( dA_2 = r^2 d\Omega_1 \), we obtain \( dA_1 d\Omega_1 = dA_2 d\Omega_2 \). Therefore, \( L_1 = L_2 \), which is the invariance of radiance along the ray in the free space.

Let us consider what condition should be satisfied by the radiance at a plane boundary between two media with indices of refraction \( n_i \) and \( n_r \), as shown in Figure 4.4. Consider that the incident power flux into a small area \( dA \) on the boundary must be equal to the sum of the reflected and transmitted fluxes, we obtain

\[
L_i dA \cos \theta_i d\Omega_i = L_i dA \cos \theta_i d\Omega_i + L_r dA \cos \theta_r d\Omega_r \quad (4.6)
\]

We note that \( d\Omega_i = d\Omega_r = \sin \theta_i d\theta_i d\phi_i \), and \( d\Omega_r = \sin \theta_r d\theta_r d\phi_r \). From Snell’s law \( n_i \sin \theta_i = n_r \sin \theta_r \), we obtain \( n_i \cos \theta_i d\theta_i = n_r \cos \theta_r d\theta_r \). Since \( d\phi_i = d\phi_r \), we obtain

\[
L_i / n_i^2 = L_r / n_r^2 + L_t / n_t^2 \quad (4.7)
\]

From Eqs. (2.41) and (2.42), we know that

\[
L_r = \left| r \right|^2 L_i \quad (4.8)
\]

where \( \left| r \right|^2 \) is equal to \( \left| r_i \right|^2 \) or \( \left| r_r \right|^2 \), depending on the polarization, or to \( \frac{1}{2}(\left| r_i \right|^2 + \left| r_r \right|^2) \) for a completely unpolarized wave.

Substituting Eq. (4.8) into Eq. (4.7), we obtain

\[
L_t = \frac{n_i^2}{n_t^2} \left( 1 - \left| r \right|^2 \right) L_i \quad (4.9)
\]

Using Eq. (2.43), we also obtain
\[ L_i = \frac{n_i^2}{n_i^2} T L_i = \frac{n_i^2}{n_i} \cos \theta_i |r_i|^2 L_i, \]

where \( T \) and \( |r_i|^2 \) consider the effects of polarization as same as \( |r|^2 \).

### 4.3. Transport Equation

Consider the radiance \( L(r, \hat{s}) \) incident upon a cylindrical volume with cross-section \( dA \) and length \( ds \), and the incident solid angle \( d\Omega \), as shown in Figure 4.5. The volumetric density of the particles inside the volume is \( \rho \). Then the lost power, \( dLdAd\Omega \), due to the scattering and absorption inside the volume, is given as the multiplication of the number of particles inside the volume, \( \rho dA ds \), and the lost power due to the scattering and absorption by one particle, \( \sigma_i L d\Omega \). That is,

\[ dLdAd\Omega = -\rho dA ds \sigma_i L d\Omega \]

or

\[ dL = -\rho ds \sigma_i L. \]  

(4.11)

![Figure 4.5: Scattering and absorption of the radiance \( L(r, \hat{s}) \) by a cylindrical volume.](image)

![Figure 4.6: Scattering from the direction \( \hat{s}' \) to the direction \( \hat{s} \).](image)
At the same time, the radiance increases because a portion of the radiance \( L(r, \hat{s}') \) incident on this volume from other direction \( \hat{s}' \), as shown in Figure 4.6. The increased power in unit solid angle due to the scattering by a single particle from the direction \( \hat{s}' \) to the direction \( \hat{s} \) is \( \frac{d\sigma}{d\Omega} L(r, \hat{s}') d\Omega' \). If we consider radiances coming from all the directions, the increased power in solid angle \( d\Omega \) is \( d\Omega \int_{4\pi} \frac{d\sigma}{d\Omega} L(r, \hat{s}') d\Omega' \). Since the increased power, \( dL dAd\Omega \), is equal to the multiplication of the number of particles inside the volume, \( \rho dAds \), and the increased power due to the scattering by a single particle from all the directions, we obtain

\[
\frac{dL dAd\Omega}{d\Omega} = \rho dAd\Omega \int_{4\pi} \frac{d\sigma}{d\Omega} L(r, \hat{s}') d\Omega'.
\]

or

\[
dL = \rho d\Omega \int_{4\pi} \frac{d\sigma}{d\Omega} L(r, \hat{s}') d\Omega'.
\]

Substituting Eq. (3.5) into (4.12), we obtain

\[
dL = ds \frac{\rho \sigma}{4\pi} \int_{4\pi} p(\hat{s}, \hat{s}') L(r, \hat{s}') d\Omega'.
\]

(4.12)

The radiance can also be increased due to the emission from the volume \( ds \). Denoting \( \varepsilon(r, \hat{s}) \) as the power radiation per unit volume per unit solid angle in the direction \( \hat{s} \), then we obtain

\[
dL dAd\Omega = ds dA \varepsilon(r, \hat{s}) d\Omega
\]

or

\[
dL = ds \varepsilon(r, \hat{s}).
\]

(4.14)

From Eqs. (4.11), (4.13) and (4.14), we obtain the transport equation as

\[
\frac{dL(r, \hat{s})}{ds} = -\rho \sigma L(r, \hat{s}) + \frac{\rho \sigma}{4\pi} \int_{4\pi} p(\hat{s}, \hat{s}') L(r, \hat{s}') d\Omega' + \varepsilon(r, \hat{s}).
\]

(4.15)

The left-hand side of this equation can also be written using a gradient of divergence operator as follows:

\[
\frac{dL(r, \hat{s})}{ds} = \hat{s} \cdot \nabla L(r, \hat{s}) = \nabla \cdot [L(r, \hat{s}) \hat{s}],
\]

(4.16)

where we have used the fact that \( \hat{s} \) is a constant vector and thus \( \nabla \cdot \hat{s} = 0 \).

In Eq. (4.15), the particle density and size can be different at different locations, and therefore \( \rho \sigma \) and \( p(\hat{s}, \hat{s}') \) are the functions of \( r \). Sometimes it is more convenient to measure the distance in terms of a non-dimensional optical distance \( \tau \) defined by

\[
\tau = \int \rho \sigma ds.
\]

(4.17)

The optical distance \( \tau = 1 \) means that over this distance the power flux diminishes by scattering and absorption to \( 1/e \) of the incident flux. Using Eq. (4.17), Eq. (4.15) can be transformed into
\[ \frac{dL(r, \hat{s})}{d\tau} = -L(r, \hat{s}) + \frac{1}{4\pi} \int_{4\pi} p(\hat{s}, \hat{s}') L(r, \hat{s}') d\Omega' + J(r, \hat{s}), \]  
(4.18)

where \( J(r, \hat{s}) = \varepsilon(r, \hat{s}) / \rho \sigma \) is called the source function.

### 4.4 Reduced Incident Radiance, Diffuse Radiance, Boundary Condition, and Source Function

It is often convenient to divide the total incident radiance \( L \) into two components: one is the reduced incident radiance, \( L_{ri} \), and another is the diffuse radiance, \( L_d \), as shown in Figure 4.7. Thus

\[ L(r, \hat{s}) = L_{ri}(r, \hat{s}) + L_d(r, \hat{s}). \]  
(4.19)

![Figure 4.7: Incident radiance \( L_i \), total radiance \( L \), reduced incident radiance \( L_{ri} \), and diffuse radiance \( L_d \).](image)

The reduced incident radiance describes the component of total incident radiance that is decreased due to the scattering and absorption in Eq. (4.11),

\[ \frac{dL_{ri}(r, \hat{s})}{ds} = -\rho \sigma_i L_{ri}(r, \hat{s}). \]  
(4.20)

Substituting Eqs. (4.19) and (4.20) into (4.15), we obtain

\[ \frac{dL_d(r, \hat{s})}{ds} = -\rho \sigma_i L_d(r, \hat{s}) + \frac{\rho \sigma_i}{4\pi} \int_{4\pi} p(\hat{s}, \hat{s}') L_d(r, \hat{s}') d\Omega' + \varepsilon_{ri}(r, \hat{s}) + \varepsilon(r, \hat{s}), \]  
(4.21)

where \( \varepsilon_{ri} \) is the equivalent source function caused from the reduced incident radiance

\[ \varepsilon_{ri}(r, \hat{s}) = \frac{\rho \sigma_i}{4\pi} \int_{4\pi} p(\hat{s}, \hat{s}') L_{ri}(r, \hat{s}') d\Omega'. \]  
(4.22)
Since the diffuse radiance is only generated inside the volume surrounded by the surface $S$, we therefore require that at the surface $S$, there should be no diffuse radiation entering the medium, and the diffuse radiance should always be pointed outward (strictly speaking, this holds only for a convex surface where the radiance does not enter the surface). Thus,

$$L_d(r, \hat{s}) = 0 \text{ on } S \text{ when } \hat{s} \text{ is pointed inward}. \tag{4.23}$$

If the medium extends to the infinity, we require that the diffuse radiance $L_d$ must diminish at infinity.

The reduced incident radiance may be **collimated or diffuse**. For example, the incident wave may be well collimated in a particular direction $\hat{s}_0$ as in the case of a laser beam or in the case of a plane wave. We call it the **collimated incident radiance**, and express it as

$$L_c(r, \hat{s}) = F_0(\hat{\Omega} - \hat{\Omega}_0), \tag{4.24}$$

where $F_0$ is the flux density, $\delta(\hat{\Omega} - \hat{\Omega}_0)$ is the solid angle delta function, and $\hat{\Omega}$ and $\hat{\Omega}_0$ are the unit vectors representing solid angles in the directions $\hat{s}$ and $\hat{s}_0$. In a spherical coordinate system,

$$\delta(\hat{\Omega} - \hat{\Omega}_0) = \frac{\delta(\theta - \theta_0)\delta(\phi - \phi_0)}{\sin \theta} \int_0^{2\pi} \delta(\hat{\Omega} - \hat{\Omega}_0) d\Omega = 1 \tag{4.25}$$

$$d\Omega = \sin \theta d\theta d\phi$$

The diffuse incident radiance comes from various directions with different magnitudes. For example, consider the radiation scattered through a cloud that is incident on an ocean surface. In this case, the wave incident on the water is not collimated, but is already diffused. We call it the diffuse incident radiance.

For the source function, if a point source is located at $r_0$ and radiates the total power $P_0$ uniformly in all directions, we can write

$$\epsilon(r) = \frac{P_0}{4\pi} \delta(r - r_0). \tag{4.26}$$

If the medium is in local thermodynamic equilibrium at temperature $T$, then Kirchoff’s law for a blackbody gives a reasonable approximation to the radiating energy $^7$. For a monochromatic wave, we approximately have

$$\epsilon(r) = \rho \sigma \frac{B(T)}{} \tag{4.27}$$

where $B(T) = (2hv^3/c^2)[\exp(hv/KT) - 1]^{-1}$, and $K$ and $h$ are the Boltzmann and Planck constants, respectively.

**4.5. Approximate Solutions for Tenuous Medium**

In most cases, the equation cannot be solved exactly and it is necessary to resort to approximate solutions. Even in cases that the exact solutions are available, the expressions are often so complex that it is desirable to have simpler approximate solutions. We consider the two extreme
cases, tenuous and dense distributions, for which relatively simple solutions can be obtained. Now we will present the approximate solutions for the tenuous media.

4.5.1. First-Order Multiple Scattering Approximation

In Eq. (4.19), the total radiance inside the medium is decomposed into the reduced incident radiance and diffuse radiance. As shown in Figure 4.7, the diffuse radiance is the sum of all the radiances scattered by the particles when illuminated by the total radiance, which is unknown. In the first approximation, we assume that this total radiance illuminating the particles is approximately equal to the known reduced incident radiance. Then we can obtain the first-order multiple scattering approximation,

\[ L_d(r,\hat{s}) \approx \int_0^\infty \exp[-(\tau - \tau_0)] \left[ \frac{\rho \sigma_t}{4\pi} \int_{4\pi} p(\hat{s},\hat{s}') L_{\psi}(r_1,\hat{s}') d\Omega' + \epsilon(r,\hat{s}) \right] ds_1. \]  (4.28)

The first-order solution is only applicable to the case where the density of the scatters is low so that the incoherent power is considerably small compared with the coherent power. This usually happens in the following two situations:

(a) For a plane wave incident upon a medium containing random particles, the first order theory applies when the optical distance is smaller than approximately 0.4 (\( \tau \leq 4 \)) and the particles are mostly absorbing (albedo \( W_0 \approx 0.5 \)). This is the case discussed in the following subsections.

(b) For a wave confined within a small angular region, such as a narrow beam from a transmitter, it is applicable to much greater distances, particularly when the particles are absorbing (\( W_0 \approx 0.9 \)). This is the situation encountered in microwave and optical propagation in the atmosphere.

Figure 4.8: Plane-parallel medium of thickness \( d \) containing randomly distributed particles illuminated by a plane wave.
4.5.2. Plane Wave Incident on a Plane-Parallel Medium

Consider a plane-parallel medium of thickness \( d \) containing random particles, as shown in Figure 4.8. A plane wave with flux density \( F_i \) is incident with the angle of incidence \( \theta \), and the x-axis is chosen so that the plane of incidence is the xz-plane. The indices of refraction for the media in regions \( z > 0 \), \( 0 < z < d \), and \( z > d \) are \( n_1 \), \( n_2 \), and \( n_3 \), respectively. For convenience, we use the optical distance \( \tau = \rho \sigma z \) in the \( z \) direction rather than the optical distance along the direction of wave propagation \( \rho \sigma z \sec \theta \). It is also convenient to use \( \mu = \cos \theta \) rather than \( \theta \) itself.

We write the reduced incident radiance as

\[
L_{\text{in}}(\tau, \mu, \phi) = F_0 \exp(-\tau/\mu_0) \delta(\hat{\Omega} - \hat{\Omega}_0),
\]

where \( F_0 = T_{12} F_i \). \( T_{12} \) is the transmission coefficient for a radiance from the medium 1 to 2, and it is given in Eq. (4.10).

Using \( \mu = \cos \theta \), Eq. (4.25) can be transformed into

\[
\delta(\hat{\Omega} - \hat{\Omega}_0) = \delta(\mu - \mu_0) \delta(\phi - \phi_0)
\]

\[
\int_0^{2\pi} \delta(\hat{\Omega} - \hat{\Omega}_0) d\Omega = \int_0^{\pi} d\phi \int_0^{2\pi} d\mu \delta(\mu - \mu_0) \delta(\phi - \phi_0) = 1
\]

From Figure 4.8, we can see that the contributions to \( L_{\text{d}} \) for \( 0 < \theta < \pi/2 \) come from the reduced incident radiance in the range \( 0 \) to \( z \); the contributions to \( L_{\text{d}} \) for \( \pi/2 < \theta < \pi \) come from the range \( z \) to \( d \). Substituting Eq. (4.29) into Eq. (4.28) and noting \( \varepsilon(r, \hat{s}) = 0 \), we obtain

\[
L_{\text{d}}(\tau, \mu, \phi) = \frac{p(\mu, \phi; \mu_0, 0)}{4\pi} \exp(-\tau/\mu_0) \exp(-\tau/\mu) \frac{1}{\mu_0 - \mu} F_0, \quad 0 < \mu \leq 1
\]

\[
L_{\text{d}}(\tau, \mu, \phi) = \frac{p(\mu, \phi; \mu_0, 0)}{4\pi} \exp(-\tau/\mu_0) \exp(-\tau/\mu) \frac{1}{(\mu_0 - \mu)} F_0, \quad -1 < \mu < 0
\]

where \( \tau_0 = \rho \sigma z \).

The diffuse radiance for \( z < 0 \) when the medium containing particles is semi-infinite, is often of special interest. The angular dependence is usually given in terms of the angle of reflection \( \theta_r = \pi - \theta \). Using \( \cos \theta_r = \mu, \rho = -\mu \), we obtain

\[
L_{\text{d}}(\mu) = \frac{p(\mu, \phi; \mu_0, 0)}{4\pi} \left( \frac{\mu_0}{\mu_0 + \mu} \right) T_{12} T_{21} F_i,
\]

which is the first order approximation.

In this analysis, we have neglected the contributions due to reflection at the back surface \( z = d \). If the surface is highly reflective and smooth with specular reflection coefficient \( R \), then the following diffuse radiance must be added to the right-handed side of Eq. (4.32):

\[
\frac{p(\mu, \phi; \mu_0, 0)}{4\pi} \exp\left(\frac{-\tau_0}{\mu_0}\right) \frac{1}{(\mu_0 + \mu)} \mu_0 R F_0 \exp\left(\frac{-\tau_0}{\mu_0}\right), \quad -1 < \mu < 0
\]
If the surface at $z = d$ is highly diffusing so that the reflected radiance is almost uniformly distributed over the solid angle $2\pi$, then the diffuse radiance

$$L'_d = \frac{R_d \mu_0 F_0}{\pi} \exp\left( -\frac{\tau_0 + \frac{\tau_0 - \tau}{\mu}}{\mu} \right), \quad -1 \leq \mu < 0$$  \hspace{1cm} (4.35)$$

must be added to Eq. (4.32), where $R_d$ is the albedo of the surface defined by the ratio $\pi L'_d / \mu_0 F_0$.

### 4.5.3. Collimated Beam Incident on A Plane-Parallel Medium

In some practical situations, the incident wave may be a narrow collimated beam. Consider a collimated beam whose radiance variation in the transverse plane is Gaussian. We assume that the beam is normally incident on a parallel-plane medium, as shown in Figure 4.9.

![Gaussian beam incident upon randomly distributed particles of thickness d.](image)

The radiance at $z = 0$ is given by

$$L(z = 0, r, \hat{s}) = F_0 \exp(-2r^2/D_0^2)\delta(\hat{\Omega}),$$  \hspace{1cm} (4.36)$$

where $r = \sqrt{x^2 + y^2}$, $D_0$ is the beam size and $F_0$ is the magnitude. The angular spread is approximated by a delta function $\delta(\hat{\Omega})$.

The reduced incident radiance $L_{ri}$ can be obtained as

$$L_{ri}(r, \hat{s}) = F_0 \exp(-2r^2/D_0^2) - \rho \sigma_i z)\delta(\hat{\Omega})$$  \hspace{1cm} (4.37)$$

and the diffuse radiance $L_d$, shown in Figure 4.9 at $(x, 0, z)$ and pointed in the direction $(\theta, 0)$ is given by

$$L_d = \int \exp[-\rho \sigma_i (z - z_i) \sec \theta] \left( \frac{\rho \sigma_i}{4\pi} \right) p(\theta, 0; 0, 0) F_0 \exp\left( -\frac{2x_i^2}{D_0^2} - \rho \sigma_i z_i \right) dz_i \sec \theta,$$  \hspace{1cm} (4.38)$$
where \( x_i = x_0 + z_i \tan \theta \).

### 4.6. Approximate Solutions for Dense Medium

The first-order multiple scattering approximation is valid only when the volume density, which is the ratio of the volume occupied by the particles to the total volume of the medium, is considerably smaller than 0.1%. When the volume density is much greater than 1%, the **diffusion approximation** gives relatively simple and good solutions. For a volume density in the neighborhood of 1%, neither the first order nor the diffuse approximation may be valid, and the complete equation must be solved.

![Figure 4.10: Diffusion radiance \( L_d(r, \hat{s}) \) for diffusion approximation.](image)

#### 4.6.1. Diffusion Approximation

In the diffusion approximation, we assume that the diffuse radiance encounters many particles and is scattered almost uniformly in all directions. Therefore the angular distribution of scattered radiance is almost uniform, as shown in Figure 4.10. The diffuse radiance should have a slightly greater magnitude in the direction of the net flux flow than in the backward direction. Mathematically the diffuse radiance is assumed to be

\[
L_d(r, \hat{s}) = U_d(r) + c F_d(r) \cdot \hat{s} ,
\]

where \( c \) is a constant, the diffuse flux vector \( F_d(r) \), whose direction is given by a unit vector \( \hat{s}_f \), is expected by

\[
F_d(r) = \int_{4\pi} L_d(r, \hat{s}) \hat{s} d\Omega = F_d(r) \hat{s}_f ,
\]

and \( U_d(r) \) is the average diffuse radiance given by

\[
U_d(r) = \frac{1}{4\pi} \int_{4\pi} L_d(r, \hat{s}) d\Omega .
\]

The constant \( c \) can be easily found by noting that
Substituting Eq. (4.39) into (4.42), we obtain $c = \frac{3}{47}$, and therefore in the diffusion approximation the diffuse radiance is given by

$$L_d(r, \hat{s}) = U_d(r) + \frac{3}{4\pi} F_d(r) \cdot \hat{s}.$$  \hspace{1cm} (4.43)

Eq. (4.43) can be regarded as the first two terms of a Taylor's expansion of $L_d$ in terms of the powers of $\hat{s} \cdot \hat{s}_f$, and therefore the second term of Eq. (4.43) must be considerably smaller than the first, that is, $U_d \gg |F_d|$. 

If we write the left term of Eq. (4.21) using the divergence form in Eq. (4.16), and integrate over all $4\pi$ of solid angle, we have

$$\nabla \cdot F_d(r) = -\rho \sigma_a \int_{4\pi} L_d(r, \hat{s}) d\Omega + \int_{4\pi} \varepsilon_a(r, \hat{s}) d\Omega + \int_{4\pi} \varepsilon(r, \hat{s}) d\Omega.$$ \hspace{1cm} (4.44)

Here we have used the expression in Eq. (3.7). If we define $U_n(r) = \frac{1}{4\pi} \int_{4\pi} L_n(r, \hat{s}) d\Omega$ and $E(r) = \int_{4\pi} \varepsilon(r, \hat{s}) d\Omega$, and use Eqs. (4.22) and (4.41), we obtain

$$\nabla \cdot F_d(r) = -4\pi \rho \sigma_a U_d(r) + 4\pi \rho \sigma_a U_n(r) + E(r).$$ \hspace{1cm} (4.45)

Substituting Eq. (4.43) into Eq. (4.21), we obtain

$$\hat{s} \cdot \nabla U_d + \frac{3}{4\pi} \hat{s} \cdot \nabla (F_d \cdot \hat{s}) = -\rho \sigma_a U_d - \frac{3}{4\pi} \rho \sigma_a F_d \cdot \hat{s} + \rho \sigma_a U_d + \frac{3}{4\pi} \rho \sigma_a F_d \cdot \hat{s} \rho + \varepsilon_n + \varepsilon.$$ \hspace{1cm} (4.46)

where $p_1$ is given by

$$p_1 = \frac{1}{4\pi} \int_{4\pi} p(\hat{s}, \hat{s}') \hat{s} \cdot \hat{s}' d\Omega'.$$ \hspace{1cm} (4.47)

and it represents the average forward scattering ($\hat{s} \cdot \hat{s}' > 0$) minus the backward scattering ($\hat{s} \cdot \hat{s}' < 0$) of a single particle.

Multiplying Eq. (4.47) with $\hat{s}$ and integrating over all $4\pi$,

$$\nabla U_d = -\frac{3}{4\pi} \rho \sigma_s (1 - p_1) F_d + \frac{3}{4\pi} \int_{4\pi} \varepsilon_n(r, \hat{s}) \hat{s} d\Omega + \frac{3}{4\pi} \int_{4\pi} \varepsilon(r, \hat{s}) \hat{s} \hat{s} d\Omega.$$ \hspace{1cm} (4.48)

We define $\sigma_s = \sigma_t (1 - p_1)$ and eliminate $F_d$ from Eqs. (4.44) and (4.48), then obtain

$$\nabla^2 U_d(r) - \kappa_d^2 U_d(r) = -3 \rho \sigma_t \rho \sigma_s U_n(r) - \frac{3}{4\pi} \rho \sigma_s E(r) + \frac{3}{4\pi} \nabla \cdot \int_{4\pi} \varepsilon_n(r, \hat{s}) \hat{s} d\Omega$$

$$+ \frac{3}{4\pi} \nabla \cdot \int_{4\pi} \varepsilon(r, \hat{s}) \hat{s} d\Omega.$$ \hspace{1cm} (4.49)

where $\kappa_d^2 = 3 \rho \sigma_a \rho \sigma_s$.

**4.6.2. Boundary Conditions**

The exact boundary condition for the diffuse radiance $L_d$ should be Eq. (4.23). However, according to the diffusion approximation, $L_d$ has a simple angular distribution given by Eq. (4.43). 

$$F_d(r) = F_d(r) \cdot \hat{s}_f = \int_{4\pi} L_d(r, \hat{s}) \hat{s} \cdot \hat{s}_f d\Omega.$$
(4.43), then the boundary condition Eq. (4.23) cannot be satisfied. We consider the approximate boundary conditions instead. One of them is the condition at a surface that the total diffuse flux directed inward must be zero:

$$\int_{2\pi} L_d(r, \hat{s})(\hat{s} \cdot n) d\Omega = 0,$$

(4.50)

where $n$ is a unit vector normally directed inward, and the integration is performed over $2\pi$ in the hemisphere $\hat{s} \cdot n > 0$.

The condition (4.50) can be expressed in terms of the average diffuse radiance $U_d$. To do this, we first write $F_d$ as a sum of the component $F_{dn}$ normal to and the component $F_{dt}$ tangential to the surface:

$$F_d = F_{dn} n + F_{dt} t.$$  (4.51)

Substituting Eqs. (4.43) and (4.51) into Eq. (4.50), we obtain

$$\frac{1}{2} U_d(r_s) + \frac{F_{dn}(r_s)}{4\pi} = 0 \quad \text{at the surface } r = r_s.$$  (4.52)

Next, we express $F_{dn}$ in Eq. (4.52) in terms of $U_d$ by using Eq. (4.48)

$$F_{dn} = n \cdot F_d = -\frac{4\pi}{3\rho\sigma_r} n \cdot \nabla U_d(r) + n \cdot Q_1(r)$$  (4.53)

where

$$Q_1(r) = \frac{1}{\rho\sigma_r} \int_{4\pi} [\varepsilon_d(r, \hat{s}) + \varepsilon(r, \hat{s})] \hat{s} d\Omega.$$  (4.54)

In the absence of sources, $\varepsilon(r, \hat{s}) = 0$, we can obtain

$$Q_1(r) = \frac{\sigma_r}{\sigma_t} \int_{4\pi} d\Omega' \left[ \frac{1}{4\pi} \int_0 \rho(\hat{s}, \hat{s}') \hat{s}' d\Omega \right] L_n(r, \hat{s}').$$  (4.55)

Note that if the scatters are isotropic, $p(\hat{s}, \hat{s}') = \text{constant}$ and $Q_1 = 0$. Therefore, $Q_1$ represents the effect of anisotropy of the scattering pattern.

Substituting Eq. (4.53) into Eq. (4.52), we obtain

$$U_d(r_s) - \frac{2}{3\rho\sigma_r} \frac{\partial}{\partial n} U_d(r_s) + \frac{2n \cdot Q_1(r_s)}{4\pi} = 0,$$  (4.56)

where $\partial/\partial n$ is the normal derivative in the direction toward the medium.

### 4.6.3. Collimated Beam Incident Upon a Slab of Particles

Let us consider a collimated beam incident normally upon a slab containing random particles, as shown in Figure 4.9. Following Eq. (4.37), the reduced incident radiance can be written as

$$L_n(r, \hat{s}) = F_0(r) \exp(-\rho\sigma_r z) \delta(\hat{\Omega} - \hat{\Omega}_z),$$  (4.57)

where $F_0(r)$ is the flux density, and $\hat{\Omega}_z$ is the unit vector pointed in the $z$ direction.
Substituting Eq. (4.57) into Eqs. (4.22) and (4.45), the source terms on the right-hand side of Eq. (4.49) can be expressed in terms of $F_0$. We note that

$$V_{1i}q_i(r, \mathbf{s}) \delta d \Omega = \frac{1}{4\pi} \frac{\rho \sigma_i}{\rho \sigma_z} p(i, z) \delta d \Omega \cdot \nabla [F_0(r) \exp(-\rho \sigma_z)]$$  

$$= - (\rho \sigma_i)^2 p_i F_0(r) \exp(-\rho \sigma_z)$$  

Substituting Eq. (4.58) into Eq. (4.49), we obtain

$$\nabla^2 U_d(r) - \kappa_d^2 U_d(r) = -Q(r), \quad \kappa_d^2 = 3(\rho \sigma_s) (\rho \sigma_n)$$

$$Q(r) = [3 \rho \sigma_s \rho \sigma_n + 3 \rho \sigma_s \rho \sigma_i \overline{\mu}] p F_0(r) \frac{\exp(-\rho \sigma_z)}{4\pi}$$

where $\mu = \mathbf{s} \cdot \mathbf{s}'$ and

$$\overline{\mu} = \left( \int_{\Omega} p(\mathbf{s}, \mathbf{s}') d \Omega' \right) / \left( \int_{\Omega} p(\mathbf{s}, \mathbf{s}') d \Omega' \right).$$

The boundary conditions are given as

$$U_d(r) - h \frac{\partial}{\partial z} U_d(r) + \frac{Q_i(r)}{2\pi} = 0 \quad \text{at} \quad z = 0$$

$$U_d(r) + h \frac{\partial}{\partial z} U_d(r) - \frac{Q_i(r)}{2\pi} = 0 \quad \text{at} \quad z = d$$

where $h = 2/3 \rho \sigma_n$ and $Q_i(r) = (\sigma_s \overline{\mu}) / \sigma_n F_0(r) \exp(-\rho \sigma_z)$.

Once $U_d$ is obtained, the diffuse radiance can be found from Eqs. (4.48) and (4.43):

$$F_d(r) = \frac{\sigma_i \overline{\mu} F_0(r)}{\sigma_n} \exp(-\rho \sigma_z) = \frac{4\pi}{3 \rho \sigma_n} \nabla U_d(r).$$

For a high density medium, $\sigma_s$ and $\sigma_n$ must be replaced by the following expressions:

$$\sigma_s \rightarrow \sigma_s (1 - H)$$

$$\sigma_n \rightarrow \sigma_n (1 - H) (1 - \overline{\mu}) + \sigma_d$$

where $H$ is the ratio of volume occupied by scatters to the total volume.

**4.6.4. A Plane Wave Incident Upon a Slab of Particles**

The case of a plane wave incident upon a slab can be analyzed by the formulation given in the preceding subsection. For a plane wave, $F_0(r)$ in Eq. (4.57) becomes constant $F_0$, and Eq. (4.59) becomes

$$\frac{\partial^2}{\partial z^2} U_d(z) - \kappa_d^2 U_d(z) = -Q_0 \exp(-\rho \sigma_z)$$

where $Q_0 = [3 \rho \sigma_s \rho \sigma_n + 3 \rho \sigma_s \rho \sigma_i \overline{\mu}] (F_0 / 4\pi)$. The boundary conditions are the same as Eq. (4.61).
The general solution of Eq. (4.64) is the sum of a particular solution \( U_{dp} \) and a complementary solution \( U_{dc} \). The solution for \( U_{dp} \) is

\[
U_{dp}(z) = -\frac{Q_0}{(\rho \sigma)^2 - \kappa_d^2} \exp(-\rho \sigma z)
\]

and the general solution for \( U_{dc} \) is

\[
U_{dc}(z) = C_1 \exp(\kappa_d z) + C_2 \exp(-\kappa_d z).
\]

\( C_1 \) and \( C_2 \) are unknown constants and are determined by applying the boundary conditions (4.61). Substituting \( U_d = U_{dp} + U_{dc} \) into Eq. (4.61), we obtain

\[
C_1(1 - \kappa_d h) + C_2(1 + \kappa_d h) = \frac{Q_0}{(\rho \sigma)^2 - \kappa_d^2} (1 + \rho \sigma, h) - \frac{Q_1(0)}{2 \pi}
\]

\[
C_1(1 + \kappa_d h) \exp(\kappa_d d) + C_2(1 - \kappa_d h) \exp(-\kappa_d d) = \left[ \frac{Q_0}{(\rho \sigma)^2 - \kappa_d^2} (1 - \rho \sigma, h) + \frac{Q_1(0)}{2 \pi} \right] \exp(-\rho \sigma d)
\]

For the semi-infinite medium \( d \to \infty \), we obtain

\[
U_d(z) = -\frac{Q_0}{(\rho \sigma)^2 - \kappa_d^2} \exp(-\rho \sigma z) + \left[ \frac{Q_0}{(\rho \sigma)^2 - \kappa_d^2} \frac{1 + \rho \sigma h}{1 + \kappa_d h} - \frac{Q_1}{2 \pi (1 + \kappa_d h)} \right] \exp(-\kappa_d z)
\]

and \( F_d \) can be obtained by Eq. (4.62).
5 Monte Carlo Simulation of Light Transport

The term "Monte Carlo" originates from a city with the name Monaco, which is famous for its casinos. "Monte Carlo" was first used to name a class of numerical methods for developing the nuclear weapons in Los Alamos National Laboratory in 1940s. The essence of these methods is to use the games of chance to study some interesting phenomena. In late 1946 Stanislaw Ulam suggested the use of random sampling to simulate the flight paths of neutrons, which late sparked to a great deal of research. With the tremendous advancement of computer technologies, Monte Carlo methods were used in many scientific and engineering fields.

The problems handled by Monte Carlo methods may be classified into two types: probabilistic and deterministic. This classification depends on whether or not a problem is concerned with the behavior and outcome of random processes. For probabilistic problems, Monte Carlo methods usually sample the observed random events in such a way that they directly simulate the physical processes of the original problems, and infer the desired solutions from the simulations. For deterministic problems, Monte Carlo methods describe the deterministic processes with some apparently unrelated random processes, and hence numerically solve the deterministic problems by simulating the random processes.

In this section, we will review the basic probability theory, sampling methods, and simulation of radiation transport. Refer to details in relevant literature.

5.1. Basic Probability Theory

5.1.1. Probability and Probability Density
A random event is an event that has a chance of happening, and probability is a numerical measure of that chance. The value of probability ranges between 0 and 1, and higher value indicates the greater chance. We write \( P(A) \) for the probability that the event \( A \) occurs; \( p(A+B+...) \) for the probability that at least one of \( A, B, ... \) occurs; \( P(AB...) \) for all the events \( A, B, ... \) occur. \( P(A \mid B) \) for the probability that the event \( A \) occurs when it is known that the event \( B \) occurs, and is therefore called the conditional probability of \( A \) given \( B \). The probability has the following properties:

\[
P(A + B + ...) \leq P(A) + P(B) + ... \tag{5.1}
\]

and

\[
P(AB) = P(A \mid B)P(B) \tag{5.2}
\]

If only one of the events \( A, B, ... \) can occur, they are called exclusive. If only at least one of the events \( A, B, ... \) must occur, they are called exhaustive. If only \( P(A \mid B) = P(A) \), \( A \) and \( B \) are independent.

Consider a set of exhaustive and exclusive events, each characterized by a number \( X \). The number \( X \) is called the random variable. The cumulative distribution function \( F(x) \) is defined as

\[
F(x) = P(X \leq x) \tag{5.3}
\]
If the set consists of continuous random events, the probability density function is defined as
\[ p(x) = \frac{dF(x)}{dx}. \] 
(5.4)

This leads to the important relationship
\[ P(\alpha < x \leq \beta) = \int_\alpha^\beta p(x)dx = F(\beta) - F(\alpha). \] 
(5.5)

For a n-dimensional random vector \((X_1, X_2, \ldots, X^n)\), the joint cumulative distribution function is defined as
\[ F(x_1, x_2, \ldots, x^n) = P(X^i \leq x^i \text{ for all } i = 1, \ldots, n) \] 
(5.6)

and the joint probability density function
\[ p(x_1, x_2, \ldots, x^n) = \frac{\partial^n F(x_1, x_2, \ldots, x^n)}{\partial x_1 \partial x_2 \ldots \partial x^n}. \] 
(5.7)

Correspondently, we have the relationship
\[ P(X \in D) = \int_D p(x_1, x_2, \ldots, x^n)dx_1dx_2\ldots dx^n \] 
(5.8)

where \(D\) is a subset of the domain \(\Omega\) of the random vector.

For a general case, given a random variable \(X\) with values in an arbitrary domain \(\Omega\), the probability measure (or probability distribution) is a measure function \(F\) such that
\[ F(D) = P(X \in D) \] 
(5.9)

for any measurable set \(D \subseteq \Omega\). Here, \(P(\Omega) = 1\). The corresponding probability density function is defined as the Radon-Nikodym derivative
\[ p(x) = \frac{dF(x)}{d\mu}, \] 
(5.10)

where \(\mu\) is the measure on \(\Omega\). Thus, the probability that \(X \in D\) can be obtained by integrating \(p(x)\) over the given region \(D\)
\[ F(D) = \int_D p(x)d\mu(x). \] 
(5.11)

Consider two random variables \(X\) and \(Y\) such that \(X \in \Omega_1\), \(Y \in \Omega_2\), \(\Omega = \Omega_1 \otimes \Omega_2\), and \((X, Y) \in \Omega\). Let \(F(D)\) be the joint probability for any measurable subset \(D \subseteq \Omega\) with \((X, Y) \in D\), the joint probability density function \(p(x, y)\) satisfies
\[ F(D) = \int_D p(x, y)d\mu_1(x)d\mu_2(y), \] 
(5.12)

where \(\mu_1\) and \(\mu_2\) are the measures on \(\Omega_1\) and \(\Omega_2\), respectively. For convenience, we drop the measure function notation, and simply write
\[ F(D) = \int_D p(x, y)dxdy. \] 
(5.13)

The marginal density function of \(X\) is defined as
\[ p(x) = \int_D p(x, y)dy, \] 
(5.14)

while the conditional density function \(p(y | x)\) is defined as
The marginal density function of $Y$ is similar to Eq. (5.14), leading to the following property
\[
p(x, y) = p(y | x) p(x) = p(x | y) p(y) .
\] (5.16)

### 5.1.2. Expected Value and Variance
If $Y = g(X)$ and $X$ is a random variable, the expected value or expectation of $Y$ is
\[
E[Y] = \sum_i P(x_i) g(x_i) ,
\] (5.17)
where its variance is
\[
V[Y] = E[((Y - E[Y])^2].
\] (5.18)

For a continuous random variable, the expectation can be expressed as
\[
E[Y] = \int_{\Omega} g(x) dF(x) = \int_{\Omega} g(x) p(x) d\mu(x) .
\] (5.19)

It is easily proved that the expectation and variance have the following properties
\[
E[aY] = aE[Y]
\]
\[
E\left[\sum_{i=1}^{N} Y_i \right] = \sum_{i=1}^{N} E[Y_i]
\] (5.20)
\[
V[aY] = a^2 V[Y]
\]

If the variables $Y_i$ are independent,
\[
V\left[\sum_{i=1}^{N} Y_i \right] = \sum_{i=1}^{N} V[Y_i].
\] (5.21)

From Eq. (5.20), we can derive a simpler expression for the variance
\[
\] (5.22)

The standard derivation of a random variable is defined as the square root of the variance
\[
\sigma[Y] = \sqrt{V[Y]},
\] (5.23)
which is also known as the RMS error.

The conditional expectation of a random variable $G$, $G = g(X, Y)$, is defined as
\[
E[G|X] = \int_{\Omega_2} g(x, y) p(y | x) dy = \frac{\int_{\Omega_2} g(x, y) p(x, y) dy}{\int_{\Omega_2} p(x, y) dy} .
\] (5.24)

We will use the notation $E_y[G]$ for $E[G|X]$ in the following for the convenience. The definition of the conditional variance $V_y[G]$ follows the same idea:
\[
\] (5.25)

It is easy to prove that
\[
V[G] = E_x[V_y[G]] + V_x[E_y[G]].
\] (5.26)
5.2. Sampling Methods

In Monte Carlo methods, it is usually required that random variables be drawn from the distribution functions that define the process. For example, to evaluate the integral
\[ \int f(x)g(x)dx , \]
values of \( x \) must be drawn from \( f(x) \) and the average value of \( g(x) \) over a set of such \( x \) calculated. The procedure of drawing values of \( x \) from \( f(x) \) is often called the “sampling \( x \) from \( f(x) \)”.

To be specific, consider some space \( \Omega \) and \( x \in \Omega \) with a probability density function \( f(x) \), where
\[ \int_\Omega f(x)dx = 1 . \] (5.27)
The sampling procedure is to produce a sequence of values of \( x \) (“random variables”) \( x_1, x_2, \ldots \) such that for any \( \Omega \in \Omega \),
\[ P(x_i \in \Omega) = \int_\Omega f(x)dx \leq 1 . \] (5.28)
For a one-dimensional distribution defined on \((0, 1)\), this means that
\[ P(x_i \in (a,b)) = \int_a^b f(x)dx , \quad 0 < a < b < 1 . \] (5.29)
For small values of \( b - a = dx \),
\[ P(x_i \in dx) = f(x)dx . \] (5.30)

In practice, it is easy to generate random numbers with a uniform distribution using a computer. The sampling procedure is equivalent to creating the random numbers with the desired distribution from the random numbers with a uniform distribution.

5.2.1. Transformation of Random Numbers

Consider a continuous non-decreasing function \( y = y(x) \) where variable \( x \) and function \( y(x) \) map to each other,
\[ y(X) \leq y(x) \text{ iff } X \leq x . \] (5.31)
The probabilities for variable \( x \) and function \( y = y(x) \) to take values of \( X \leq x \) and \( y(X) \leq y(x) \) have the following relation:
\[ P(y(X) = Y \leq y(x)) = P(X \leq x) . \] (5.32)
Correspondently, the cumulative distribution function \( F_y(y) \) of function \( y = y(x) \) has the relation with the cumulative distribution function \( F_x(x) \) of variable \( x \) as
\[ F_y(y) = F_x(x) . \] (5.33)
Using the definition of probability density given in Eq. (5.4), we can obtain
\[ p_y(y) \frac{dy}{dx} = p_x(x) . \] (5.34)
Suppose that \( y(x) \) is a non-increasing function, then
\[
P\left(y(X) = Y \leq y(x)\right) = P\left(X \geq x\right) = 1 - P(X < x) .
\]
(5.35)
Since \( P(X \geq x) + P(X < x) = 1 \), \( F_y(y) \) and \( F_x(x) \) have the following relation
\[
F_y(y) = 1 - F_x(x) .
\]
(5.36)
Therefore,
\[
p_y(y) \frac{dy}{dx} = -p_x(x) .
\]
(5.37)
In both cases, we obtain
\[
p_y(y) \left| \frac{dy}{dx} \right| = p_x(x) .
\]
(5.38)
Furthermore,
\[
p_y(y) = p_x(x) \left| \frac{dy}{dx} \right|^{-1} = p_x \left( x(y) \right) \left| \frac{dx}{dy} \right|
\]
(5.39)
and
\[
F_y(y) = \int_0^y p_y(u) du .
\]
(5.40)
Let us consider how to sample function \( y = y(x) \). The cumulative distribution function can be determined from Eq. (5.33). If \( \xi \) is uniform, its cumulative distribution function is
\[
F_\xi(\xi) = \begin{cases} 
0, & \xi < 0 \\
\xi, & 0 \leq \xi \leq 1 \\
1 & \xi \geq 1
\end{cases}
\]
(5.41)
Therefore on \((0,1)\), the cumulative distribution function for \( y \) is determined by solving the equation
\[
F_y(y) = \xi
\]
(5.42)
for \( y \). It is easy to generate the random number \( \xi \) with a computer.

5.2.2. Numerical Transformation
In practice, it might be impossible to obtain the analytic solution of Eq. (5.42) for some functions \( y = y(x) \). In this case, we have to seek for an approximation solution. A typical approximation method is to calculate the cumulative distribution function for some specific values of \( y \) in advance,
\[
F(y_n) = \int_{y_0}^{y_n} f_y(y) dy = \frac{n}{N}, \quad n = 0, 1, ..., N,
\]
(5.43)
where \( y_0 = 0 \) and \( y_n = Y \). To solve Eq. (5.42) approximately, we can find \( n \) such that
\[
\frac{n}{N} < \xi < \frac{n+1}{N} .
\]
(5.44)
The value for \( y(\xi) \) may be calculated by the linear interpolation
\[ y(\xi) = y_n + (y_{n+1} - y_n)u, \]

where \( u = N\xi - n \) and \( 0 < u < 1 \).

### 5.2.3. Sampling Discrete Distributions

Suppose that we have a class of events \( E_k \) with probabilities \( f_k \), and we wish to sample the events at random. Since \( \sum f_k = 1 \), it is possible to take the interval \((0,1)\) and exhaust it by dividing it into segments, each with a length equal to some \( f_i \) (Figure 5.1). Then we generate a uniform variable \( \xi \), and find the smallest \( i \) so that the sum of \( f_k \) is greater than the random number, i.e.

\[
\sum_{k=0}^{i-1} f_k < \xi \leq \sum_{k=0}^{i} f_k.
\]

![Figure 5.1: Dividing the interval (0,1) into segments of length \( f_i \).](image)

### 5.2.4. Composition of Random Variables

Transforming or mapping random variables may lead to unpleasantly complicated equations to solve. Another technique for generating random variables with the desired distribution is to take two or more different (usually independent) random variables drawn from known distributions and combine them in a proper way.

First, consider a sample of sampling the sum of two uniform random variables. Let \( x \) and \( y \) be uniform on \((0,1)\), and \( z = x + y \); then

\[
F(z) = P(x + y < z)
\]

is the area under the line \( z = x + y \) within the unit square, as shown in Figure 5.2.

![Figure 5.2: The sum of two random variables.](image)
For $x + y < 1$, as shown in Figure 5.3(a), $F(z)$ is geometrically seen to be the area of the triangle with sides equaling $z$:

$$F(z) = \frac{z^2}{2}.$$  \hspace{1cm} (5.48a)

For $x + y > 1$, as shown in Figure 5.3(b), $F(z)$ is

$$F(z) = 1 - \frac{1}{2}(2 - z)^2.$$  \hspace{1cm} (5.48b)

The corresponding probability density function is

$$f(z) = \begin{cases} z, & 0 < z < 1 \\ 2 - z, & 1 \leq z < 2 \end{cases}.$$  \hspace{1cm} (5.49)

Second, consider a sample of sampling a random variable raised to a power. Let $x_1, x_2, \ldots, x_n$ be drawn independently from the cumulative distribution functions $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$. Set $z$ to be the largest of the $x_i$,

$$z = \max\{x_1, x_2, \ldots, x_n\}.$$  \hspace{1cm} (5.50)

Then the following statement holds:

$$Z = \max\{x_1, x_2, \ldots, x_n\} \leq z$$  \hspace{1cm} (5.51)

if and only if $x_1 \leq z$ and $x_2 \leq z$ ... and $x_n \leq z$. Since $x_i$ are all independently distributed,

$$P\{Z \leq z\} = P\{x_1 \leq z\} P\{x_2 \leq z\} \ldots P\{x_n \leq z\}$$  \hspace{1cm} (5.52a)

or

$$P\{Z \leq z\} = \prod_{i=1}^{n} F_i(z).$$  \hspace{1cm} (5.52b)

Suppose that $z$ equals the maximum of $k$ uniform random variables

$$z = \max\{\xi_1, \xi_2, \ldots, \xi_n\},$$  \hspace{1cm} (5.53)

then

$$F(z) = \prod_{i=1}^{k} F_i(z) = z^k.$$  \hspace{1cm} (5.54)
Correspondently, the probability density function is
\[ f(z) = k z^{k-1}, \quad 0 < z < 1. \]  
(5.55)

Third, consider a sample of sampling the distribution \( f(z) = z(1-z) \). Let \( z \) equal the middle value of three random numbers,
\[ z = \text{mid}(\xi_1, \xi_2, \xi_3), \]  
(5.56)
then the probability density function for \( z \) is
\[ f(z) = 6z(1-z). \]  
(5.57)
This result can be proved as follows. Assume that \( \xi_1 < \xi_2 < \xi_3 \) so that \( z = \xi_2 \). The probability that \( \xi_2 \) is in the range \( d\xi_2 \) and that \( \xi_1 < \xi_2 \) and \( \xi_2 < \xi_3 \) is
\[ f(\xi_2) d\xi_2 = d\xi_2 P(\xi_1 < \xi_2) P(\xi_3 > \xi_2) = d\xi_2 \xi_2 (1-\xi_2). \]  
(5.58)
Since
\[ \int_0^1 f(\xi) d\xi = \int_0^1 \xi (1-\xi) d\xi = \frac{1}{6}, \]  
(5.59)
then
\[ f(z) = 6z(1-z). \]  
(5.60)

Finally, consider a sample of sampling the sum of several arbitrary distributions. The expression is given as
\[ f_x(x) = \sum_n \alpha_n g_n(x) \]
\[ \alpha_n \geq 0, \quad g_n(x) \geq 0 \]  
(5.61)
and \( f_x(x) \) satisfies the condition
\[ \int f_x(x) dx = 1. \]  
(5.62)
Eq. (5.61) can be renormalized so that
\[ f_x(x) = \sum_n \alpha_n \left[ \int g_n(u) du \right] \frac{g_n(x)}{\int g_n(w) dw} = \sum_n \beta_n h_n(x) \]  
(5.63)
where
\[ \beta_n = \alpha_n \int g_n(u) du \]  
(5.64)
and
\[ h_n(x) = \frac{g_n(x)}{\int g_n(w) dw}. \]  
(5.65)
Now \( \beta_n \) and \( h_n(x) \) satisfy
Here $\beta_n$ is the probability for the choice of an event $n$. Let us select event $m$ with probability $\beta_m$. Then sample $X$ from $h_m(x)$ for that $m$. The probability that $m$ is chosen and $X \leq x$ is 

$$P_{\beta}{X \leq x} = \sum_{m} \beta_m \int_{0}^{x} h_m(t)dt.$$ 

(5.67)

5.2.5. Rejection Technique

The rejection technique has the advantage that it works for sampling any probability distribution, but it has the disadvantage that the computational efficiency may be low. Let a random variable $Z$ have a probability density function $g(z)$. This $Z$ is accepted if we create a random variable $\xi_2$ and $\xi_2 \leq h(Z) < 1$; otherwise sample another $Z$. There are two discrete events: success ($Z$ is accepted) and not success ($Z$ is rejected). The joint probability that $Z < z$ and $\xi_2 \leq h(Z)$ is 

$$P{Z < z \text{ and } \xi_2 \leq h(Z)} = \int_{-\infty}^{z} h(t)g(t)dt,$$ 

(5.68)

where $h(t)$ is the probability of success given $t$ and $g(t)dt$ is the probability that $t$ is in interval $dt$.

We might write the joint probability as the product of a marginal probability for success and a conditional probability that $Z < z$:

$$P{Z < z \text{ and success}} = P{\text{success}}P{Z < z | \text{success}}.$$ 

(5.69)

Since $P{Z < \infty} = 1$, we obtain

$$P{\text{success}} = \int_{-\infty}^{\infty} h(z)g(z)dz.$$ 

(5.70)

The rejection technique yields a $Z$ only when a success occurs, that is,

$$P{Z < z | \text{success}} = \frac{\int_{-\infty}^{z} h(t)g(t)dt}{\int_{-\infty}^{\infty} h(z)g(z)dz}.$$ 

(5.71)

Therefore, the probability density distribution resulting from the rejection technique is

$$f(z) = \frac{h(z)g(z)}{\int_{-\infty}^{\infty} h(t)g(t)dt}.$$ 

(5.72)

If we define a priori probability of a success $\varepsilon$ as

$$\varepsilon = P{\text{success}} = \int_{-\infty}^{\infty} h(z)g(z)dz,$$ 

(5.73)

the expected number of trials up to and including a success per accepted random variable is

$$\sum_{k=0}^{\infty} (k+1)(1-\varepsilon)^k \varepsilon = \frac{1}{\varepsilon}.$$ 

(5.74)
If the arbitrary probability density function \( f(z) \) is given, we can set function \( h(z) \) to be
\[
    h(z) = \frac{f(z)}{g(z)} \frac{1}{B_h}, \tag{5.75}
\]
where \( B_h \) is an upper bound for \( f(z)/g(z) \). Therefore, \( h(z) \leq 1 \). To sample \( f(z) \), we create a random variable \( \xi_1 \), and let \( z_0 = \xi_1 \). Then we create another random variable \( \xi_2 \), if \( \xi_2 \leq h(z_0) \), we accept \( \xi_1 \); else, we reject it. The efficiency for this choice of \( h(z) \) becomes
\[
    \varepsilon = \frac{1}{B_h} \int_{-\infty}^{\infty} \left( \frac{f(z)}{g(z)} \right) g(z) \, dz = \frac{1}{B_h}. \tag{5.76}
\]

In implementation, \( B_h \) should be the least upper bound for \( f(z)/g(z) \) to maximize the efficiency.

### 5.2.6. \textit{M(RT)}\textsuperscript{2} Algorithm

This sampling method is an advanced technique introduced by Metropolis et al.\textsuperscript{12} In statistical mechanics, the equilibrium is achieved when the statistical properties are independent of the kinetics of the system. The statistical properties are the averages of the macroscopic mechanical parameters over all the possible states. One condition that a system evolves toward equilibrium mechanics, the equilibrium is achieved when the statistical properties are independent of the kinetics of the system. The statistical properties are the averages of the macroscopic mechanical parameters over all the possible states. One condition that a system evolves toward equilibrium is known, and one has the task of finding \( f(X) \). The \textit{M(RT)}\textsuperscript{2} algorithm reverses this: one has the task of finding a convenient and correct kinetics that will equilibrate the system so that the given \( f(X) \) turns out to be the chance of observing the system in the state \( X \).

In treating a physical system, one usually assumes that \( K(X \mid Y) \) is known, and one has the task of finding \( f(X) \). The \textit{M(RT)}\textsuperscript{2} algorithm reverses this: one has the task of finding a convenient and correct kinetics that will equilibrate the system so that the given \( f(X) \) turns out to be the chance of observing the system in the state \( X \).

This turns out to be extremely easy given the elegant device suggested by \textit{M(RT)}\textsuperscript{2}. Transitions are proposed from, say, \( Y \) to \( X' \) using essentially any distribution \( T(X' \mid Y) \). Then on comparing \( f(X') \) with \( f(Y) \) and taking into account \( T \) as well, the system is either moved to \( X' \) (move "accepted") or returned to \( Y \) (move "rejected"). Acceptance of the move occurs with the probability \( A(X' \mid Y) \), which must be calculated so as to satisfy detailed balance. We then have
\[
    K(X \mid Y) = A(X \mid Y) T(X \mid Y) \tag{5.78}
\]

Detailed balance requires
\[
    A(X \mid Y) T(X \mid Y) f(Y) = A(Y \mid X) T(Y \mid X) f(X) \tag{5.79}
\]

We expect that the ratio
\[
    q(X \mid Y) = \frac{T(Y \mid X) f(X)}{T(X \mid Y) f(Y)} \tag{5.80}
\]
will play a significant role in determining \( A \).
At each step in the random walk, we calculate the quantity $q(X_1)$. The technique establishes a random walk whose steps are designed so that when repeated again and again, the asymptotic distribution of $X$'s is $f(X)$. Suppose that $X_1, X_2, \ldots, X_n$ are the steps in a random walk. Each of the $X$'s is a random variable and has an associated probability $\phi(Y)$, where $\phi(Y)$ can be any distribution for $X$. The $\phi_n(Y)$ has the property that asymptotically

$$\lim_{n \to \infty} \phi_n(Y) = f(Y).$$

(5.81)

At each step in the random walk, we calculate the quantity $q(X \mid Y)$. From $q(X \mid Y)$, the probability of accepting a move can be calculated. One typical method to calculate $A(X \mid Y)$ is

$$A(X \mid Y) = \min(1, q(X \mid Y)).$$

(5.82)

Then the algorithm can now be described concretely. At step $n$ of the random walk, the value of $X$ is $X_n$; a possible next value for $X$, $X_{n+1}$, is sampled from $T(X_{n+1} \mid X_n)$, and the probability of accepting $X_{n+1}$ is computed. If $q(X_{n+1} \mid X_n) > 1$ then $A(X_{n+1} \mid X_n) = 1$; if $q(X_{n+1} \mid X_n) < 1$, then $A(X_{n+1} \mid X_n) = q(X_{n+1} \mid X_n)$, where

$$q(X_{n+1} \mid X_n) = \frac{T(X_{n+1} \mid X_n) f(X_{n+1})}{T(X_{n+1} \mid X_n) f(X_n)}.$$ 

(5.83)

We create the random variable $\xi$, if $A(X_{n+1} \mid X_n) > \xi$, we set $X_{n+1} = X_{n+1}$; otherwise, we set $X_{n+1} = X_n$. This procedure contains an element of rejection; however, if a $X_{n+1}$ is not accepted, we use the previous value rather than sample a new value.

In a Monte Carlo calculation, we are often trying to evaluate quantities of the form

$$G = \int g(X) f(X) dX \quad \left/ \int f(X) dX \right.$$

(5.84)

Using the M(RT)$^2$ algorithm, the integral can be evaluated as

$$G = \int g(X) f(X) dX = \sum_{n=L}^{L+N-1} \frac{g(X_n)}{N}.$$ 

(5.85)

Here the evaluation begins from the $L$ step. This is due to the fact that M(RT)$^2$ technique reaches the desired distribution asymptotically, as shown in Eq. (5.81). We discard the previous $L-1$ steps for accurate calculations. In practice, the value of $L$ is determined experimentally.

### 5.2.7. Importance Sampling

The integral evaluated by Monte Carlo method usually has the following form

$$G = \int_{x_0} g(X) f(X) dX,$$

(5.86)

where

$$f(X) \geq 0, \int_{x_0} f(X) dX = 1.$$
The function $f(X)$ is not necessarily the best probability density function to use in the Monte Carlo calculation even though it appears in the integrand. A different probability density function, $\tilde{f}(X)$, can be introduced into the integral as follows:

$$G = \int \left[ \frac{g(X)f(X)}{\tilde{f}(X)} \right] \tilde{f}(X) dX,$$

where

$$\tilde{f}(X) \geq 0, \quad \int \tilde{f}(X) dX = 1,$$

and $g(X)f(X)/\tilde{f}(X) < \infty$ except perhaps on a (countable) set of points. The variance of $G$ when $\tilde{f}(X)$ is used becomes

$$\text{var}[G] = \int \left[ \frac{g^2(X)f^2(X)}{\tilde{f}^2(X)} \right] \tilde{f}(X) dX - G^2.$$

Since $G^2$ is fixed, we can minimize $\text{var}[G]$ by finding the appropriate $\tilde{f}(X)$ so that the quantity $\int [g^2(X)f^2(X)/\tilde{f}(X)] dX$ is minimized. Using Lagrange multiplier, we obtain

$$\tilde{f}(X) = \frac{g(X)f(X)}{G}.$$

If we already know the correct answer $G$, the Monte Carlo calculation will certainly give it back with zero variance! However, if $G$ is already known, we do not need to evaluate the integral. In practical application, we always try to find the similar function to reduce the variance instead of Eq. (5.89). Here is an example,

$$G = \int \cos \left( \frac{\pi x}{2} \right) dx.$$

A straightforward Monte Carlo algorithm would be to sample $x$ uniformly on $(0,1)$, ($f_1(x) = 1$), and to sum the quantity $\cos(\pi x/2)$. The variance is evaluated as 0.0947. By expanding $\cos(\pi x/2)$ in a power series, a better choice for the importance function may be found,

$$\cos \left( \frac{\pi x}{2} \right) = 1 - \frac{\pi^2 x^2}{8} + \frac{\pi^4 x^4}{2^4 4!} - \ldots$$

and we can let

$$\tilde{f}(x) = \frac{1}{2} (1 - x^2).$$

Then the estimator for $G$ is now

$$\tilde{g} = \frac{g}{\tilde{f}} = \frac{2 \cos(\pi x/2)}{3 \cdot 1 - x^2}$$

and the variance associated with a single sample is 0.00099.

**5.2.8. Splitting and Russian Roulette**

This method is often used in the deep-penetration problems. Suppose we have an optically thick slab with the thickness $T$, with a source of particles prescribed either at or near the plane $z = 0$. At the specified plane boundaries $z = z_i$ in the increasing $z$ direction, assume that the particle
splits into \( \nu \) identical particles, each of weight \( W/\nu \), where \( W \) is the incoming weight. The weight is always preserved, and we process more particles with smaller weights. Although \( \nu \) may frequently be an integer, it need not be. All that is necessary is that the expected number of split particles be \( \nu \); if \( n < \nu < n+1 \), a common recipe is to choose \( n \) particles with the probability \( n+1-\nu \), and \( n+1 \) particles with the probability \( \nu - n \).

If it is desirable to split particles when they penetrate deeper into a shield, or, in general, when they enter a more important region of the problem, then it is usually sound practice to decrease the number of particles followed when the they enter a less desirable region. This can be achieved by Russian roulette. In our example, if a particle proceeds across the boundary \( z = z_i \) in the direction of decreasing \( z \), Russian roulette allows the particle to survive with probability \( \nu^{-1} \) and its weight to be increased by a factor \( \nu \). The particle is killed with probability \( 1-\nu^{-1} \).

In complex geometries each cell of geometric region of the problem can be assigned an importance \( I \). Then, when a particle enters cell \( n+1 \) from cell \( n \), the ratio \( I_{n+1}/I_n \) is examined. If \( I_{n+1}/I_n > 1 \), so that the particle is entering cell of greater importance, then the particle is split into \( \nu = I_{n+1}/I_n \) identical particles, each of weight \( W\nu^{-1} \), where \( W \) is the incoming weight. If \( \nu = I_{n+1}/I_n < 1 \), then Russian roulette is played, and the particle survives with probability \( \nu \) with the weight \( W\nu^{-1} \).

### 5.3. Simulation of Radiation Transport

The transport of radiation is a natural stochastic process that is amenable to Monte Carlo modeling. The particular simulation is generally a simplification of the naturally radiation process, such as the radiations of light, X-rays, neutrons, and neutrinos. Generally seven or eight variables may be used to specify uniquely the state of radiation. The position coordinates may be taken as the Cartesian coordinates \((x, y, z)\). The direction of radiation can be specified by the direction cosines, \((\Omega_x, \Omega_y, \Omega_z)\). An energy or equivalent variable must be given. In scattering visible light by an atom, the energy of the photon hardly changes. Also, in dental X-rays, it is often the total dosage of radiation is of interest, not the dosage as a function of time.

Once the system to be simulated has been clearly defined and the type of information wanted from the simulation is known, the structure of an appropriate code can be devised. For the simulation of dental X-rays, a possible structure is as follows:
1. Pick a set of source variable (initial state of system).
2. Follow the X-ray until it interacts with an atom.
3. Determine whether the X-ray scatters.
   a. If so, repeat from step 2.
   b. If not, terminate the history.
Steps 2 and 3 are repeated until the X-ray photon is absorbed or is no longer capable of affecting the answer to any appreciable extent.
4. Repeat the whole process from step 1 as many times as necessary to achieve the accuracy needed for the solution.
5. Take arithmetic average of answers of all the histories.

5.3.1. Characterization of the Source

In general, there exists a probability density function for the production of radiation per unit volume, per unit direction, per unit energy, and per unit time:

\[ S(x, y, z, \Omega_x, \Omega_y, \Omega_z, E, t). \quad (5.94) \]

We shall assume here that the source probability density function can be factored into a product of the source's position, direction, energy, and time:

\[ S = S_r(x, y, z)S_\Omega(\Omega)xS_E(E)S_t(t). \quad (5.95) \]

In the simulation, we usually assume that the source is a point source at the position \((x_0, y_0, z_0)\), and each photon has the same energy \(E_0\). The variable \((x, y, z)\) and \(E\) are then perfectly determined. The appropriate source's position and energy are given as

\[ S_r = \delta(x-x_0)\delta(y-y_0)\delta(z-z_0) \]
\[ S_E = \delta(E-E_0) \quad (5.96) \]

The source can be assumed to be time independent, so \(S_r(t)\) is a constant. If we consider that radiation is uniform to all the directions, then \(S_\Omega(\Omega) = 1/4\pi\).

5.3.2. Tracing a Path

Since photons travel along straight lines until interactions, the change in position coordinates and time can be written down immediately as

\[
\begin{align*}
x &= x_0 + \Omega_x \cdot S \\
y &= y_0 + \Omega_y \cdot S \\
z &= z_0 + \Omega_z \cdot S \\
t &= t_0 + S/v
\end{align*}
\]

where \(v\) is the radiation velocity, and \(S\) is the distance to the next interaction with an atom. The probability per unit path length of having an interaction is \(\rho\sigma_i\), where \(\rho\) is the scatter density, and \(\sigma_i\) is the scattering cross-section. The probability \(U(S | \Omega, E, r)\) is a cumulative distribution for the first interaction at \(S' > S\), upon the initial values of \(\Omega, E,\) and \(r\). The value of \(U\) at some \(S_i\) can be written as

\[ U(S_i) = U(S_i) + P\{ S_i \geq S > S_1 \} \text{ for } S_i > S_1, \quad (5.98) \]

and \(P\{\ldots\}\) is the probability that an interaction occurred between \(S_i\) and \(S_2\). Eq. (5.98) may be rewritten as

\[ U(S_i) - U(S_2) = U(S_i) + P\{ S_2 \geq S > S_1 | S > S_i \}, \quad (5.99) \]

where \(P\{\ldots\}\) is now the conditional probability that an interaction occurred if \(S > S_i\). For small values of \(S_2 - S_i\),

\[ U(S_i) - U(S_2) = U(S_i)\rho(S_i)\sigma_i(S_i)(S_2 - S_i) + O(S_2 - S_i)^2. \quad (5.100) \]

Upon taking the derivative of Eq. (5.100) with respect to \(S_2\), we find

\[ -U''(S) = U(S)\rho(S)\sigma_i(S). \quad (5.101) \]
Therefore, we can obtain

\[ U(S) = \exp \left[ -\int S \rho(S') \sigma(S') dS' \right]. \tag{5.102} \]

In a homogeneous medium,

\[ U(S) = \exp \left[ -\rho \sigma S \right]. \tag{5.103} \]

For the purposes of the simulation, we need to decide how to sample the next event. We can sample for a next event by equating a uniform random variable to a distribution function:

\[ 1 - U(S) = \xi = 1 - \xi. \tag{5.104} \]

If the variable \( \xi \) is uniformly distributed, then \( 1 - \xi \) is also uniformly distributed. From Eq. (5.104), we can obtain

\[ S = -\log \frac{\xi}{\rho \sigma}. \tag{5.105} \]

### 5.3.3. Modeling Collision Events

Suppose only two events can occur upon collision: the radiation absorbed by the scatter can be converted into the heat energy, or it is dissipated locally. The second is the scattering of the radiation by the scatter. We can generate a random variable \( \xi \). If

\[ \xi < \text{prob} \{ \text{absorption} \}, \] absorption occurs; \( \tag{5.106} \)

otherwise a scattering event occurs. If scattering happens, then the path of the radiation is continued as before but the direction of flight might be changed, allowing the collisions to occur at random until the radiation is absorbed or leaves the area of interest.

![Figure 5.4: The coordinates for the scattering.](image)

The new direction of the radiation is chosen by sampling \( \cos \theta \) from the scattering phase function and then sampling \( \phi = 2\pi \xi \), as shown in Figure 5.4. The old direction cosines were \((\Omega'_x, \Omega'_y, \Omega'_z)\); the new direction cosines are given by

\[
\Omega'_x = \frac{\sin \theta}{\sqrt{1 - \Omega'^2}} \left[ \Omega'_y \sin \phi - \Omega'_z \Omega'_x \cos \phi \right] + \Omega'_x \cos \theta
\]

\[
\Omega'_y = \frac{\sin \theta}{\sqrt{1 - \Omega'^2}} \left[ -\Omega'_x \sin \phi - \Omega'_z \Omega'_y \cos \phi \right] + \Omega'_y \cos \theta \tag{5.107}
\]

\[
\Omega'_z = \sin \theta \sqrt{1 - \Omega'^2} \cos \phi + \Omega'_z \cos \theta
\]

The set \((\Omega_x, \Omega_y, \Omega_z)\) is not unique. Eq. (5.107) results from a particular choice of the origin of \( \phi \), but does satisfy
Eqs. (5.107) are not stable numerically, and the normalization given in Eq. (5.108b) tends to drift from 1 after repeated usage of Eqs. (5.107). The \((\mathbf{r}, \mathbf{S}, \mathbf{Z})\) must be periodically renormalized.

The von Neumann rejection technique can be profitably used to choose \(\sin \phi\) and \(\cos \phi\), especially since they occur in association with a square root. If by chance \(\Omega' = 1\), Eq. (5.107) becomes indeterminate; this can be overcome by cyclic permutations of \(x, y, z\) in the equations. This change is also worth carrying out if \(\Omega'_z\) is close enough to 1 to produce significant round-off error.
6 Illumination Models of Subsurface Scattering

In computer graphics, subsurface scattering and light transport in volumetric media are two major factors in creating realistic images. Examples of common real-world translucent media include biological materials such as leaves, skins, and fish, and non-biological materials such as clouds, water, marbles.

Accurate illumination models for translucent media (media that contain scattering particles in subsurface layers or volume) should be based on the physics of the optical scattering and transport process. The modeling and validation often involves the components light scattering theory, radiation and transport theory, and Monte Carlo methods. We have reviewed these components in earlier sections.

In this section, our focus will be on the graphics application of light scattering and transport. We will first introduce a Monte-Carlo method called path tracing that simulates multiple-scattering processes. Then we will present a special BRDF model for multi-layered media accounting for subsurface scattering and transport of light. Following that, we will introduce a simulation model where light transport is handled using a volumetric photon mapping technique. After that, we will consider a new model with BSSRDF technique, especially for translucent materials. Finally, we will discuss several analytical approximation models on subsurface scattering and light transport. These techniques and models were developed to describe a soft and smooth appearance of participating media that contains scattering particles such as fog, clouds, and translucent materials.

6.1. Monte-Carlo and Bidirectional Path Tracing

To simulate light transport in a participating medium, the simplest way is to consider direct illumination or single scattering only and ignore multiple scattering. This approximation works for optically thin media such as atmosphere where multiple scattering is not important. For the optically thick media like clouds, Monte-Carlo method is widely used to deal with multiple scattering.

6.1.1. Monte-Carlo Path Tracing

Path tracing was first proposed by Kajiya. It is an extension of the distributed ray-tracing algorithm developed by Cook which uses stochastic sampling to create some special effects such as motion blurs and soft shadows. The path tracing method can simulate all possible light bounces in a model by integrating over the involved space such as light from an area and indirect light reflected from a diffuse surface.

In mathematical sense, path tracing is a continuous Markov Chain random walk for solving the rendering equation, which addresses that the outgoing radiance $L_o$ at any location $x$ in direction $\hat{\Omega}$ on the surface is the sum of the emitted radiance $L_e$ and the reflected radiance $L_r$:

$$L_o(x, \hat{\Omega}) = L_e(x, \hat{\Omega}) + L_r(x, \hat{\Omega}) \quad (6.1)$$

The reflected radiance $L_r$ can be obtained by considering incident radiance $L_i$ from all directions on the sphere around $x$ scaled by as
where $f_{r}$ is the surface BRDF and $\hat{\Omega} \cdot \hat{n}$ is the cosine term. Given a generic scene, it is not feasible to solve analytically the radiance distribution over surfaces. Thus, either an assumption for simplification has to be made or a numerical integration technique has to be used.

The numerical solution can be obtained by Monte Carlo sampling of the Neumann series expressed as

$$L = L_{e} + \Gamma L = L_{e} + \Gamma (L_{e} + \Gamma L) = \cdots = L_{e} + \Gamma^{2} L_{e} + \Gamma^{3} L_{e} + \cdots = \sum_{n=0}^{\infty} \Gamma^{n} L_{e},$$

where the integral operation of the rendering equation is denoted by $\int_{4\pi}$,

$$f_{r} = R f_{r,s} + T f_{r,v} = R f_{r,s} + (1-R) f_{r,v} < \Gamma g > (x, \hat{\Omega}) = \int f_{r}(x, \hat{\Omega}', \hat{\Omega}) g(x, \hat{\Omega}') (\hat{\Omega} \cdot \hat{n}) d\Omega'.$$

The Neumann-series solution can be interpreted as a sum of light reflected of 0, 1, 2, 3... times. (Reflection with 0 time refers to light from a self-emitted surface.) In path tracing, the light distribution is sampled by tracing rays randomly in all light paths that are physically possible. For example, the integral of light over a pixel can be estimated by averaging a set of sample rays that pass through that pixel.

The algorithm is shown in Figure 6.1. There are two major characteristics associated with this algorithm. First, it only involves light rays of one-bounce reflection to estimate the indirect illumination. Second, it considers other relevant factors such as the time and lens position of each ray.
render()
for each pixel
    color = 0
for each sample
    create a ray from the eye through a random position in pixel
    choose a random time and lens position for the ray
    color = color + trace(ray)
pixel_color = color / number of samples

trace(ray)
find the nearest intersection between the ray and the scene and obtain the position and normal vector at the intersection point
color = shade(point, normal)
return color

shade(point, normal)
    color = 0
    for each light source
        if light is visible
            color += direct illumination
            color += trace(a randomly reflected ray)
    return color

Figure 6.1: Pseudocode of the Monte-Carlo path tracing algorithm.

The algorithm of path tracing treats every pixel and every point on surfaces independently. This is particularly useful for a specular surface where the light intensity varies rapidly over the surface. Meanwhile, this approach is limited for diffuse illumination which varies slowly over surfaces. So the first problem of path tracing is that extra effort has to be made to obtain a good estimation for direct illumination and diffuse illumination. The cost of eliminating noise is another problem. As shown in Figure 6.2, the left image is rendered using 10 paths per pixel and the right image using 100 paths per pixel. Although a greater path/pixel ratio provides better rendering than a lower ratio, using 100 paths/pixel is still not adequate to eliminate noise. Moreover, using more paths will increase the complexity and decrease the rendering speed. For example, this issue is important when rendering caustics – the effect of light reflection and refraction by some material where focused and bright light spot is generated and noise is a major issue in modeling.
Figure 6.2: The left image shows the Cornell box scene rendered using 10 paths/pixel and the right shows the same box scene using 100 paths/pixel. (Courtesy of Jensen)

6.1.2. Monte-Carlo Bidirectional Path Tracing

As an extension of path-tracing algorithm, Lafortune and Willems\textsuperscript{22} and Veach and Guibas\textsuperscript{23} independently proposed a bidirectional path tracing method (Fig 6.3). This method traces paths from both the eye and the light sources. It is motivated by the observation that some paths are more easily traced from the eye while others more easily from the light sources. By evaluating the visibility of every vertex-vertex pair (one is from eye path and another from light path), the contributions from these two types of paths can be combined. The weighted sum of contributions from all paths can be determined by

\[
L_p = \sum_{i=0}^{N_i} \sum_{j=0}^{N_j} w_{i,j} I_{r(i,j)}
\]

\[
w_{i,j} = \frac{p_{i,j}^\beta}{\sum_{k=0}^{i+j-k} p_{k,i+j-k}^\beta} \quad \beta = 2
\]

(6.4)

Figure 6.3: Bidirectional path tracing.
where $L_{r(i,j)}$ is the reflected radiance at $x_i$ due to $y_j$, $w_{i,j}$ is the weight, and $p_{i,j}$ is the probability density for generating the path $x_0, x_i, y_0, ..., y_l$. The algorithm is shown in Figure 6.4.

```python
render()
    for each pixel
        color = 0
        for each sample
            pos = random position in pixel
            trace_path(pos)

trace_path(pos)
    trace a ray from the eye through pixel position
    generate an eye path and a light path
    combine(eye path, light path)

combine(eye path, light path)
    for each vertex $y_l$ on the light path
        for each vertex $x_i$ on the eye path
            if $x_i$, $y_l$ are mutually visible
                compute the weight for the $(x_i, y_l)$ path
                add the weighted contribution to the pixel
```

Figure 6.4: Algorithm of bidirectional path tracing.

Generally speaking, when the scene contains small sources and has strong indirect illumination, bidirectional path tracing is more effective than simple path tracing. Simple path tracing works better for the cases when the light sources are easily reached from the eye.

In summary, Monte-Carlo methods are accurate, easy to implement and require less memory. However, this approach is computationally expensive. For example, path tracing needs a significant amount of time to generate sufficient paths to eliminate noise. This makes the Monte-Carlo methods less applicable to rendering caustic phenomena and effects of participating media such as translucent materials.

### 6.2. Hanrahan-Krueger Multi-layers BRDF Model

Hanrahan-Krueger model\(^{14}\) applies to a surface with a series of layers as shown in Figure 6.5. Each layer has a different set of parameters, including the index of refraction, the absorption and scattering coefficients, the layer thickness, and the phase function.
Hanrahan-Krueger model explicitly evaluates the reflection and transmission of light at medium boundaries, and the scattering of light within each layer. The total BRDF $f_r$ is described by a combination of the reflection function $f_{r,s}$ on the outer surface and the internal subsurface scattering $f_{r,v}$, modulated by Fresnel coefficients $R$ and $T$, respectively:

$$f_r = R f_{r,s} + T f_{r,v} = R f_{r,s} + (1 - R) f_{r,v}. \quad (6.5)$$

The algorithm is based on 1D transport model where the light propagation in a participating medium is described by a linear integro-differential equation considering energy balance within a differential volume element: (Note that in this section, $\sigma$ and $\sigma'$ are equivalent to $\rho\sigma$ and $\rho\sigma'$ in Eq. 4.15, respectively.)

$$\frac{\partial L(x, \theta, \phi)}{\partial s} = -\sigma_s L(x, \theta, \phi) + \sigma_s \int p(x; \theta, \phi; \theta', \phi') L(x; \theta', \phi') d\theta' d\phi'. \quad (6.6)$$

The equation shows that the change of radiance along a particular infinitesimal direction $ds$ contains two components: the decreased radiance due to absorption and scattering, and the total radiance scattered in the direction $ds$ from all other directions. For layered media, an assumption is made that all components depend on depth $z$ instead of the path length $s$ ($z = s \cos \theta$). Then the equation can be simplified as

$$L(z; \theta, \phi) = \int e^{-\int_0^z \sigma_s ds'} \int p(z'; \theta, \phi; \theta', \phi') L(z'; \theta', \phi') d\theta' d\phi' \frac{dz'}{\cos \theta}. \quad (6.7)$$

tone may solve this equation using Neumann series

$$L = \sum_{i=0}^\infty L^{(i)}. \quad (6.8)$$

The interpretation is that the total radiance consists of contributions from scattering of 0, 1, 2, … times.
When the layer is very thin or the scattered light energy is much less than the light energy that is absorbed in the layer, it is sufficient to use the low-order approximation to represent the radiance and ignore the high-order scattering. For example, the first-order approximation method may be applied to find the solution to the integro-differential equation in Eq. (6.6).

The first-order approximation assumes that light is attenuated by absorption only, i.e. no scattering. So the intensity $L_{\nu}$ is given by the equation

$$L_{\nu}(z) = L_{\nu}(z = 0)e^{\cos \theta}, \tau(z) = \int_0^z \sigma_t dz.$$  \hspace{1cm} (6.9)

Combining with boundary conditions to measure how much light will pass through the outermost surface:

$$L_{\nu}(z = 0) = T_{12}L_{\nu}(\theta_t, \phi_t).$$ \hspace{1cm} (6.10)

The transmitted intensity $L_{\nu, t}^{(0)}(\theta_t, \phi_t)$ is

$$L_{\nu, t}^{(0)}(\theta_t, \phi_t) = T_{23}L_{\nu}^{(0)}(z) = T_{23}L_{\nu}(z = 0)e^{\cos \theta} = T_{12}T_{23}e^{-\tau_d} L_{\nu}(\theta_t, \phi_t).$$ \hspace{1cm} (6.11)

If $\sigma_t$ is constant, $\tau_d = \sigma_t d$. Substituting the $0^{th}$-order solution into the integral equation, the $1^{st}$-order solution for both forward and backward scattering could be written as

$$L_{\nu, t}^{(1)}(\theta_t, \phi_t) = WT_{12}T_{31} p(\pi - \theta_t, \phi_t; \theta_t, \phi_t) \frac{\cos \theta_t}{\cos \theta_t + \cos \theta_r} (1 - e^{\frac{-\tau_d}{\cos \theta_t + \cos \theta_r}}) L_{\nu}(\theta_t, \phi_t),$$

$$L_{\nu, b}^{(1)}(\theta_t, \phi_t) = WT_{12}T_{31} p(\theta_t, \phi_t; \theta_t, \phi_t) \frac{\cos \theta_t}{\cos \theta_t - \cos \theta_r} (e^{\frac{-\tau_d}{\cos \theta_t} - e^{\frac{-\tau_d}{\cos \theta_r}}) L_{\nu}(\theta_t, \phi_t),$$  \hspace{1cm} (6.12)

where $W$ is the albedo.
1 Initialize: A particle enters the layer at the origin. Initialize \( \vec{p} \) to the origin and the direction \( \vec{s} \) to the direction at which the ray enters the layer. Set the weight \( w = 1 \).

2 Events: Repeat the following steps until the ray weight drops below some threshold or the ray exits the layer.

2A Step: First, estimate the distance to the next interaction:

\[
    d = \frac{-\log r}{\sigma_t}
\]

Where \( r \) in this and the following formulas is a uniformly distributed random number between 0 and 1. Then, compute the new position:

\[
    \vec{p} = \vec{p} + d \vec{s}
\]

And, finally set the particle weight to

\[
    w = w - \frac{\sigma_s}{\sigma_a + \sigma_s}
\]

Note: If \( d \) causes the particle to leave the layer, break from the repeat loop and adjust the weight using the distance to the boundary.

2B Scatter: First, estimate the cosine of the scattering angle for the Henyey-Greenstein phase function using the following formula.

\[
    \cos j = \frac{1}{|2g|} \left( 1 + g^2 - (\frac{1 - g^2}{1 - g + 2gr})^2 \right)
\]

and \( \cos \phi \) and \( \sin \phi \) with \( \phi = 2\pi r \). Then, compute the new direction:

\[
    \vec{t} = \left( \frac{\vec{s} \cdot x \cos \phi \cos \theta - \vec{s} \cdot y \sin \phi}{\sin \theta} \right) /
    \left( \frac{\vec{s} \cdot y \cos \phi \cos \theta + \vec{s} \cdot x \sin \phi}{\sin \theta} \right)
\]

\[
    \vec{s} = \vec{s} \cos j + \vec{t} \sin j
\]

Here, \( \cos \theta = \vec{s} \cdot z \) and \( \sin \theta = \sqrt{1 - \vec{s} \cdot z^2} \). Note: Care must be taken if \( \sin \theta = 0 \).

3 Score: Divide the sphere into regions of equal solid angle and add the weight of the particle to the weight associated with the bin in which it is contained.

Figure 6.6: Monte-Carlo Algorithm for Layered media
(Courtesy of Hanrahan and Kruger).

For multiple layers, the total first-order solution will be the sum of the first-order scattering from each layer, weighted by the percentage of light making it to the layer and returning from the
layer. A Monte Carlo sampling scheme is employed for computing light transport in layered media. The algorithm is described in Figure 6.6.

Figure 6.7 shows an example of a head rendered by Lambert shading and Harahan-Krueger subsurface model. The skin is assumed to have two layers: the outer layer has tissue and pigment particles containing melanin which selectively absorbs light, producing a brown-to-black appearance and scatters light strongly in the forward direction. The inner layer contains blood, absorbs green and blue light, and offers substantial isotropic scattering. The left two columns were rendered using Lambert shading and the two middle columns using the subsurface model. The right columns show the relative differences between the two models. The red color indicates more reflection from the new model and the blue color indicates less reflection.

Different from Hanrahan-Krueger model, Stam et. al.\textsuperscript{24} presented an illumination model accounting for multiple anisotropic scattering in a skin layer bounded by two rough surfaces. To handle rough surfaces, using BRDF and BTDF, the model used a suitable “diagonalization” of the “transfer matrix” to solve the problem. Fourier transformation and eigenanalysis were used to perform the task.

Figure 6.8 shows a human head rendered with Lambert model (left), Hanrahan-Krueger Model (middle) and Stam model (right). The Stam’s model seems to be a blend between Lambert model and Hanrahan-Krueger model. Besides, the differences between this model and the simple Lambert model are noticeable but not very obvious.

Jensen and Christensen\textsuperscript{15} described a new global illumination algorithm to simulate light transport in participating media based on Monte Carlo bidirectional ray tracing and volumetric photon mapping. This algorithm extends the original photon mapping method\textsuperscript{25} to reduce the impact of noise. It is the first method that implemented a full simulation of subsurface scattering inside the media.

Photon mapping is a two-pass method, originally developed for global illumination in scenes without participating media. It decouples the illumination solution from the geometry and the solution is represented in a spatial data structure called the photon map. In the first pass, two photon maps are constructed to store the rendering equation terms calculated separately by emitting photons from the light sources (as shown in Figure 6.9) and tracing these photons through the scene by photon tracing.

The first photon map is a high-quality caustic photon map, including all photons that have been traced from the light source through a number of specular reflections and transmissions before intersecting with a diffuse surface. The second photon map is the global photon map which is less accurate but contains all photons representing indirect illumination on a non-specular surface.
In the second pass, the scene is rendered using the pre-computed data stored in the photon maps based on distributed ray tracing\(^\text{20}\). The photon map is used to estimate the radiance \(L_r\) at a given position \(x\) in the direction \(\Omega\) on the surface using the flux \(\Delta \Phi_p\) carried by each photon \(p\) in the direction \(\hat{\Omega}_p\). By searching a number of photons with the shortest distance to \(x\), the reflected radiance \(L_r\) from a surface can be directly estimated by

\[
L_r(x, \hat{\Omega}) = \sum_{p=1}^{n} f_r(x, \hat{\Omega}, \hat{\Omega}_p) \frac{\Delta \Phi_p(x, \hat{\Omega}_p)}{\pi r^2},
\]

where \(f_r\) is BRDF function and \(r\) is the distance to the \(n\)th nearest photon.

When the participating media effects (such as subsurface scattering for rendering translucent materials or volume caustics) are considered, the in-scattered radiance has two component: direct radiance \(L_{i,d}\) and indirect radiance \(L_{i,i}\). The direct part can be computed by sampling the light source using ray tracing while the indirect part is estimated using volume photon map. To implement this, several techniques need to be employed.

The first technique is concerned with storing the photons. To keep the details such as volume caustics, the photons will be stored in a volume map to separate from the photons on surfaces. The volume map is used to compute the illumination inside a participating medium, while the global photon map is used to compute the illumination on surfaces. Since the direct illumination of participating media is easy to compute using traditional ray tracing techniques and the photon maps are used to calculate the indirect illumination, we only account for those photons which have been reflected or transmitted by surfaces before interacting with the media.

In the first pass, photon maps are constructed by the second technique called photon tracing. It works exactly the same way as ray tracing, except that the flux is propagated by photons whereas in ray tracing the radiance is gathered by rays. When a photon hits an object, it can be absorbed, reflected or transmitted. If the photon interacts with the medium and does not come directly from a light source, it will be stored in the photon map. A stochastic technique called Russian roulette\(^\text{26}\) is used to decide whether the photon is scattered or absorbed in order to remove unimportant photons. The probability of the photon being scattered is given by the scattering albedo \(W = \sigma_s(x) / \sigma_r(x)\). Hence, the in-scattered radiance can be described as

\[
L_i(x, \hat{\Omega}) = L_{i,d}(x, \hat{\Omega}) + \frac{\sigma_s(x)}{\sigma_r(x)} L_{i,i}(x, \hat{\Omega}).
\]

Photons stored in the volume photon map are used to estimate the radiance. However, the radiance cannot be computed directly using the radiance estimate for surface as shown in Eq. (6.13). In view of the relation between the scattered flux \(\Phi\) and radiance \(L\) in a participating medium

\[
L(x, \hat{\Omega}) = \frac{d^2 \Phi(x, \hat{\Omega})}{\sigma_s(x) d\Omega dV}.
\]

Substituting the radiance with Eq. (6.13), the in-scattered radiance can be obtained by
where $dV$ is the differential volume containing the photons which is approximated by $\frac{4}{3} \pi r^3$, corresponding to the smallest bounding sphere containing the $n$ nearest photons. Then the indirect radiance can be computed by

$$L_i(x, \hat{\Omega}) = \int f(x, \hat{\Omega}, \hat{\Omega}) L(x, \hat{\Omega}) d\Omega'$$

$$= \int f(x, \hat{\Omega}, \hat{\Omega}) \frac{d^2 \Phi(x, \hat{\Omega})}{\sigma_s(x) d\Omega'} d\Omega'$$

$$= \frac{1}{\sigma_s(x) \pi} \int f(x, \hat{\Omega}, \hat{\Omega}) \frac{d^2 \Phi(x, \hat{\Omega})}{dV}$$

$$\approx \frac{1}{\sigma_s(x)} \sum_{p=1}^{n} f_r(x, \hat{\Omega}_p, \hat{\Omega}) \frac{\Delta \Phi_{p,i}(x, \hat{\Omega}_p)}{4 \pi r^3}$$

(6.16)

where $\Delta \Phi_{p,i}$ is the flux carried by the photons corresponding to the indirect illumination. With this formula the indirect in-scattered radiance at any given point inside a participating medium can be computed.

In the second pass, the surfaces are rendered using the same approach in the original photon mapping method. In addition, the rays that pass through the medium will be considered to integrate the result of participating media into this method by an adaptive ray marching algorithm which iteratively computes the radiance at points along the ray.

Translucent media with subsurface scattering is a class of participating media. Hence, rendering these materials proceeds in a similar way as rendering participating media. When a ray intersects with a subsurface scattering material, it is refracted into the medium. The contribution due to in-scattered radiance can be evaluated along the refracted ray as shown in Figure 6.10 and computed as a sum of two terms: a direct single-scattering term and an indirect multiple-scattering term. The indirect term is computed using the volume radiance estimation, and the direct single term can be computed by ray tracing.
Figure 6.10: Rendering material with subsurface scattering using photon map.

Figure 6.11 shows the result of Dorsey et. al. using volumetric photon mapping to render the weathering stones. The photon mapping technique is simple, efficient and significantly faster than the Monte-Carlo based algorithm. It provides a general framework for handling both homogeneous and inhomogeneous media with isotropic or anisotropic scattering. However, it is still very costly when rendering highly scattering materials such as milk and skin. In terms of this observation, Jensen et. al. developed a new technique based on a diffusion approximation as explained in Section 6.4.

Figure 6.11: Weathering simulation of a granite sphinx, the left image is a fresh granite, and the right image shows the effect of weathering (Courtesy of Dorsey et al.)

6.4. Jensen BSSRDF Model
Jensen et. al. introduced a new approach which extends from the traditional point-based reflection model (BRDF) and uses bidirectional subsurface scattering distribution function (BSSRDF). In this model Jensen et. al. also presented a computationally efficient analytical
The diffuse approximation to multiple scattering, which could handle the materials that exhibit considerable subsurface scattering such as milk and skin where the computations are very expensive with photon mapping techniques.

BSSRDF describes light transport between two rays that hit a surface which can be represented by

\[ dL_o(x_o, \hat{\Omega}_o) = S(x_i, \hat{\Omega}_i; x_o, \hat{\Omega}_o) d\Phi_i(x_i, \hat{\Omega}_i). \quad (6.18) \]

The complete BSSRDF model is a sum of the diffusion approximation \( S_d \) for multiple scattering and the single scattering term \( S^{(1)} \) for single scattering.

\[ S(x_i, \hat{\Omega}_i; x_o, \hat{\Omega}_o) = S_d(x_i, \hat{\Omega}_i; x_o, \hat{\Omega}_o) + S^{(1)}(x_i, \hat{\Omega}_i; x_o, \hat{\Omega}_o). \quad (6.19) \]

After integrating the radiative transport equation 4.15, the relation between the scalar radiance or fluence, \( \phi(x) = \int_{4\pi} L(x, \hat{\Omega}) d\Omega \) and the vector irradiance, \( E(x) = \int_{4\pi} L(x, \hat{\Omega}) \hat{\Omega} d\Omega \) can be represented by

\[ \nabla \cdot E(x) = -\sigma_o \phi(x) + Q_o(x). \quad (6.20) \]

The diffusion approximation is based on the observation that the light distribution tends to be isotropic in highly scattering media. In this case, the radiance can be approximated by a composition of radiant fluence and the vector irradiance as

\[ L(x, \hat{\Omega}) = \frac{1}{4\pi} \phi(x) + \frac{3}{4\pi} \hat{\Omega} \cdot E(x). \quad (6.21) \]

Using a dipole method\textsuperscript{28,29}, the diffuse component can be computed as the approximation of the volumetric source distribution which is shown in Figure 6.12. The dipole method consists of positioning two point sources near the surface in a way to satisfy the required boundary condition. One point source, as the positive real light source is located at the distance \( z_r \) beneath the surface and the other as the negative virtual light source is located above the surface at a distance \( z_v = z_r + 4AD \), where

\[ A = \frac{1 + F_{dr}}{1 + F_{dr}}, \quad D = \frac{1}{3\sigma_t}, \quad F_{dr} = \frac{1.440}{\eta} + \frac{0.710}{\eta} + 0.668 + 0.0636\eta. \quad (6.22) \]
The resulting fluence is
\[ \phi_r(x) = \frac{\Phi}{4\pi D} \left( e^{-\sigma_v d_r} \right), \quad \phi_s(x) = -\frac{\Phi}{4\pi D} \left( e^{-\sigma_s d_s} \right), \] where \( d_r = \| x - x_r \|, d_s = \| x - x_s \|. \] (6.23)

The diffuse reflectance due to dipole sources can be computed as
\[ R_d(r) = \frac{\alpha'}{4\pi} \left( \left( \sigma_v d_r + 1 \right) e^{-\sigma_v d_r} + \left( \sigma_s d_s + 1 \right) e^{-\sigma_s d_s} \right) \sigma_v = \sqrt{3\sigma_s \sigma_i} \] (6.24)
where \( \sigma_i', \sigma_s' \) are reduced extinction coefficients, given by \( \sigma_i' = \sigma_s' + \sigma_i, \sigma_s' = \sigma_s(1 - g). g \) is the mean cosine of the scattering angle. Accordingly, taking the Fresnel reflection at the boundary into consideration, the diffuse term of the BSSRDF is
\[ S_d(x_i, \hat{\Omega}_i; x_o, \hat{\Omega}_o) = \frac{1}{\pi} F_i(\eta, \hat{\Omega}_i) R_d(\| x_i - x_o \|) F_i(\eta, \hat{\Omega}_o) \] (6.25)

Accounting for local variations in lighting over the surface, the single scattering term is implicitly expressed in equation
\[ L_{0}^{1} (x_o, \hat{\Omega}_o) = \sigma_s(x_o) \int_{2\pi} p(\hat{\Omega}_i; \hat{\Omega}_o) \int_{0}^{\pi} e^{-\sigma_{\nu} L_i(x_i, \hat{\Omega}_i) d\Omega_i} ds d\Omega_i \]
\[ = \int_{4} \int_{2\pi} S_{0}^{1} (x_i, \hat{\Omega}_i; x_o, \hat{\Omega}_o) L_i(x_i, \hat{\Omega}_i) d\Omega_i dA(x_i), \] (6.26)
where \( \hat{\Omega}_i, \hat{\Omega}_o \) are the refracted incoming and outgoing directions, the combined extinction coefficient \( \sigma_c \) is given by
\[ \sigma_c = \sigma_i(x_o) + G\sigma_s(x_i), \quad G = \frac{|n_i \cdot \hat{\Omega}_o'|}{|n_i \cdot \hat{\Omega}_i|}. \]

Figure 6.13 shows several different simulations of subsurface scattering in a marble bust. The left image was rendered with BRDF approximation. The middle image was rendered with BSSRDF model and the right one was rendered with full Monte-Carlo simulation. The results show that the BSSRDF simulation is very similar to the full Monte-Carlo simulation but it is significantly faster than Monte-Carlo simulation.