Chapter 3

Radiometry and Light Transport

In this chapter, we describe the domains, quantities, and equations that are used for light transport calculations. Many of these concepts have their origins in *radiometry*, a field that studies the measurement of electromagnetic radiation. Radiometry is a natural foundation for graphics, because light is part of the electromagnetic spectrum.

We start by discussing the mathematical representation of the scene model. We then discuss the *phase space* and *trajectory space*, and show how radiometric quantities can be defined in terms of *photon events*. Next we give definitions of the quantities that are needed for light transport calculations, including power, irradiance, radiance, and spectral radiance. We also discuss the concepts of incident and exitant radiance functions.

We then describe how the light transport problem is formulated mathematically. This starts with the definition of the *bidirectional scattering distribution function* (BSDF), which gives a mathematical description of the way that light is scattered by a surface. We show how the BSDF is used to define the basic light transport equations, and we give a brief introduction to adjoint methods and bidirectional algorithms. We also explain why non-symmetric BSDF's require special treatment in bidirectional algorithms, and we define the useful concept of an *adjoint BSDF*. These ideas will be of central importance for the next several chapters.

Appendix 3.A discusses field and surface radiance functions [Arvo 1995], and compares them with the incident and exitant radiance functions that we use instead. Finally, Appendix 3.B gives the details of our measure-theoretic radiometry framework, in which we apply the tools of measure theory to define radiometric concepts more precisely. The main task is to define and use suitable measure functions, extending the work of Arvo [1995, Chapter 2].

A good introduction to radiometry is the book by McCluney [1994]. Other good references include [Nicodemus 1976], [Arvo 1995], [Cohen & Wallace 1993], and [Glassner 1995]. Note that our development is quite different than the standard treatments, due to the emphasis on measure theory.

3.1 Domains and measures

We assume that the scene geometry consists of a finite set of surfaces in $\mathbb{R}^3$, whose union is denoted $\mathcal{M}$. Formally, each surface is a piecewise differentiable two-dimensional manifold, possibly with boundary. For technical reasons, we require each manifold to be a closed set; that is, every manifold M must include its boundary ∂M. This prevents gaps between abutting surfaces (e.g. consider a cube formed from six squares). Note that $\mathcal{M}$ itself is not necessarily a manifold. For example, consider two spheres that touch at a point, or a box sitting on a table.

The surfaces divide $\mathbb{R}^3$ into a number of connected cells, each filled with a non-participating medium with a constant refractive index (i.e. volume absorption, emission, and scattering are not allowed).[1] It is possible that some surfaces do not belong to any cell boundary (e.g., a polygon floating in space).

We define an area measure A on $\mathcal{M}$ in the obvious way,[2] so that $A(D)$ denotes the area

[1] With this convention, all objects are hollow inside; a "solid" object is simply an empty cell with an opaque boundary. This representation is actually used by many rendering systems. Alternatively, a cell could be allowed to contain a perfectly absorbing medium. However, this would require some extra care with definitions, for example when defining the visibility and ray-casting functions used in Chapter 4.

[2] Given that $\mathcal{M}$ is the union of manifolds $\mathcal{M}_1, \ldots, \mathcal{M}_N$, we define $A(D)$ as the sum of the areas $A_i(D \cap \mathcal{M}_i)$, where A_i is the usual area measure on the manifold $\mathcal{M}_i$. The measurable sets $D \subset \mathcal{M}$ are defined by the requirement that all $D \cap \mathcal{M}_i$ are measurable. We also require that the intersection between any pair of surfaces $\mathcal{M}_i$ and $\mathcal{M}_j$ is a set of measure zero. In practice, this means that when the intersection between two surfaces has non-zero area (e.g. a cube sitting on a table), the rendering system must arbitrarily choose one surface over the other. This ensures that almost every point of $\mathcal{M}$ (up to a set of area measure zero) has a unique set of surface properties.

of a region $D \subset \mathcal{M}$. The notation

$$\int_{\mathcal{M}} f(\mathbf{x})\, dA(\mathbf{x})$$

denotes the Lebesgue integral of the function $f : \mathcal{M} \to \mathbb{R}$ with respect to surface area.

Directions are represented as unit-length vectors $\omega \in \mathbb{R}^3$. The set of all directions is denoted $\mathcal{S}^2$, the unit sphere in $\mathbb{R}^3$. Let σ be the usual surface area measure on $\mathcal{S}^2$. Given a set of directions $D \subset \mathcal{S}^2$, the *solid angle* occupied by D is simply $\sigma(D)$. Similarly, the solid angle *subtended* by a surface P from a point $\mathbf{x}$ is determined by projecting P onto the unit sphere centered at $\mathbf{x}$, and computing the measure of the resulting set of directions.

Another useful concept is the *projected solid angle* [Nicodemus 1976, p. 70], which arises in determining the irradiance (power per unit area) received by surface. Given a point $\mathbf{x} \in \mathcal{M}$, let $\mathbf{N}(\mathbf{x})$ be the surface normal at $\mathbf{x}$. Given a set of directions $D \subset \mathcal{S}^2$, the projected solid angle measure $\sigma_{\mathbf{x}}^{\perp}$ is defined by

$$\sigma_{\mathbf{x}}^{\perp}(D) \;=\; \int_D |\omega \cdot \mathbf{N}(\mathbf{x})|\, d\sigma(\omega)\,. \tag{3.1}$$

The factor $\omega \cdot \mathbf{N}(\mathbf{x})$ is often written as $\cos\theta$, where θ is the polar angle of ω (i.e. the angle between ω and the surface normal).

The name *projected solid angle* arises from the following geometric interpretation. Let $T_{\mathcal{M}}(\mathbf{x})$ be the *tangent space* at the point $\mathbf{x}$, i.e. the space of vectors in $\mathbb{R}^3$ that are perpendicular to the surface normal:

$$T_{\mathcal{M}}(\mathbf{x}) \;=\; \{\mathbf{y} \in \mathbb{R}^3 \mid \mathbf{y} \cdot \mathbf{N}(\mathbf{x}) = 0\}\,.$$

(Unlike the more familiar *tangent plane*, the tangent space passes through the origin. Thus it is a linear space rather than an affine one.) The tangent space divides $\mathcal{S}^2$ into two hemispheres, namely the *upward hemisphere*

$$\mathcal{H}^2_+(\mathbf{x}) \;=\; \{\omega \in \mathcal{S}^2 \mid \omega \cdot \mathbf{N}(\mathbf{x}) > 0\} \tag{3.2}$$

and the *downward hemisphere*

$$\mathcal{H}^2_-(\mathbf{x}) \;=\; \{\omega \in \mathcal{S}^2 \mid \omega \cdot \mathbf{N}(\mathbf{x}) < 0\}\,.$$

Now given a set of directions D contained by just one hemisphere, the projected solid angle can be obtained by simply projecting D orthogonally onto the tangent space, and then finding the area of the resulting planar region. For example, suppose that D is the entire upward hemisphere $\mathcal{H}^2_+$. The corresponding projected region is a unit disc, so we have $\sigma^\perp_{\mathbf{x}}(\mathcal{H}^2_+) = \pi$.

3.2 The phase space

Radiometric quantities can be defined within the more general framework of *transport theory*, which studies the motion of particles in an abstract setting. Each particle is characterized by a small number of parameters, which vary as a function of time. Typical particles such as neutrons or gas molecules can be represented by their position and velocity, for a total of 6 degrees of freedom. The state of a system of N particles is then represented as a $6N$-dimensional vector, which can be thought of as a point in the $6N$-dimensional *phase space* containing all possible system states.[3] The evolution of the system over time corresponds to a one-dimensional curve in phase space.

We now consider how this applies to light transport. Under the assumption that light is unpolarized and perfectly incoherent, the state of each photon can be represented by its position $\mathbf{x}$, direction of motion ω, and wavelength λ [Nicodemus 1976, p. 8]. Thus for a system of N photons, the phase space would be $6N$-dimensional.

However, for particles that do not interact with each other (such as photons), it is more useful to let the phase space correspond to the state of a single particle. With this convention, the phase space ψ is only 6-dimensional, and can be expressed as

$$\psi \;=\; \mathbb{R}^3 \times \mathcal{S}^2 \times \mathbb{R}^+ \,,$$

where $\mathbb{R}^+$ denotes the positive real numbers (corresponding to the range of allowable wavelengths). A system of N photons is represented as a set of N points in this 6-dimensional space, whose positions vary as a function of time.

Radiometric quantities can then be defined by counting the number of photons in a given

[3]For many problems the natural phase space is not really $6N$-dimensional, since physical laws may cause certain properties of the initial state to be preserved for all time (e.g., the total energy). This restricts the phase space to be a lower-dimensional manifold within the $6N$-dimensional Euclidean space defined above.

region of the phase space, or measuring their density with respect to one or more parameters. The most basic of these quantities is the *photon number* N_{p}, which simply measures the number of photons in a given phase space region [McCluney 1994, p. 26]. For example, we could count the number of photons in a given spatial volume $\Omega \subset \mathbb{R}^3$ at a fixed time t_0, with no restrictions on the direction or wavelength parameters. This corresponds to the region $\Omega \times \mathcal{S}^2 \times \mathbb{R}^+$ of the phase space ψ.

3.3 The trajectory space and photon events

We generalize the notion of a radiometric measurement further, by considering the time dimension explicitly. If the phase space positions of all photons are graphed over time, we obtain a set of one-dimensional curves in the *trajectory space*

$$\Psi \;=\; \mathbb{R} \times \psi \,,$$

where the first parameter represents time. Radiometric measurements are defined by specifying a set of *photon events* along these curves, and then measuring the distribution of these events in various ways.

A photon event is a single point in the trajectory space Ψ. Some events have natural definitions; for example, each emission, absorption, or scattering event corresponds to a single point along a photon trajectory.[4] Other events can be defined artificially, usually by specifying a *surface* in Ψ that intersects the photon trajectories at a set of points. For example, we could define the events to be the photon states at a particular time t_0. This corresponds to intersecting the trajectories with the plane $t = t_0$ in the trajectory space Ψ. Similarly, given an arbitrary plane P in $\mathbb{R}^3$, we could define a photon event to be a crossing of P, corresponding to an intersection with the surface $\mathbb{R} \times P \times \mathcal{S}^2 \times \mathbb{R}^+$ in trajectory space.

Once the photon events have been defined, we are left with a set of points in the trajectory

[4] In fact, each scattering event corresponds to *two* points along the photon trajectory, since the ω parameter has different values before and after the collision (corresponding to a discontinuity in the trajectory). Similarly, the wavelength parameter λ could change discontinuously in a fluorescent material. Thus, we must distinguish between *in-scattering* and *out-scattering* events, according to whether we measure the photon state before or after the collision. This is the basis for distinguishing between incident and exitant radiance functions, discussed in Section 3.5.

space Ψ. These points may be distributed throughout the whole space Ψ, or they may lie on some lower-dimensional manifold (e.g. if the photon events were defined as an intersection with a surface). To define a radiometric quantity, we then measure the distribution of these events with respect to a suitable geometric measure.

For this purpose, it is convenient to assume that the events are so numerous that their density can be modeled by continuous distributions rather than discrete ones. Rather than counting the photon events in a given region of the trajectory space, for example, we determine their total *radiant energy* Q (measured in joules $[\mathrm{J}]$). We will ignore the discrete nature of photons and assume that Q can take on any non-negative real value. (Note that each photon has an energy of $h\nu$, where h is Planck's constant, and $\nu = 1/\lambda$ is frequency.)

3.4 Radiometric quantities

We now discuss some of the most important radiometric quantities. Each of these is defined by measuring the distribution of energy with respect to one or more parameters. The discussion here is informal; a more detailed development is given in Appendix 3.B.

3.4.1 Power

Radiant power is defined as energy per unit time,

$$\Phi \;=\; \frac{dQ}{dt}\,, \tag{3.3}$$

and is measured in watts $[\mathrm{W} = \mathrm{J} \cdot \mathrm{s}^{-1}]$. For example, this is the quantity used to describe the rate at which energy is emitted or absorbed by a finite surface $S \subset \mathbb{R}^3$.

The notation (3.3) could be written more precisely as

$$\Phi(t) \;=\; \frac{dQ(t)}{dt}\,,$$

which makes it clear that Φ and Q are functions of time. Obviously Q *must* be defined as a function of time, in order for the idea of differentiating it to make sense. In general, this is done by defining $Q(t)$ to measure the energy of the photon events in some region $D(t)$ of trajectory space, where the region $D(t)$ grows with time. For example, suppose that we are counting emission events, and consider the region

$$D(t) \;=\; [0, t] \times S \times \mathcal{S}^2 \times \mathbb{R}^+ \,,$$

where $S \subset \mathbb{R}^3$ is a finite surface. In this case, $Q(t)$ represents the total energy emitted by S over the time interval $[0, t]$, so that $\Phi(t) \,=\, dQ(t)/dt$ measures the energy emission per unit time (at each time t).

However, we will usually ignore these subtleties. Most often we are concerned with systems in equilibrium, so that the density of photon events in phase space does not change with time. In this case, the t parameter can be omitted from the notation, as in equation (3.3).

3.4.2 Irradiance

Continuing with our discussion of radiometric quantities, *irradiance* is defined as power per unit surface area:

$$E(\mathbf{x}) \;=\; \frac{d\Phi(\mathbf{x})}{dA(\mathbf{x})} \,, \tag{3.4}$$

with units of $[\mathrm{W} \cdot \mathrm{m}^{-2}]$. It is always defined with respect to a point $\mathbf{x}$ on a surface S (either real or imaginary), with a specific normal $\mathbf{N}(\mathbf{x})$. The term *irradiance* also generally implies the measurement of incident radiation, on one side of the surface only (i.e. light incident from the upward hemisphere $\mathcal{H}^2_+(\mathbf{x})$). When light is leaving the surface, through either emission or scattering, the preferred term is *radiant exitance* (denoted by the symbol M) [Nicodemus 1978, p. 11]. Another common term is *radiosity*, which was introduced by Moon [1936] and popularized in the heat transfer literature (cf. Heckbert [1992]).

3.4.3 Radiance

For light transport calculations, by far the most important quantity is *radiance*, defined by

$$L(\mathbf{x}, \omega) \;=\; \frac{d^2\Phi(\mathbf{x}, \omega)}{dA^\perp_\omega(\mathbf{x})\, d\sigma(\omega)} \,, \tag{3.5}$$

where $A^\perp_\omega$ is the *projected area measure*, which measures area on a hypothetical surface perpendicular to ω. That is, to measure the radiance at $(\mathbf{x}, \omega)$, we count the number of photons per unit time passing through a small surface $dA^\perp_\omega(\mathbf{x})$ perpendicular to ω, whose directions are contained in a small solid angle $d\sigma(\omega)$ around ω. Radiance is defined as the limiting ratio of the power $d\Phi$ represented by these photons, divided by the product $dA^\perp_\omega(\mathbf{x})\, d\sigma(\omega)$. The corresponding units are $[\mathrm{W} \cdot \mathrm{m}^{-2} \cdot \mathrm{sr}^{-1}]$.

When measuring the radiance leaving a real surface S, a more convenient equation is given by

$$L(\mathbf{x}, \omega) \;=\; \frac{d^2\Phi(\mathbf{x}, \omega)}{|\omega \cdot \mathbf{N}(\mathbf{x})|\, dA(\mathbf{x})\, d\sigma(\omega)}\,, \tag{3.6}$$

where as before A is the area measure on S, and $\mathbf{N}(\mathbf{x})$ is the surface normal at $\mathbf{x}$. This relates the projected area $dA^\perp_\omega$ to the ordinary area dA, according to[5]

$$dA^\perp_\omega(\mathbf{x}) \;=\; |\omega \cdot \mathbf{N}(\mathbf{x})|\, dA(\mathbf{x})\,. \tag{3.7}$$

Alternatively, the $|\omega \cdot \mathbf{N}(\mathbf{x})|$ factor can be absorbed into the projected solid angle measure defined above, leading to

$$L(\mathbf{x}, \omega) \;=\; \frac{d^2\Phi(\mathbf{x}, \omega)}{dA(\mathbf{x})\, d\sigma^\perp_{\mathbf{x}}(\omega)}\,. \tag{3.8}$$

This is the most useful definition when dealing with radiance on real surfaces, because it uses the natural area measure A.

3.4.4 Spectral radiance

Carrying this one step further, *spectral radiance* L_λ is defined by

$$L_\lambda(\mathbf{x}, \omega, \lambda) \;=\; \frac{d^3\Phi(\mathbf{x}, \omega, \lambda)}{dA(\mathbf{x})\, d\sigma^\perp_{\mathbf{x}}(\omega)\, d\lambda}\,, \tag{3.9}$$

that is, $L_\lambda = dL/d\lambda$. The units are typically given as $[\mathrm{W} \cdot \mathrm{m}^{-2} \cdot \mathrm{sr}^{-1} \cdot \mathrm{nm}^{-1}]$, where the use of nanometers for wavelength helps to avoid confusion with the spatial variables [Nicodemus 1976, p. 49]. Other spectral quantities can be defined similarly, e.g. spectral

[5] More precisely, the projected area measure is defined by $A^\perp_\omega(D) = \int_D |\omega \cdot \mathbf{N}(\mathbf{x})|\, dA(\mathbf{x})$, where D is an arbitrary region of S.

power is defined by $\Phi_\lambda = d\Phi/d\lambda$.

Spectral radiance is often considered to be the fundamental radiometric quantity, in that many other common quantities can be derived from it. For example, radiance is given by

$$L(\mathbf{x}, \omega) \;=\; \int_0^\infty L_\lambda(\mathbf{x}, \omega, \lambda)\, d\lambda\,,$$

from which irradiance can be obtained by

$$E(\mathbf{x}) \;=\; \int_{\mathcal{H}^2_+(\mathbf{x})} L(\mathbf{x}, \omega)\, d\sigma^\perp_{\mathbf{x}}(\omega)\,.$$

In this dissertation, we will most often deal with spectral radiance L_λ. However, for conciseness we will usually just refer to this as "radiance" and use the symbol L. This is a slight abuse of terminology, but it is common practice in computer graphics.

Many other radiometric quantities have been defined, but we will not need them here. The manual by Nicodemus [1976] is an excellent reference on this topic, although some of the notation has been superceded by the *USA Standard Nomenclature and Definitions for Illuminating Engineering* [American National Standards Institute 1986].

3.5 Incident and exitant radiance functions

A *radiance function* is simply a function whose values correspond to radiance measurements.[6] Most often, we will work with functions of the form

$$L : \mathcal{M} \times \mathcal{S}^2 \to \mathbb{R}\,,$$

where $\mathcal{M}$ is the set of scene surfaces (Section 3.1). Occasionally, radiance functions of the form

$$L : \mathbb{R}^3 \times \mathcal{S}^2 \to \mathbb{R}$$

will also be useful. Note that we allow negative values for $L(\mathbf{x}, \omega)$ (which have no physical meaning), to ensure that the set of all radiance functions is a vector space.

[6] As mentioned in Section 3.4, we will often use the terms *radiance* and *spectral radiance* interchangeably, ignoring the extra λ parameter.

We will distinguish between *incident* and *exitant*[7] radiance functions, according to the interpretation of the ω parameter. An incident function $L_{\mathrm{i}}(\mathbf{x}, \omega)$ measures the radiance *arriving* at $\mathbf{x}$ from the direction ω, while an exitant function $L_{\mathrm{o}}(\mathbf{x}, \omega)$ measures the radiance *leaving* from $\mathbf{x}$ in the direction ω. In free space, these quantities are related by

$$L_{\mathrm{i}}(\mathbf{x}, \omega) \;=\; L_{\mathrm{o}}(\mathbf{x}, -\omega)\,. \tag{3.10}$$

However, at surfaces the distinction is more fundamental: L_{i} and L_{o} measure different sets of photon events, corresponding to the photon states just before their arrival at the surface, or just after their departure respectively. The relation between L_{i} and L_{o} can be quite complex, since it depends on the scattering properties of the surface.

The difference between incident and exitant radiance can be understood more precisely in terms of the trajectory space Ψ. Recall that each photon traces out a one-dimensional curve in this space, namely the graph of the function $(\mathbf{x}_i, \omega_i, \lambda_i)(t)$ over all values of t. To measure radiance, we define a photon event to be an intersection of one of these curves with the surface $\mathbb{P} = \mathbb{R} \times \mathcal{M} \times \mathcal{S}^2 \times \mathbb{R}^+$ in trajectory space. Our key observation is that this curve is not continuous at $\mathbb{P}$, since scattered photons instantaneously change their direction and/or wavelength. (A continuous curve would correspond to a photon that passes through $\mathcal{M}$ without any change.) Similarly, the curves for emitted and absorbed photons are discontinuous, since they are defined on only one side of $\mathbb{P}$.

We now observe that L_{i} and L_{o} measure events that are limit points of trajectories on *opposite sides* of the surface $\mathbb{P}$. Each event $(t_i, \mathbf{x}_i, \omega_i, \lambda_i)$ measured by L_{i} is the limit of a trajectory defined for $t < t_i$, while an event measured by L_{o} is the limit of a trajectory defined for $t > t_i$. This gives a simple and precise way to differentiate between incident and exitant radiance.

Note that incident and exitant radiance functions are quite similar to the *field* and *surface radiance functions* proposed by Arvo [1995] (the main difference is that the direction of ω is reversed for field radiance as compared to incident radiance). Appendix 3.A discusses these two approaches and explains the advantages of incident and exitant radiance functions.

[7] Nicodemus prefers the spelling *exitent*, and states that this term was coined by Richmond (cf. [Nicodemus 1976, p. 25]). Our use of *exitant* stems from [Christensen et al. 1993], where the term appears to have been re-invented.

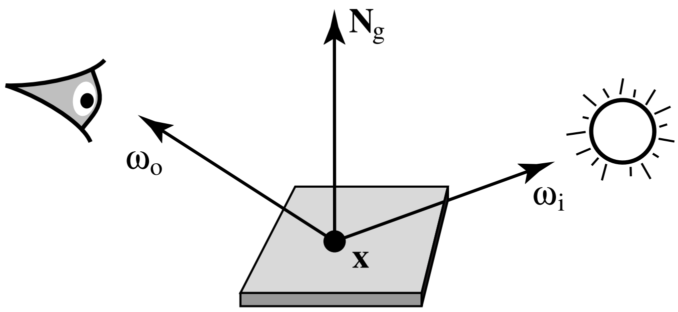


Figure 3.1: Geometry for defining the bidirectional scattering distribution function (BSDF).

3.6 The bidirectional scattering distribution function

The *bidirectional scattering distribution function* (BSDF) is a mathematical description of the light-scattering properties of a surface. Let $\mathbf{x} \in \mathcal{M}$ be a fixed point on the scene surfaces, and consider the radiance leaving $\mathbf{x}$ in a particular direction ω_o (see Figure 3.1). We will denote this $L_\mathrm{o}(\omega_\mathrm{o})$, dropping $\mathbf{x}$ from our notation. In general, the radiance $L_\mathrm{o}(\omega_\mathrm{o})$ depends on the radiance arriving at $\mathbf{x}$ from all directions. For now, we fix a particular direction ω_i, and consider the incident light from an infinitesimal cone around ω_i, where the cone occupies a solid angle of $d\sigma(\omega_\mathrm{i})$. This light strikes the surface at the point $\mathbf{x}$, and generates an irradiance equal to

$$dE(\omega_\mathrm{i}) \;=\; L_\mathrm{i}(\omega_\mathrm{i})\, d\sigma^\perp(\omega_\mathrm{i})\,.$$

The light is then scattered by the surface in all directions; we let $dL_\mathrm{o}(\omega_\mathrm{o})$ represent the contribution made to the radiance leaving in direction ω_o.

It can be observed experimentally that as $dE(\omega_\mathrm{i})$ is increased (by increasing either L_i or $d\sigma(\omega_\mathrm{i})$), there is a proportional increase in the observed radiance $dL_\mathrm{o}(\omega_\mathrm{o})$:

$$dL_\mathrm{o}(\omega_\mathrm{o}) \;\propto\; dE(\omega_\mathrm{i})\,.$$

This corresponds to the fact that light behaves linearly under normal circumstances (recall Section 1.5.3).

The BSDF $f_s(\omega_i \to \omega_o)$ is now simply defined to be this constant of proportionality:

$$f_s(\omega_i \to \omega_o) \;=\; \frac{dL_o(\omega_o)}{dE(\omega_i)} \;=\; \frac{dL_o(\omega_o)}{L_i(\omega_i)\, d\sigma^\perp(\omega_i)}\,. \tag{3.11}$$

In words, $f_s(\omega_i \to \omega_o)$ is the observed radiance leaving in direction ω_o, per unit of irradiance arriving from ω_i. The notation $\omega_i \to \omega_o$ symbolizes the direction of light flow.

3.6.1 The scattering equation

By integrating the relationship

$$dL_o(\omega_o) \;=\; L_i(\omega_i)\, f_s(\omega_i \to \omega_o)\, d\sigma^\perp(\omega_i)$$

over all directions, we can now predict $L_o(\omega_o)$. This is summarized by the *(surface) scattering equation*,[8]

$$L_o(\omega_o) \;=\; \int_{\mathcal{S}^2} L_i(\omega_i)\, f_s(\omega_i \to \omega_o)\, d\sigma^\perp(\omega_i)\,. \tag{3.12}$$

This equation can be used to predict the appearance of the surface, given a description of the incident illumination.

3.6.2 The BRDF and BTDF

The BSDF is not a standard concept in radiometry.[9] More typically, the scattered light is subdivided into reflected and transmitted components, which are treated separately. This leads to the definition of the *bidirectional reflectance distribution function* (BRDF), and the *bidirectional transmittance distribution function* (BTDF), denoted f_r and f_t respectively.

The BRDF is obtained by simply restricting f_s to a smaller domain:

$$f_r : \mathcal{H}_i^2 \times \mathcal{H}_r^2 \to \mathbb{R}\,,$$

[8] The corresponding equation for one-sided, opaque surfaces is called the *reflectance equation* [Cohen & Wallace 1993, p. 30].

[9] The name appears to have been introduced by Heckbert [Heckbert 1991, p. 26]. Previously, he used the term *bidirectional distribution function* (BDF) [Heckbert 1990], however we feel that this term is more appropriate for a category of such functions, containing the various B*DF's as members.

where $\mathcal{H}^2_{\mathrm{i}}$ and $\mathcal{H}^2_{\mathrm{r}}$ are often called the *incident* and *reflected* hemispheres respectively. In fact, both symbols refer to the same set of directions ($\mathcal{H}^2_{\mathrm{i}} = \mathcal{H}^2_{\mathrm{r}}$), which can be either the upward hemisphere $\mathcal{H}^2_+$, or its complement $\mathcal{H}^2_-$.

The BTDF is defined similarly to the BRDF, by restricting f_{s} to a domain of the form

$$f_{\mathrm{t}} : \mathcal{H}^2_{\mathrm{i}} \times \mathcal{H}^2_{\mathrm{t}} \to \mathbb{R}\,,$$

where the transmitted hemisphere $\mathcal{H}^2_{\mathrm{t}} = -\mathcal{H}^2_{\mathrm{i}}$ is the complement of $\mathcal{H}^2_{\mathrm{i}}$. As before $\mathcal{H}^2_{\mathrm{i}}$ can represent either the upward hemisphere $\mathcal{H}^2_+$, or its complement $\mathcal{H}^2_-$.

Thus, we see that the BSDF is the union of two BRDF's (one for each side of the surface), and two BTDF's (one for light transmitted in each direction). Its main advantage is convenience: we only need to deal with one function, rather than four. The BSDF allows us to write equations that are simple and yet general, capable of describing the scattering from any kind of surface. Surfaces that are purely reflective or transmissive are simply special cases of this formulation. In addition, the BSDF is actually easier to define, since we do not need to specify the hemispherical domains needed by the BRDF and BTDF.

Properties of the BRDF. The BRDF's that describe real surfaces are known to have a number of basic properties. For example, they are *symmetric*:

$$f_{\mathrm{r}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}}) \;=\; f_{\mathrm{r}}(\omega_{\mathrm{o}} \to \omega_{\mathrm{i}}) \qquad \text{for all } \omega_{\mathrm{i}}, \omega_{\mathrm{o}}\,. \tag{3.13}$$

Because of the symmetry, the notation $f_{\mathrm{r}}(\omega_{\mathrm{i}} \leftrightarrow \omega_{\mathrm{o}})$ is often used. Another property shared by physical BRDF's is *energy conservation*, as embodied by the condition

$$\int_{\mathcal{H}^2_{\mathrm{o}}} f_{\mathrm{r}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}})\, d\sigma^{\perp}(\omega_{\mathrm{o}}) \;\le\; 1 \qquad \text{for all } \omega_{\mathrm{i}} \in \mathcal{H}^2_{\mathrm{i}}\,. \tag{3.14}$$

Further explanation of BRDF's and their properties can be found in [Nicodemus et al. 1977, p. 5] or [Cohen & Wallace 1993, p. 28].

Note that these simple conditions are unique to reflection, and do not always apply to surfaces that transmit light. Thus, it cannot be assumed that BSDF's or BTDF's satisfy the simple rules above. We will investigate the correct generalization of these properties to arbitrary surfaces in Chapter 6.

3.6.3 Angular parameterizations of the BSDF

It is common to write BSDF's in terms of polar and azimuthal angles, rather than unit direction vectors. We will use this parameterization later in this chapter, to derive the scaling of radiance at a refractive interface (Section 5.2). We show how the two parameterizations are related, and summarize the advantages of the unit vector form.

In the angular parameterization, a direction $\omega \in \mathcal{S}^2$ is represented as a pair of angles (θ, ϕ). The polar angle θ measures the angle between ω and the normal $\mathbf{N}$, while the azimuthal angle ϕ measures the angle between ω and a fixed direction $\mathbf{T}$ lying in the tangent space at $\mathbf{x}$. The angular and vector representations are thus related by

$$\begin{aligned} \cos\theta &= \omega \cdot \mathbf{N}\,, \\ \cos\phi &= \omega \cdot \mathbf{T}\,. \end{aligned}$$

To use this parameterization, we must also know how the angle measures σ and $\sigma^{\perp}$ are represented. The solid angle σ corresponds to

$$\begin{aligned} d\sigma(\omega) &\equiv \sin\theta\, d\theta\, d\phi \\ &\equiv d!\cos\theta\, d\phi\,, \end{aligned} \tag{3.15}$$

while the projected solid angle $\sigma^{\perp}$ can be written in a number of forms:

$$\begin{aligned} d\sigma^{\perp}(\omega) &\equiv |\cos\theta|\, \sin\theta\, d\theta\, d\phi \\ &\equiv |\cos\theta|\, d!\cos\theta\, d\phi \\ &\equiv \sin\theta\, d\!\sin\theta\, d\phi \\ &\equiv (1/2)\, d!\cos^2\theta\, d\phi \\ &\equiv (1/2)\, d\!\sin^2\theta\, d\phi\,. \end{aligned} \tag{3.16}$$

With the angular parameterization, the scattering equation (3.12) thus becomes

$$L_{\mathrm{o}}(\theta_{\mathrm{o}}, \phi_{\mathrm{o}}) = \int_0^{2\pi}\int_0^{\pi} L_{\mathrm{i}}(\theta_{\mathrm{i}}, \phi_{\mathrm{i}})\, f_{\mathrm{s}}(\theta_{\mathrm{i}}, \phi_{\mathrm{i}}, \theta_{\mathrm{o}}, \phi_{\mathrm{o}})\, |\cos\theta_{\mathrm{i}}|\, \sin\theta_{\mathrm{i}}\, d\theta_{\mathrm{i}}\, d\phi_{\mathrm{i}}\,, \tag{3.17}$$

where the other representations (3.16) for the projected solid angle could also be used.

Although the angular representation is common, there are good reasons to prefer the unit vector representation $\omega \in \mathcal{S}^2$. First, note that (θ, ϕ) is a *local* representation of direction, since the angles θ and ϕ depend on the surface normal. When more than one surface point is involved, it is much more convenient to work with direction vectors. Second, the (θ, ϕ) representation creates the impression that many formulas involve trigonometric functions, when in fact they are usually implemented with dot products. Finally, the angular representation depends on an extra parameter (the tangent vector $\mathbf{T}$), which must be chosen arbitrarily since it has no physical significance.

3.7 Introduction to light transport

This section reviews the main concepts of light transport, without getting into too much detail. (These ideas will be defined more precisely in the next chapter, where we reformulate light transport in terms of linear operators.) We discuss the measurement, light transport, and importance transport equations. We also outline the ideas of bidirectional methods for the light transport problem, and explain why they are often the most efficient methods for its solution. Finally, we explain why bidirectional algorithms need to evaluate BSDF's with special care, and we define the useful concept of an *adjoint BSDF*.

3.7.1 The measurement equation

The goal of light transport is to compute a set of real-valued *measurements* $I_1, \ldots, I_M$. For example, in a light transport algorithm that computes an image directly, each measurement I_j represents the value of a single pixel, and M is the number of pixels in the image.

Each measurement corresponds to the output of a hypothetical *sensor* that responds to the radiance $L_{\mathrm{i}}(\mathbf{x}, \omega)$ incident upon it. The response may vary according to the position and direction at which light strikes the sensor; this is characterized by the sensor *responsivity* $W_{\mathrm{e}}(\mathbf{x}, \omega)$. The total response is determined by integrating the product $W_{\mathrm{e}} L_{\mathrm{i}}$, according to

$$I \;=\; \int_{\mathcal{M} \times \mathcal{S}^2} W_{\mathrm{e}}(\mathbf{x}, \omega)\, L_{\mathrm{i}}(\mathbf{x}, \omega)\, dA(\mathbf{x})\, d\sigma_{\mathbf{x}}^{\perp}(\omega)\,. \tag{3.18}$$

This is called the *measurement equation*. Note that there is actually one equation for each measurement I_j, each with a different responsivity function $W_{\mathrm{e}}^{(j)}$ (although we will usually drop the superscript). Also note that we have assumed that the sensors are modeled as part of the scene $\mathcal{M}$, in order that we can integrate over their surface.

3.7.2 The light transport equation

Generally, we are most interested in measuring the *steady-state* or *equilibrium* radiance function for the given scene.[10] It is conventional to solve for the exitant version of this quantity, L_{o}, from which the incident radiance L_{i} can be obtained using

$$L_{\mathrm{i}}(\mathbf{x}, \omega) \;=\; L_{\mathrm{o}}(\mathbf{x}_{\mathcal{M}}(\mathbf{x}, \omega), -\omega)\,.$$

Here $\mathbf{x}_{\mathcal{M}}(\mathbf{x}, \omega)$ is the *ray-casting function*, which returns the first point of $\mathcal{M}$ visible from $\mathbf{x}$ in direction ω.

We can express L_{o} as the sum of *emitted radiance* L_{e}, and *scattered radiance* $L_{\mathrm{o,s}}$:

$$L_{\mathrm{o}} \;=\; L_{\mathrm{e}} + L_{\mathrm{o,s}}\,.$$

The emitted radiance function $L_{\mathrm{e}}(\mathbf{x}, \omega)$ is provided as part of the scene description, and represents all of the light sources in the scene. On the other hand, $L_{\mathrm{o,s}}$ is determined using the scattering equation (3.12), according to

$$L_{\mathrm{o,s}}(\mathbf{x}, \omega_{\mathrm{o}}) \;=\; \int_{\mathcal{S}^2} L_{\mathrm{i}}(\mathbf{x}, \omega_{\mathrm{i}})\, f_{\mathrm{s}}(\mathbf{x}, \omega_{\mathrm{i}} \!\rightarrow\! \omega_{\mathrm{o}})\, d\sigma^{\perp}_{\mathbf{x}}(\omega_{\mathrm{i}})\,.$$

By putting these equations together, we get a complete specification of the light transport problem. The most interesting feature is that L_{o} and L_{i} have been defined in terms of each other; commonly their definitions are combined to obtain

$$L_{\mathrm{o}}(\mathbf{x}, \omega_{\mathrm{o}}) \;=\; L_{\mathrm{e}}(\mathbf{x}, \omega_{\mathrm{o}}) + \int_{\mathcal{S}^2} L_{\mathrm{o}}(\mathbf{x}_{\mathcal{M}}(\mathbf{x}, \omega_{\mathrm{i}}), -\omega_{\mathrm{i}})\, f_{\mathrm{s}}(\mathbf{x}, \omega_{\mathrm{i}} \!\rightarrow\! \omega_{\mathrm{o}})\, d\sigma^{\perp}_{\mathbf{x}}(\omega_{\mathrm{i}})\,, \tag{3.19}$$

which is known as the *light transport equation*. Since L_{i} does not appear in this equation, the subscript on L_{o} is usually dropped. The form of this equation naturally leads to recursive solutions (the essence of traditional Monte Carlo methods).

[10]Since light travels so much faster than the everyday objects around us, equilibrium is achieved very quickly after any changes to the environment. Effectively, the world we perceive is always in equilibrium (with respect to light transport).

3.7.3 Importance and adjoint methods

As we have presented them, the transport rules apply to the scattering of light, as emitted by the sources. However, the transport rules can be applied equally well to the sensors, by treating the responsivity $W_{\mathrm{e}}(\mathbf{x}, \omega)$ as an emitted quantity. In this context, $W_{\mathrm{e}}(\mathbf{x}, \omega)$ is called an *emitted importance function*, since W_{e} specifies the "importance" of the light arriving along each ray to the corresponding measurement I.

This idea is the basis of *adjoint* methods, which apply the transport rules to importance rather than radiance. These methods start with the emitted importance $W_{\mathrm{e}}(\mathbf{x}, \omega)$, and solve for the *equilibrium importance function* $W(\mathbf{x}, \omega)$, according to the *importance transport equation*

$$W(\mathbf{x}, \omega) \;=\; W_{\mathrm{e}}(\mathbf{x}, \omega) + \int_{\mathcal{S}^2} W(\mathbf{x}_{\mathcal{M}}(\mathbf{x}, \omega_{\mathrm{i}}), -\omega_{\mathrm{i}})\, f_{\mathrm{s}}(\mathbf{x}, \omega_{\mathrm{o}} \!\rightarrow\! \omega_{\mathrm{i}})\, d\sigma^{\perp}_{\mathbf{x}}(\omega_{\mathrm{i}})\,. \qquad (3.20)$$

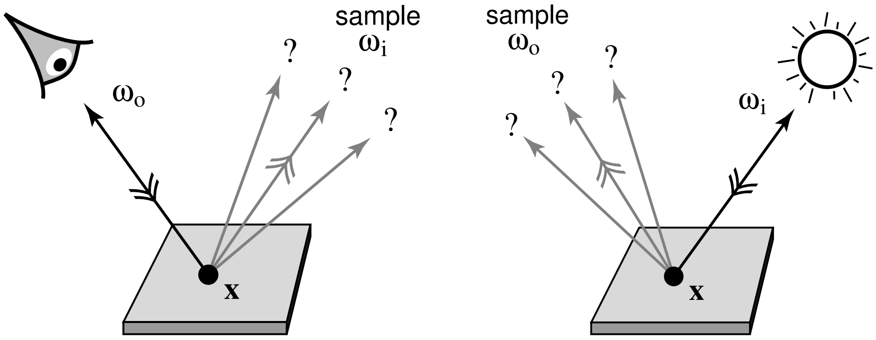
This equation is virtually identical to the light transport equation (3.19), except that the directional arguments to the BSDF have been exchanged.

Given the equilibrium importance W, measurements are computed by integrating the product $W L_{\mathrm{e}}$ (similar to (3.18)). Note that while there is only one equilibrium radiance function, there can be many different equilibrium importance functions (one for each sensor). This is an important difference between direct and adjoint methods.

3.7.4 Bidirectional methods

Many recent algorithms combine features from both of these approaches, leading to *bidirectional* light transport methods. The computation is guided by the viewing information (sensors), as well as the lighting information (sources). This allows these algorithms to be more efficient, since they can do less work in regions that are dark or that are not visible. This concept is similar to certain planning problems in artificial intelligence, where the objective is to get from an initial state to a goal, given some set of possible actions. It is possible to reduce the search complexity by simultaneously working forward from the initial state, and backward from the goal, until the two searches meet somewhere in the middle.

Bidirectional algorithms can appear in a number of different forms. *Importance-driven*



(a) Path tracing **(b)** Particle tracing

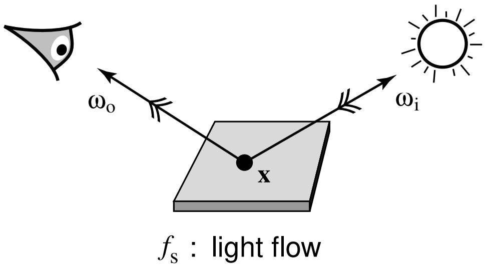
Figure 3.2: Path tracing and particle tracing sample the BSDF in different ways. **(a)** For path tracing, the direction ω_o is given (it points toward the previous vertex on a path leading to a sensor). The path is extended by sampling a direction ω_i according to the BSDF. **(b)** For particle tracing, the direction ω_i is given (pointing along a path toward a light source), and the path is extended by sampling a direction ω_o.

methods use viewing information to guide mesh refinement, by increasing the mesh resolution in regions where the equilibrium importance is high (since these regions have the greatest influence on the desired set of measurements). With Monte Carlo approaches, bidirectional methods often combine *path tracing*, where the transport equation is sampled starting from the sensors, and *particle tracing*, where sampling begins at the light sources.

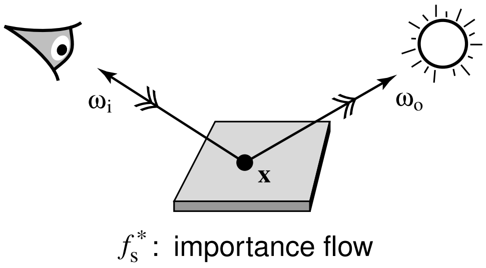
In one way or another, almost all recent light transport algorithms have taken a bidirectional approach. These include finite element approaches [Smits et al. 1992, Schröder & Hanrahan 1994, Christensen et al. 1996], multi-pass methods [Chen et al. 1991, Zimmerman & Shirley 1995], particle tracing algorithms [Heckbert 1990, Pattanaik & Mudur 1995, Shirley et al. 1995, Jensen 1996], and bidirectional path tracing [Lafortune & Willems 1993, Veach & Guibas 1994, Veach & Guibas 1995].

3.7.5 Sampling and evaluation of non-symmetric BSDF's

Scene models often contain materials whose BSDF is not symmetric, i.e. for which $f_s(\omega_i \to \omega_o) \neq f_s(\omega_o \to \omega_i)$. Great care must be taken when such materials are used with bidirectional algorithms, because in this case the transport rules for light and importance are



(**a**) The normal BSDF f_s.



(**b**) The adjoint BSDF f_s^*.

Figure 3.3: By adopting the convention that ω_i is always the sampled direction, the BSDF f_s and its adjoint f_s^* are used for different purposes. (**a**) The BSDF $f_s(\omega_i \to \omega_o)$ is used for radiance evaluation, and to scatter importance particles. (**b**) The adjoint BSDF $f_s^*(\omega_i \to \omega_o)$ is used for importance evaluation, and to scatter light particles.

different. Formally, this can be seen by noting that the light transport equation (3.19) and the importance transport equation (3.20) are identical, except that the directional arguments to the BSDF have been exchanged. Thus if the BSDF is not symmetric, then light and importance satisfy different transport equations. From another point of view, recall that the BSDF was defined in terms of light propagation: light flows from the incoming direction ω_i to the outgoing direction ω_o. Thus importance flows from ω_o to ω_i, since it is transported in the opposite direction as light. Similarly, different scattering rules must be used for particle tracing and path tracing to obtain correct results when non-symmetric BSDF's are present (see Figure 3.2). Thus, bidirectional algorithms must take care when evaluating or sampling the BSDF, to ensure that ω_i and ω_o are ordered correctly.

In the next few chapters, we will study non-symmetric BSDF's and their consequences for bidirectional algorithms in detail.

3.7.6 The adjoint BSDF

Given an arbitrary BSDF f_s, the *adjoint BSDF* f_s^* is defined by

$$f_s^*(\omega_i \to \omega_o) \;=\; f_s(\omega_o \to \omega_i) \qquad \text{for all } \omega_i, \omega_o \in \mathcal{S}^2 \,. \tag{3.21}$$

The main advantage of the adjoint BSDF is that it lets the importance transport equation (3.20) have the same form as the light transport equation (3.19). Recall that the only difference between these two equations is that the arguments to the BSDF are exchanged ($f_s(\omega_o \to \omega_i)$ instead of $f_s(\omega_i \to \omega_o)$). By using the adjoint BSDF f_s^* in the importance transport equation, this difference is eliminated: the two equations have exactly the same form, but they use different BSDF's (see Figure 3.3.)

The adjoint BSDF also provides a useful convention for sampling. Recall that in path tracing, we sample the BSDF to determine the incident direction ω_i (since ω_o is given). We extend this idea, by adopting the convention that ω_i is *always* the sampled direction during a random walk. We refer to the opposite situation (where ω_i is provided, and ω_o is sampled) as *sampling the adjoint BSDF*. For example, according to this convention the adjoint BSDF is used to scatter light particles.

We also mention two other techniques that can be used in bidirectional algorithms. The first of these is *importance particle tracing*, in which particles are emitted from the sensors and scattered throughout the environment, in order to obtain a set of samples that represent the equilibrium importance. This process is similar to ordinary particle tracing, except that importance is used instead of light. This implies that importance particles should be scattered using the ordinary BSDF f_s. The second technique is *importance evaluation*, in which the equilibrium importance on a ray $(\mathbf{x}, \omega)$ is estimated by recursively sampling the importance transport equation. This is similar to the evaluation of radiance using path tracing, except that the adjoint BSDF f_s^* is used instead of f_s.

To summarize, the adjoint BSDF is used for importance evaluation and for scattering light particles (i.e. sampling processes that start at a light source), while the normal BSDF is used for radiance evaluation and for scattering importance particles (sampling processes that start at a sensor). These rules will be justified formally in Chapter 4.

Appendix 3.A Field and surface radiance functions

In this appendix we consider the *field* and *surface* radiance functions defined by Arvo [1995, p. 28], and compare them with the incident and exitant radiance functions described in Section 3.5. Basically, field radiance L_{f} is similar to incident radiance L_{i}, while surface radiance L_{s} is similar to exitant radiance L_{o}. The main difference is that L_{f} and L_{s} are defined only in the context of one-sided (reflective) surfaces, which allows them to be defined as two halves of a single radiance distribution L.

To define field and surface radiance precisely, let S be a surface bounding an opaque object, and consider the radiance distribution $L(\mathbf{x}, \omega)$ at a point $\mathbf{x} \in S$. Arvo [1995] observes that since scattering occurs on only one side of S, the direction of ω can be used to distinguish incident photons from exitant ones: if ω is in the upward hemisphere $\mathcal{H}^2_+(\mathbf{x})$, then $L(\mathbf{x}, \omega)$ refers to radiance leaving the surface, and otherwise $L(\mathbf{x}, \omega)$ refers to radiance arriving at the surface. Applying this observation, he proposed that $L(\mathbf{x}, \omega)$ is naturally partitioned into *surface radiance* $L_{\mathrm{s}}(\mathbf{x}, \omega)$ and *field radiance* $L_{\mathrm{f}}(\mathbf{x}, \omega)$, according to whether $\omega \cdot \mathbf{N}(\mathbf{x})$ is positive or negative respectively.

However, there are several important differences between incident/exitant radiance and field/surface radiance. First, the sense of the direction parameter ω is reversed for L_{f} as compared to L_{i}:

$$L_{\mathrm{f}}(\mathbf{x}, \omega) \;=\; L_{\mathrm{i}}(\mathbf{x}, -\omega)\,.$$

The field radiance definition L_{f} would appear to be more natural, since ω corresponds to the direction of travel of the photons. However, the L_{i} definition has two important advantages. At reflective surfaces, it corresponds to the convention assumed by most BRDF formulas, where ω_{i} and ω_{o} both point outward. More significantly, the L_{i} definition causes certain natural transport operators to become self-adjoint (namely the $\mathbf{G}$ and $\mathbf{K}$ operators defined in Section 4.3), which increases the symmetry between the equations governing light and importance transport.

A second difference is that L_{i} and L_{o} are defined for *two-sided* surfaces, e.g. those that allow both reflection and transmission. For these surfaces, ω cannot be used to distinguish between incident and exitant photons, since L_{i} and L_{o} are both defined for all $\omega \in \mathcal{S}^2$. Instead, the two sets of photon events must be distinguished using the time dimension, as we have outlined above.

Finally, field and surface radiance are defined only at surfaces, while incident and exitant radiance are defined in space as well. (The distinction between L_{i} and L_{o} is still useful in this context, since it can be used to define self-adjoint operators for volume scattering.)

Appendix 3.B Measure-theoretic radiometry

Typically, radiometric quantities are defined using "infinitesimals" and limit arguments. Arvo [Arvo 1995, Chapter 2] has taken a different approach, by proposing a set of axioms that correspond to the observable behavior of photons, and then deriving radiometric quantities using the tools of measure theory. His analysis focused on the spatial distribution of steady-state, monochromatic radiation, and led to a measure-theoretic definition of the *phase space density* (defined below). In this section we show how to extend his techniques to a more general class of radiometric quantities: for example, we give measure-theoretic definitions of *spectral radiance* and *spectral radiant sterisent*.

3.B.1 Measure spaces

A *measure space* is a triple $(\mathbb{P}, \mathcal{P}, \varrho)$, where $\mathbb{P}$ is a set (the *underlying set* of the measure space), $\mathcal{P}$ is a collection of subsets of $\mathbb{P}$ (the *measurable sets*), and $\varrho : \mathcal{P} \to [0, \infty]$ is a non-negative, countably additive set function (the *measure function*, or simply the *measure*). The countably additive property means that

$$\varrho\left(\bigcup_{i=1}^{\infty} D_i\right) \;=\; \sum_{i=1}^{\infty} \varrho(D_i)$$

whenever the D_i are mutually disjoint measurable sets.

The measurable sets form a *σ-algebra*, meaning that $\mathcal{P}$ contains $\mathbb{P}$, and is closed under the operations of complementation and countable unions. For technical reasons, $\mathcal{P}$ is generally a proper subset of 2^{P}, that is, some sets are not measurable. However, for the measure spaces we are interested in (those constructed as the product of Lebesgue measures), the unmeasurable sets represent pathological situations that can be ignored in practice.

Sometimes, the measures we consider will not be finite; that is, $\varrho(\mathbb{P}) = \infty$. However, they will always have the weaker property of being *σ-finite*, meaning that there is an infinite sequence $D_1, D_2, \ldots$ of measurable sets such that

$$\bigcup_{i=1}^{\infty} D_i \;=\; \mathbb{P}\,,$$

and $\varrho(D_i)$ is finite for all i. That is, a σ-finite measure space is one that can be decomposed into countably many regions, each with finite measure.

3.B.2 The photon event space

To define a radiometric quantity, we first choose an appropriate *photon event space*. This is the subset $\mathbb{P} \subset \Psi$ of the trajectory space containing all possible locations of the photon events we wish to count. Thus, $\mathbb{P}$ depends on the definition of a photon event; by defining photon events in different ways, we will obtain different radiometric quantities. For example, consider the case of *volume emission* (e.g. the light emitted by a fire). Without knowledge of the specific scene geometry, we must assume that a photon could be emitted from any point in $\mathbb{R}^3$, in any direction, at any wavelength, at any time; thus we would set $\mathbb{P} = \Psi$ (the whole trajectory space). On the other hand, if photon events were defined as crossings of a hypothetical surface $S \subset \mathbb{R}^3$, then the photon event space would be

$$\mathbb{P} \;=\; \mathbb{R} \times S \times \mathcal{S}^2 \times \mathbb{R}^+ \,.$$

In this example, $\mathbb{P}$ is a 6-dimensional manifold within the 7-dimensional trajectory space Ψ.

3.B.3 The geometric measure

Next we define a measure ϱ on the photon event space, called the *geometric measure*, which will be used to measure the density of photon events. It will normally be defined as product of the natural Lebesgue measures on the components of $\mathbb{P}$. For example, in the case of volume emission ϱ is given by

$$\varrho \;=\; l \times v \times \sigma \times l^+ \,,$$

where l and l^+ are the usual length measures on $\mathbb{R}$ and $\mathbb{R}^+$ respectively, and v is the usual volume measure on $\mathbb{R}^3$. Note that this definition also establishes the geometrically measurable sets $\mathcal{P}$, according to the usual rules for product measures [Halmos 1950, p. 140].[11]

3.B.4 The energy measure

To count the photon events in various regions of $\mathbb{P}$, we also define an *energy content function*

$$Q : \mathcal{P} \to [0, \infty] \,.$$

[11]Technically, we work with the *completion* of the product measure, which augments $\mathcal{P}$ to include sets of the form $D \triangle N$, where $\triangle$ denotes the symmetric difference of two sets, D is a measurable set, and N is an arbitrary subset of a set of measure zero.

To each measurable set D of the photon event space, it assigns a non-negative real number $Q(D)$ that measures the total energy of the photon events in D. The function Q is assumed to obey the following physically plausible axioms (see [Arvo 1995, p. 19]):

(A1) $Q : \mathcal{P} \to [0, \infty]$

(A2) $Q\left(\bigcup_{i=1}^{\infty} D_i\right) = \sum_{i=1}^{\infty} Q(D_i)$ for mutually disjoint $\{D_i\} \subset \mathcal{P}$

(A3) $\varrho(D) < \infty \implies Q(D) < \infty$

(A4) $\varrho(D) = 0 \implies Q(D) = 0$

Axiom (A1) states that every region contains a non-negative quantity of energy. Axiom (A2) states that Q is countably additive; that is, if we consider a countable set of disjoint regions D_i, the energy contained their union is simply the sum of their individual energies. Together, (A1) and (A2) imply that Q is a non-negative, countably additive set function, so that by definition Q is a measure (on the same measurable sets for which ϱ is defined).

Axiom (A3) states that every region with finite ϱ-measure contains a finite quantity of energy. Intuitively, this says that the energy density is finite everywhere, a concept that will be made more precise below. From a measure-theoretic point of view, it ensures that the σ-finite property of ϱ carries over to Q.

Finally, (A4) states that Q is *continuous* with respect to ϱ, meaning that every set with zero ϱ-measure also has zero Q-measure. This important property allows the "ratio" of two measures to be defined rigorously, as we shall see below.

By translating these axioms into the language of measure theory, we obtain the following theorem (cf. Arvo, Theorem 1 [Arvo 1995, p. 22]):

Theorem 3.1 (Existence of Energy Measures). *Given a photon event space $\mathbb{P}$ with geometric measure ϱ, and an energy content function Q satisfying axioms (A1), (A2), (A3), and (A4), then Q defines a positive σ-finite measure over $\mathbb{P}$, and Q is continuous with respect to ϱ.*

Thus, we will now refer to Q as the *energy measure* on $\mathbb{P}$.

3.B.5 Defining radiometric quantities as a ratio of measures

Loosely speaking, a radiometric quantity can now be defined by measuring the density of Q with respect to ϱ, i.e. the ratio $dQ/d\varrho$ for a region D that becomes arbitrarily small. This idea can be made precise by means of the Radon-Nikodym theorem [Halmos 1950, p. 128]:[12]

Theorem 3.2 (Radon-Nikodym). *If* $(\mathbb{P}, \mathcal{P}, \varrho)$ *is a* σ*-finite measure space, and if a* σ*-finite measure* Q *on* $\mathcal{P}$ *is continuous with respect to* ϱ*, then there exists a non-negative, real-valued,* ϱ*-measurable function* f *on* $\mathbb{P}$ *such that*

$$Q(D) \;=\; \int_D f \, d\varrho$$

for every measurable set $D \in \mathcal{P}$*. The function* f *is unique up to a set of* ϱ*-measure zero.*

The function f is called the *Radon-Nikodym derivative* of Q with respect to ϱ, denoted

$$f \;=\; \frac{dQ}{d\varrho}\,. \tag{3.22}$$

This notation emphasizes its similarity with ordinary differentiation, with which it shares many properties.

Using the Radon-Nikodym theorem, we can thus define a function f corresponding to the density of photon events. The meaning of this density obviously depends on how the events are defined. However, we can summarize the fact of its existence as follows (cf. Arvo, Theorem 3 [Arvo 1995, p. 23]):

Theorem 3.3 (Existence of Energy Density). *Given a photon event space* $\mathbb{P}$ *with geometric measure* ϱ*, and an energy content function* Q *satisfying axioms (A1), (A2), (A3), and (A4), then there exists a* ϱ*-measurable function* $f : \mathbb{P} \to (0, \infty)$*, which is unique to within a set of* ϱ*-measure zero, satisfying*

$$Q(D) \;=\; \int_D f \, d\varrho\,,$$

where $D \in \mathcal{P}$ *is a measurable subset of* $\mathbb{P}$*.*

3.B.6 Examples of measure-theoretic definitions

We now give several examples showing how these concepts can be applied.

3.B.6.1 Spectral radiant sterisent

Consider again the case of volume emission. Recall from Section 3.B.2 that the photon event space is the whole trajectory space $\mathbb{P} = \Psi$, while the geometric measure is $\varrho = l \times v \times \sigma \times l^+$. By taking

[12]Notice that we have restricted our definition of a measure space to *positive, total* measures, which simplifies the statement of the theorem somewhat.

the derivative $dQ/d\varrho$, we obtain a quantity

$$L_\lambda^* \;=\; \frac{dQ}{d\varrho} \;=\; \frac{dQ}{dl\,dv\,d\sigma\,dl^+}\,.$$

This quantity is called *spectral radiant sterisent* [Nicodemus 1978, p. 55], and has units of power per unit area per unit solid angle per unit wavelength $[\mathrm{W}\cdot\mathrm{m}^{-3}\cdot\mathrm{sr}^{-1}\cdot\mathrm{nm}^{-1}]$. It is used for the measurement of emission, scattering, and absorption within volumes.

3.B.6.2 Spectral phase space density

As another example, consider the events defined by intersecting the photon trajectories with the surface $t = t_0$. This allows us to measure the instantaneous spatial distribution of the photons, a concept that is particularly useful for steady-state systems. This was the situation studied by Arvo [1995], who developed a measure-theoretic *phase space density* for photons distributed in $\mathbb{R}^3 \times \mathcal{S}^2$.

In our framework, the event space for this situation is

$$\mathbb{P} \;=\; \{t_0\} \times \psi\,,$$

where $\{t_0\}$ denotes the set containing the single value t_0, and recalling that ψ is the phase space $\psi = \mathbb{R}^3 \times \mathcal{S}^2 \times \mathbb{R}^+$. The geometric measure ϱ is just the natural measure on the phase space ψ, with a slight technical modification to account for presence of the fixed time t_0:

$$\varrho \;=\; \Lambda_{t_0} \times v \times \sigma \times l^+\,,$$

where $\Lambda_{t_0}(D) = 1$ if $t_0 \in D$ and $\Lambda_{t_0}(D) = 0$ otherwise. Then the quantity

$$u_\lambda \;=\; \frac{dQ}{d\varrho} \;=\; \frac{dQ}{dv\,d\sigma\,dl^+} \tag{3.23}$$

measures the density of energy with respect to volume, direction, and wavelength $[\mathrm{J}\cdot\mathrm{m}^{-3}\cdot\mathrm{sr}^{-1}\cdot\mathrm{nm}^{-1}]$. We call u_λ the *spectral phase space density*. It is similar to the phase space density u described by Arvo [1995], except that we have also taken the derivative with respect to wavelength.

3.B.6.3 Spectral radiance

As a final example, define the photon events as crossings of a surface S. The event space is

$$\mathbb{P} \;=\; \mathbb{R} \times S \times \mathcal{S}^2 \times \mathbb{R}^+\,,$$

with the geometric measure defined by

$$\varrho \;=\; l \times A \times \sigma_{\mathbf{x}}^{\perp} \times l^{+} \,.$$

Notice that $A \times \sigma_{\mathbf{x}}^{\perp}$ is simply the measure that was used to define radiance. Thus the density

$$L_\lambda \;=\; \frac{dQ}{d\varrho} \;=\; \frac{dQ}{dl\,dA\,d\sigma_{\mathbf{x}}^{\perp}\,dl^{+}}$$

corresponds to *spectral radiance* as defined earlier (3.9). Notice that although this definition is valid only on the surface S, the choice of S was arbitrary. Thus we can use this equation to define L_λ anywhere in the trajectory space Ψ.

3.B.7 Discussion: fundamental vs. derived quantities

It is sometimes claimed that spectral radiance is the "fundamental" radiometric quantity, from which all others can be derived. As we have seen, this is not so. All of the three quantities defined in Section 3.B.6 are fundamental, because they measure *different sets of photon events*. It is not possible to obtain one from another by integration. Each quantity must be defined independently, by first specifying the photon events, and then describing their density using a Radon-Nikodym derivative.[13]

There is not even a unique geometric space that we can use, since different kinds of photon events require different geometric measures. In some cases, the measure can be defined on all of Ψ (as with L_λ^*), while in other cases it must be defined on a lower-dimensional subset of Ψ (as with u_λ and L_λ).

Note that many "derived" quantities (i.e. one that is obtained by integrating a fundamental quantity, as we did in Section 3.4 to obtain radiance from spectral radiance) can be interpreted directly as Radon-Nikodym derivatives, by reducing the dimension of the underlying measure space. For example, to interpret radiance as a Radon-Nikodym derivative, we could redefine the trajectory space to be $\mathbb{R} \times \mathbb{R}^3 \times \mathcal{S}^2$ (omitting the wavelength parameter), and then proceed as for spectral radiance

[13]Note that by making additional assumptions, it is often possible to express one fundamental quantity in terms of another. For example, Arvo [1995, p. 26] shows how radiance can be defined in terms of the phase space density u, by assuming that all photons travel at the same speed c. He then observes that radiance and phase space density are related according to $L_\lambda = c\,u_\lambda$, where c is the speed of light. (A similar observation appears in [Milne 1930, p. 76].)

Note that this relationship is only true in a vacuum, since in general photons travel at the speed c/η (where η is the local refractive index, which may vary with position). It is even possible that photons at same point in space will travel at different speeds (i.e. if they have different wavelengths, in a dispersive medium). Thus in general, u_λ and L_λ cannot be derived from each other without additional assumptions, so that we consider both of them to be fundamental quantities.

(Section 3.B.6). This technique can be used to give rigorous meaning to the various derivative notations we used in Section 3.4, such as $E = d^2Q \,/\, (dt\, dA)$.