Chapter 4

A General Operator Formulation of Light Transport

The goal of this chapter is to develop a rigorous theoretical basis for bidirectional light transport algorithms. Current frameworks do not adequately describe the relationships between light and importance transport; between finite element, recursive evaluation, and particle tracing approaches; or between incident and exitant transport quantities, especially when materials with non-symmetric BSDF's are used. As a result, given a bidirectional algorithm that uses some combination of these features, it can be difficult to verify whether it actually solves the original transport equations. This can lead to significant mistakes when bidirectional algorithms are implemented, as we will see in Chapter 5.

To remedy these problems, we need a better theoretical framework for light transport calculations. This theory should clearly state the relationships between the various solution techniques mentioned above, using only a small number of basic concepts. It should also show how these techniques are affected by non-symmetric scattering, and specify a set of rules that allow correct results to be obtained. All components of the framework should be expressed in terms of standard mathematical concepts, and the notation should be concise and yet rigorous.

In this chapter, we develop a light transport framework that addresses these goals. It concisely expresses the relationships between light and importance transport, in both their incident and exitant forms, and also their relationship to particle tracing. The fundamental building blocks used are measures, function spaces, inner products, and linear operators. Our work builds directly on the elegant formulation of Arvo [1995], who considered light transport among reflective surfaces with symmetric BRDF's. We also incorporate ideas from Christensen et al. [1993], Schröder & Hanrahan [1994], and Pattanaik & Mudur [1995].

However, many aspects of our framework are new. Most importantly, we do not make any assumptions about the symmetry of BSDF's. This leads to a framework with a richer structure than previous approaches. There are four distinct transport quantities L_{i}, L_{o}, W_{i}, W_{o}, corresponding to incident/exitant radiance/importance. For each of these quantities, there is a distinct transport operator and measurement equation. All of these are related in a simple way, since they are constructed from just two basic elements: the *scattering* and *propagation* operators, which describe independent aspects of the light transport process. This additional structure actually helps to clarify the relationships among transport quantities, since we can see which relationships are fundamental, and which depend on the symmetry of the BSDF.

There are several other contributions. We characterize particle tracing in a new and more useful way, as a condition on the probability distribution of a set of weighted sample rays. We also introduce the *ray space* abstraction, which simplifies the notation and clarifies the structure of light transport calculations. Finally, we point out that *incident* rather than *field* radiance functions must be used to make certain transport operators self-adjoint.

This chapter is organized as follows. We start by defining the ray space and reviewing some useful properties of functions on ray space. Next, we describe the scattering and propagation operators, and we show how they can be used to represent light transport. We then consider sensors and measurements, and show that the scattering and propagation operators can also be used for importance transport. In Section 4.7, we give a summary of the complete transport framework.

Appendix 4.A considers particle tracing algorithms, and describes a new condition that can be used to verify their correctness. Finally, Appendix 4.B gives an analysis of the inverses, adjoints, and norms of the operators we have defined.

4.1 Ray space

We define the ray space and throughput measure, which together form a natural basis for light transport calculations. We show that it is possible to represent the ray space in more than one way, and we also discuss the advantages of defining the ray space abstractly, as opposed to using an explicit representation of rays.

The *ray space* $\mathcal{R}$ consists of all rays that start at points on the scene surfaces. Formally, $\mathcal{R}$ is the Cartesian product

$$\mathcal{R} \;=\; \mathcal{M} \times \mathcal{S}^2\,, \tag{4.1}$$

where as usual, $\mathcal{M}$ is the set of surfaces in the scene, and $\mathcal{S}^2$ is the set of all unit direction vectors. The ray $\mathbf{r} = (\mathbf{x}, \omega)$ has origin $\mathbf{x}$ and direction ω. The reason for requiring the origin to lie on a surface is that in the absence of participating media, the radiance along a given ray is constant. Thus instead of representing the radiance at every point in an environment, it is sufficient to represent the radiance leaving surfaces.

The throughput measure. We define a measure μ on $\mathcal{R}$, called the *throughput measure*, that is used to integrate functions on ray space. Consider a small bundle of rays around a central ray $\mathbf{r} = (\mathbf{x}, \omega)$, such that the origins of these rays occupy an area dA, and their directions lie within a solid angle of $d\sigma$. Then the throughput of this small bundle is defined as

$$d\mu(\mathbf{r}) \;=\; d\mu(\mathbf{x}, \omega) \;=\; dA(\mathbf{x})\, d\sigma^{\perp}_{\mathbf{x}}(\omega)\,, \tag{4.2}$$

that is, μ is simply the product of the area and projected solid angle measures. This is known as the *differential form* of the throughput measure. Note that μ is invariant under Euclidean transformations, which makes it unique up to a constant factor [Ambartzumian 1990, p. 51].

To define $\mu(D)$ for a general set of rays $D \subset \mathcal{R}$, we integrate the differential measure (4.2) over the domain D:

$$\mu(D) \;=\; \int_D dA(\mathbf{x})\, d\sigma^{\perp}_{\mathbf{x}}(\omega)\,,$$

which can be written more explicitly as

$$\mu(D) \;=\; \int_{\mathcal{M}} \sigma^{\perp}_{\mathbf{x}}(D_{\mathbf{x}})\, dA(\mathbf{x}) \qquad \text{where} \qquad D_{\mathbf{x}} \;=\; \{\omega \mid (\mathbf{x}, \omega) \in D\}\,.$$

The quantity $\mu(D)$ measures the light-carrying capacity of a bundle of rays D, and corresponds to the classic radiometric concept of *throughput* [Steel 1974, Nicodemus 1976, Cohen & Wallace 1993]. The measure μ is also similar to the usual measure on lines in $\mathbb{R}^3$ (see [Ambartzumian 1990]). However, note that the measures on line space and ray space are not the same, since unlike line space, the ray space $\mathcal{R}$ can contain distinct rays that are colinear (corresponding to lines that intersect $\mathcal{M}$ at more than one point).

The differential form (4.2) of the throughput measure can be written in several alternative forms that are sometimes useful. By expanding the definition (3.1) of projected solid angle, we get

$$\begin{aligned} d\mu(\mathbf{x}, \omega) &= |\omega \cdot \mathbf{N}_g(\mathbf{x})| \, dA(\mathbf{x}) \, d\sigma(\omega) \,, &(4.3)\\ &= dA^{\perp}_{\omega}(\mathbf{x}) \, d\sigma(\omega) \,, &(4.4)\end{aligned}$$

where $A^{\perp}$ is the projected area measure (3.7). All of these definitions are equivalent.

The throughput measure also allows us to define radiance in a simpler and more natural way, namely as power per unit throughput:

$$L(\mathbf{r}) = \frac{d\Phi(\mathbf{r})}{d\mu(\mathbf{r})} \,, \tag{4.5}$$

It is easy to check that this definition is equivalent to the ones given in Section 3.4.3.

Other representations of ray space. Although we will most often use the representation $\mathbf{r} = (\mathbf{x}, \omega)$ for a ray, it is possible to represent the ray space in other ways. For example, we could define $\mathcal{R}$ as

$$\mathcal{R} = \mathcal{M} \times \mathcal{M} \,, \tag{4.6}$$

so that each ray is a pair $\mathbf{r} = \mathbf{x} \to \mathbf{x}'$ (where the arrow notation denotes the direction of the ray). Notice that there is some redundancy in this representation, since the rays $\mathbf{x} \to \mathbf{x}'$ and $\mathbf{x} \to \mathbf{x}''$ are equivalent whenever $\mathbf{x}'$ and $\mathbf{x}''$ lie in the same direction from $\mathbf{x}$. However, this redundancy is sometimes useful: for example, it allows us to construct a basis for functions on ray space as a tensor product of bases defined on the scene surfaces. Also notice that with this representation, there is no way to represent light that radiates out to infinity: thus, it is most useful when $\mathcal{M}$ is a closed environment, and we are only interested in light transport between elements of $\mathcal{M}$.

Even when $\mathcal{R}$ is represented in different ways, the throughput measure μ should be understood to have the same meaning. For example, with the representation above, μ is defined by

$$d\mu(\mathbf{x} \to \mathbf{x}') \;=\; V(\mathbf{x} \leftrightarrow \mathbf{x}') \, \frac{\cos(\theta) \, \cos(\theta')}{\|\mathbf{x} - \mathbf{x}'\|^2} \, dA(\mathbf{x}) \, dA(\mathbf{x}') \,. \tag{4.7}$$

Here θ and θ' are the angles between the segment $\mathbf{x} \leftrightarrow \mathbf{x}'$ and the surface normals at $\mathbf{x}$ and $\mathbf{x}'$ respectively, while $V(\mathbf{x} \leftrightarrow \mathbf{x}')$ is the visibility function, which is 1 if $\mathbf{x}$ and $\mathbf{x}'$ are mutually visible and 0 otherwise.[1] As usual, the notation $\mathbf{x} \to \mathbf{x}'$ indicates the direction of a ray, and $f(\mathbf{x} \leftrightarrow \mathbf{x}')$ indicates a symmetric function.

Advantages of the ray space abstraction. There are several reasons to use the abstract representation $\mathbf{r} \in \mathcal{R}$ for rays, rather than writing $(\mathbf{x}, \omega)$ explicitly. First, it clarifies the structure of radiometric formulas, by hiding the details of the ray representation. Second, it emphasizes that the representation is a superficial decision that can easily be changed. Finally, it allows us to define concepts whose meanings do not depend on how the rays are represented, e.g. the throughput measure μ.

4.2 Functions on ray space

The distribution of radiance or importance in a given scene can be represented as a real-valued function on ray space, i.e. a function of the form

$$f : \mathcal{R} \to \mathbb{R} \,.$$

In this section, we study the properties of such functions (e.g. norms and inner products), and review some terminology related to *function spaces* (i.e. collections of functions that all have some specified property). These ideas will be used later to analyze the properties of light transport operators.

[1] That is, $V(\mathbf{x} \leftrightarrow \mathbf{x}') = 1$ if the open line segment between $\mathbf{x}$ and $\mathbf{x}'$ does not intersect $\mathcal{M}$. Note that the visibility factor can be removed from the definition (4.7), by restricting $\mathcal{R}$ to contain only those rays where $V(\mathbf{x} \leftrightarrow \mathbf{x}') = 1$.

Norms. We restrict our attention to the L_p *norms*, which are defined by

$$\|f\|_p \;=\; \left(\int_{\mathcal{R}} |f(\mathbf{r})|^p \, d\mu(\mathbf{r}) \right)^{1/p} , \tag{4.8}$$

where p is a positive integer. In the limit as $p \to \infty$, we obtain the L_∞ *norm*:

$$\|f\|_\infty \;=\; \operatorname*{ess\,sup}_{\mathbf{r} \in \mathcal{R}} |f(\mathbf{r})| , \tag{4.9}$$

where $\operatorname{ess\,sup}$ denotes the *essential supremum*, i.e. the smallest real number m such that $f(\mathbf{r}) \le m$ almost everywhere.[2] The most commonly used norms are the L_1, L_2, and L_∞ norms, which measure the average, root-mean-square (RMS), and maximum absolute values of a function respectively. When the particular norm being used is not important, we will simply write $\|f\|$.

For the purposes of analysis, it is convenient to consider only the functions whose L_p norm is finite. The collection of all such functions (for a given value of p) is called an L_p *space*, which we will denote by $L_p(\mathcal{R})$ (to emphasize the domain $\mathcal{R}$ of these functions). These spaces have desirable analytic properties (which depend on the assumption of finite norms).

There are a variety of terms that are used to describe L_p spaces, corresponding to the various properties that they possess. At the most basic level, they are *vector spaces*, since each space $L_p(\mathcal{R})$ is closed under the operations of addition and scalar multiplication. Vectors spaces are also known as *linear spaces*, and in this context, as *function spaces* (since each element of $L_p(\mathcal{R})$ is a function).

The L_p spaces are also *complete*, meaning that all Cauchy sequences converge[3] (this property is useful for analysis). Thus, $L_p(\mathcal{R})$ is a complete, normed, linear space; in the terminology of functional analysis, this is called a *Banach space*.

---

[2] *Almost everywhere* means that the rays for which $f(\mathbf{r}) > m$ form a set of measure zero (with respect to the throughput measure μ). Thus according to this definition, the essential supremum ignores values of f that are attained only at isolated points, etc.

[3] A sequence of functions $f_1, f_2, \ldots$ is a *Cauchy sequence* if for any $\epsilon > 0$, there is an index N such that $\|f_i - f_j\| < \epsilon$ for all $i, j > N$. Such a sequence *converges* if there is a function $f \in L_p(\mathcal{R})$ such that $\lim_{N\to\infty} \|f_i - f\| = 0$.

Inner products. Another useful operation is the *inner product* of two functions on ray space, defined by

$$\langle f, g \rangle \;=\; \int_{\mathcal{R}} f(\mathbf{r})\, g(\mathbf{r})\, d\mu(\mathbf{r})\,. \tag{4.10}$$

Notice that the inner product notation is more concise than writing the integral explicitly, and yet it also imparts more information (since it can immediately be recognized as an inner product, rather than some other kind of integral). A linear space $\mathcal{F}$ equipped with an inner product is called an *inner product space*.

Every inner product has an *associated norm* defined by

$$\|f\| \;=\; \langle f, f \rangle^{1/2}\,,$$

which in this case is identical to the L_2 norm. Thus, the space $L_2(\mathcal{R})$, together with the inner product (4.10), is an example of an inner product space that is complete with respect to its associated norm: this is called a *Hilbert space*.

It is also possible to define weighted inner products between functions on ray space, by multiplying the integrand of (4.10) by a positive weighting function $w(\mathbf{r})$. This technique can also be used to define other norms. In this chapter, however, we will only have need for the unweighted versions defined above.

4.3 The scattering and propagation operators

From a physical standpoint, we can consider light transport to be an alternation of two steps. The first is *scattering*, which describes the interaction of photons with surfaces. The other is *propagation*, in which photons travel in straight lines through a fixed medium. Following Arvo et al. [1994], we will define each of these steps as a linear operator acting on radiance functions.

A *linear operator* is simply a linear function $\mathbf{A} : \mathcal{F} \to \mathcal{F}$ whose domain is a vector space $\mathcal{F}$. In our case, $\mathcal{F}$ is a space of radiance functions, as defined above. The notation $\mathbf{A}f$ denotes the application of an operator to a function, whose result is another function.

The local scattering operator. We begin with the *local scattering operator*, defined by[4]

$$(\mathbf{K}h)(\mathbf{x},\omega_\mathrm{o}) \;=\; \int_{\mathcal{S}^2} f_\mathrm{s}(\mathbf{x},\omega_\mathrm{i}\!\rightarrow\!\omega_\mathrm{o})\, h(\mathbf{x},\omega_\mathrm{i})\, d\sigma^\perp_\mathbf{x}(\omega_\mathrm{i})\,. \tag{4.11}$$

When this operator is applied to an incident radiance function L_i, it returns the exitant radiance $L_\mathrm{o} = \mathbf{K}L_\mathrm{i}$ that results from a single scattering operation. Equation (4.11) is similar to the scattering equation (3.12), except that $\mathbf{K}$ operates on entire radiance functions, rather than being restricted to a single point $\mathbf{x}$. It maps one function L into another function $\mathbf{K}L$, where each function is defined over the whole ray space $\mathcal{R}$.

The propagation operator. To define the propagation operator, we first give a more precise definition of the *ray-casting function* $\mathbf{x}_\mathcal{M}(\mathbf{x},\omega)$ mentioned in Section 3.7. First, let

$$d_\mathcal{M}(\mathbf{x},\omega) \;=\; \inf\{d > 0 \mid \mathbf{x} + d\omega \in \mathcal{M}\}\,, \tag{4.12}$$

which is called the *boundary distance function* [Arvo 1995, p. 136]. We then define the ray-casting function as

$$\mathbf{x}_\mathcal{M}(\mathbf{x},\omega) \;=\; \mathbf{x} + d_\mathcal{M}(\mathbf{x},\omega)\,\omega\,, \tag{4.13}$$

so that $\mathbf{x}_\mathcal{M}(\mathbf{x},\omega)$ represents the first point of $\mathcal{M}$ that is visible from $\mathbf{x}$ in the direction ω. When the ray $(\mathbf{x},\omega)$ does not intersect $\mathcal{M}$, we have $d_\mathcal{M}(\mathbf{x},\omega) = \infty$, and $\mathbf{x}_\mathcal{M}$ is not defined.[5]

The propagation of light in straight lines is now represented by the *geometric* or *propagation operator* $\mathbf{G}$, defined by

$$(\mathbf{G}h)(\mathbf{x},\omega_\mathrm{i}) \;=\; \begin{cases} h(\mathbf{x}_\mathcal{M}(\mathbf{x},\omega_\mathrm{i}),-\omega_\mathrm{i}) & \text{if } d_\mathcal{M}(\mathbf{x},\omega_\mathrm{i}) < \infty\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{4.14}$$

This operator expresses the incident radiance L_i in terms of the exitant radiance L_o leaving the other surfaces of the environment, according to $L_\mathrm{i} = \mathbf{G}L_\mathrm{o}$.

These definitions of $\mathbf{G}$ and $\mathbf{K}$ are slightly different than those of Arvo [1995]. First, we

[4] Although this definition seems to depend on the particular ray representation $\mathbf{r} = (\mathbf{x},\omega)$, in fact it can be used with any representation. To do this, simply replace the argument on the left-hand side by a single parameter $\mathbf{r}$, and replace the symbols $\mathbf{x}$ and ω_o by functions $\mathbf{x} = \mathbf{x}(\mathbf{r})$ and $\omega_\mathrm{o} = \omega_\mathrm{o}(\mathbf{r})$ (whose definitions depend on the representation used).

[5] With respect to equation (4.12), we have used the convention that $\inf \emptyset = \infty$, where $\emptyset$ is the empty set.

have considered transmission as well as reflection, by using the BSDF in the definition of $\mathbf{K}$. Second, we have used incident and exitant radiance rather than field and surface radiance (see Section 3.5), so that the direction of ω_{i} is reversed compared to [Arvo 1995]. The main advantage of this convention is that $\mathbf{G}$ and $\mathbf{K}$ are both self-adjoint when f_{s} is symmetric, which greatly increases the symmetry between light and importance transport (as we will see in Section 4.6). On the other hand, the $\mathbf{G}$ and $\mathbf{K}$ defined by Arvo are not self-adjoint. He handles this by introducing an isomorphism $\mathbf{H}$ between surface and field radiance functions, such that $\mathbf{HG}$ and $\mathbf{KH}$ are equivalent to the $\mathbf{G}$ and $\mathbf{K}$ defined here [Arvo 1995, p. 151].

Locality. Notice that to evaluate the radiance scattered along a given ray $(\mathbf{x}, \omega)$, we only need to know the incident radiance at the same point $\mathbf{x}$. In other words, the evaluation of $(\mathbf{K}h)(\mathbf{x}, \omega)$ only requires the evaluation of h on rays of the form $(\mathbf{x}, \omega')$. This property of the scattering operator $\mathbf{K}$ is called *locality*.

In general, we say that a transport operator $\mathbf{A}$ is *local* if the evaluation of $(\mathbf{A}h)(\mathbf{r})$ only requires the evaluation of h on a small set of rays $\mathbf{r}'$. In this sense, the propagation operator $\mathbf{G}$ is also local, since to evaluate $(\mathbf{G}h)(\mathbf{x}, \omega)$ we only need the value of h on a single ray $(\mathbf{x}', -\omega)$. In fact we could say that $\mathbf{G}$ is more local than $\mathbf{K}$, since $(\mathbf{G}h)(\mathbf{r})$ depends on the value of h on a single ray, while $(\mathbf{K}h)(\mathbf{r})$ depends on the value of h on a two-dimensional subset of $\mathcal{R}$.

Locality is important, since it dictates how much of the domain of h must be examined in order to compute $(\mathbf{A}h)(\mathbf{r})$ for a given ray $\mathbf{r}$. This type of locality has been successfully exploited in radiosity calculations, in order to handle textures more efficiently [Gershbein et al. 1994].

4.4 The light transport and solution operators

The composition of the scattering and propagation operators is called the *light transport operator*,

$$\mathbf{T} \;=\; \mathbf{K}\mathbf{G}\,.$$

This operator maps an exitant radiance function L_{o} into the exitant function $\mathbf{T}L_{\mathrm{o}}$ that results after a single scattering step. (When there is no ambiguity, we will drop the subscript on exitant functions and simply write L.)

Recall that our goal is to measure the equilibrium radiance L. The condition that must be satisfied in equilibrium is that

$$L \;=\; L_{\mathrm{e}} + \mathbf{T}L\,, \tag{4.15}$$

where $L_{\mathrm{e}}(\mathbf{r})$ is the emitted radiance function (specified as part of the scene model). This is called the *light transport equation*. It is simply a reformulation of (3.19), which says that at equilibrium, the exitant radiance must be the sum of emitted and scattered radiance.

The solution operator. Formally, the solution can be obtained by inverting the operator equation (4.15):

$$\begin{aligned}(\mathbf{I}-\mathbf{T})\,L &= L_{\mathrm{e}}\\ L &= (\mathbf{I}-\mathbf{T})^{-1}\,L_{\mathrm{e}}\,,\end{aligned}$$

where $\mathbf{I}$ is the identity operator. It is convenient to rewrite this equation in terms of the *solution operator*[6]

$$\mathbf{S} \;=\; (\mathbf{I}-\mathbf{T})^{-1}\,, \tag{4.16}$$

in which case the solution is simply $L = \mathbf{S}L_{\mathrm{e}}$.

Conditions for invertibility. These formal manipulations are valid only if the operator $\mathbf{I}-\mathbf{T}$ is invertible. A sufficient condition is that $\|\mathbf{T}\| < 1$, where $\|\mathbf{T}\|$ is the standard *operator norm*

$$\|\mathbf{T}\| \;=\; \sup_{\|f\|\le 1} \|\mathbf{T}f\|\,, \tag{4.17}$$

[6]Note that $\mathbf{S}$ is closely related to the *resolvent operator* $\mathbf{R}_\lambda$ used in spectral analysis, except that $\mathbf{R}_\lambda$ has a parameter λ, and does not have a universally accepted definition (e.g. compare [Delves & Mohamed 1985, p. 74], [Taylor & Lay 1980, p. 272]). It is also closely related to the "GRDF" of Lafortune & Willems [1994], which is simply a new name for the kernel of the solution operator $\mathbf{S}$.

where the norms on the right are function norms.[7] Given that $\|\mathbf{T}\| < 1$, the inverse of $\mathbf{I} - \mathbf{T}$ exists and is given by

$$\mathbf{S} \;=\; (\mathbf{I} - \mathbf{T})^{-1} \;=\; \sum_{i=0}^{\infty} \mathbf{T}^i \;=\; \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots \,. \tag{4.18}$$

This is called the *Neumann series* (after C. Neumann, though the method goes back as far as Liouville [Taylor & Lay 1980, p. 191]). This expansion has a physical interpretation when applied to $L = \mathbf{S} L_{\mathrm{e}}$, since

$$L \;=\; L_{\mathrm{e}} + \mathbf{T} L_{\mathrm{e}} + \mathbf{T}^2 L_{\mathrm{e}} + \cdots$$

expresses L as the sum of emitted light, plus light that has been scattered once, twice, etc.

The validity of (4.18) raises the issue of whether $\|\mathbf{T}\| < 1$. In general, this depends on physical assumptions about the scene model, as well as the norm used for radiance functions. We will consider several cases.

For (one-sided) reflective surfaces, Arvo has shown that $\|\mathbf{G}\|_p \le 1$ for any $1 \le p \le \infty$ [Arvo 1995, Theorem 14]. Furthermore, he has shown that $\|\mathbf{K}\|_p \le 1$, as long as all BRDF's in the scene are energy-conserving and symmetric. By making the additional assumption that no surface is perfectly reflective, he obtains $\|\mathbf{K}\|_p < 1$ [Arvo 1995, Theorem 13], and thus

$$\|\mathbf{T}\|_p \;=\; \|\mathbf{K}\mathbf{G}\|_p \;\le\; \|\mathbf{K}\|_p \, \|\mathbf{G}\|_p \;<\; 1 \,.$$

In the case of general scattering (i.e. transmission as well as reflection), things are slightly more complicated. Arvo's proof that $\|\mathbf{G}\| \le 1$ requires some modifications, because it depends on the fact that $\mathbf{G}^2 = \mathbf{I}$ when $\mathcal{M}$ forms an enclosure (which does not hold under the more general assumptions considered here). We will give a different proof below (Appendix 4.B). As for $\mathbf{K}$, it is no longer true that $\|\mathbf{K}\| < 1$. In fact, it is only true that

$$\|\mathbf{K}\| \;<\; \frac{\eta_{\max}^2}{\eta_{\min}^2} \,,$$

[7] Each function norm induces a distinct operator norm. The notation $\|\cdot\|_p$ can mean either the L_p norm on functions (4.8), or the corresponding operator norm, depending on the type of its argument.

where $\eta_{\min}$ and $\eta_{\max}$ denote the minimum and maximum refractive indices in the environment. This corresponds to the fact that radiance can increase during scattering, due to refraction. Putting these facts together, it is possible that $\|\mathbf{T}\|_p > 1$.

However, the condition $\|\mathbf{T}\| < 1$ is not strictly necessary, since the inverse of $\mathbf{I} - \mathbf{T}$ exists whenever the series given by (4.18) converges [Taylor & Lay 1980, p. 192]. A weaker yet sufficient condition for convergence is that $\|\mathbf{T}^k\| < 1$ for some $k \geq 1$.[8] In Section 4.B.3, we will show that this condition is satisfied for any physically valid scene model, and therefore the Neumann series converges (which makes $\mathbf{S}$ well-defined).

4.5 Sensors and measurements

The goal of light transport algorithms is to estimate a finite number of measurements of the equilibrium radiance L. For example, if the algorithm computes an image directly, then the measurements consist of many pixel values $I_1, \ldots, I_M$, where M is the number of pixels in the image. If the algorithm computes a finite-element solution, on the other hand, then the measurements I_j would simply be the basis function coefficients (with one measurement for each basis function).

Each measurement can be thought of as the response of a hypothetical sensor placed somewhere in the scene. For example, we can imagine that each pixel is a small piece of film within a virtual camera, and that the pixel value is proportional to the radiant power that it receives. Of course, most of the time the camera and lens system are not modeled explicitly. However, for any given pixel it is still possible to identify the set of rays in world space that contribute to its value, and assume that there is an imaginary sensor that responds to the radiance along these rays.

The sensor response can vary according to the position and direction of the incident radiance. We will only deal with linear sensors, in which case the response is characterized by a function

$$W_{\mathrm{e}}(\mathbf{x}, \omega) \;=\; \frac{dS(\mathbf{x}, \omega)}{d\Phi(\mathbf{x}, \omega)} \qquad (4.19)$$

[8] Note that this condition implies that perfectly reflective mirrors are allowable, as long as it is not possible for light to continue bouncing indefinitely between these mirrors without some light escaping to another (more absorptive) portion of the scene.

that specifies the sensor response per unit of power arriving at $\mathbf{x}$ from direction ω. For real sensors, $W_{\rm e}$ is called the *flux responsivity* of the sensor [Nicodemus 1978, p. 59]. The corresponding units are $[S \cdot \mathrm{W}^{-1}]$, where S is the unit of sensor response. Depending on the sensor, S could represent a voltage, current, change in photographic film density, deflection of a meter needle, etc.

For the hypothetical sensors used in graphics, $W_{\rm e}$ is called an *exitant importance function* (we think of the sensor as emitting importance).[9] The corresponding sensor response is unitless, and thus importance has units of $[\mathrm{W}^{-1}]$. However, the symbol S is a convenient reminder that something is being measured. We assume that $W_{\rm e}$ is defined over the entire ray space $\mathcal{R}$, although it will be zero over most of this domain for typical sensors. In the case where measurements represent pixel values, note that $W_{\rm e}$ can model arbitrary lens systems used to form the image, as well as any linear filters used for anti-aliasing.

The measurement equation. To compute a measurement, we integrate the response

$$dS(\mathbf{r}) \;=\; W_{\rm e}(\mathbf{r})\, d\Phi(\mathbf{r}) \;=\; W_{\rm e}(\mathbf{r})\, L_{\rm i}(\mathbf{r})\, d\mu(\mathbf{r})$$

for all the incident radiance falling on the sensor. This is summarized by Nicodemus' *measurement equation* [Nicodemus 1978, p. 85], expressed in our notation as

$$I \;=\; \langle W_{\rm e}, L_{\rm i} \rangle \;=\; \int_{\mathcal{R}} W_{\rm e}(\mathbf{r})\, L_{\rm i}(\mathbf{r})\, d\mu(\mathbf{r})\,, \tag{4.20}$$

where I is a measurement, $W_{\rm e}$ is the emitted importance, and $L_{\rm i}$ is the incident radiance.[10]

Notice that we have defined both $L_{\rm e}$ and $W_{\rm e}$ as exitant quantities. This is natural, since it lets us define their values at points *on* the source or sensor. It would not be intuitive to define

[9] This follows the terminology of [Lewins 1965, p. 7, p. 21], where each importance function pertains to a single "meter reading" (measurement). The alternative term *potential function* [Pattanaik & Mudur 1995] is undesirable because it has a well-known, different meaning in physics (a function satisfying Poisson's equation, e.g. the electric or gravitational potential functions).

It is also allowable for an importance function to represent the average of a set of measurements (e.g. the average of all pixel values in an image). This is the case with importance-driven radiosity methods [Smits et al. 1992, p. 275], where the importance function is used only to guide the solution (and the value of the corresponding "measurement" is irrelevant).

[10] Strictly speaking, the measurement equation should also integrate over frequency (since $L_{\rm i}$ and $W_{\rm e}$ are spectral quantities, defined separately at each frequency ν). However, to simplify the notation we will usually ignore this detail.

$L_e(\mathbf{x}, \omega)$ as the amount of light *arriving* at $\mathbf{x}$ from direction ω, since the actual emission takes place somewhere else. Similarly, it is more natural to define sensor response in terms of radiance arriving at the sensor, rather than radiance leaving points elsewhere in the scene (e.g. as with the "visual potential" proposed by Pattanaik & Mudur [1995]). The definitions of L_e and W_e as exitant quantities also increases the symmetry between light and importance transport [Christensen et al. 1993], as we discuss below.

Also notice that although we have defined the equilibrium solution $L = \mathbf{S}L_e$ as an exitant quantity, the measurement equation (4.20) requires an incident function. This problem can be solved with the $\mathbf{G}$ operator, by using the relationship $L_i = \mathbf{G}L$.[11] Each measurement now has the form

$$I \;=\; \langle W_e, L_i \rangle \;=\; \langle W_e, \mathbf{G}L \rangle \;=\; \langle W_e, \mathbf{G}\mathbf{S}L_e \rangle\,. \tag{4.21}$$

Notice that it is the explicit inclusion of $\mathbf{G}$ in this equation that allows us to use the exitant forms of both L_e and W_e.

4.6 Importance transport via adjoint operators

Adjoint operators are a powerful tool for understanding light transport algorithms. They allow us to evaluate measurements in a variety of ways, which can lead to new insights and rendering algorithms.

The *adjoint* of an operator $\mathbf{H}$ is denoted $\mathbf{H}^*$, and is defined by the property that[12]

$$\langle \mathbf{H}^* f, g \rangle \;=\; \langle f, \mathbf{H}g \rangle \quad \text{for all} \quad f, g\,. \tag{4.22}$$

An operator is *self-adjoint* if $\mathbf{H} = \mathbf{H}^*$. This corresponds to the familiar concept of a symmetric matrix in real linear algebra.

To show how the adjoint can be used, we apply the identity (4.22) to the measurement

[11] In order for $L_i = \mathbf{G}L$ to represent the radiance arriving at the sensors, the sensors must be modeled as part of the domain $\mathcal{M}$. For the purposes of this framework, the sensors can be modeled without affecting light transport in the rest of the scene by making them completely transparent.

[12] The adjoint of an operator depends on the inner product used. In this chapter, we always use the inner product (4.10).

equation (4.21), yielding

$$I \;=\; \langle W_{\rm e}, \mathbf{G}\mathbf{S} L_{\rm e} \rangle \;=\; \langle (\mathbf{G}\mathbf{S})^* W_{\rm e}, L_{\rm e} \rangle\,. \tag{4.23}$$

This suggests that we can evaluate I by transporting importance in some way. To determine exactly what this means, we must express the operator $(\mathbf{G}\mathbf{S})^*$ in terms of the known operators $\mathbf{G}$ and $\mathbf{K}$.

We start by examining the adjoints of $\mathbf{G}$ and $\mathbf{K}$. It is relatively straightforward to show that $\mathbf{G} = \mathbf{G}^*$. On the other hand, the adjoint of $\mathbf{K}$ is given by

$$(\mathbf{K}^* h)(\mathbf{x}, \omega_{\rm o}) \;=\; \int_{\mathcal{S}^2} f_{\rm s}^*(\mathbf{x}, \omega_{\rm i} \!\rightarrow\! \omega_{\rm o})\, h(\mathbf{x}, \omega_{\rm i})\, d\sigma_{\mathbf{x}}^{\perp}(\omega_{\rm i}) \tag{4.24}$$

(see Appendix 4.B for proofs of these results). Notice that $\mathbf{K}^*$ is the same as $\mathbf{K}$, except that it uses the adjoint BSDF

$$f_{\rm s}^*(\mathbf{x}, \omega_{\rm i} \!\rightarrow\! \omega_{\rm o}) = f_{\rm s}(\mathbf{x}, \omega_{\rm o} \!\rightarrow\! \omega_{\rm i})\,.$$

For now, let us suppose that $f_{\rm s}$ is symmetric at every point $\mathbf{x} \in \mathcal{M}$, so that $\mathbf{K} = \mathbf{K}^*$. Putting these facts together with standard identities (Appendix 4.B), it is easy to show that

$$(\mathbf{G}\mathbf{S})^* \;=\; \mathbf{G}\mathbf{S}\,, \tag{4.25}$$

i.e. the operator $(\mathbf{G}\mathbf{S})$ is self-adjoint as well.

Thus according to (4.23), measurements can be evaluated using either

$$I \;=\; \langle W_{\rm e}, \mathbf{G}\mathbf{S} L_{\rm e} \rangle \qquad \text{or} \qquad I \;=\; \langle \mathbf{G}\mathbf{S} W_{\rm e}, L_{\rm e} \rangle\,. \tag{4.26}$$

The only difference between these two expressions is that $W_{\rm e}$ and $L_{\rm e}$ have been exchanged.

Importance transport. The significance of this symmetry is that any algorithms that apply to light transport may also be used for importance transport. There is an exact correspondence between the concepts, quantities, and equations in the two cases. In particular, the *equilibrium importance function* is given by $W \;=\; \mathbf{S} W_{\rm e}$, and satisfies the *importance transport equation*

$$W \;=\; W_{\rm e} + \mathbf{T} W\,.$$

Similarly, the relationship $L_{\mathrm{i}} = \mathbf{G} L_{\mathrm{o}}$ for incident radiance becomes $W_{\mathrm{i}} = \mathbf{G} W_{\mathrm{o}}$ for incident importance. Rewriting (4.26) using these definitions, we get the symmetric measurement equations

$$I = \langle W_{\mathrm{e}}, L_{\mathrm{i}} \rangle \qquad \text{or} \qquad I = \langle W_{\mathrm{i}}, L_{\mathrm{e}} \rangle \, .$$

However, if the scene model contains any surface with a non-symmetric BSDF, then $\mathbf{K} \neq \mathbf{K}^*$. This does not affect the light transport operator $\mathbf{T} = \mathbf{K}\mathbf{G}$, which we will rename $\mathbf{T}_L$, but the importance transport operator becomes

$$\mathbf{T}_W = \mathbf{K}^* \mathbf{G}$$

(see Appendix 4.B). This means that in general, light and importance obey different transport equations.

Furthermore, we have not yet considered the transport operators for the corresponding incident quantities, L_{i} and W_{i}. This leads to a multitude of possibilities for evaluating measurements, all with different transport equations. Fortunately, all of these equations share the same general structure, as described in the next section.

4.7 Summary of the operator framework

We consider the four basic transport quantities L_{o}, L_{i}, W_{o}, and W_{i}, corresponding to exitant radiance, incident radiance, exitant importance, and incident importance respectively. The propagation operator $\mathbf{G}$ maps exitant quantities to incident ones, according to the relationships

$$L_{\mathrm{i}} = \mathbf{G} L_{\mathrm{o}} \qquad \text{and} \qquad W_{\mathrm{i}} = \mathbf{G} W_{\mathrm{o}} \, . \tag{4.27}$$

Similarly, the local scattering operator $\mathbf{K}$ maps incident quantities to exitant ones:

$$L_{\mathrm{o}} = \mathbf{K} L_{\mathrm{i}} \qquad \text{and} \qquad W_{\mathrm{o}} = \mathbf{K}^* W_{\mathrm{i}} \, . \tag{4.28}$$

Recall that the operators $\mathbf{K}$ and $\mathbf{K}^*$ differ only in the ordering of the BSDF arguments ω_{i} and ω_{o} (see (4.24)).

By putting these relationships together in various ways, we obtain a different transport operator $\mathbf{T}_X$ for each quantity X, where X is one of $L_{\rm i}$, $L_{\rm o}$, $W_{\rm i}$, or $W_{\rm o}$. These operators are summarized in the table below:

	Exitant	Incident
Light	$\mathbf{T}_{L_{\rm o}} = \mathbf{K}\mathbf{G}$	$\mathbf{T}_{L_{\rm i}} = \mathbf{G}\mathbf{K}$
Importance	$\mathbf{T}_{W_{\rm o}} = \mathbf{K}^*\mathbf{G}$	$\mathbf{T}_{W_{\rm i}} = \mathbf{G}\mathbf{K}^*$

To solve for the equilibrium value of any of these quantities, we use the transport equation

$$X \;=\; X_{\rm e} + \mathbf{T}_X X\,,$$

where $X_{\rm e}$ is the given emission function for X. The formal solution to this equation is

$$X \;=\; \mathbf{S}_X X_{\rm e}\,,$$

where $\mathbf{S}_X \;=\; (\mathbf{I} - \mathbf{T}_X)^{-1}$ is called the *solution operator* for X. Finally, a measurement I can be computed using any of the following expressions:

$$\begin{aligned} I \;&=\; \langle W_{\rm e}, L_{\rm i}\rangle \;&=\; \langle W_{\rm i}, L_{\rm e}\rangle \\ &=\; \langle W_{\rm e,i}, L_{\rm o}\rangle \;&=\; \langle W_{\rm o}, L_{\rm e,i}\rangle\,. \end{aligned} \tag{4.29}$$

To apply these equations, recall that we are initially given two emission functions, one that describes the emitted radiance, and one that describes the emitted importance (i.e. the sensor responsivity). Most often, both are given as exitant functions ($L_{\rm e}$ or $W_{\rm e}$), but for some problems, the incident form is more natural ($L_{\rm e,i}$ or $W_{\rm e,i}$). For example, suppose that we wish to project the equilibrium radiance $L_{\rm o}$ onto a set of orthonormal basis functions B_j (e.g. as with finite element approaches). In this situation, the coefficient of each basis function B_j is given by the inner product $\langle B_j, L_{\rm o}\rangle$. Comparing this "measurement" against the templates above (4.29), we see that B_j is considered to be an incident importance function ($W_{\rm e,i}$). This is because B_j specifies the response to radiance *leaving* the corresponding surface, rather than radiance arriving at it.

If the emitted radiance and importance are supplied in opposite forms (one incident and one exitant), the equations above can be applied in a straightforward manner by solving for the equilibrium function of one quantity (e.g. L_{o}), and computing an inner product with the emitted function of the other (e.g. $W_{\mathrm{e,i}}$). On the other hand, if two exitant functions are supplied (L_{e} and W_{e}), one of them must be converted to an incident quantity using the relationships $L_{\mathrm{i}} = \mathbf{G}L_{\mathrm{o}}$ or $W_{\mathrm{i}} = \mathbf{G}W_{\mathrm{o}}$(4.27), before a measurement can be computed (e.g. one possibility is $\langle \mathbf{G}W_{\mathrm{e}}, L_{\mathrm{o}} \rangle$). Similarly, if two incident emission functions are provided ($L_{\mathrm{e,i}}$ and $W_{\mathrm{e,i}}$), one of them must be converted to an exitant function using the relationships $L_{\mathrm{o}} = \mathbf{K}L_{\mathrm{i}}$ or $W_{\mathrm{o}} = \mathbf{K}W_{\mathrm{i}}$ (e.g. a measurement of the form $\langle W_{\mathrm{e,i}}, \mathbf{S}_{L_{\mathrm{o}}}\mathbf{K}L_{\mathrm{e,i}} \rangle$).

Together, these equations specify many ways in which measurements can be made. Everything is constructed from only two basic operators, $\mathbf{G}$ and $\mathbf{K}$, which represent independent components of the light transport process. It is clear which relationships are fundamental, and which depend on the assumption of symmetric BSDF's (i.e. $\mathbf{K} = \mathbf{K}^*$). The notation has been chosen to simplify the structure as much as possible, by using the concepts of ray space, measures, inner products, and linear operators.

Previous authors have stated special cases of these results. For example, Christensen et al. [1993] show that L_{o} and W_{o} satisfy the same transport equation, provided that all BSDF's are symmetric. Similarly, Pattanaik & Mudur [1995] show that L_{o} and W_{i} satisfy adjoint transport equations (i.e. $\mathbf{T}_{W_{\mathrm{o}}} = \mathbf{T}^*_{L_{\mathrm{o}}}$), although their arguments are not rigorous. Arvo [1995] derives the adjoints of $\mathbf{G}$ and $\mathbf{K}$ for one-sided, reflective surfaces, but does not discuss their significance. Adjoint relationships have also been used extensively in the field of neutron transport [Spanier & Gelbard 1969, Lewins 1965], but those results apply to volume scattering rather than surfaces.

Appendix 4.A A new characterization of particle tracing

Intuitively, particle tracing consists of following the paths of "photons" as they are emitted, scattered, and eventually absorbed. These *particle histories* are then used to approximate the equilibrium radiance function, either as a discrete set of measurements (e.g. pixel values, basis function coefficients), or in some other way (e.g. density estimation [Shirley et al. 1995]). This simple idea can be made quite sophisticated by applying different estimators to the particle histories, or by sampling the particles in clever ways [Spanier & Gelbard 1969].

Several explanations of particle tracing have been proposed in computer graphics. Most often these methods are justified intuitively, by appealing to the notion that each particle carries a certain amount of "energy" (e.g. [Shirley et al. 1995]). Pattanaik & Mudur [1995] propose a different approach, by interpreting particle tracing as a random walk solution of the importance transport equation. Our goal is to relate particle-based methods to the transport framework of this chapter, and study the conditions that must be satisfied to ensure that particle tracing algorithms are correct.

Our approach. We present a new characterization of particle tracing that addresses these issues. We define a particle tracing algorithm as a method for generating a set of N weighted sample rays

$$(\alpha_i, \mathbf{r}_i)\,,$$

where each $\mathbf{r}_i$ is a ray, and α_i is its corresponding weight. These samples must be an unbiased representation of the equilibrium radiance L, such that the estimate

$$E\left[\frac{1}{N}\sum_{i=1}^{N}\alpha_i W_{\mathrm{e}}(\mathbf{r}_i)\right] \;=\; \langle W_{\mathrm{e}}, L\rangle \tag{4.30}$$

holds for any importance function W_{e}. Essentially, this identity states that an arbitrary linear measurement can be estimated by taking a weighted sum over the given set of random sample rays.

Formally, this is a condition on the joint density function $p(\alpha, \mathbf{r})$ of the weighted sample rays:

$$\int_{\mathbb{R}} \alpha\, p(\alpha, \mathbf{r})\, d\alpha \;=\; L(\mathbf{r})\,, \tag{4.31}$$

since this ensures that

$$\begin{aligned} E\left[\frac{1}{N}\sum_{i=1}^{N}\alpha_i\, W_{\mathrm{e}}(\mathbf{r}_i)\right] &= \int_{\mathcal{R}}\int_{\mathbb{R}} \alpha\, W_{\mathrm{e}}(\mathbf{r})\, p(\alpha, \mathbf{r})\, d\alpha\, d\mu(\mathbf{r}) \\ &= \int_{\mathcal{R}} W_{\mathrm{e}}(\mathbf{r})\, L(\mathbf{r})\, d\mu(\mathbf{r}) \;=\; \langle W_{\mathrm{e}}, L\rangle\,. \end{aligned}$$

Using these ideas, rendering algorithms based on particle tracing can be decomposed into two independent steps. First, there must be a method for generating a set of sample rays that satisfy the conditions above. The simplest way of doing this involves tracing n independent particle histories starting from the light sources, as described below. Second, the algorithm must use these samples in some way to compute the desired set of measurements. In addition to unbiased estimators of the form (4.30), other (biased) possibilities include density estimation [Shirley et al. 1995] and various forms of interpolation (e.g. see [Jensen 1996]).

The conditions (4.30) and (4.31) are a specification of the *interface* between these two components of a rendering algorithm. On the one hand, there are a variety of ways to prove them for specific particle generation schemes. On the other hand, they concisely state the properties that the generated particles possess, i.e. the properties that higher-level rendering algorithms are allowed to depend on. Essentially, these conditions are a rigorous interpretation of the "energy packets" approach: they preserve the idea that particles represent the equilibrium radiance itself, independent of any particular sensor.

Generating the particles. The simplest way to generate a set of weighted ray samples satisfying condition (4.30) is to follow a random walk. This process can be summarized as follows:

1. Choose a random ray $\mathbf{r}_0 = (\mathbf{x}_0, \omega_0)$ starting on a light source, and let its weight be

$$\alpha_0 \;=\; \frac{L_\mathrm{e}(\mathbf{r}_0)}{p_0(\mathbf{r}_0)}$$

   where $p_0(\mathbf{r})$ is the density from which $\mathbf{r}_0$ was sampled. The initial state of the particle is defined to be $(\alpha_0, \mathbf{r}_0)$.

2. Given the current state $(\alpha_i, \mathbf{r}_i)$, decide whether to continue the random walk. We let q_{i+1} denote the probability with which the random walk is continued (where q_{i+1} depends on the current path in some way). If the walk is terminated (which happens with probability $1-q_{i+1}$), we let $k = i$ denote its length.

3. Otherwise, let $\mathbf{x}_{i+1}$ be the first intersection point of the ray $\mathbf{r}_i = (\mathbf{x}_i, \omega_i)$ with a surface (see Figure 4.1). Choose a random scattering direction ω_{i+1} according to some density function p_{i+1} that approximates the BSDF there. The particle weight α_{i+1} is then computed from α_i using the formula

$$\alpha_{i+1} \;=\; \alpha_i \, \frac{1}{q_{i+1}} \, \frac{f_\mathrm{s}^*(\mathbf{x}_{i+1}, \omega_{i+1} \to -\omega_i)}{p_{i+1}(\omega_{i+1})} \,, \tag{4.32}$$

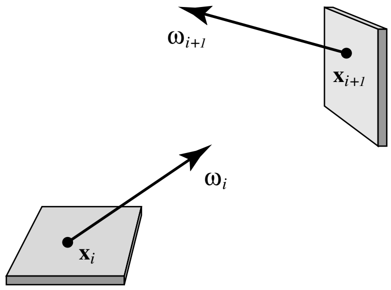

Figure 4.1: A scattering step in particle tracing.

where the density $p_{i+1}(\omega)$ is measured with respect to projected solid angle. (If this density is measured with respect to ordinary solid angle instead, the corresponding update formula is

$$\alpha_{i+1} \;=\; \alpha_i \, \frac{1}{q_{i+1}} \, \frac{f_s^*(\mathbf{x}_{i+1}, \omega_{i+1} \to -\omega_i) \, |\omega_{i+1} \cdot \mathbf{N}(\mathbf{x}_{i+1})|}{p_{i+1}(\omega_{i+1})} \,, \tag{4.33}$$

where $\mathbf{N}(\mathbf{x})$ is the normal at $\mathbf{x}$.) Notice that we have used the adjoint BSDF in this expression, according to the sampling conventions of Section 3.7.6.

4. Return to step 2.

This process yields a set of ray samples $\mathbf{r}_0, \ldots \mathbf{r}_k$, where k is the length of the random walk. Each ray $\mathbf{r}_i = (\mathbf{x}_i, \omega_i)$ is assigned the weight

$$\alpha_i \;=\; \frac{L_e(\mathbf{x}_0, \omega_0)}{p_0(\mathbf{x}_0, \omega_0)} \prod_{j=0}^{i-1} \frac{1}{q_{j+1}} \, \frac{f_s^*(\mathbf{x}_{j+1}, \omega_{j+1} \to -\omega_j)}{p_{j+1}(\omega_{j+1})} \,, \tag{4.34}$$

which was obtained by expanding the recursion (4.32). There are a variety of ways to show that these samples satisfy the desired condition (4.30), either directly from the light transport equation (similar to [Spanier & Gelbard 1969, p. 62]), or from the importance transport equation (as we discuss below).

There are many other possibilities for generating a suitable set of particles, by modifying and extending the basic particle tracing technique. For example, different density functions could be used for sampling, Russian roulette or splitting could be applied, the samples could be resampled to make the weights all equal, and so on. Our new characterization (4.30) applies to all of these possibilities: this makes it clear that the essence of particle tracing is not how the samples are obtained, but what they represent.

Particle tracing as importance transport. For comparison, we summarize the approach of Pattanaik & Mudur [1995] (in our notation, and in a more general form). Given a set of basis functions

$$B_1, \ldots, B_M,$$

they consider the problem of approximating the equilibrium radiance as a linear combination

$$L_{\mathrm{o}} \approx \sum_{i=1}^{M} I_j B_j,$$

where $I_j = \langle B_j, L_{\mathrm{o}} \rangle$ is the coefficient of the j-th basis function. Notice that since the inner product $\langle B_j, L_{\mathrm{o}} \rangle$ measures the radiance *leaving* the corresponding surface, we can interpret B_j as an incident importance function $W_{\mathrm{e,i}}^{(j)}$, and rewrite the expression for each coefficient I_j as

$$I_j = \langle W_{\mathrm{e,i}}^{(j)}, L_{\mathrm{o}} \rangle . \tag{4.35}$$

To evaluate this expression, they first rewrite it as an importance transport problem:

$$I_j = \langle W_{\mathrm{i}}^{(j)}, L_{\mathrm{e}} \rangle . \tag{4.36}$$

This equation is then evaluated by recursively sampling the importance transport equation. This leads to a random walk that is very similar to the one that we have already described above.

They start by letting $\mathbf{r}_0 = (\mathbf{x}_0, \omega_0)$ be a random ray sampled on a light source, from which I_j can be estimated using

$$I_j = E\left[\frac{W_{\mathrm{i}}^{(j)}(\mathbf{r}_0)\, L_{\mathrm{e}}(\mathbf{r}_0)}{p_0(\mathbf{r}_0)}\right] \tag{4.37}$$

where $p_0(\mathbf{r}_0)$ is the density for sampling $\mathbf{r}_0$. Notice that this expression is just the usual Monte Carlo estimate $f(x)/p(x)$, applied to equation (4.36).

The factors L_{e} and p_0 in (4.37) can easily be evaluated, leaving only the equilibrium importance $W_{\mathrm{i}}^{(j)}(\mathbf{r}_0)$. This factor is evaluated recursively, using the transport equation

$$W_{\mathrm{i}}^{(j)} = W_{\mathrm{e,i}}^{(j)} + \mathbf{G}\mathbf{K}^* W_{\mathrm{i}}^{(j)} . \tag{4.38}$$

Starting with $i = 0$, this is done by casting the ray $\mathbf{r}_i = (\mathbf{x}_i, \omega_i)$ to find the first intersection point $\mathbf{x}_{i+1}$ with a surface. Next, a new ray direction ω_{i+1} is chosen, according to a density function p_{i+1} that approximates the BSDF at $\mathbf{x}_{i+1}$ (recall Figure 4.1). It is then easy to check that $W_{\mathrm{i}}^{(j)}(\mathbf{r}_i)$ can be estimated as

$$
\begin{aligned}
W_{\mathrm{i}}^{(j)}(\mathbf{r}_i) &= W_{\mathrm{e,i}}^{(j)}(\mathbf{r}_i) + (\mathbf{G}\mathbf{K}^* W_{\mathrm{i}}^{(j)})(\mathbf{r}_i) \\
&= W_{\mathrm{e,i}}^{(j)}(\mathbf{r}_i) + \begin{cases} 0 & \text{if } d(\mathbf{r}_i) = \infty \text{ or } \omega_{i+1} = \Lambda\,, \\ E\left[\dfrac{f_{\mathrm{s}}^*(\mathbf{x}_{i+1}, \omega_{i+1} \to -\omega_i)\, W_{\mathrm{i}}^{(j)}(\mathbf{r}_{i+1})}{q_{i+1}\, p_{i+1}(\omega_{i+1})}\right] \\ \qquad \text{otherwise}\,, \end{cases}
\end{aligned}
\tag{4.39}
$$

where p_{i+1} is measured with respect to projected solid angle. The recursion stops when a ray fails to intersect a surface ($d(\mathbf{r}_i) = \infty$), or when the walk is randomly terminated (as indicated by setting $\mathbf{x}_{i+1} = \Lambda$), which happens with probability $1 - q_{i+1}$.

Discussion. This process is superficially quite similar to particle tracing: it generates a random walk that starts at a light source, is scattered at each surface, and is eventually absorbed. Furthermore, it is theoretically well-founded, since it computes an unbiased estimate of the measurement I.

However, this view of particle tracing also has several disadvantages. First, there is no obvious relationship to the concepts of photons or particle energies. Furthermore, the sampling process seems to depend on which particular measurement I_j we evaluate (since equation (4.38) describes the equilibrium importance for a specific sensor). This would seem to imply that different particle histories are needed to estimate each measurement, which would be very inefficient. In contrast, in practice the random walks are chosen *not* to depend on any particular I_j, since this allows each random walk to be used for all measurements simultaneously. The formal dependence on a particular measurement I_j is rather non-intuitive.

Relationship to particle weights. Finally, observe that the importance transport process generates a set of ray samples $\mathbf{r}_0, \ldots, \mathbf{r}_k$. Furthermore, if we expand the recursion (4.39) and multiply together the factors that weight the emitted importance $W_{\mathrm{e,i}}(\mathbf{r}_i)$, we obtain exactly the same weights α_i that were used for particle tracing (see equation (4.34)). Thus, if we insert this set of weighted ray samples into equation (4.30), we obtain the same estimate for a given measurement I that would have been obtained by recursively sampling the importance transport equation (we have simply rearranged the calculations). Since we have already shown that the importance transport scheme gives an unbiased estimate (for any importance function $W_{\mathrm{e,i}}$), it follows that this set of weighted ray samples satisfies the desired condition (4.30). This validates the expression for the particle weights given earlier.

Appendix 4.B Properties of the operators G, K, T, and S

This appendix gives some formal results on the invertibility, adjoints, and norms of the operators $\mathbf{G}$, $\mathbf{K}$, $\mathbf{T}_X$, and $\mathbf{S}_X$. In some cases, these results are parallel to those of Arvo [1995], who considered the case of purely reflective, one-sided surfaces. However, different proof techniques are required for the general case considered here, and in most cases the specific results are different as well.

The proofs are not difficult; however, it is surprisingly tricky to state the results correctly. There are often exceptions to "intuitively obvious" properties that require careful thought. We have also emphasized some of the more subtle issues that arise in the proofs.

We will need various properties of the scene $\mathcal{M}$, so it is important to clarify exactly what is allowed. Recall that in Section 3.1, we defined $\mathcal{M}$ to be the union of a finite number of closed, piecewise differentiable, two-dimensional manifolds (possibly with boundary). In particular, we allow manifolds to be unbounded (e.g. a plane), to have any number of handles (e.g. a torus), and to have any number of holes (e.g. an annulus). A manifold can even be disconnected; for example, a single manifold may represent several spheres (this could arise naturally as an implicit surface). On the other hand, true fractals are not allowed (since they are not manifolds), and $\mathcal{M}$ must always be representable as a *finite* union of manifolds.[13] Note that despite this restriction, $\mathcal{M}$ may still contain an infinite number of connected surfaces, since one manifold could represent an infinite stack of planes, or an infinite grid of spheres.

4.B.1 Invertibility

Invertibility is an important property for computer vision problems. For example, suppose that we wish to determine the incident radiance L_{i} at a surface point, given the exitant radiance L_{o}. This situation is described by the inverse of the operator that applies to the corresponding graphics problem.

In this section, we show that $\mathbf{G}$ and $\mathbf{K}$ are not invertible in general. We also show that it is possible to construct special scene models where their inverses do exist.[14]

[13] If countable unions were allowed, then $\mathcal{M}$ could be a dense subset of $\mathbb{R}^3$. For example, let $q_1, q_2, \ldots$ be an enumeration of the rational numbers, and define $\mathcal{M}$ as the union of the planes $x = q_i$.

[14] In the case of one-sided, reflective surfaces that form an enclosure, $\mathbf{G}$ is invertible [Arvo 1995]. In fact, $\mathbf{G}$ is its own inverse ($\mathbf{G}^2 = \mathbf{I}$). However, for this to be true, the space of radiance functions f must be defined carefully. In particular, the domain of these functions is not $\mathcal{M} \times \mathcal{S}^2$; it is the set of rays $(\mathbf{x}, \omega)$ where $\mathbf{x} \in \mathcal{M}$ and $\omega \in \mathcal{H}^2_+(\mathbf{x})$ (the upward hemisphere at $\mathbf{x}$).

Recall the definition of the *boundary distance function*

$$d_{\mathcal{M}}(\mathbf{x}, \omega) \;=\; \inf\{d > 0 \mid \mathbf{x} + d\omega \in \mathcal{M}\}\,,$$

and the *ray casting function*

$$\mathbf{x}_{\mathcal{M}}(\mathbf{x}, \omega) \;=\; \mathbf{x} + d_{\mathcal{M}}(\mathbf{x}, \omega)\, \omega\,.$$

For brevity, we will omit the $\mathcal{M}$ subscripts on these functions in this appendix.

We begin with the following definitions:

1. A ray $\mathbf{r}$ is *reversible* if $d(\mathbf{r}) < \infty$.
2. The *reversible ray space* $\tilde{\mathcal{R}}$ is the set of all reversible rays in $\mathcal{R}$.
3. The *reversal map* is a function $M : \tilde{\mathcal{R}} \to \mathcal{R}$ whose value is given by

$$M(\mathbf{x}, \omega) \;=\; (\mathbf{x}(\mathbf{x}, \omega), -\omega)\,.$$

The following lemma implies that the range of M is actually the reversible ray space $\tilde{\mathcal{R}}$ (rather than all of $\mathcal{R}$, as we defined it above), and that furthermore $M : \tilde{\mathcal{R}} \to \tilde{\mathcal{R}}$ is a bijection.

Lemma 4.1. *If* $\mathbf{r}$ *is reversible, then* $M(\mathbf{r})$ *is also reversible, and* $M(M(\mathbf{r})) = \mathbf{r}$.

Proof. Let $\mathbf{r} = (\mathbf{x}, \omega)$ and $\mathbf{r}' = M(\mathbf{r})$. Also let $d = d(\mathbf{x}, \omega)$ and $\mathbf{x}' = \mathbf{x} + d\omega$, so that $\mathbf{r}' = (\mathbf{x}', -\omega)$. Now by definition,

$$d(\mathbf{r}') \;=\; \inf\{d > 0 \mid \mathbf{x}' - d\omega \in \mathcal{M}\}\,.$$

Since $\mathbf{x}' - d\omega = \mathbf{x} \in \mathcal{M}$, we have $d(\mathbf{r}') \le d$, and so $\mathbf{r}'$ is reversible.

Now assume that $d(\mathbf{r}') = d'$, where $0 < d' < d$. Then $\mathbf{x}' - d'\omega = \mathbf{x} + (d - d')\omega$ is a point of $\mathcal{M}$, which contradicts the fact that $d(\mathbf{x}, \omega) = d$. ∎

With respect to these definitions, the propagation operator (4.14) is defined by

$$(\mathbf{G}h)(\mathbf{r}) \;=\; \begin{cases} h(M(\mathbf{r})) & \text{for } \mathbf{r} \in \tilde{\mathcal{R}}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Theorem 4.2. *Whenever* $\mathcal{M}$ *is non-empty and bounded, then* $\mathbf{G}$ *is not invertible. However,* $\mathbf{G}$ *is invertible for some unbounded scenes.*

Proof. Assuming that $\mathcal{M}$ is non-empty and bounded, we first show that some rays are not reversible, i.e. $\mu(\mathcal{R} - \tilde{\mathcal{R}}) > 0$. (These are the "outward-pointing" rays that do not intersect $\mathcal{M}$.) This may seem obvious, but since there are scenes for which all rays are reversible (see below), we must be a bit careful.

Consider the non-reversible rays $(\mathbf{x}, \omega)$ for a fixed direction $\omega \in \mathcal{S}^2$. The origins of these rays are the maximal points of $\mathcal{M}$ along lines in the direction ω. Thus, there is an exact correspondence between non-reversible rays, and directed lines that intersect $\mathcal{M}$. Given that ω occupies an infinitesimal solid angle $d\sigma(\omega)$, the measure of these rays is

$$A_\omega(\Pi_\omega(\mathcal{M}))\, d\sigma(\omega)\,,$$

where Π_ω denotes orthogonal projection onto the plane perpendicular to ω, and A_ω is the area measure on that plane. Thus, the total measure of the non-reversible rays is exactly

$$\mu(\mathcal{R} - \tilde{\mathcal{R}}) \;=\; \int_{\mathcal{S}^2} A_\omega(\Pi_\omega(\mathcal{M}))\, d\sigma(\omega)\,.$$

This is simply 4π times the average projected surface area of $\mathcal{M}$, which is positive.[15]

Now let L_1, L_2 be any two radiance functions such that

$$L_1(\mathbf{r}) \;=\; L_2(\mathbf{r}) \qquad \text{for } \mathbf{r} \in \tilde{\mathcal{R}}\,.$$

Then clearly $\mathbf{G}L_1 = \mathbf{G}L_2$, no matter what values the L_i have for $r \in \mathcal{R} - \tilde{\mathcal{R}}$. Since $\mu(\mathcal{R} - \tilde{\mathcal{R}}) > 0$, the functions L_1 and L_2 can be distinct, which shows that $\mathbf{G}$ is not invertible in general.

For an example where $\mathbf{G}$ is invertible, let $\mathcal{M}$ be an infinite stack of planes, or an infinite set of concentric spheres. In this case, all rays are reversible (up to a set of μ-measure zero). This can also be achieved with a single infinite surface that is diffeomorphic to a plane. In general, $\mathbf{G}$ is invertible if and only if $\mathcal{M}$ has a non-empty intersection with the interior of every infinite cone in $\mathbb{R}^3$ (or if $\mathcal{M}$ is empty). ∎

Note especially that $\mathbf{G}$ is not its own inverse, so that the relationship $L_{\mathrm{i}} = \mathbf{G}L_{\mathrm{o}}$ does not imply $L_{\mathrm{o}} = \mathbf{G}L_{\mathrm{i}}$.

[15]Technically, this is also a bit tricky. It is possible make the average projected surface area of $\mathcal{M}$ arbitrarily small, while keeping the same total area, by making $\mathcal{M}$ very convoluted. However, we can use the assumption that $\mathcal{M}$ is a finite union of piecewise differentiable manifolds. This implies that for almost every point $\mathbf{x} \in \mathcal{M}$, we can find a neighborhood $U_{\mathbf{x}}$ that is arbitrarily close to being a disc (of very small radius). The average projected surface area of any such $U_{\mathbf{x}}$ is positive (because it is almost a disc), and this is a lower bound on the average projected surface area of $\mathcal{M}$.

Theorem 4.3. *The operator* $\mathbf{K}$ *is not invertible in general. However,* $\mathbf{K}$ *is invertible for some special scene models.*

Proof. Suppose that every point of $\mathcal{M}$ has a constant BSDF (with regard to both reflection and transmission):

$$f_{\rm s}(\mathbf{x}, \omega_{\rm i} \to \omega_{\rm o}) \;=\; g(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathcal{M} \text{ and } \omega_{\rm i}, \omega_{\rm o} \in \mathcal{S}^2 \,.$$

In this case, $\mathbf{K}$ reduces to

$$(\mathbf{K}L)(\mathbf{x}, \omega_{\rm o}) \;=\; \int_{\mathcal{S}^2} g(\mathbf{x})\, L(\mathbf{x}, \omega_{\rm i})\, d\sigma^{\perp}_{\mathbf{x}}(\omega_{\rm i}) \,.$$

Since the right-hand side does not depend on $\omega_{\rm o}$, $\mathbf{K}L$ is a function of position only. That is, for every L there is a function $h_0 : \mathcal{M} \to \mathbb{R}$ such that $(\mathbf{K}L)(\mathbf{x}, \omega) = h_0(\mathbf{x})$. However, it is clear that infinitely many distinct functions L map to each such h_0, and thus $\mathbf{K}$ is not invertible.

It is easy to see that this argument still holds if any part of the scene has a diffuse BRDF or BTDF, thus $\mathbf{K}$ is not invertible for most graphics models.

On the other hand, suppose that every point of $\mathcal{M}$ is a mirror. In this case, it is easy to see that $\mathbf{K}$ is an isomorphism. There are also less trivial examples. For example, $f_{\rm s}$ could be chosen so that $\mathbf{K}$ encodes the two-dimensional Fourier transform of the input signal. ∎

With regard to the other operators we have defined, it is easy to see that the transport operators $\mathbf{T}_X$ are not invertible in general, since they are compositions of $\mathbf{K}$ and $\mathbf{G}$. For the operators $\mathbf{I} - \mathbf{T}_X$, on the other hand, invertibility depends on the norms of $\mathbf{K}$ and $\mathbf{G}$ (to be discussed in Section 4.B.3). These operators must be invertible in order for the solution operators $\mathbf{S}_X = (\mathbf{I} - \mathbf{T}_X)^{-1}$ to exist.

4.B.2 Adjoints

We derive the adjoints of $\mathbf{G}$ and $\mathbf{K}$, and use them to prove the relationship

$$\langle W_{\rm e}, \mathbf{G}\mathbf{S}_{L_{\rm o}} L_{\rm e} \rangle \;=\; \langle \mathbf{G}\mathbf{S}_{W_{\rm o}} W_{\rm e}, L_{\rm e} \rangle \,.$$

(The operators $\mathbf{T}_X$ and $\mathbf{S}_X$ are defined in Section 4.7.) Our approach is unique in that we use this identity to *define* the equilibrium importance, according to

$$W_{\rm o} \;=\; \mathbf{S}_{W_{\rm o}} W_{\rm e} \,.$$

From this we derive the transport equation satisfied by W_{o}, and we derive similar results for the incident quantities L_{i} and W_{i}.

We begin by showing that $\mathbf{G}$ is self-adjoint.[16] The following lemma states that M preserves the measure μ.

Lemma 4.4. *Let μ be the throughput measure (4.2), and let $D \subset \tilde{\mathcal{R}}$ be a measurable set. Then*

$$\mu(M(D)) \;=\; \mu(D)\,,$$

where $M(D) = \{M(\mathbf{x}) \mid \mathbf{x} \in D\}$.

Proof. Since M is a bijection, it is sufficient to show that μ is preserved locally. Let $dA(\mathbf{x}) \times d\sigma(\omega)$ be an infinitesimal neighborhood of the ray $(\mathbf{x}, \omega) \in D$, and define

$$(\mathbf{x}', \omega') \;=\; M(\mathbf{x}, \omega) \;=\; (\mathbf{x}(\mathbf{x}, \omega), -\omega)\,.$$

We must show that $d\mu(\mathbf{x}, \omega) = d\mu(\mathbf{x}', \omega')$. Recall that one expression for μ is given by

$$d\mu(\mathbf{x}, \omega) \;=\; dA^{\perp}_{\omega}(\mathbf{x})\, d\sigma(\omega)\,.$$

We immediately have $d\sigma(\omega') = d\sigma(-\omega) = d\sigma(\omega)$. As for the other factor, recall that $A^{\perp}_{\omega}$ measures projected surface area on a plane perpendicular to ω. By definition of $\mathbf{x}(\mathbf{x}, \omega)$, however, $\mathbf{x}' - \mathbf{x}$ is always parallel to ω. Thus two corresponding areas $dA(\mathbf{x}), dA(\mathbf{x}')$ on $\mathcal{M}$ will always have the same projected area,

$$dA^{\perp}_{\omega}(\mathbf{x}') \;=\; dA^{\perp}_{\omega}(\mathbf{x}(\mathbf{x}, \omega)) \;=\; dA^{\perp}_{\omega}(\mathbf{x})\,. \quad \blacksquare$$

In terms of the composition measure notation (5.29), the preceding lemma states that $\mu \circ M = \mu$ (restricted to reversible rays).

Theorem 4.5. *The operator $\mathbf{G}$ is self-adjoint (for any scene model).*

Proof. Given $f, g \in L_2$, we have

$$\begin{aligned} \langle f, \mathbf{G}g \rangle &= \int_{\tilde{\mathcal{R}}} f(\mathbf{r})\, g(M(\mathbf{r}))\, d\mu(\mathbf{r}) \\ &= \int_{\tilde{\mathcal{R}}} f(\mathbf{r})\, g(M(\mathbf{r}))\, \frac{d\mu(\mathbf{r})}{d(\mu \circ M)}\, d\mu(M(\mathbf{r})) \end{aligned}$$

[16] Our proof is necessarily different than the one in [Arvo 1995], which assumed that $\mathbf{G}^2 = \mathbf{I}$ when $\mathcal{M}$ forms an enclosure.

$$\begin{aligned} &= \int_{\tilde{\mathcal{R}}} f(M(\mathbf{r}'))\, g(\mathbf{r}')\, d\mu(\mathbf{r}') \\ &= \langle \mathbf{G} f, g \rangle\,, \end{aligned}$$

where we have used the fact that M is a bijection, that $\mu \circ M = \mu$, and (5.32). ∎

The following theorem and proof are similar to [Arvo 1995, Theorem 16].

Theorem 4.6. *The adjoint of* $\mathbf{K}$ *is given by*

$$(\mathbf{K}^* h)(\mathbf{x}, \omega_\mathrm{o}) \;=\; \int_{\mathcal{S}^2} f_\mathrm{s}^*(\mathbf{x}, \omega_\mathrm{i} \to \omega_\mathrm{o})\, h(\mathbf{x}, \omega_\mathrm{i})\, d\sigma_{\mathbf{x}}^{\perp}(\omega_\mathrm{i})\,.$$

In particular, $\mathbf{K}$ *is self-adjoint if and only if* f_s *is symmetric for almost every* $\mathbf{x} \in \mathcal{M}$ *(i.e. except on a set of* A*-measure zero).*

Proof. We have

$$\begin{aligned} \langle f, \mathbf{K} g \rangle &= \int_{\mathcal{R}} f(\mathbf{x}, \omega) \int_{\mathcal{S}^2} f_\mathrm{s}(\mathbf{x}, \omega' \to \omega)\, g(\mathbf{x}, \omega')\, d\sigma_{\mathbf{x}}^{\perp}(\omega')\, d\mu(\mathbf{x}, \omega) \\ &= \int_{\mathcal{M}} \int_{\mathcal{S}^2} \int_{\mathcal{S}^2} f(\mathbf{x}, \omega)\, f_\mathrm{s}(\mathbf{x}, \omega' \to \omega)\, g(\mathbf{x}, \omega')\, d\sigma_{\mathbf{x}}^{\perp}(\omega')\, d\sigma_{\mathbf{x}}^{\perp}(\omega)\, dA(\mathbf{x}) \\ &= \int_{\mathcal{M}} \int_{\mathcal{S}^2} \int_{\mathcal{S}^2} f(\mathbf{x}, \omega)\, f_\mathrm{s}(\mathbf{x}, \omega' \to \omega)\, g(\mathbf{x}, \omega')\, d\sigma_{\mathbf{x}}^{\perp}(\omega)\, d\sigma_{\mathbf{x}}^{\perp}(\omega')\, dA(\mathbf{x}) \\ &= \int_{\mathcal{R}} \int_{\mathcal{S}^2} f_\mathrm{s}^*(\mathbf{x}, \omega \to \omega') f(\mathbf{x}, \omega)\, d\sigma_{\mathbf{x}}^{\perp}(\omega)\, g(\mathbf{x}, \omega')\, d\mu(\mathbf{x}, \omega') \\ &= \langle \mathbf{K}^* f, g \rangle\,, \end{aligned}$$

where we have used Fubini's theorem to change the order of integration. ∎

We are now in a position to study importance transport, which usually proceeds by writing down a formula for the equilibrium importance W and verifying that it has the desired properties. We will take the opposite approach, by starting with the fundamental relationship (4.21) for light transport,

$$I \;=\; \langle W_\mathrm{e}, \mathbf{G}\mathbf{S}_{L_\mathrm{o}} L_\mathrm{e} \rangle\,,$$

and then deriving the equations for importance based on the principle that we wish to compute the same value for the measurement I.

We start with the identity

$$I \;=\; \langle (\mathbf{G}\mathbf{S}_{L_\mathrm{o}})^* W_\mathrm{e}, L_\mathrm{e} \rangle\,,$$

which follows from the definition (4.22) of an adjoint operator.

Lemma 4.7. $(\mathbf{G}\mathbf{S}_{L_\mathrm{o}})^* \;=\; \mathbf{G}\mathbf{S}_{W_\mathrm{o}}$, *provided that* $\mathbf{S}_{L_\mathrm{o}}$ *and* $\mathbf{S}_{W_\mathrm{o}}$ *exist.*

Proof. We will make use of the following simple identities, which follow directly from the definition (4.22):

$$\begin{aligned}
\mathbf{I}^* &= \mathbf{I} \\
(\mathbf{A}+\mathbf{B})^* &= \mathbf{A}^*+\mathbf{B}^* \\
(\mathbf{A}\mathbf{B})^* &= \mathbf{B}^*\mathbf{A}^* \\
(\mathbf{A}^{-1})^* &= (\mathbf{A}^*)^{-1}
\end{aligned}$$

Provided that the operators $(\mathbf{I}-\mathbf{K}\mathbf{G})^{-1}$ and $(\mathbf{I}-\mathbf{K}^*\mathbf{G})^{-1}$ exist, we now have

$$\begin{aligned}
(\mathbf{G}\mathbf{S}_{L_\mathrm{o}})^* &= (\mathbf{G}(\mathbf{I}-\mathbf{K}\mathbf{G})^{-1})^* &&= (\mathbf{I}-\mathbf{G}\mathbf{K}^*)^{-1}\mathbf{G} \\
&= \textstyle\sum_{i=0}^{\infty}(\mathbf{G}\mathbf{K}^*)^i\mathbf{G} &&= \textstyle\sum_{i=0}^{\infty}\mathbf{G}(\mathbf{K}^*\mathbf{G})^i \\
&= \mathbf{G}(\mathbf{I}-\mathbf{K}^*\mathbf{G})^{-1} &&= \mathbf{G}\mathbf{S}_{W_\mathrm{o}}\,. \quad \blacksquare
\end{aligned}$$

We have thus proven that $I = \langle W_\mathrm{e}, \mathbf{G}\mathbf{S}_{L_\mathrm{o}} L_\mathrm{e}\rangle = \langle \mathbf{G}\mathbf{S}_{W_\mathrm{o}} W_\mathrm{e}, L_\mathrm{e}\rangle$, which is the basis for the following definition:

Definition 4.8. *The* exitant equilibrium importance function W_o *is defined by*

$$W_\mathrm{o} \;=\; \mathbf{S}_{W_\mathrm{o}} W_\mathrm{e}\,.$$

Theorem 4.9. W_o *satisfies the transport equation* $W_\mathrm{o} \;=\; W_\mathrm{e} + \mathbf{T}_{W_\mathrm{o}} W_\mathrm{o}$, *where* $\mathbf{T}_{W_\mathrm{o}} \;=\; \mathbf{K}^*\mathbf{G}$. *In particular,* L_o *and* W_o *obey the same transport equation when* $\mathbf{K} = \mathbf{K}^*$.

Proof. This follows directly from Lemma 4.7, Definition 4.8, and the definition $\mathbf{S}_{W_\mathrm{o}} \;=\; (\mathbf{I} - \mathbf{K}^*\mathbf{G})^{-1}$ from Section 4.7. $\blacksquare$

Theorem 4.10. *The incident equilibrium quantities* L_i *and* W_i *satisfy*

$$\begin{aligned}
L_\mathrm{i} &= L_{\mathrm{e,i}} + \mathbf{G}\mathbf{K}L_\mathrm{i}\,, \\
W_\mathrm{i} &= W_{\mathrm{e,i}} + \mathbf{G}\mathbf{K}^*W_\mathrm{i}\,.
\end{aligned}$$

Proof. We have

$$\begin{aligned} L_{\mathrm{o}} &= L_{\mathrm{e}} + \mathbf{K}\mathbf{G}L_{\mathrm{o}} \\ \mathbf{G}L_{\mathrm{o}} &= \mathbf{G}L_{\mathrm{e}} + \mathbf{G}\mathbf{K}\mathbf{G}L_{\mathrm{o}} \\ L_{\mathrm{i}} &= L_{\mathrm{e,i}} + \mathbf{G}\mathbf{K}L_{\mathrm{i}}\,, \end{aligned}$$

where we have used the definition $L_{\mathrm{i}} = \mathbf{G}L_{\mathrm{o}}$, and we have assumed that emitted radiance is specified as an incident quantity $L_{\mathrm{e,i}}$. A similar relationship holds for W_{i}. ∎

The complete results are summarized in Section 4.7.

4.B.3 Norms

In this section, we prove conditions on the norms of $\mathbf{G}$ and $\mathbf{K}$ that are sufficient to ensure that the various solution operators $\mathbf{S}_X$ are well-defined. This is possible despite the fact that we can have $\|\mathbf{K}\| > 1$ for physically valid scene models.

The following theorem is similar to [Arvo 1995, Theorem 14]. However, our proof holds for two-sided surfaces, and involves only geometric concepts (Arvo's proof requires the principle that radiance is constant along straight lines in free space).

Theorem 4.11. *For any $1 \le p \le \infty$, we have $\|\mathbf{G}\|_p \le 1$. Furthermore, $\|\mathbf{G}\|_p = 1$ unless $\mathcal{M}$ is contained by a plane in $\mathbb{R}^3$, in which case $\|\mathbf{G}\|_p = 0$.*

Proof. For any $1 \le p < \infty$ and any $L \in L_p$, we have

$$\begin{aligned} \|\mathbf{G}L\|_p &= \left(\int_{\mathcal{R}} |(\mathbf{G}L)(\mathbf{r})|^p \, d\mu(\mathbf{r}) \right)^{1/p} \\ &= \left(\int_{\tilde{\mathcal{R}}} |L(M(\mathbf{r}))|^p \, d\mu(\mathbf{r}) \right)^{1/p} \\ &= \left(\int_{\tilde{\mathcal{R}}} |L(M(\mathbf{r}))|^p \, \frac{d\mu(\mathbf{r})}{d(\mu \circ M)} \, d\mu(M(\mathbf{r})) \right)^{1/p} \\ &= \left(\int_{\tilde{\mathcal{R}}} |L(\mathbf{r}')|^p \, d\mu(\mathbf{r}') \right)^{1/p} \\ &\le \left(\int_{\mathcal{R}} |L(\mathbf{r}')|^p \, d\mu(\mathbf{r}') \right)^{1/p} \\ &= \|L\|_p\,, \end{aligned}$$

where we have used the fact that M is a measure-preserving bijection on $\tilde{\mathcal{R}}$ (Lemma 4.4). The case $p = \infty$ is similar, but only needs the fact that M is measure-preserving on sets of measure zero. Thus $\|\mathbf{G}\|_p \le 1$ for all $1 \le p \le \infty$.

Now we consider the conditions for which $\|\mathbf{G}\|_p = 1$ exactly. If $\mu(\tilde{\mathcal{R}}) > 0$ (i.e. there are reversible rays), then consider the function L defined by

$$L(\mathbf{r}) = \begin{cases} 1 & \text{for } \mathbf{r} \in \tilde{\mathcal{R}}\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Then $\|\mathbf{G}L\|_p = \|L\|_p > 0$, and so $\|\mathbf{G}\|_p = 1$. On the other hand, if $\mu(\tilde{\mathcal{R}}) = 0$, then $\|\mathbf{G}L\|_p = 0$ for all L, and we have $\|\mathbf{G}\|_p = 0$.

Thus we must show that $\mu(\tilde{\mathcal{R}}) = 0$ if and only if $\mathcal{M}$ is contained by a plane in $\mathbb{R}^3$. The "if" direction is clear. For the converse, recall that $\mathcal{M}$ is a union of piecewise differentiable manifolds. Choose points $\mathbf{x}, \mathbf{x}' \in \mathcal{M}$ each lying in the differentiable interior of some manifold, and such that $\mathbf{x}'$ does not lie in the tangent plane at $\mathbf{x}$. (This is possible since $\mathcal{M}$ is not contained by any plane.) Because $\mathcal{M}$ is differentiable at $\mathbf{x}$ and $\mathbf{x}'$, we can choose small disk-shaped regions $dA(\mathbf{x}), dA(\mathbf{x}') \subset \mathcal{M}$ that are arbitrarily close to lying in the tangent planes at $\mathbf{x}$ and $\mathbf{x}'$ respectively. The set of rays leaving $dA(\mathbf{x})$ toward $dA(\mathbf{x}')$ now has positive μ-measure (even if $\mathbf{x}$ happens to lie in the tangent plane of $\mathbf{x}'$). Furthermore, these rays are reversible (even if there are other surfaces between $\mathbf{x}$ and $\mathbf{x}'$). ∎

The following results will be proven in Section 7.B.2. They are stated here for completeness.

Theorem 4.12. *For any physically valid scene, and for any* $1 \le p \le \infty$*,*

$$\|\mathbf{K}\|_p < \frac{\eta_{\max}^2}{\eta_{\min}^2}\,,$$

where $\eta_{\min}$ *and* $\eta_{\max}$ *denote the minimum and maximum refractive indices in the environment.*

Theorem 4.13. *For any physically valid scene, the solution operators* $\mathbf{S}_X$ *exist and are well-defined.*