Non-symmetric Scattering in Light Transport Algorithms

Eric Veach
Computer Science Department, Stanford University

Abstract

Non-symmetric scattering is far more common in computer graphics than is generally recognized, and can occur even when the underlying scattering model is physically correct. For example, we show that non-symmetry occurs whenever light is refracted, and also whenever shading normals are used (e.g. due to interpolation of normals in a triangle mesh, or bump mapping [5]).

We examine the implications of non-symmetric scattering for light transport theory. We extend the work of Arvo et al. [4] into a complete framework for light, importance, and particle transport with non-symmetric kernels. We show that physically valid scattering models are not always symmetric, and derive the condition for arbitrary models to obey Helmholtz reciprocity. By rewriting the transport operators in terms of optical invariants, we obtain a new framework where symmetry and reciprocity are the same.

We also consider the practical consequences for global illumination algorithms. The problem is that many implementations indirectly assume symmetry, by using the same scattering rules for light and importance, or particles and viewing rays. This can lead to incorrect results for physically valid models. It can also cause different rendering algorithms to converge to different solutions (whether the model is physically valid or not), and it can cause shading artifacts. If the non-symmetry is recognized and handled correctly, these problems can easily be avoided.

1 Introduction

The equations governing the transport and measurement of light energy can be written in two equivalent forms, depending on whether we solve for radiance or importance. Most current global illumination algorithms take advantage of this duality. For example, importance is often used to guide mesh refinement in finite-element approaches, and traditional ray tracing is the dual of *particle tracing*, which simulates the emission and scattering of photons.

When these algorithms are implemented, it is very common to assume that light and importance obey the same scattering rules. Equivalently, the *bidirectional scattering distribution function* (BSDF) for every surface is assumed to be symmetric. This is usually considered reasonable, since it is well-known that reflective surfaces have symmetric BRDF's. The symmetry allows many simplifications, such as using the same code to trace light particles and viewing rays, or to solve for radiance and importance.

As mentioned in the abstract, we show that non-symmetric scattering is actually quite common. This can cause problems for any algorithm that uses both lighting and viewing information, including importance-driven finite element approaches [33, 30, 8], particle tracing algorithms [15, 28, 31], and bidirectional path-tracing [18, 37, 38]. However, we show that these problems are easy to fix.

2 Light transport with non-symmetric scattering

We develop a consistent framework for light, importance, and particle transport, building on the work of Arvo et al. [4], Christensen et al. [7], and Pattanaik and Mudur [28]. Our approach is new in several ways. Due to the absence of symmetry assumptions, our framework has a richer structure: for each of the four basic transport quantities (exitant/incident radiance/importance), there is a distinct transport operator and measurement equation. Each of these is useful for different rendering algorithms, and they are all related in a simple way (see Sec. 2.5 for a summary). Other contributions include the *ray space* abstraction, and the use of *incident* rather than *field* radiance functions.

2.1 Domains and measures

Most light transport calculations are naturally defined over the space $\mathcal{R}$ of all *rays*. By equipping $\mathcal{R}$ with a suitable measure and topology, the details of ray parameterization can be hidden, while clarifying the essential structure of the theory.

The ray space $\mathcal{R}$ can be parameterized in many ways. We will represent it as the Cartesian product $\mathcal{R} = \mathcal{M} \times \mathcal{S}^2$, where $\mathcal{M}$ is the set of surfaces in the scene, and $\mathcal{S}^2$ is the set of all unit direction vectors.[1] The ray $\mathbf{r} = (\mathbf{x}, \omega)$ has origin $\mathbf{x}$ and direction ω.

[1] In closed environments, another common parameterization is $\mathcal{R} = \mathcal{M} \times \mathcal{M}$, so that each ray is a pair $(\mathbf{x}, \mathbf{x}')$. To avoid multiple representations of the same ray, we must include a visibility test $V(\mathbf{x}, \mathbf{x}')$ in the measure μ, or restrict $\mathcal{R}$ to pairs where $\mathbf{x}$ and $\mathbf{x}'$ are mutually visible. We used this representation in [37]. Note that $\mathcal{R}$ and μ are independent of the parameterization, which is really just a way to assign *labels* to rays.

To integrate real-valued functions, we define a measure function μ. This allows us to define the *inner product* of two functions on $\mathcal{R}$,

$$\langle f, g \rangle = \int_{\mathcal{R}} f(\mathbf{r})\, g(\mathbf{r})\, d\mu(\mathbf{r}) \quad . \tag{1}$$

The measure μ is called *throughput* [35, 25, 9], and is defined[2] by

$$d\mu(\mathbf{r}) = d\mu(\mathbf{x}, \omega) = dA(\mathbf{x})\, d\sigma_{\mathbf{x}}^{\perp}(\omega) \quad , \tag{2}$$

where A is the area measure on $\mathcal{M}$, and $\sigma_{\mathbf{x}}^{\perp}$ is the *projected solid angle* measure on $\mathcal{S}^2$ [25, p.70]. The projected solid angle of a set $D \subset \mathcal{S}^2$ is defined by

$$\sigma_{\mathbf{x}}^{\perp}(D) = \int_D |\omega \cdot \mathbf{N}_{\mathrm{g}}(\mathbf{x})| \; d\sigma(\omega) \quad ,$$

where $\mathbf{N}_{\mathrm{g}}(\mathbf{x})$ is the geometric surface normal, and σ is the usual solid angle measure.

2.2 Scattering and transport operators

The scattering of light at a surface is typically divided into reflected and transmitted components (the BRDF f_{r} and the BTDF f_{t})[3]. It is often more convenient to combine these into a single *bidirectional scattering distribution function* (BSDF),[4] defined as

$$f_{\mathrm{s}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}}) = dL(\omega_{\mathrm{o}}) \,/\, dE(\omega_{\mathrm{i}}) \quad ,$$

so that $f_{\mathrm{s}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}})$ is the observed radiance leaving in direction ω_{o}, per unit of irradiance arriving from direction ω_{i} (see [27, p.5], [9, p.28]).

[2]The measure μ is similar to the usual geometric measure on lines in $\mathbb{R}^3$. Formally, it can be defined as a composition of Lebesgue measures on fibers, where the fibers of $\mathcal{R}$ are sets of rays with the same origin $\mathbf{x}$ (similar to [2, p.11]). This measure is invariant under Euclidean transformations, which makes it unique up to a constant factor [2, p.51].

[3]The "r" in f_r is usually assumed to be a label, denoting "reflectance". However, Nicodemus et al. [27] derive the BRDF from the more general BSSRDF, which allows light to strike and leave the surface at different points (e.g. due to sub-surface scattering). Just before the definition of the BRDF, r is used to denote the distance between these points, so that the r in f_r could be a parameter (whose value is usually zero in computer graphics). This has been interpreted both ways (e.g. [12], [13, p.665], [13, p.722]). In his other writings, however, Nicodemus uses r only as a label [24, 23, 22]. For the "distance" interpretation of r, the notation $S(\omega_{\mathrm{i}}, \omega_{\mathrm{r}}; r)$ for the BSSRDF is used [27, p.28]. These ambiguities can be avoided by using different typefaces for labels and parameters (Roman vs. italics), as we do in this paper.

[4]This name was introduced by Heckbert [16, p.26]. Previously he used the term *bidirectional distribution function* (BDF) [15], but this sounds more like a category (containing the B*DF's as members).

Following [4], we define light transport in terms of operators acting on radiance functions. We distinguish between *incident* radiance functions $L_\mathrm{i}(\mathbf{x}, \omega)$, which give the radiance arriving at $\mathbf{x}$ from direction ω, and *exitant* radiance functions $L_\mathrm{o}(\mathbf{x}, \omega)$, which measure the radiance leaving $\mathbf{x}$ in direction ω.

The *local scattering operator* is now defined by

$$(\mathbf{K}L)(\mathbf{x}, \omega_\mathrm{o}) = \int_{\mathcal{S}^2} f_\mathrm{s}(\mathbf{x}, \omega_\mathrm{i} \!\rightarrow\! \omega_\mathrm{o})\, L(\mathbf{x}, \omega_\mathrm{i})\, d\sigma^\perp_\mathbf{x}(\omega_\mathrm{i}) \ . \tag{3}$$

When applied to incident radiance L_i, it returns the exitant radiance $L_\mathrm{o} = \mathbf{K}L_\mathrm{i}$ after one scattering step. The *geometric* or *propagation operator* $\mathbf{G}$ completes the cycle, by expressing incident radiance in terms of the exitant radiance of the surrounding environment, according to $L_\mathrm{i} = \mathbf{G}L_\mathrm{o}$. It is defined by

$$(\mathbf{G}L)(\mathbf{x}, \omega_\mathrm{i}) = L(\mathbf{x}_\mathcal{M}(\mathbf{x}, \omega_\mathrm{i}), -\omega_\mathrm{i}) \ ,$$

where $\mathbf{x}_\mathcal{M}(\mathbf{x}, \omega)$ is the first point of $\mathcal{M}$ visible from $\mathbf{x}$ in direction ω.

The *field radiance* of [4] is related to incident radiance by

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

so the direction of ω_i is reversed (compared to [4]) in the equations above. This has two advantages: $\mathbf{G}$ and $\mathbf{K}$ become self-adjoint (when f_s is symmetric),[5] and ω_i and ω_o both point outward for reflective surfaces (as assumed by most BRDF formulas). Notice also that $\mathbf{K}$ acts locally with respect to position, while $\mathbf{G}$ acts locally with respect to direction (since ω_i is simply flipped around the origin).

The *light transport* or *rendering equation* is now

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

where $\mathbf{T} = \mathbf{K}\mathbf{G}$ is the *light transport operator*, $L_\mathrm{e}(\mathbf{x}, \omega)$ is the emitted radiance, and $L(\mathbf{x}, \omega)$ is the equilibrium radiance (the desired steady-state solution).

By inverting this operator equation, we obtain the formal solution $L = \mathbf{S}L_\mathrm{e}$, where

$$\mathbf{S} = (\mathbf{I} - \mathbf{T})^{-1}$$

is called the *solution operator*.[6] The solution exists and is unique provided that $\|\mathbf{T}^k\| < 1$ for some $k \geq 1$.

---

[5]The $\mathbf{G}$ and $\mathbf{K}$ of [4] are not self-adjoint. Arvo handles this with an isomorphism $\mathbf{H}$ between surface and field functions, such that $\mathbf{H}\mathbf{G}$ and $\mathbf{K}\mathbf{H}$ are equivalent to the $\mathbf{G}$ and $\mathbf{K}$ defined above [3, p.151].

[6]The "GRDF" described by Eric Lafortune [19] is simply the kernel of $\mathbf{S}$. 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 [10, p.74], [36, p.272]).

2.3 Sensors and measurements

Global illumination algorithms compute a finite number of measurements of the solution L. Each measurement is expressed as the response of a hypothetical sensor, e.g. a small area of film within a virtual camera. Linear sensors are characterized by a weighting function $W_{\mathrm{e}}(\mathbf{x}, \omega)$, which specifies the sensor response per unit power arriving at $\mathbf{x}$ from direction ω. Nicodemus calls this the *flux responsivity* of the sensor [26, p.59], with units $[S \cdot W^{-1}]$, where the unit S of sensor response depends on the sensor.

To make a measurement, we integrate the radiance falling on the sensor, weighted according to W_{e}. This is expressed by Nicodemus' *measurement equation* [26, p.85],

$$I = \langle W_{\mathrm{e}}, L_{\mathrm{i}} \rangle = \int_{\mathcal{R}} W_{\mathrm{e}}(\mathbf{r})\, L_{\mathrm{i}}(\mathbf{r})\, d\mu(\mathbf{r}) \quad , \tag{4}$$

where I is the measurement (e.g. a pixel value), W_{e} is the weighting function for this measurement, and L_{i} is the incident radiance. Each W_{e} is called an *exitant importance function* (we think of the sensor as emitting importance[7]).

We have defined both L_{e} and W_{e} as exitant quantities. This is natural, since it lets us define their values at points *on* the source or sensor (unlike the "visual potential" of [28]), and it reveals many similarities between light and importance transport [7].

The equilibrium solution $L = \mathbf{S} L_{\mathrm{e}}$ is also an exitant quantity. However, the measurement equation (4) requires an incident function. This problem is solved using the $\mathbf{G}$ operator, since $L_{\mathrm{i}} = \mathbf{G} L$.[8] Each measurement now has the form

$$I = \langle W_{\mathrm{e}}, L_{\mathrm{i}} \rangle = \langle W_{\mathrm{e}}, \mathbf{G} L \rangle = \langle W_{\mathrm{e}}, \mathbf{G}\mathbf{S} L_{\mathrm{e}} \rangle \quad . \tag{5}$$

2.4 Adjoint operators and importance transport

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 algorithms.

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

$$\langle \mathbf{H}^* f, g \rangle = \langle f, \mathbf{H} g \rangle$$

[7] We follow the terminology of [20, p.7, p.21], where an importance function pertains to a single "meter reading" (measurement). Since the average of several measurements is itself a measurement, importance functions may correspond to a set of measurements (e.g. all pixels in an image). In the case of importance-driven methods, the measurement may even be hypothetical [33, p.275]. The alternative term *potential function* [28] is undesirable because it has a well-known, different meaning in physics (a function satisfying Poisson's equation, e.g. the electric or gravitational potentials).

[8] We assume that the sensors are modeled as part of the domain $\mathcal{M}$. If not, we can always add them to $\mathcal{M}$ for theoretical purposes by assuming that they are completely transparent.

for all f, g.[9] Applying this identity to (5), we get

$$I = \langle W_\mathrm{e}, \mathbf{GS} L_\mathrm{e} \rangle = \langle (\mathbf{GS})^* W_\mathrm{e}, L_\mathrm{e} \rangle \ , \tag{6}$$

which suggests that we can evaluate I by transporting *importance* in some way.

To do this, we must evaluate $(\mathbf{GS})^*$. It is straightforward to show that $\mathbf{G}^* = \mathbf{G}$, i.e. $\mathbf{G}$ is self-adjoint [3, p.152]. Similarly, $\mathbf{K}$ is self-adjoint if we assume that the BSDF is symmetric [3]. Using standard identities, this implies that $\mathbf{GS}$ is self-adjoint as well.

Applying this to (6), measurements can be evaluated using either

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

Because of the symmetry, any equations or algorithms that apply to light transport may also be used for importance transport. In particular,

$$W = \mathbf{S} W_\mathrm{e}$$

is called the *equilibrium importance function*, and satisfies the *importance transport equation*

$$W = W_\mathrm{e} + \mathbf{T} W \ .$$

Incident importance is defined by $W_\mathrm{i} = \mathbf{G} W$, and finally measurements can be evaluated using either

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

These symmetries are very useful in algorithm design and implementation.

However, if the BSDF is not symmetric, then $\mathbf{K} \neq \mathbf{K}^*$. The adjoint is given by

$$(\mathbf{K}^* L)(\mathbf{x}, \omega_\mathrm{o}) = \int_{\mathcal{S}^2} f_\mathrm{s}(\mathbf{x}, \omega_\mathrm{o} \!\rightarrow\! \omega_\mathrm{i}) \, L(\mathbf{x}, \omega_\mathrm{i}) \, d\sigma^\perp_\mathbf{x}(\omega_\mathrm{i}) \ , \tag{7}$$

and thus $\mathbf{K} \neq \mathbf{K}^*$ (the arguments to f_s have been exchanged). This does not affect the light transport operator $\mathbf{T} = \mathbf{KG}$, which we will rename $\mathbf{T}_L$, but the importance transport operator becomes $\mathbf{T}_W = \mathbf{K}^* \mathbf{G}$. Details are given in the next section.

[9]The adjoint of an operator depends on the inner product used. We always use the inner product (1).

2.5 Summary of the transport rules

We consider the four basic transport quantities L_o, L_i, W_o, and W_i (exitant/incident radiance/importance). The incident quantities are obtained from the exitant ones by $L_\mathrm{i} = \mathbf{G} L_\mathrm{o}$ and $W_\mathrm{i} = \mathbf{G} W_\mathrm{o}$. Each quantity X has a transport operator $\mathbf{T}_X$, given by the following table:

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

The transport equation for X is given by

$$X = X_\mathrm{e} + \mathbf{T}_X X \ .$$

Its formal solution is $X = \mathbf{S}_X X_\mathrm{e}$, where

$$\mathbf{S}_X = (\mathbf{I} - \mathbf{T}_X)^{-1}$$

is the solution operator. Finally, measurements are made using any of[10]

$$I \;=\; \langle W_\mathrm{e}, L_\mathrm{i} \rangle \;=\; \langle W_\mathrm{i}, L_\mathrm{e} \rangle \;=\; \langle W_\mathrm{e,i}, L \rangle \;=\; \langle W, L_\mathrm{e,i} \rangle$$

(where we have dropped the "o" subscript on exitant quantities).

The operator $\mathbf{K}^*$ has the same form as $\mathbf{K}$, except that the arguments $\omega_\mathrm{i}, \omega_\mathrm{o}$ to the BSDF have been interchanged. We describe this by saying that $\mathbf{K}^*$ uses the *adjoint BSDF* f_s^*, which is defined by

$$f_\mathrm{s}^*(\omega_\mathrm{i} \!\rightarrow\! \omega_\mathrm{o}) = f_\mathrm{s}(\omega_\mathrm{o} \!\rightarrow\! \omega_\mathrm{i}) \ .$$

Note that only one of f_s or f_s^* is correct for any given transport situation. If we use the wrong one, we will compute the wrong answer. Even worse, if a non-symmetric BSDF is not recognized as such, then we will inadvertently use f_s where f_s^* should be used. This is inconsistent; it is like using a *different* BSDF for transport situations involving the adjoint. The errors will vary depending on how the BSDF is used by a specific transport algorithm. For example, particle tracing will converge to a different result than path tracing, since these two algorithms must sample the BSDF in opposite directions for consistent results (as discussed below).

[10]The forms involving $L_\mathrm{e,i}$ and $W_\mathrm{e,i}$ are useful when the emitted radiance/importance is naturally defined as an incident function. For example, to project L onto a set of orthonormal basis functions B_j, we compute inner products of the form $\langle B_j, L \rangle$ rather than $\langle B_j, L_\mathrm{i} \rangle$. Each B_j is thus an incident importance function.

2.5.1 Particle tracing.

Particle tracing algorithms have been explained in terms of "energy packets" (e.g. [31]), or as random walk solutions of the importance transport equation [28]. We explain how these algorithms are affected by non-symmetric BSDF's.

Recall that the BSDF arguments are labeled according to the direction of light flow (from ω_i to ω_o). For particle scattering, we are given ω_i and must sample ω_o. This is reversed compared to ray or path tracing, so we must be careful when f_s is not symmetric; otherwise incorrect and inconsistent results will be obtained, as described above.

Rather than swapping directions, it is often more convenient to use the adjoint BSDF. We adopt the convention that ω_i is always the *sampled* direction during a random walk (i.e. ω_o has already been chosen). The other situation is described as "sampling the adjoint BSDF". With this convention, f_s^* applies to light particles and importance evaluation, while f_s is used for "importance particles" and radiance evaluation.

3 Sources of non-symmetric scattering

Most reflection models in computer graphics are symmetric. One notable exception is the original Phong model for glossy reflection [29].[11] Although his "shading formula" is symmetric, the corresponding BRDF has an extra factor of $1/\cos(\theta_i)$.

However, some sources of non-symmetry are not so obvious. We discuss two such examples: refraction and shading normals.

3.1 Refraction

When light is transmitted between media with different refractive indices, the corresponding BTDF is not symmetric. Radiance crossing the interface is scaled by an extra factor of $(\eta_t/\eta_i)^2$ that is not required for importance (or light particles). If this extra scaling is ignored, there can be substantial errors.

For example, consider a light source shining on a swimming pool with a diffuse bottom and sides. Suppose that we use particle tracing to accumulate the caustic pattern on the bottom of the pool, and then we render an image using ray tracing. If the radiance for the viewing rays is not scaled, the caustics in the image will be *too bright* by a factor of $(\eta_t/\eta_i)^2$ (about 1.78 for water).

We describe Helmholtz reciprocity and other general properties of optical systems, show how they apply in the case of pure specular refraction, and finally we examine ways to "fix" the theoretical framework to increase its symmetry.

[11] Phong also proposed the use of shading normals, which is another source of non-symmetry.

3.1.1 Reciprocities.

The Helmholtz reciprocity principle can be found in his famous treatise on physiological optics, first published in 1856 [40, p.231]. It is formulated as a theorem in geometric optics, and states that if we follow a beam of light on any path through an optical system, the loss in intensity[12] is the same as for a beam traveling in the reverse direction.[13]

With respect the symmetry of BRDF's, an observation of Lord Rayleigh is actually more relevant (cf. [6, p.177]). Consider a small reflective surface, exposed to a small light source and a small irradiance detector. His reciprocity principle states that if the positions of the source and detector are exchanged, the reflected irradiance measured by the sensor will be the same. This implies the symmetry of the corresponding BRDF.

Many reciprocity relationships, including these, can be derived from thermodynamic principles (e.g. [32, p.65], [11, p.505]). Such arguments consider an isothermal, black enclosure containing an arbitrary test surface (reflective or transmissive). According to Kirchoff's law,[14] if the (spectral) radiance in such an enclosure is divided by the index of refraction squared (which may vary with position), the result is purely a function of temperature. In other words, L/η^2 does not vary with position or direction.

Using this fact, we can derive a reciprocity principle that is satisfied by all physically valid BSDF's. By considering the energy exchange between two small surface elements of the enclosure, it is straightforward to show that

$$f_{\mathrm{s}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}}) \,/\, \eta_{\mathrm{o}}^2 = f_{\mathrm{s}}(\omega_{\mathrm{o}} \to \omega_{\mathrm{i}}) \,/\, \eta_{\mathrm{i}}^2 \tag{8}$$

(see [32, p.65] for a similar argument). This is clearly a generalization of the usual symmetry condition for BRDF's. However, we see that any BSDF involving refraction is not symmetric. The ratio of f_{s} to f_{s}^* is $(\eta_{\mathrm{o}}/\eta_{\mathrm{i}})^2$, so that radiance and importance are scaled differently when they are transmitted through the surface.

[12] Helmholtz phrased this law in terms of "quantities of light" (flux) rather than "brightness" (radiance), because he was aware of the change in radiance due to the index of refraction [40, p.233].

[13] Helmholtz reciprocity is not universally valid. There are some situations where optical paths are not reversible, i.e. light flowing in the reverse direction follows a different path. This can happen when electromagnetic waves are transmitted into metals [39]. Also, reciprocity can fail for polarized light in the presence of an external magnetic field [40, p.231]. Interestingly, Helmholtz did not provide a proof of his principle, because "anybody who is at all familiar with the laws of optics can easily prove it for himself" [40, p.231].

[14] The dependence of blackbody radiation on η^2 is often falsely attributed to Clausius (cf. [11, p.504]).

3.1.2 Perfect specular refraction.

We show how these principles apply in the case of perfect specular refraction, where the interface between media is optically smooth. This is by far the most common example of refraction in graphics.[15] We also give explicit formulas for the corresponding BTDF and its adjoint. For simplicity, we will ignore reflection and assume that all light is transmitted through the interface.

Intuitively, when light enters a medium with a higher refractive index, the same light energy is squeezed into a smaller volume. To see this, consider a small patch dA exposed to uniform radiance in the hemisphere of incident directions Ω_{i}, and assume that $\eta_{\mathrm{i}} < \eta_{\mathrm{t}}$. Since $\sin\theta_{\mathrm{t}} < \eta_{\mathrm{i}}/\eta_{\mathrm{t}}$ by Snell's law, the transmitted light does not fill the entire hemisphere Ω_{t}. Thus radiance must increase, by conservation of energy.

In fact, the incident and transmitted radiance are related by

$$L_{\mathrm{i}}/\eta_{\mathrm{i}}^2 = L_{\mathrm{t}}/\eta_{\mathrm{t}}^2 \ . \tag{9}$$

This can be shown using Snell's law (e.g. see [21], [14, p.30]). These arguments first prove a relationship between the throughputs of the incident and transmitted beams:

$$\eta_{\mathrm{i}}^2 \, d\mu(\mathbf{r}_{\mathrm{i}}) = \eta_{\mathrm{t}}^2 \, d\mu(\mathbf{r}_{\mathrm{t}}) \ , \tag{10}$$

where μ is the throughput measure (2). They also need the fact that $d\Phi_{\mathrm{i}} = d\Phi_{\mathrm{t}}$, by conservation of energy. Finally, equation (9) follows from the relationship $d\Phi = L\, d\mu$ between power and radiance.

Thus radiance is scaled by $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$ when it crosses the interface. The BTDF is

$$f_{\mathrm{t}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{t}}) = (\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2 \delta_{\sigma^\perp}(\omega_{\mathrm{i}} - \tau(\omega_{\mathrm{t}})) \ ,$$

where $\omega_{\mathrm{i}} = \tau(\omega_{\mathrm{t}})$ is the mapping between ω_{i} and ω_{t} determined by Snell's law, and $\delta_{\sigma^\perp}$ denotes the Dirac delta distribution with respect to the projected solid angle measure.[16]

According to (8), the adjoint BTDF is

$$f_{\mathrm{t}}^*(\omega_{\mathrm{i}} \to \omega_{\mathrm{t}}) = (\eta_{\mathrm{i}}/\eta_{\mathrm{t}})^2 f_{\mathrm{t}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{t}}) = \delta_{\sigma^\perp}(\omega_{\mathrm{i}} - \tau(\omega_{\mathrm{t}})) \ .$$

The $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$ factor is *not* present in the adjoint, so that importance (and light particles) are not scaled when they cross the interface.

Hall [14] pointed out the $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$ scaling for radiance, but many ray tracers ignore this. We are not aware of any system that implements different rules for radiance and importance/light particles. This is easy to do, and essential for correctness.

---

[15] Equation (8) applies to more general BTDF's as well, e.g. frosted glass.

[16] For our purposes, the following is sufficient. Given some measure μ, δ_μ is defined by the property that $\int_\Omega g(\mathbf{x})\, \delta_\mu(\mathbf{x} - \mathbf{x}_0)\, d\mu(\mathbf{x}) = g(\mathbf{x}_0)$ (see [1], p.28–32).

3.1.3 Optical invariants.

We have already mentioned several *optical invariants*, quantities that are preserved as a beam of radiation propagates through an optical system. To simplify things, we will assume that the beam follows a single path (no partial reflection, etc.) and that there are no losses due to absorption.

The first quantity is L/η^2, which Nicodemus calls *basic radiance* [21] [25, p.29]. Its invariance is known as *Abbe's law* [17, p.195], or *radiance invariance* [25, p.26].[17]

From this fact and conservation of energy, we can also show the invariance of $\eta^2\, d\mu$. This quantity is known as *basic throughput* [25, p.37], or *etendue* [35].[18]

3.1.4 An invariant transport framework.

By writing the transport equations in terms of these invariants, we obtain a framework where light and importance obey the same scattering rules. First, we use basic radiance

$$L' = L/\eta^2$$

for all lighting calculations. Next, we redefine the inner product on $\mathcal{R}$ to be

$$\langle L', W \rangle = \int_{\mathcal{R}} L'(\mathbf{r})\, W(\mathbf{r})\, d\mu'(\mathbf{r}) \ ,$$

where $d\mu' = \eta^2\, d\mu$ is the basic throughput measure. This is used for all measurements. Finally, we replace the BSDF by

$$f_{\mathrm{s}}'(\omega_{\mathrm{i}} \!\rightarrow\! \omega_{\mathrm{o}}) = f_{\mathrm{s}}(\omega_{\mathrm{i}} \!\rightarrow\! \omega_{\mathrm{o}})/\eta_{\mathrm{o}}^2 \ ,$$

which is symmetric according to (8). With these changes, $\mathbf{G}$ and $\mathbf{K}$ are both self-adjoint for physically valid models, and thus basic radiance and importance obey the same scattering rules.

From an implementation standpoint, we work with L/η^2 instead of L, and compensate by multiplying by η^2 at the point where light and importance interact. To do this, we must know the refractive index of the surrounding medium.[19]

[17] *Basic spectral radiance* L_ν/η^2 is invariant as well. If spectral radiance is parameterized by wavelength, then L_λ/η^3 is invariant [25, p.52], since wavelengths (unlike frequencies) are modified at the interface.

[18] This was first derived (in a simpler linear form) by Smith in 1738 (cf. [40, p.74]). The linear invariant is called the *Lagrange invariant* or the *Smith-Helmholtz invariant*.

[19] Another way to make the framework symmetric is to work with the quantities L/η and W/η (leaving the BSDF and inner product alone). This gives correct results as long as all sources and sensors are in the same medium. However, given a system that allows BSDF's to be non-symmetric (as required to handle shading normals), there is little reason not to handle light and importance differently.

3.2 Shading normals

We show that shading normals lead to non-symmetric BSDF's. This also causes problems with conservation of energy, and shading discontinuities.

Let $\mathbf{x} \in \mathcal{M}$ be a fixed point, so that we can omit $\mathbf{x}$ from the notation. Let $\mathbf{N}_\mathrm{g}$ be the geometric normal, let $\mathbf{N}_\mathrm{s}$ be the shading normal, and let $f_{\mathrm{s},\mathbf{N}}(\omega_\mathrm{i} \to \omega_\mathrm{o})$ denote the BSDF, rotated as though the surface had normal $\mathbf{N}$. The shading normal is used in lighting calculations according to

$$L_\mathrm{o}(\omega_\mathrm{o}) = (\mathbf{K}L)(\omega_\mathrm{o}) = \int_{\mathcal{S}^2} L(\omega_\mathrm{i})\, f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{i} \to \omega_\mathrm{o})\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{s}|\, d\sigma(\omega_\mathrm{i}) \ . \tag{11}$$

Of course, this does not actually change the normal. Some calculations will detect this (e.g. if we compute the solid angle occupied by this surface from some other point), so we should expect inconsistencies. Really, the shading normal is a parameter that modifies the *BSDF* to change the surface appearance. Even if the original BSDF is symmetric, the modified BSDF is not. To see this, we write (11) in the standard form (3):

$$(\mathbf{K}L)(\omega_\mathrm{o}) = \int_{\mathcal{S}^2} L(\omega_\mathrm{i})\, f'_\mathrm{s}(\omega_\mathrm{i} \to \omega_\mathrm{o})\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|\, d\sigma(\omega_\mathrm{i}) \ , \tag{12}$$

where

$$f'_\mathrm{s}(\omega_\mathrm{i} \to \omega_\mathrm{o}) = f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{i} \to \omega_\mathrm{o})\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{s}| \,/\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}| \ .$$

The modified BSDF f'_s is exactly equivalent to the original shading formula (11). However, since f'_s is not symmetric, importance is scattered according to

$$(\mathbf{K}^*W)(\omega_\mathrm{o}) = \int_{\mathcal{S}^2} W(\omega_\mathrm{i})\, f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{o} \to \omega_\mathrm{i})\, \frac{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{s}|}{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}|}\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|\, d\sigma(\omega_\mathrm{i}) \ .$$

This also applies to particle transport (see Sec. 2.5). The directions ω_i and ω_o are labeled with respect to importance transport, and so particles go from ω_o to ω_i. Scattered particles weights are thus multiplied by

$$\alpha(\omega_\mathrm{i}) = \frac{f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{o} \to \omega_\mathrm{i})}{p_{\omega_\mathrm{o}}(\omega_\mathrm{i})}\, \frac{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{s}|}{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}|}\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}| \ , \tag{13}$$

where $p_{\omega_\mathrm{o}}(\omega)\, d\sigma(\omega)$ is the distribution that ω_i was sampled from. If particles are weighted in this way, the results will be consistent with shading formula (11).

3.2.1 Examples.

For a diffuse surface, the BRDF is a constant K_d. Inserting this in (13), we see that particles are scattered according to $|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|$, with a weight that depends on the direction they *arrived* from. This is very different than radiance sampling, where the samples are distributed according to $|\omega_i \cdot \mathbf{N}_s|$.

As another example, consider a perfect mirror, whose BRDF is a delta function ([27, p.44], [9, p.31]). Applying (13) to this BRDF, we find that reflected particles should be weighted by

$$\alpha = |\omega_i \cdot \mathbf{N}_g| \,/\, |\omega_o \cdot \mathbf{N}_g| \ .$$

We give pseudocode below that shows how to evaluate the scattering kernel (including the factor of $|\omega_i \cdot \mathbf{N}_g|$ hidden in the $\sigma^\perp$ notation). The *adjoint* flag controls whether $\mathbf{K}$ or $\mathbf{K}^*$ is evaluated. We also show how to prevent light from "leaking" through the surface [34], by checking that each direction lies on the same side of the surface with respect to both normals. The easiest solution is just to return zero when this happens.

EVAL-KERNEL($\omega_i \to \omega_o$, *adjoint*)
  assert $\mathbf{N}_g \cdot \mathbf{N}_s \geq 0$ (if not, flip $\mathbf{N}_g$)
  if $(\omega_i \cdot \mathbf{N}_g)(\omega_i \cdot \mathbf{N}_s) \leq 0$ **or** $(\omega_o \cdot \mathbf{N}_g)(\omega_o \cdot \mathbf{N}_s) \leq 0$
    then **return** 0
  if *adjoint*
    then **return** $f_{s,\mathbf{N}_s}(\omega_o \to \omega_i)\, |\omega_o \cdot \mathbf{N}_s|\, |\omega_i \cdot \mathbf{N}_g| \,/\, |\omega_o \cdot \mathbf{N}_g|$
    else **return** $f_{s,\mathbf{N}_s}(\omega_i \to \omega_o)\, |\omega_i \cdot \mathbf{N}_s|$

3.2.2 No conservation of energy.

For energy to be conserved, we must have

$$\int_{\mathcal{S}^2} f_s'(\omega_i \to \omega_o)\, d\sigma^\perp(\omega_o) \leq 1 \quad \text{for all } \omega_i \ .$$

However, in (12), the factor $|\omega_i \cdot \mathbf{N}_s| / |\omega_i \cdot \mathbf{N}_g|$ can be arbitrarily large. For intuition about this, consider Fig. 1(a). Even though the surfaces are nearly perpendicular to the light they receive, they are *shaded* as though they were facing directly toward the source. However, their total area is much larger than could be achieved with a surface facing the source (and occupying the same solid angle). The total power reflected by these surfaces can thus be far greater than that emitted by the source.

3.2.3 Shading discontinuities.

The adjoint BSDF is essential for the smooth shading of polygonal meshes, when particle tracing is used. Consider Fig. 1(b). Light of uniform intensity is arriving from direction ω_o at the polygonal surface shown (as before, particles go from ω_o to ω_i). The shading normal is assumed to be continuous across the boundary between

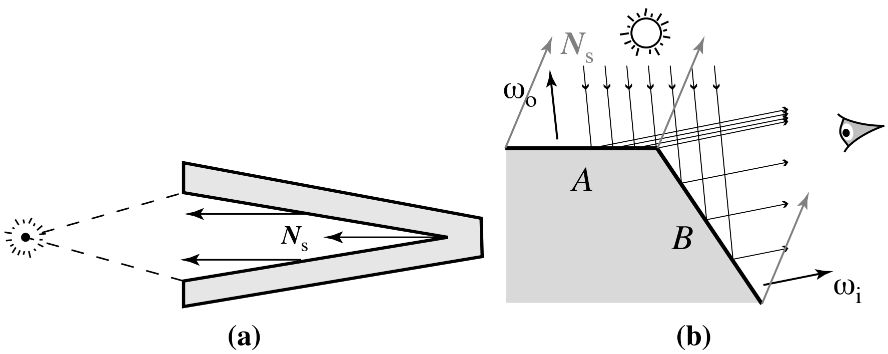

Figure 1: **(a)** Two surfaces whose shading normals point toward a light source — they receive more power than the light source emits. **(b)** The density of particles seen by the eye is discontinuous, even though the shading normal is continuous (in fact, constant) across the polygon boundary.

polygons A and B, so the shading of the mesh should appear smooth. However, the geometric normals of A and B are different, so that fewer particles per unit area are received by B than by A. The irradiances at A and B are in the ratio

$$\frac{E_A}{E_B} = \frac{|\omega_{\mathrm{o}} \cdot \mathbf{N}_A|}{|\omega_{\mathrm{o}} \cdot \mathbf{N}_B|} \ .$$

If we render this surface using a particle-tracing algorithm, there will be a discontinuity in the apparent brightness.

Furthermore, suppose that we use an image-space splatting algorithm, where each particle makes a dot at the appropriate point in the image. The image brightness depends on the spacing between the particles measured perpendicular to the viewing direction ω_{i}, since this determines how many particles will strike each pixel. The spacing perpendicular to ω_{i} is simply the spacing on the surface, divided by $|\omega_{\mathrm{i}} \cdot \mathbf{N}|$. The image intensities at A and B are thus in the ratio

$$\frac{I_A}{I_B} = \frac{|\omega_{\mathrm{o}} \cdot \mathbf{N}_A|\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_B|}{|\omega_{\mathrm{o}} \cdot \mathbf{N}_B|\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_A|} \ .$$

Yet if we weight the particles according to (13), the shading will be smooth (assuming that $\mathbf{N}_{\mathrm{s}}$ is continuous across the boundary). The weight includes a factor of $|\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}| / |\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}|$, which exactly compensates for the discontinuity in particle density.

Many algorithms use particle tracing to render at least some component of the lighting on directly visible surfaces (e.g. caustics). If the adjoint BSDF is not used, this can cause false discontinuities in the image.

3.3 Results

Plate 1 (see color section) shows a bump-mapped teapot, and a polygonalized sphere with smooth shading normals. The images are simulations of a splatting algorithm: when a particle strikes a surface, a splat is made at the corresponding point in the image. The splat intensity depends on how much light is reflected toward the eye. Plate 1a shows the correct result (using the adjoint BSDF), while Plate 1b shows what happens if particles are scattered just like viewing rays (i.e. the non-symmetry caused by shading normals is not recognized). Both images use the same shading normals.

Plate 2 shows a pool of water with small waves, illuminated by two area light sources, and rendered with a particle tracing algorithm. Plate 2a shows the correct result (where radiance is scaled by $(\eta_t/\eta_i)^2$, but particle weights are not[20]). In Plate 2b, neither radiance nor particle weights are scaled (the non-symmetry of the BTDF is not recognized), leading to caustics that are too bright by a factor of $(\eta_t/\eta_i)^2$ (see Sec. 3.1).

Acknowledgments

Jim Arvo helped with terminology and notation, and pointed out many heat transfer references. Pat Hanrahan suggested the possibility of an invariant framework. Thanks to Leo Guibas for encouragement and discussions, and to Matt Pharr and Luanne Lemmer for valuable comments and last-minute help. This research was supported by the National Science Foundation (CCR-9215219).

References

[1] AL-GWAIZ, M. A. *Theory of Distributions*. Marcel Dekker, Inc., New York, 1992.

[2] AMBARTZUMIAN, R. V. *Factorization Calculus and Geometric Probability*. Cambridge University Press, 1990.

[3] ARVO, J. *Analytic Methods for Simulated Light Transport*. PhD thesis, Yale University, Dec. 1995.

[20]Note that when shading normals are used on a refractive interface, the results of both Sec. 3.1 and Sec. 3.2 apply. Thus radiance is scaled by $(\eta_t/\eta_i)^2$, while particle weights are scaled by $|\omega_t \cdot \mathbf{N}_s|\,|\omega_i \cdot \mathbf{N}_g| \,/\, |\omega_t \cdot \mathbf{N}_g|\,|\omega_i \cdot \mathbf{N}_s|$, where particles go from ω_t to ω_i.

[4] ARVO, J., TORRANCE, K., AND SMITS, B. A framework for the analysis of error in global illumination algorithms. In *Proceedings of SIGGRAPH '94* (July 1994), ACM Press, pp. 75–84.

[5] BLINN, J. F. Simulation of wrinkled surfaces. In *Computer Graphics (SIGGRAPH '78 Proceedings)* (Aug. 1978), vol. 12, pp. 286–292.

[6] CHANDRASEKAR, S. *Radiative Transfer*. Dover Publications, New York, 1960.

[7] CHRISTENSEN, P. H., SALESIN, D. H., AND DEROSE, T. D. A continuous adjoint formulation for radiance transport. In *Fourth Eurographics Workshop on Rendering* (Paris, June 1993), pp. 95–104.

[8] CHRISTENSEN, P. H., STOLLNITZ, E. J., SALESIN, D. H., AND DEROSE, T. D. Global illumination of glossy environments using wavelets and importance. *ACM Trans. Graphics 15* (1996), 37–71.

[9] COHEN, M. F., AND WALLACE, J. R. *Radiosity and Realistic Image Synthesis*. Academic Press Professional, San Diego, CA, 1993.

[10] DELVES, L. M., AND MOHAMED, J. L. *Computational Methods for Integral Equations*. Cambridge University Press, New York, 1985.

[11] DRUDE, P. *The Theory of Optics*. Dover Publications, New York, 1959. Original date of publication 1900. Translated from the German by C. R. Mann and R. A. Millikan.

[12] FOURNIER, A. From local to global illumination and back. In *Rendering Techniques '95* (New York, 1995), P. M. Hanrahan and W. Purgathofer, Eds., Springer-Verlag, pp. 127–136.

[13] GLASSNER, A. *Principles of Digital Image Synthesis*. Morgan Kaufmann, New York, 1995.

[14] HALL, R. *Illumination and Color in Computer Generated Imagery*. Springer-Verlag, New York, 1989.

[15] HECKBERT, P. S. Adaptive radiosity textures for bidirectional ray tracing. *Computer Graphics 24*, 4 (July 1990), 145–154.

[16] HECKBERT, P. S. *Simulating Global Illumination Using Adaptive Meshing*. PhD thesis, University of California, Berkeley, June 1991.

[17] KEITZ, H. A. E. *Light Calculations and Measurements*. The Macmillan Company, New York, 1971.

[18] LAFORTUNE, E. P., AND WILLEMS, Y. D. Bi-directional path tracing. In *Proceedings of CompuGraphics* (Alvor, Portugal, Dec. 1993), pp. 145–153.

[19] LAFORTUNE, E. P., AND WILLEMS, Y. D. A theoretical framework for physically based rendering. *Computer Graphics Forum 13*, 2 (June 1994), 97–107.

[20] LEWINS, J. *Importance, The Adjoint Function: The Physical Basis of Variational and Perturbation Theory in Transport and Diffusion Problems*. Pergamon Press, New York, 1965.

[21] NICODEMUS, F. E. Radiance. *Am. J. Phys. 31* (1963), 368–377.

[22] NICODEMUS, F. E. Directional reflectance and emissivity of an opaque surface. *Applied Optics 4* (1965), 767–773.

[23] NICODEMUS, F. E. Reflectance nomenclature and directional reflectance and emissivity. *Applied Optics 9* (1970), 1474–1475.

[24] NICODEMUS, F. E. Comment on "current definitions of reflectance". *J. Opt. Soc. Am. 66* (1976), 283–285.

[25] NICODEMUS, F. E., Ed. *Self-Study Manual on Optical Radiation Measurements: Part I—Concepts, Chapters 1 to 3*. No. 910-1 in Technical Note. National Bureau of Standards (US), Mar. 1976.

[26] NICODEMUS, F. E., Ed. *Self-Study Manual on Optical Radiation Measurements: Part I—Concepts, Chapters 4 and 5*. No. 910-2 in Technical Note. National Bureau of Standards (US), Feb. 1978.

[27] NICODEMUS, F. E., ET AL. Geometric considerations and nomenclature for reflectance. Monograph 161, National Bureau of Standards (US), Oct. 1977.

[28] PATTANAIK, S. N., AND MUDUR, S. P. Adjoint equations and random walks for illumination computation. *ACM Trans. Graphics 14* (Jan. 1995), 77–102.

[29] PHONG, B. T. Illumination for computer generated pictures. *Communications of the ACM 18*, 6 (June 1975), 311–317.

[30] SCHRÖDER, P., AND HANRAHAN, P. Wavelet methods for radiance computations. In *Fifth Eurographics Workshop on Rendering* (Darmstadt, Germany, June 1994), pp. 303–311.

[31] SHIRLEY, P., WADE, B., HUBBARD, P. M., ZARESKI, D., WALTER, B., AND GREENBERG, D. P. Global illumination via density-estimation. In *Rendering Techniques '95* (New York, 1995), P. M. Hanrahan and W. Purgathofer, Eds., Springer-Verlag, pp. 219–230.

[32] SIEGEL, R., AND HOWELL, J. R. *Thermal Radiation Heat Transfer*, second ed. Hemisphere Publishing Corp., New York, 1981.

[33] SMITS, B. E., ARVO, J. R., AND SALESIN, D. H. An importance-driven radiosity algorithm. In *Computer Graphics (SIGGRAPH '92 Proceedings)* (July 1992), vol. 26, pp. 273–282.

[34] SNYDER, J. M., AND BARR, A. H. Ray tracing complex models containing surface tessellations. In *Computer Graphics (SIGGRAPH '87 Proceedings)* (July 1987), vol. 21, pp. 119–128.

[35] STEEL, W. H. Luminosity, throughput, or etendue? *Applied Optics 13* (1974), 704–705.

[36] TAYLOR, A. E., AND LAY, D. C. *Introduction to Functional Analysis*, second ed. John Wiley & Sons, New York, 1980.

[37] VEACH, E., AND GUIBAS, L. Bidirectional estimators for light transport. In *Fifth Eurographics Workshop on Rendering* (Darmstadt, Germany, June 1994), pp. 147–162.

[38] VEACH, E., AND GUIBAS, L. J. Optimally combining sampling techniques for monte carlo rendering. In *SIGGRAPH '95 Conference Proceedings* (Aug. 1995), Addison Wesley, pp. 419–428.

[39] VON FRAGSTEIN, C. Ist eine lichtbewegung stets umkehrbar? *Opt. Acta 2* (1955), 16–22.

[40] VON HELMHOLTZ, H. *Helmholtz's Treatise on Physiological Optics*, vol. 1. Dover Publications, New York, 1962. James P. C. Southall, Ed. Original date of publication 1856. Translated from the third German edition.