Chapter 5

The Sources of Non-Symmetric Scattering

In this chapter, we study two examples of non-symmetric scattering that have not previously been recognized. Specifically, we show that non-symmetric scattering occurs whenever light is refracted, and also whenever shading normals are used. We show how to handle these situations correctly in bidirectional light transport algorithms, by deriving and using the corresponding adjoint BSDF's.

It is important to note that these sources of non-symmetry are not obvious, and that shading normals and refraction are widely assumed to be described by symmetric BSDF's. We show that this can cause significant problems when bidirectional algorithms are used. For example, it can cause rendered images to have large errors, even when the scene model is physically valid. It can also cause rendering algorithms that are supposedly equivalent to converge to different solutions (whether the scene model is physically valid or not). Finally, it can cause shading artifacts that should not be present, such as brightness discontinuities. These problems can occur whenever a non-symmetric BSDF is used without recognizing it (i.e. when it is handled as though it were symmetric).

We show that there are two distinct situations where non-symmetric BSDF's can arise. First, some scattering models in computer graphics are not physically valid. A good example of this is the use of shading normals (which are commonly applied to make polygonal surfaces look smooth, or to add detail to coarse geometric models). Although shading normals do not have any well-defined physical basis, they are very convenient and useful for many graphics applications. Thus it is important to be able to handle these non-physical materials correctly when bidirectional light transport algorithms are used.

The other situation where non-symmetric BSDF's arise is the refraction of light between two different media. Notice that in this case, the BSDF describes a real physical effect. This implies that even when a scene model is physically valid, it is sometimes necessary to use different transport rules for light and importance (or path tracing and particle tracing) in order for bidirectional algorithms to converge to the correct result.

This chapter is organized as follows. In Section 5.1 we explain why non-symmetric BSDF's are sometimes difficult to recognize, and we describe the significant problems that this can cause. We also discuss several elementary sources of non-symmetric scattering that are well-known in graphics. We then analyze in detail two sources of non-symmetry described above: namely refraction (Section 5.2) and the use of shading normals (Section 5.3). We also present test cases demonstrating the errors that occur in computed images when these BSDF's are not handled correctly.

Another contribution is the idea of Dirac distributions with respect to general measures, which can be used to model specular scattering and transport singularities in general. This concept is described in Appendix 5.A, along with several identities that can be used to manipulate and evaluate them in a consistent way. (Although this idea seems quite basic, we are unable to give a reference for it.)

5.1 Introduction

5.1.1 The problems caused by non-symmetric scattering

We explain the problems that arise when non-symmetric BSDF's are treated as though they were symmetric. This provides some motivation for the rest of this chapter, where we study the various reasons that non-symmetric scattering occurs.

The main problem with non-symmetric BSDF's is that they are sometimes difficult to recognize. Most often this occurs when a scattering model is defined procedurally (rather than by giving the BSDF as an explicit function). For example, refraction is generally implemented as a procedure that maps an incident direction ω_{i} into a transmitted direction ω_{t}. The BSDF itself is not evaluated or even represented (because as we will see, it is a *Dirac distribution* rather than an ordinary function). Another example of a procedurally defined scattering model is the use of shading normals. With this technique, the true surface normal $\mathbf{N}_{\mathrm{g}}$ is replaced by a different vector $\mathbf{N}_{\mathrm{s}}$ (the *shading normal*) when the BSDF is evaluated. This can be interpreted as a procedural modification of an existing BSDF.

In these cases, it is easy to use a non-symmetric BSDF without realizing it. When this happens, the same BSDF f_{s} is used in all situations (even those where the adjoint BSDF f_{s}^* should be used instead). This creates problems, because the BSDF is not used consistently. For example, consider a bidirectional algorithm that uses particle tracing in one phase, and path tracing in another. Recall that in order to get correct results with such an algorithm, the adjoint BSDF f_{s}^* must be used during the particle tracing phase (see Section 3.7.5). Therefore, using the ordinary BSDF f_{s} for particle tracing is equivalent to solving a light transport equation that uses the adjoint BSDF f_{s}^*. By using the same BSDF f_{s} in both phases, we get results that are almost certainly wrong: they could converge to the solution of a light transport equation that uses f_{s}, f_{s}^*, or any combination of the two. This has a number of practical consequences:

- Computed images can have substantial errors (even when the scene model is physically correct). The computed radiance values can easily be wrong by a factor of two or more.

- Rendering algorithms that are supposed to be equivalent may in fact converge to different answers. This can happen whether the model is physically valid or not. For example, path tracing might converge to a different result than particle tracing, because particle tracing must use the adjoint BSDF to get results that are consistent with path tracing.

- Computed images can have spurious, visually objectionable artifacts. For example, if the adjoint BSDF for shading normals is not used correctly, there can be false discontinuities in the image reminiscent of flat-shaded polygons. (This will be explained in Section 5.3.)

Note that these errors can occur with all types of bidirectional algorithm. For example, this includes importance-driven finite element methods, multi-pass algorithms, particle tracing approaches, and bidirectional path tracing.[1]

Fortunately, these problems are easy to fix. It is only necessary to recognize non-symmetric BSDF's whenever they exist, and make appropriate use of the corresponding adjoint BSDF.

5.1.2 Elementary sources of non-symmetric scattering

There are several reasons why the BSDF's used in graphics are sometimes not symmetric. One reason is that shading models are sometimes derived empirically, without regard for the laws of physics. Another reason is that shading models are sometimes approximated, to make them faster to evaluate. These sources of non-symmetry are well-known, and they are relatively easy to recognize and handle correctly.

On the other hand, some sources of non-symmetry are not so easily recognized. This category includes non-symmetry due to refraction and the use of shading normals, which are discussed separately in Sections 5.2 and 5.3.

5.1.2.1 Empirical shading models

The most obvious source of non-symmetric scattering is that some shading models are derived empirically, using formulas that are convenient to calculate and happen to give interesting visual results. Probably the best-known example of this is the original Phong model for glossy reflection [Phong 1975].[2] Translating his formula into our terminology, the reflected radiance from a glossy surface is computed according to

$$L_{\mathrm{o}}(\omega_{\mathrm{o}}) \;=\; \int_{\mathcal{H}_{\mathrm{i}}^2} C_{\mathrm{r}} \, \max(0, \omega_{\mathrm{i}} \cdot M_{\mathbf{N}(\mathbf{x})}(\omega_{\mathrm{o}}))^n \, L_{\mathrm{i}}(\omega_{\mathrm{i}}) \, d\sigma(\omega_{\mathrm{i}}) \,, \tag{5.1}$$

[1] Note that with importance-driven algorithms, the adjoint BSDF is only used to compute importance. Thus if we use the wrong BSDF, it will not cause errors in the solution (provided that importance is only used to guide mesh refinement). However, any error *estimates* that depend on importance will be wrong.

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

where C_r controls the color and intensity of the glossy highlights, n controls the apparent specularity of the surface, and $M_\mathbf{N}(\omega_\mathrm{o})$ is the mirror direction (see Section 5.2.1.2).

Although this "shading formula" is symmetric, notice that the integration is with respect to solid angle σ, whereas in the scattering equation (3.12) the projected solid angle $\sigma^\perp_\mathbf{x}$ is used. Thus when this shading formula is expressed as a BRDF, it has an extra factor of

$$1 \,/\, |\omega_\mathrm{i} \cdot \mathbf{N}(\mathbf{x})|$$

that makes it non-symmetric. (This factor is required to cancel the factor of $|\omega_\mathrm{i} \cdot \mathbf{N}(\mathbf{x})|$ that is hidden by the projected solid angle notation (3.1).) It is easy to fix the non-symmetry, of course, by changing the definition (5.1) to use integration with respect to projected solid angle.

5.1.2.2 Approximations of symmetric BSDF's

Non-symmetric BSDF's can also arise when physically valid scattering models are approximated (to make their computation more efficient). For example, the Cook-Torrance model [Cook & Torrance 1982] has the form

$$f_\mathrm{s}(\omega_\mathrm{i} \to \omega_\mathrm{o}) \;=\; \frac{DGF}{\cos\theta_\mathrm{i}\cos\theta_\mathrm{o}}\,,$$

where D, G, and F are symmetric functions of ω_i and ω_o. This BSDF is clearly symmetric. However, notice that when it is inserted into the scattering equation (3.12), the factor of $\cos\theta_\mathrm{i}$ is canceled by the corresponding factor in the projected solid angle notation. When this formula is implemented in hardware, it is common to throw away the factor of $\cos\theta_\mathrm{o}$ as well. This saves a division operation, but destroys the symmetry of the corresponding BSDF. For example, this is the approach taken by the OpenGL specification [OpenGL Architecture Review Board 1992].

5.2 Non-symmetry due to refraction

We show that when light is refracted, the corresponding BSDF is not symmetric. In particular, we show that radiance crossing the interface must be scaled by a factor of $(\eta_\mathrm{t}/\eta_\mathrm{i})^2$, but

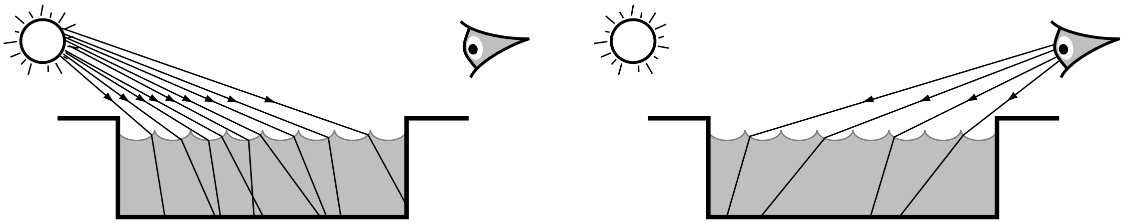


Figure 5.1: Two-pass rendering of caustics on a pool bottom. **(a)** First pass: particles are traced from the light sources, and their distribution on the pool bottom is recorded. **(b)** Second pass: an image is rendered using ray tracing. When rays intersect the pool bottom, the light distribution recorded by the first pass is sampled. To obtain correct results in this example, it is essential to handle particles and viewing rays differently at the air-water interface: radiance is scaled by $(\eta_t/\eta_i)^2$, while particle weights are left unchanged.

that no scaling is required for importance. (For simplicity we will ignore partial reflection in this section; in other words, we assume that all light is transmitted through the interface.) We derive explicit formulas for corresponding BSDF and its adjoint, and we discuss the implications for bidirectional rendering algorithms.

Note that there can be substantial errors if the $(\eta_t/\eta_i)^2$ scaling factor for radiance is ignored. For example, consider a light source shining on a swimming pool with a diffuse bottom and sides (see Figure 5.1). Suppose that particle tracing is used to accumulate the caustic pattern on the bottom of the pool, followed by a ray tracing pass to render the final image. If the radiance of the viewing rays is not scaled at the air-water interface, the caustics in the image will be too bright by a factor of $(\eta_t/\eta_i)^2$ (about 1.78 for water). In particular, the caustics will be brighter than they would be in a path-traced image. On the other hand, if the $(\eta_t/\eta_i)^2$ scaling factor is applied to *both* the viewing rays and the particles, the caustics will be too dim by a factor of $(\eta_t/\eta_i)^2$.

The main point of this section is not that radiance must be scaled when it enters a medium with a different refractive index; this fact is well-known in radiometry and optics, although it does not seem to have been implemented in many graphics systems. Rather, the point is that the adjoint BSDF does *not* involve any such scaling. Thus the BSDF is not symmetric, and so different rules must be used for radiance and importance (or particle tracing and path tracing) in order to obtain correct results.

5.2.1 Background material

We review two results that will be used to derive the BSDF for refraction (and its adjoint). First, we show that when light is refracted, radiance is scaled by a factor of $(\eta_\mathrm{t}/rii)^2$ as it crosses the interface. Second, we show how to write the BSDF for reflection from a perfect mirror, using a new notation involving a Dirac distribution with respect to the projected solid angle measure.

5.2.1.1 Radiance scaling at a refractive 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 $dA(\mathbf{x})$ that is exposed to uniform radiance over the incident hemisphere $\mathcal{H}^2_\mathrm{i}$, and assume that this light is transmitted into a medium with a higher refractive index (Figure 5.2). Then the transmitted light does not fill the entire hemisphere $\mathcal{H}^2_\mathrm{t}$, since by Snell's law, the angle of refraction satisfies

$$\sin\theta_\mathrm{t} \;<\; \frac{\eta_\mathrm{i}}{\eta_\mathrm{t}}\,.$$

Thus radiance must increase as light crosses the interface (at least on some subset of the rays), by conservation of energy.

In fact, the incident and transmitted radiance are related by

$$L_\mathrm{t} \;=\; \frac{\eta_\mathrm{t}^2}{\eta_\mathrm{i}^2}\,L_\mathrm{i}\,. \tag{5.2}$$

This can be shown using Snell's law (e.g. [Milne 1930, p. 74], [Nicodemus 1963], [Hall 1989, p. 30]). We repeat this argument here, since we will need some of the intermediate results.

Consider a beam of light that strikes small surface patch $dA(\mathbf{x})$, and occupies a solid angle $d\sigma(\omega_\mathrm{i})$ (see Figure 5.3). Let ω_t be the direction of the refracted beam (determined using Snell's law), and suppose that it occupies a solid angle $d\sigma(\omega_\mathrm{t})$. The power carried by

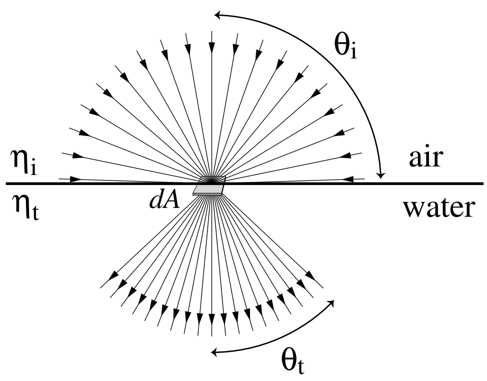


Figure 5.2: When light enters a medium with a higher refractive index, the same light energy is squeezed into a smaller volume. This causes the radiance along each ray to increase.

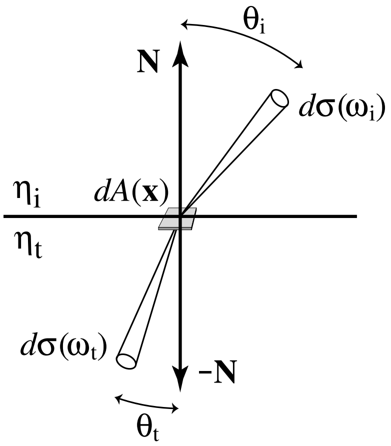


Figure 5.3: Geometry for deriving the $(\eta_{\rm t}/\eta_{\rm i})^2$ scaling of radiance due to refraction.

the incident beam is

$$d\Phi_{\mathrm{i}} \;=\; L_{\mathrm{i}}\, dA(\mathbf{x})\, d\sigma^{\perp}(\omega_{\mathrm{i}})\,,$$

where L_{i} is the radiance of the incident beam (see (3.8)), and similarly

$$d\Phi_{\mathrm{t}} \;=\; L_{\mathrm{t}}\, dA(\mathbf{x})\, d\sigma^{\perp}(\omega_{\mathrm{t}})\,.$$

Thus by conservation of energy,

$$L_{\mathrm{t}} \;=\; \frac{d\sigma^{\perp}(\omega_{\mathrm{i}})}{d\sigma^{\perp}(\omega_{\mathrm{t}})}\, L_{\mathrm{i}}\,. \tag{5.3}$$

We now turn to the angular parameterization $\omega \equiv (\theta, \phi)$, for which we have

$$d\sigma^{\perp}(\omega) = \cos\theta \sin\theta \, d\theta \, d\phi$$

(see Section 3.6.3). Using this relationship, we can relate $d\sigma^{\perp}(\omega_{\mathrm{i}})$ and $d\sigma^{\perp}(\omega_{\mathrm{t}})$ by differentiating Snell's law:

$$\begin{aligned} \eta_{\mathrm{i}} \sin\theta_{\mathrm{i}} &\;=\; \eta_{\mathrm{t}} \sin\theta_{\mathrm{t}} \qquad \text{(Snell's law)} \\ \eta_{\mathrm{i}} \cos\theta_{\mathrm{i}}\, d\theta_{\mathrm{i}} &\;=\; \eta_{\mathrm{t}} \cos\theta_{\mathrm{t}}\, d\theta_{\mathrm{t}}\,. \end{aligned}$$

Similarly, the relationship $\phi_{\mathrm{t}} = \phi_{\mathrm{i}} \pm \pi$ implies

$$d\phi_{\mathrm{i}} \;=\; d\phi_{\mathrm{t}}\,.$$

Multiplying these three equations together and using (3.16), we get

$$\eta_{\mathrm{i}}^2\, d\sigma^{\perp}(\omega_{\mathrm{i}}) \;=\; \eta_{\mathrm{t}}^2\, d\sigma^{\perp}(\omega_{\mathrm{t}})\,, \tag{5.4}$$

which together with (5.3) gives the desired relationship

$$L_{\mathrm{t}} \;=\; \frac{\eta_{\mathrm{t}}^2}{\eta_{\mathrm{i}}^2}\, L_{\mathrm{i}}\,.$$

To be precise, this equation only applies to spectral radiance that is measured with respect to frequency (L_{ν}). For spectral radiance that is measured with respect to wavelength (L_{λ}), the correct relationship is $L_{\mathrm{t}} \;=\; (\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^3 L_{\mathrm{i}}$ (as we will discuss in Section 6.2).

5.2.1.2 The BSDF for specular reflection

We will study in detail the BSDF that describes specular reflection, i.e. a perfect (two-sided) mirror. This BSDF sometimes causes confusion, because it involves Dirac distributions. We will introduce a new, simpler notation for this BSDF, using Dirac distributions defined on the unit sphere. The concepts developed here will be needed below, when we study perfect specular refraction. (Further information on the mirror BSDF can be found in Nicodemus et al. [1977], Cohen & Wallace [1993], and Glassner [1995].)

For a perfect mirror, the desired relationship between L_{i} and L_{o} is that

$$L_{\mathrm{o}}(\omega_{\mathrm{o}}) \;=\; L_{\mathrm{i}}(M_{\mathbf{N}}(\omega_{\mathrm{o}}))\,. \tag{5.5}$$

Here $M_{\mathbf{N}}(\omega_{\mathrm{o}})$ is the *mirror direction*, obtained by reflecting ω_{o} around the normal $\mathbf{N}$. (Algebraically, the mirror direction is defined by $M_{\mathbf{N}}(\omega_{\mathrm{o}}) = 2(\omega_{\mathrm{o}} \cdot \mathbf{N})\mathbf{N} - \omega_{\mathrm{o}}$.)

We would like to find a BSDF that produces the relationship (5.5) when it is inserted into the scattering equation (3.12). We will show how to define this BSDF in terms of a special *Dirac distribution* $\delta_{\sigma^\perp}$, which is defined by the property that

$$\int_{\mathcal{S}^2} f(\omega)\, \delta_{\sigma^\perp}(\omega - \omega')\, d\sigma^\perp(\omega) \;=\; f(\omega')$$

for any function f that is continuous at ω'.

Our notion of a Dirac distribution is slightly more general than the one usually encountered. Often, the Dirac distribution (or *delta function*) is understood to be a "function" $\delta(x)$ defined on $\mathbb{R}$ with the following properties:

1. $\delta(x) = 0$ for all $x \neq 0$.

2. $\int_{\mathbb{R}} \delta(x)\, dx = 1$.

These imply the more useful identity that

$$\int_{\mathbb{R}} f(x)\, \delta(x - x_0)\, dx \;=\; f(x_0)\,, \tag{5.6}$$

provided that f is continuous at x_0.

Our notation simply extends the identity (5.6) to integration on more general domains. Given a domain Ω, a measure μ on Ω, and a function $f : \Omega \to \mathbb{R}$ that is continuous at $\mathbf{x}_0$, the notation $\delta_\mu(\mathbf{x} - \mathbf{x}_0)$ refers to a "function" with the following property:

$$\int_\Omega f(\mathbf{x})\, \delta_\mu(\mathbf{x} - \mathbf{x}_0)\, d\mu(\mathbf{x}) \;=\; f(\mathbf{x}_0)\,. \tag{5.7}$$

The rigorous meaning of this notation is discussed in Appendix 5.A.

Given this background, we can write the BSDF for a perfect mirror as

$$f_\mathrm{s}(\omega_\mathrm{i} \!\rightarrow\! \omega_\mathrm{o}) \;=\; \delta_{\sigma^\perp}(\omega_\mathrm{i} - M_\mathbf{N}(\omega_\mathrm{o}))\,. \tag{5.8}$$

It is easy to see that this is a correct representation of the mirror BSDF, by inserting these definitions into the scattering equation (3.12) to obtain (5.5).

This BSDF can be written in several other equivalent ways. For example, suppose that we write the scattering equation (3.12) in the form

$$L_\mathrm{o}(\omega_\mathrm{o}) \;=\; \int_{\mathcal{S}^2} L_\mathrm{i}(\omega_\mathrm{i})\, f_\mathrm{s}(\omega_\mathrm{i} \!\rightarrow\! \omega_\mathrm{o})\, |\omega_\mathrm{i} \cdot \mathbf{N}|\, d\sigma(\omega_\mathrm{i}) \tag{5.9}$$

(by expanding the definition of the projected solid angle measure). From this equation, it is clear that the mirror BSDF could also be written as

$$f_\mathrm{s}(\omega_\mathrm{i} \!\rightarrow\! \omega_\mathrm{o}) \;=\; \frac{\delta_\sigma(\omega_\mathrm{i} - M_\mathbf{N}(\omega_\mathrm{o}))}{|\omega_\mathrm{i} \cdot \mathbf{N}|}\,. \tag{5.10}$$

Note that expressions containing Dirac distributions must be evaluated with great care. This is particularly true when the measure function associated with the Dirac distribution is different than the measure function used for integration (e.g. suppose that we are given the form (5.8) of the mirror BSDF, together with the form (5.9) of the scattering equation). In Appendix 5.A, we derive several identities that allow such expressions to be evaluated correctly and easily.

5.2.2 The BSDF for refraction

We will write the BSDF for refraction using the Dirac distribution notation developed in Section 5.2.1.2 and Appendix 5.A. It can also be written with ordinary δ-functions, using the (θ, ϕ) parameterization of BSDF's, as we will discuss in Appendix 5.C.

For a fixed point $\mathbf{x} \in \mathcal{M}$, let Ω_R be the set of directions $\omega_\mathrm{i} \in \mathcal{S}^2$ that are not subject to total internal reflection. We now define a mapping

$$R : \Omega_R \to \Omega_R\,,$$

such that $R(\omega_{\mathrm{i}})$ is the transmitted direction corresponding to the incident direction ω_{i}:[3]

$$\omega_{\mathrm{t}} \;=\; R(\omega_{\mathrm{i}})\,.$$

Note that R is easily seen to be its own inverse:

$$R(R(\omega)) \;=\; \omega \qquad \text{for all } \omega \in \Omega_R\,. \tag{5.11}$$

Given this mapping, the relationship between L_{i} and L_{t} due to refraction can be expressed as

$$L_{\mathrm{t}}(\omega_{\mathrm{t}}) \;=\; \frac{\eta_{\mathrm{t}}^2}{\eta_{\mathrm{i}}^2}\, L_{\mathrm{i}}(R(\omega_{\mathrm{t}}))\,, \tag{5.12}$$

where we have used the self-inverse property (5.11) to obtain $\omega_{\mathrm{i}} = R(\omega_{\mathrm{t}})$. The corresponding BSDF is thus

$$f_{\mathrm{s}}(\omega_{\mathrm{i}} \!\to\! \omega_{\mathrm{t}}) \;=\; \frac{\eta_{\mathrm{t}}^2}{\eta_{\mathrm{i}}^2}\, \delta_{\sigma^\perp}(\omega_{\mathrm{i}} - R(\omega_{\mathrm{t}}))\,, \tag{5.13}$$

where $\delta_{\sigma^\perp}$ is the Dirac distribution with respect to $\sigma^\perp$, as defined in Section 5.2.1.2. Inserting this BSDF into the scattering equation (3.12), it is easy to check that we get the desired relationship (5.12).

This is the simplest way to write the BSDF from a conceptual point of view; it expresses the desired relationship between ω_{i} and ω_{t}, and also the fact that radiance is scaled by a factor of $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$, with a minimum of extra clutter.

---

[3] Algebraically, R is defined by

$$R(\theta_{\mathrm{i}}, \phi_{\mathrm{i}}) \;=\; (\theta_{\mathrm{t}}, \phi_{\mathrm{t}}) \;=\; (\sin^{-1}(\frac{\eta_{\mathrm{t}}}{\eta_{\mathrm{i}}} \sin\theta_{\mathrm{i}}), \phi_{\mathrm{i}} \pm \pi)\,,$$

where θ_{t} is chosen to lie on the opposite side of the surface as θ_{i}, i.e. $\cos\theta_{\mathrm{i}}\ \cos\theta_{\mathrm{t}} \le 0$. The symbols η_{i} and η_{t} denote the refractive indices on the side of the surface containing ω_{i} and ω_{t} respectively. (Since ω_{i} can lie on either side of the surface, this means that η_{i} is actually a function of θ_{i}.)

Notice that we have used the angular parameterization $\omega \equiv (\theta, \phi)$ to define R. It is possible to define $R(\omega)$ directly in terms of the unit vector ω, but the result is relatively complicated. The vector form is commonly used in implementations, for example see [Glassner 1989, p. 298]).

5.2.3 The adjoint BSDF for refraction

We now derive the adjoint BSDF for refraction. We will make use of the following general symmetry relationship, which holds for any physically valid BSDF:[4]

$$\frac{f_{\rm s}(\omega_{\rm i} \to \omega_{\rm o})}{\eta_{\rm o}^2} = \frac{f_{\rm s}(\omega_{\rm o} \to \omega_{\rm i})}{\eta_{\rm i}^2} \,. \tag{5.14}$$

This fact will be derived in Chapter 6. For now, we can use it to immediately obtain the adjoint BSDF for refraction:

$$\begin{aligned} f_{\rm s}^*(\omega_{\rm i} \to \omega_{\rm t}) &= f_{\rm s}(\omega_{\rm t} \to \omega_{\rm i}) \\ &= (\eta_{\rm i}/\eta_{\rm t})^2 f_{\rm s}(\omega_{\rm i} \to \omega_{\rm t}) \\ &= \delta_{\sigma^\perp}(\omega_{\rm i} - R(\omega_{\rm t}))\,, \end{aligned} \tag{5.15}$$

where we have used (5.14) and (5.13) in the second and third lines respectively.

The main thing to notice is that the $(\eta_{\rm t}/\eta_{\rm i})^2$ factor is *not* present in the adjoint BSDF. Thus, importance and light particles are not scaled when they cross the interface. (Notice that this corresponds to the intuitive idea that light particles carry "power", since power (unlike radiance) is conserved when light enters a different medium.)

In an implementation, the difference between $f_{\rm s}$ and $f_{\rm s}^*$ must be represented explicitly. It is not possible to evaluate the adjoint BSDF by just exchanging the directional arguments, since there is no way to evaluate the BSDF at all. Specular BSDF's contain Dirac distributions, which means that the only allowable operation is sampling: there must be an explicit procedure that generates a sample direction and a weight. When the specular BSDF is not symmetric, the direction and/or weight computation for the adjoint is different, and thus there must be two different sampling procedures, or an explicit flag that specifies whether the direct or adjoint BSDF is being sampled.

In Appendix 5.B, we give a different derivation of the adjoint BSDF for refraction. The problem with the derivation given here is that it depends on the laws of physics, by way of the symmetry condition (5.14). Since the adjoint BSDF is a purely mathematical concept,

[4]Equation (5.14) applies to the BSDF's of a much larger class of surfaces than we consider here, including frosted glass for example. Perfect refraction corresponds to the special case of an optically smooth interface between two dielectric media.

it should be possible to derive it mathematically, and this is what is done in Appendix 5.B.

5.2.4 Results

Figure 5.4 shows a pool of water with small waves, illuminated by two area light sources. This image simulates the results of a two-pass rendering algorithm, consisting of a particle tracing pass followed by a ray tracing pass (Figure 5.1).[5] The waves (and the floor) were modeled using bump mapping.

Figure 5.4(a) shows the correct image, whose computation requires that viewing rays and particles be handled differently at the air-water interface. The radiance along viewing rays is scaled by $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$ when they cross the interface, but the particle weights are left unchanged.

Figure 5.4(b) shows the errors that occur when the BSDF at the air-water interface is assumed to be symmetric, i.e. when the same scattering rules are used for viewing rays and particles (for this image, neither one was scaled). This leads to caustics that are too bright, by a factor of $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$.

5.2.5 Discussion

Hall [1989] pointed out that radiance should be scaled by $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$ at a refractive interface, but this fact has been ignored by most ray tracing systems. Our results take this one step further, by showing that the $(\eta_{\mathrm{t}}/\eta_{\mathrm{i}})^2$ scaling should not be applied to importance or light particles. We are not aware of any system (other than ours) that implements different scattering rules for radiance vs. importance or path tracing vs. particle tracing in this way. As we have shown, this is easy to do, and essential for the correctness of bidirectional algorithms.[6]

---

[5] Both of these images were actually computed using the Metropolis light transport algorithm (Chapter 11), with modifications that simulate the results of two-pass algorithms such as [Shirley et al. 1995, Jensen 1996].

[6] The radiance scaling is not as important for ray tracing or path tracing, since when a path enters and exits a given medium, the two factors cancel out. However, the results of these algorithms will be incorrect when the sources and sensors are in different media (e.g. underwater lights). Here we have assumed that the underwater lights are modeled as direct emitters, rather than as a filament surrounded by a glass shell (since most rendering algorithms would be very inefficient if this representation were used). However, note that errors occur in both cases when bidirectional methods are used (e.g. recall the pool example, for which the source and sensor were in the same medium).

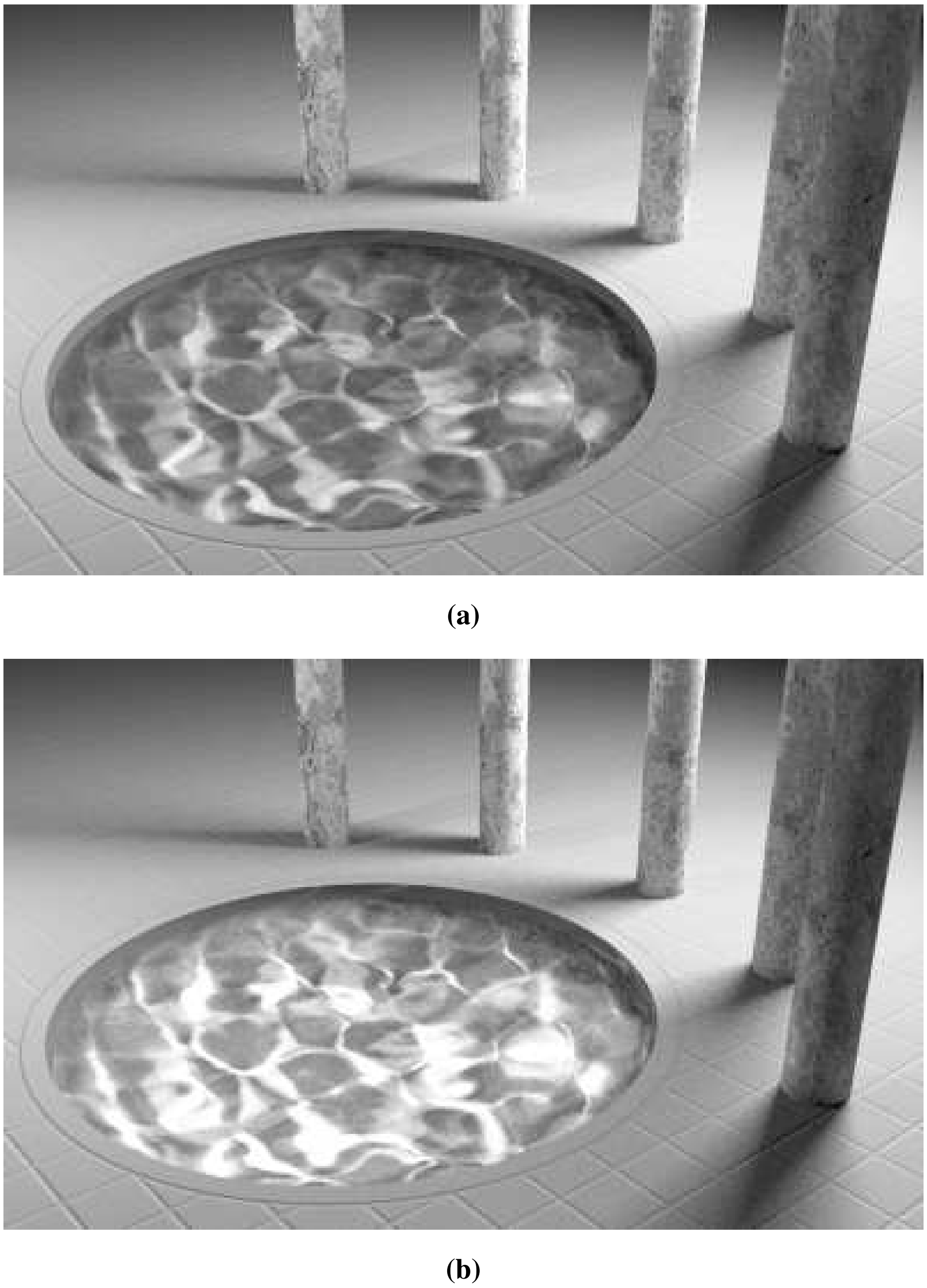


Figure 5.4: **(a)** A pool of water as it would be rendered by a particle tracing algorithm (reference image). **(b)** Incorrect caustics (too bright), caused by assuming that refraction between air and water is modeled by a symmetric BSDF.

Another approach to correctly handling refraction is to find a way to represent it by a symmetric function. This seems plausible, given the existence of the general symmetry condition (5.14). This topic is explored further in Chapter 7, where we derive a framework for light transport in which light, importance, and particles all obey the same scattering rules.

5.3 Non-symmetry due to shading normals

Shading normals are used to change the apparent orientation of a surface without changing its geometry. The mechanism is simple: when the surface is shaded (e.g. using the scattering equation (3.12)), the surface normal $\mathrm{N}(\mathbf{x})$ is replaced by a different, arbitrary direction vector. The new direction is called the *shading normal* $\mathrm{N}_{\mathrm{s}}(\mathbf{x})$, and corresponds to the desired orientation of the surface. To avoid confusion, we will refer to the true surface normal as the *geometric normal* $\mathrm{N}_{\mathrm{g}}(\mathbf{x})$.

Shading normals are useful tool for many graphics applications. For example, their original purpose was to make polygonal surfaces appear more smooth [Phong 1975]. To do this, a *vertex normal* is defined at each vertex of a polygonal mesh. Shading normals are obtained by linearly interpolating the vertex normals across each polygonal face, to give the appearance of a smoothly changing surface orientation. This simple technique is still widely in use today, because computer models of smooth surfaces are usually converted to polygons before they are rendered.

Shading normals are also used for *bump mapping* [Blinn 1978]. This is a technique for adding detail to surfaces that are otherwise smooth and uninteresting. By perturbing the surface normal, it is possible to create the impression of high geometric complexity; for example, a flat rectangle can be given the appearance of a stucco wall.

However, there is a "catch". As we will explain, shading normals modify the BSDF of the material to which they are applied. (In some sense this is obvious, since shading normals change the surface appearance, and surface appearance is completely determined by the BSDF.) Unfortunately, the modified BSDF does not possess the same properties as the original: in general it is not symmetric, and it does not conserve energy. It should not be surprising that shading normals cause problems, since they do not have any physical basis.

Nevertheless, shading normals are a useful tool for many graphics applications, and we observe that there is still a well-defined set of equations to be solved. By deriving and using the correct adjoint BSDF, we can ensure that different bidirectional light transport algorithms all converge to the same mathematically correct solution. On the other hand, we show that if the non-symmetry due to shading normals is not recognized (i.e. the adjoint BSDF is not used), then different rendering algorithms will converge to different results. This is clearly undesirable.

One way to see the problems caused by shading normals is to observe that some certain calculations still depend on the geometric normal. By changing the normal used in some calculations, but not in others, we get an inconsistent representation of the scene model. In a particle tracing simulation, for example, consider the number of particles received by a given polygon A. Clearly this depends on the geometric normal of A, rather than its shading normal. Similarly, observe that the solid angle subtended by a polygon is not affected by its shading normal. These inconsistencies between geometric and shading normals can cause problems, unless the correct adjoint BSDF is used.

5.3.1 How shading normals modify the BSDF

Let $\mathbf{x} \in \mathcal{M}$ be a fixed point, so that we can omit $\mathbf{x}$ from our notation. When shading normals are not considered, recall that the radiance leaving $\mathbf{x}$ can be evaluated using the scattering equation (5.9),

$$\begin{aligned} L_{\mathrm{o}}(\omega_{\mathrm{o}}) &= \int_{\mathcal{S}^2} L_{\mathrm{i}}(\omega_{\mathrm{i}})\, f_{\mathrm{s},\mathbf{N}_{\mathrm{g}}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}})\, d\sigma^{\perp}(\omega_{\mathrm{i}}) \\ &= \int_{\mathcal{S}^2} L_{\mathrm{i}}(\omega_{\mathrm{i}})\, f_{\mathrm{s},\mathbf{N}_{\mathrm{g}}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}})\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|\, d\sigma(\omega_{\mathrm{i}})\,. \end{aligned}$$

With shading normals, however, the following equation is used instead (see Figure 5.5):

$$L_{\mathrm{o}}(\omega_{\mathrm{o}}) = \int_{\mathcal{S}^2} L_{\mathrm{i}}(\omega_{\mathrm{i}})\, f_{\mathrm{s},\mathbf{N}_{\mathrm{s}}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}})\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{s}}|\, d\sigma(\omega_{\mathrm{i}})\,, \qquad (5.16)$$

i.e. the shading normal is used when evaluating the BSDF and the projected solid angle.

This formula is very effective at changing the apparent orientation of a surface (from $\mathbf{N}_{\mathrm{g}}$ to $\mathbf{N}_{\mathrm{s}}$). However, it does not actually change the surface geometry. Instead, the shading normal should be thought of as a parameter that modifies the BSDF. We can write an explicit

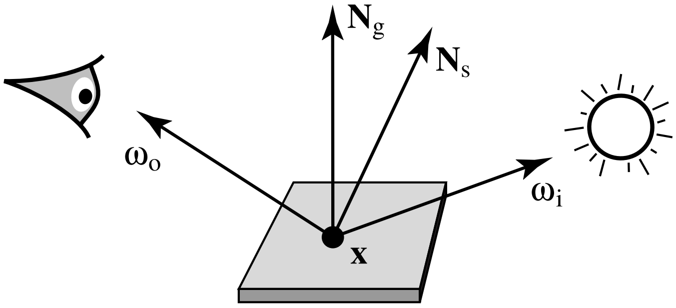


Figure 5.5: Geometry for defining the effect of shading normals.

formula for this modified BSDF, by converting the shading formula (5.16) to the standard form (3.12). In other words, we would like to write it in the form

$$L_o(\omega_o) \;=\; \int_{\mathcal{S}^2} L_i(\omega_i)\, \bar{f}_s(\omega_i \to \omega_o)\, |\omega_i \cdot \mathbf{N}_g|\, d\sigma(\omega_i)\,.$$

This can be achieved by letting $\bar{f}_s$ be the *modified BSDF* given by

$$\bar{f}_s(\omega_i \to \omega_o) \;=\; f_{s,\mathbf{N}_s}(\omega_i \to \omega_o)\, \frac{|\omega_i \cdot \mathbf{N}_s|}{|\omega_i \cdot \mathbf{N}_g|}\,. \tag{5.17}$$

It is easy to verify that the formula obtained by using this BSDF is indistinguishable from the original shading formula (5.16) (notice that the two $|\omega_i \cdot \mathrm{N}_g|$ factors cancel each other). We have simply interpreted the calculation in a different way.

5.3.2 The adjoint BSDF for shading normals

However, even if the original BSDF was symmetric, the modified BSDF is not. Its adjoint $\bar{f}_s^*$ is given by

$$\begin{aligned} \bar{f}_s^*(\omega_i \to \omega_o) \;&=\; \bar{f}_s(\omega_o \to \omega_i) \\ &=\; f_{s,\mathbf{N}_s}(\omega_o \to \omega_i)\, \frac{|\omega_o \cdot \mathbf{N}_s|}{|\omega_o \cdot \mathbf{N}_g|}\,. \end{aligned} \tag{5.18}$$

This is the BSDF that must be used for importance transport and particle tracing.

For example, the scattering equation for importance is given by

$$\begin{aligned} W_{\mathrm{o}}(\omega_{\mathrm{o}}) &= \int_{\mathcal{S}^2} W_{\mathrm{i}}(\omega_{\mathrm{i}})\, \bar{f}_{\mathrm{s}}^{*}(\omega_{\mathrm{i}} \!\to\! \omega_{\mathrm{o}})\, d\sigma^{\perp}(\omega_{\mathrm{i}}) \\ &= \int_{\mathcal{S}^2} W_{\mathrm{i}}(\omega_{\mathrm{i}})\, f_{\mathrm{s},\mathbf{N}_{\mathrm{s}}}(\omega_{\mathrm{o}} \!\to\! \omega_{\mathrm{i}})\, \frac{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{s}}|}{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}|}\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|\, d\sigma(\omega_{\mathrm{i}})\,. \end{aligned}$$

This will give results that are consistent with the formula (5.16) for radiance evaluation. Notice that the formula for evaluating importance is more complex than the formula for evaluating radiance, because there is no cancelation of the $|\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|$ factors.

Similarly, recall that the adjoint BSDF is used in particle tracing (see Sections 3.7.5 and 4.A). Given a particle that arrives from direction ω_{o} and is scattered in direction ω_{i}, the particle weight should be multiplied by

$$\begin{aligned} \alpha(\omega_{\mathrm{i}}) &= \frac{\bar{f}_{\mathrm{s}}^{*}(\omega_{\mathrm{i}} \!\to\! \omega_{\mathrm{o}})\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|}{P_{\sigma}(\omega_{\mathrm{i}})} \\ &= \frac{f_{\mathrm{s},\mathbf{N}_{\mathrm{s}}}(\omega_{\mathrm{o}} \!\to\! \omega_{\mathrm{i}})}{P_{\sigma}(\omega_{\mathrm{i}})}\, \frac{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{s}}|}{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}|}\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|\,, \end{aligned} \tag{5.19}$$

where $P_{\sigma}(\omega_{\mathrm{i}})$ is the density with respect to solid angle for sampling direction ω_{i} (see equations (4.33) and (5.18)). If particles are weighted in this way, the results will be consistent with the desired shading formula (5.16).

5.3.3 Examples of shading normal BSDF's and their adjoints

We show how these results apply to diffuse surfaces (i.e. Lambertian), and perfect specular surfaces (i.e. mirrors).

For a diffuse surface, we will show that the importance sampling techniques needed for ray tracing and particle tracing are different. We start with the constant BRDF:

$$f_{\mathrm{r}}(\omega_{\mathrm{i}} \!\to\! \omega_{\mathrm{o}}) \;=\; K_{\mathrm{d}}\,.$$

For radiance evaluation (ray tracing), we insert this in (5.16) to obtain

$$L_{\mathrm{o}}(\omega_{\mathrm{o}}) \;=\; \int_{\mathcal{S}^2} K_{\mathrm{d}}\, L_{\mathrm{i}}(\omega_{\mathrm{i}})\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{s}}|\, d\sigma(\omega_{\mathrm{i}})\,.$$

For particle tracing, according to (5.19) scattered particle weights should be multiplied by

$$\alpha(\omega_{\mathrm{i}}) \;=\; \frac{K_{\mathrm{d}}}{p_{\omega_{\mathrm{o}}}(\omega_{\mathrm{i}})} \, \frac{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{s}}|}{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}|} \, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}| \,,$$

where we have used the convention that particles go from ω_{o} to ω_{i}.

These equations imply that the importance sampling techniques needed for ray tracing and particle tracing are completely different. In both cases, the task is to sample an appropriate direction ω_{i}, when ω_{o} is already given. For ray tracing, ω_{i} should be chosen with probability proportional to $|\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{s}}|$, since this is the factor by which incoming radiance is weighted. However, for particles, ω_{i} should be chosen according to the cosine with the *geometric* normal, $|\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|$. (The weighting factors involving ω_{o} are irrelevant, since they depend only on the direction the particle arrived from.) The fact that two different density functions are needed for sampling can have important implications for rendering system design (see Section 5.3.4).

As another example, consider a perfect (two-sided) mirror. We will show that reflected particles (or importance) must be weighted by an extra factor of

$$\alpha \;=\; \frac{|\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|}{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}|}$$

to get correct results.

To show this, recall that the BSDF for a perfect mirror was derived in Section 5.2.1.2 as

$$f_{\mathrm{s},\mathbf{N}}(\omega_{\mathrm{i}} \!\rightarrow\! \omega_{\mathrm{o}}) \;=\; \frac{\delta_\sigma(\omega_{\mathrm{i}} - M_{\mathbf{N}}(\omega_{\mathrm{o}}))}{|\omega_{\mathrm{i}} \cdot \mathbf{N}|} \,.$$

To apply equation (5.19) to this BSDF, we must also know the density function $p_{\omega_{\mathrm{o}}}(\omega_{\mathrm{i}})$ that the reflected particles are sampled from. For a perfect mirror, the direction ω_{i} is chosen deterministically, as represented by the density function

$$p_{\omega_{\mathrm{o}}}(\omega_{\mathrm{i}}) \;=\; \delta_\sigma(\omega_{\mathrm{i}} - M_{\mathbf{N}}(\omega_{\mathrm{o}})) \,.$$

Plugging these into equation (5.19), and noting that $\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{s}} = \omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{s}}$, we get

$$\alpha(\omega_{\mathrm{i}}) \;=\; \frac{f_{\mathrm{s},\mathbf{N}_{\mathrm{s}}}(\omega_{\mathrm{o}} \!\rightarrow\! \omega_{\mathrm{i}})}{p_{\omega_{\mathrm{o}}}(\omega_{\mathrm{i}})} \, \frac{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{s}}|}{|\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}|} \, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}|$$

$$\begin{aligned} &= \frac{|\omega_{\rm o} \cdot \mathbf{N}_{\rm s}|}{|\omega_{\rm o} \cdot \mathbf{N}_{\rm g}|} \frac{|\omega_{\rm i} \cdot \mathbf{N}_{\rm g}|}{|\omega_{\rm i} \cdot \mathbf{N}_{\rm s}|} \\ &= \frac{|\omega_{\rm i} \cdot \mathbf{N}_{\rm g}|}{|\omega_{\rm o} \cdot \mathbf{N}_{\rm g}|} \,. \end{aligned}$$

As a final example, recall the pool test case of Section 5.2.4. The waves in this scene were modeled using bump mapping, so that the air-water interface involves both refraction and shading normals. This means that both sets of results apply: radiance is scaled by a factor of $(\eta_{\rm o}/\eta_{\rm i})^2$ when it crosses the interface, while particle weights are scaled by

$$\frac{|\omega_{\rm o} \cdot \mathbf{N}_{\rm s}|\,|\omega_{\rm i} \cdot \mathbf{N}_{\rm g}|}{|\omega_{\rm o} \cdot \mathbf{N}_{\rm g}|\,|\omega_{\rm i} \cdot \mathbf{N}_{\rm s}|} \,,$$

where as usual particles go from $\omega_{\rm o}$ to $\omega_{\rm i}$ (since this is a representation of the adjoint BSDF).

5.3.4 Pseudocode for the correct use of shading normals

In Figure 5.6, we give pseudocode to evaluate the factor

$$K(\omega_{\rm i} \to \omega_{\rm o}) \;=\; \bar{f}_{\rm s}(\omega_{\rm i} \to \omega_{\rm o})\, |\omega_{\rm i} \cdot \mathbf{N}_{\rm g}| \tag{5.20}$$

that appears in the scattering equation (5.9), and also the adjoint factor

$$K^*(\omega_{\rm i} \to \omega_{\rm o}) \;=\; \bar{f}^*_{\rm s}(\omega_{\rm i} \to \omega_{\rm o})\, |\omega_{\rm i} \cdot \mathbf{N}_{\rm g}| \tag{5.21}$$

that is used for importance evaluation and light particles. We will call these quantities the *scattering kernel* and the *adjoint scattering kernel* respectively. In Figure 5.6, the *adjoint* flag controls which of these is returned.

It may seem redundant to provide both the direct and adjoint kernels (because $\bar{f}^*_{\rm s}(\omega_{\rm i} \to \omega_{\rm o}) = \bar{f}_{\rm s}(\omega_{\rm o} \to \omega_{\rm i})$). However, this is actually quite useful. First, it allows higher-level rendering algorithms to always have the same form, whether they use radiance, importance, or particles. By supplying the appropriate *adjoint* flag, the BSDF is "transposed" appropriately for sampling or evaluation.[7] Second, it allows different density functions to be used for sampling in the direct and adjoint cases, as was shown to be necessary in Section 5.3.3. Given a

[7] In any case, the adjoints of specular BSDF's must always be represented explicitly, as mentioned in Section 5.2.5.

function EVAL-KERNEL$(\omega_\mathrm{i} \to \omega_\mathrm{o}$, *adjoint*$)$

assert $\mathrm{N_g} \cdot \mathrm{N_s} \geq 0$ (if not, flip $\mathrm{N_g}$)

if $(\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g})(\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{s}) \leq 0$ **or** $(\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g})(\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{s}) \leq 0$

then return 0

if *adjoint*

then return $f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{o} \to \omega_\mathrm{i})\, |\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{s}|\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}| \,/\, |\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}|$

else return $f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{i} \to \omega_\mathrm{o})\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{s}|$

Figure 5.6: Evaluation of the scattering kernel K when the geometric and shading normals are different. The *adjoint* flag controls whether K or K^* is returned (these are used for ray tracing and particle tracing respectively). The return value includes the factor of $|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|$ that is hidden by the projected solid angle notation.

direction ω_o, the ideal density function for ω_i is proportional to either K or K^* (depending on the *adjoint* flag). These density functions can be attached to the BSDF and returned as a single object during ray casting. Notice that for materials with symmetric BSDF's, we have $K = K^*$ and the *adjoint* flag can be ignored by the material implementation.

5.3.4.1 The prevention of "light leaks"

The pseudocode in Figure 5.6 also shows how to prevent light from "leaking" through the surface [Snyder & Barr 1987]. The problem is that an opaque surface can actually transmit light when shading normals are used. This happens when ω_i and ω_o lie geometrically on opposite sides of the surface, and yet they are on the same side of the surface according to the shading normal (see Figure 5.7(a)), so that the BSDF is evaluated as though light were "reflected" from one side of the surface to the other.

The simplest way to prevent this is to check that ω_i lies on the same side of the surface with respect to both normals, i.e. that $\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}$ and $\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{s}$ have the same sign. We also perform this test on ω_o, and if either test fails, we return a zero value for the BSDF (see Figure 5.6). This technique is effective in preventing the "light leaks" described above. However, it can also cause ordinary surfaces to appear completely black. This happens when the shading normal faces away from the viewing direction, i.e. when ω_o lies on opposite sides of the

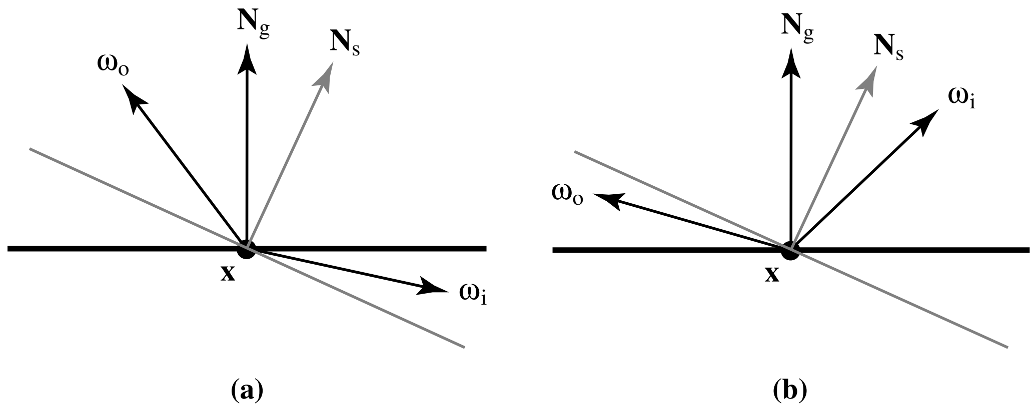


Figure 5.7: **(a)** The directions ω_i and ω_o are on opposite sides of the surface geometrically, yet they are on the same side with respect to the shading normal. Thus if we simply evaluate the BSDF using the shading normal, light can be "reflected" from one side of the surface to the other. The gives the visual impression that light is somehow leaking through the surface. **(b)** The easiest solution to this problem is check if ω_i lies on opposite sides of the surface with respect to $\mathbf{N}_\mathrm{g}$ and $\mathbf{N}_\mathrm{s}$. A similar test is performed on ω_o, and if either test fails, we return a zero value for the BSDF. However, this creates a new problem: if the test for ω_o fails (as shown in the diagram), then the BSDF is zero for all ω_i. This leads to sporadic "black patches" on the rendered surface.

surface with respect to $\mathbf{N}_\mathrm{g}$ and $\mathbf{N}_\mathrm{s}$ (see Figure 5.7(b)).

We now describe a way to solve both the light leak and black surface problems. To do this, we represent reflection and transmission by separate functions:

$$f_{\mathrm{r},\mathbf{N}} : \mathcal{S}^2 \to \mathcal{S}^2 \qquad \text{and} \qquad f_{\mathrm{t},\mathbf{N}} : \mathcal{S}^2 \to \mathcal{S}^2 \,.$$

Notice that both of these functions are defined for all $\omega_\mathrm{i}, \omega_\mathrm{o} \in \mathcal{S}^2$, i.e. they can be thought of as *extensions* of the BRDF and BTDF.

To evaluate the BSDF with shading normals, we proceed as follows (see Figure 5.8). If ω_i and ω_o lie on the same (geometric) side of the surface, then f_r is used, and otherwise f_t is used. Then, the chosen function f is evaluated with respect to the shading normal N_s (correcting for the adjoint if necessary).

With respect to this framework, f_r specifies how much light *would* be reflected between any two directions ω_i and ω_o, even if ω_i and ω_o lie on opposite sides of the surface. Similarly,

function EVAL-KERNEL-EXTENDED$(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}}, \textit{adjoint})$
  if $(\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}})(\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}) \le 0$
    then $f \leftarrow f_{\mathrm{r}}$
    else $f \leftarrow f_{\mathrm{t}}$
  if *adjoint*
    then return $f_{\mathbf{N}_{\mathrm{s}}}(\omega_{\mathrm{o}} \to \omega_{\mathrm{i}})\, |\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{s}}|\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{g}}| \,/\, |\omega_{\mathrm{o}} \cdot \mathbf{N}_{\mathrm{g}}|$
    else return $f_{\mathbf{N}_{\mathrm{s}}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}})\, |\omega_{\mathrm{i}} \cdot \mathbf{N}_{\mathrm{s}}|$

Figure 5.8: This pseudocode shows a different way to prevent light from leaking through a surface, by extending the BRDF and BTDF to be functions defined for all directions. We use the extended BRDF when ω_{i} and ω_{o} lie on the same side of the surface (with respect to the geometric normal), and otherwise we use the extended BTDF.

f_{t} is extended to describe transmission between directions on the same side of the surface. This extra information is used only when the shading and geometric normals give conflicting information.

For example, a diffuse surface would be represented by

$$f_{\mathrm{r}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}}) \;=\; K_{\mathrm{d}} \qquad \text{and} \qquad f_{\mathrm{t}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}}) \;=\; 0\,,$$

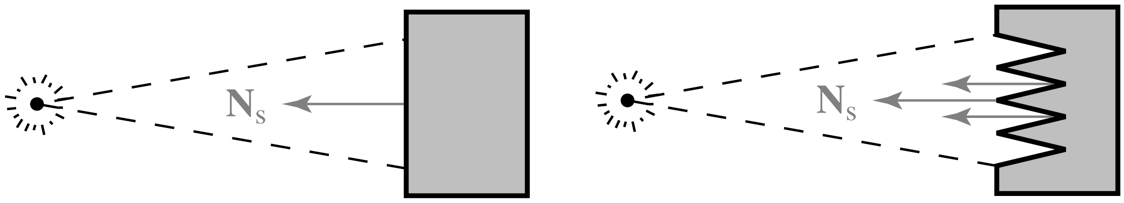
for all $\omega_{\mathrm{i}}, \omega_{\mathrm{o}} \in \mathcal{S}^2$. With these definitions, no light will leak through the surface in Figure 5.7(a), and yet the surface will not appear to be black in Figure 5.7(b).

Many other BRDF's, such as those based on microfacet theory, can naturally be extended to a function defined over all directions. Thus, this idea can be applied quite generally to solve the problem of light leaks. However, it is important to note that these are not the only artifacts associated with shading normals; see [Snyder & Barr 1987] for further examples.

5.3.5 Shading normals violate conservation of energy

In Section 6.3, we will show that any energy-conserving BSDF must satisfy

$$\int_{\mathcal{S}^2} f_{\mathrm{s}}(\omega_{\mathrm{i}} \to \omega_{\mathrm{o}})\, d\sigma^{\perp}(\omega_{\mathrm{o}}) \;\le\; 1 \qquad \text{for all } \omega_{\mathrm{i}}\,.$$



(a) **(b)**

Figure 5.9: **(a)** A flat, diffuse surface facing toward a point light source, with $\mathbf{N}_\mathrm{s} = \mathbf{N}_\mathrm{g}$. The surface is assumed to not to absorb any light, so that the incident and reflected power is the same. **(b)** A ridged surface with shading normals that point toward the light. It receives the same power as (a), but reflects far more due to its larger surface area.

When shading normals are used, this condition applies to the modified BSDF $\bar{f}_\mathrm{s}$ defined by (5.17), leading to

$$\int_{\mathcal{S}^2} f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{i} \to \omega_\mathrm{o}) \, \frac{|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{s}|}{|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|} \, d\sigma^\perp(\omega_\mathrm{o}) \;\leq\; 1 \qquad \text{for all } \omega_\mathrm{i} \,.$$

However, if $\mathbf{N}_\mathrm{g}$ and $\mathbf{N}_\mathrm{s}$ are different, then the factor $|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{s}| \,/\, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|$ can be arbitrarily large (by choosing ω_i nearly perpendicular to $\mathbf{N}_\mathrm{g}$, but not $\mathbf{N}_\mathrm{s}$). Notice that this factor can be taken outside the integral, since it does not depend on ω_o. Thus energy is not conserved (for some values of ω_i).

For intuition about this, consider Figure 5.9(a), which shows a point light source shining on a flat, perfectly reflective, diffuse surface. To determine whether energy is conserved, we compare the power received by the surface to the power that is reflected. In Figure 5.9(a), these two quantities are equal.

In Figure 5.9(b), the surface is covered with steep ridges, but the shading normals point toward the light source as though the surface were flat. This surface receives the same total power as (a), since it occupies the same solid angle with respect to the light source. It also has the same apparent brightness as (a) at every point, because it has the same shading normal. In other words, the reflected radiance at every point and in every direction is the same in both cases, so that (b) reflects the same power per unit area as (a). However, the total surface area of (b) is much larger than (a). Thus, surface (b) reflects far more power than it receives.

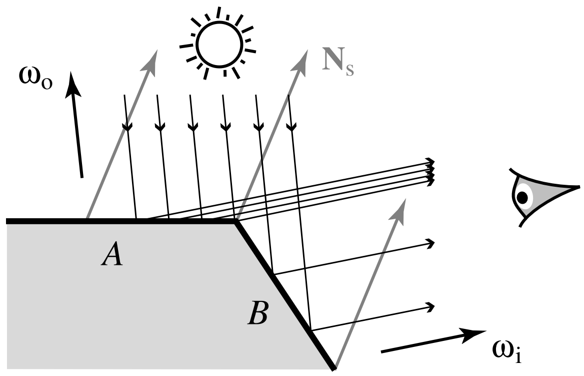


Figure 5.10: Light of uniform intensity (represented by equally spaced particles) arrives from direction ω_o at the polygonal surface shown. The shading normal $\mathbf{N}_s$ is continuous across the boundary between polygons A and B, implying that the shading should be continuous as well. However, suppose that an image is computed by estimating the apparent particle density from some other direction ω_i. This density is discontinuous at the boundary, as shown; to get continuous shading, the particles must be weighted according to equation (5.19).

This lack of energy conservation has an important consequence for particle tracing algorithms: namely, that sometimes particle weights will increase during a scattering operation. This is especially important for algorithms that use unweighted particles (e.g. density estimation [Shirley et al. 1995]), since *splitting* of particles may be required. That is, rather than multiplying the current particle's weight by α, we replace it by $\lfloor \alpha \rfloor$ new particles, plus an extra particle with probability $\alpha - \lfloor \alpha \rfloor$.

5.3.6 Shading normals can cause brightness discontinuities

It is very important to use the adjoint BSDF $\bar{f}_s^*$ in particle tracing algorithms. Otherwise, there can be noticeable artifacts in the shading of polygonal meshes.

Consider Figure 5.10. Light of uniform intensity is arriving from direction ω_o at the polygonal surface shown. The shading normal N_s is continuous (in fact, constant) across the boundary between polygons A and B, implying that the shading of the mesh should appear smooth (no matter what rendering algorithm is used).

However, suppose that a particle tracing algorithm is used, and that an image is computed directly by making a dot for each particle collision at the corresponding point in the image. (This is an example of an *image space rendering algorithm*, as discussed in Section 1.4.3.) We let ω_o denote the direction that particles arrive from, while ω_i denotes the direction toward the viewer, following our convention that ω_i is always the sampled direction in a random walk (Section 3.7.5).

We will show that the computed brightness of the polygons A and B is not the same, implying that there is a discontinuity at the boundary between A and B. To see this, observe that the brightness of the image regions corresponding to A and B is proportional to the number of dots per pixel made there. In turn, this is proportional to the apparent density of particles on A and B, measured perpendicular to the viewing direction ω_i.

To compute this apparent density, first note that the geometric normals of A and B are different, so that fewer particles per unit area are received by B than by A. It is easy to show that the particle densities on the polygons A and B are in the ratio

$$\frac{E_A}{E_B} \;=\; \frac{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}(A)|}{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}(B)|}\,.$$

(This is also the ratio of the irradiances on A and B.) From this, we can now compute the apparent particle density as seen from the viewpoint. This is simply the density of particles on the surface, divided by $|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|$. (Observe that if $|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}|$ is small, then we are looking at the surface edge-on, and thus the particles will appear much more dense.)

Putting this all together, the image brightnesses of A and B are in the ratio

$$\frac{I_A}{I_B} \;=\; \frac{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}(A)|}{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}(B)|}\,\frac{|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}(B)|}{|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}(A)|}\,, \tag{5.22}$$

and so there is a discontinuity in the image brightness at the boundary between A and B.

Our original goal was that this boundary should appear smooth (since $\mathbf{N}_\mathrm{s}$ is continuous there). We will show that if the particles are weighted according to the adjoint BSDF $\bar{f}_\mathrm{s}^*$ (as they should be), this will be achieved. Referring to (5.19), the particles striking A are weighted by a factor of

$$\alpha_A \;=\; \frac{|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}(A)|}{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}(A)|}\,,$$

where we have ignored weighting factors that are the same for particles on A and B. A similar weighting factor applies to particles striking B, and the ratio of these weights is

$$\frac{\alpha_A}{\alpha_B} = \frac{|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}(A)|}{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}(A)|} \frac{|\omega_\mathrm{o} \cdot \mathbf{N}_\mathrm{g}(B)|}{|\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}(B)|} .$$

Comparing this with (5.22), we see that the change in particle weight exactly compensates for the change in particle density, resulting in smooth shading across the boundary.

Particle tracing techniques are not often used to compute direct illumination, as we have supposed here, because other techniques are usually more efficient. However, it is quite common that particle tracing is used to render at least *some* component of the lighting on visible surfaces; for example, particle tracing is often used to render caustics. If the adjoint BSDF is not used for these particles, there will be false discontinuities in the image as we have outlined above.

5.3.7 Results

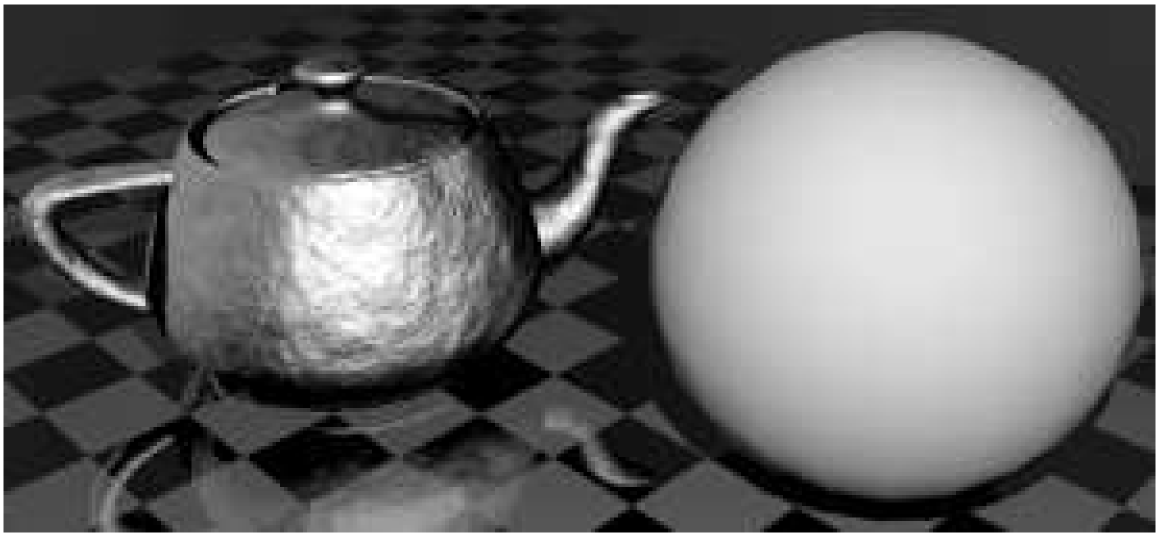
Figure 5.11 shows a bump-mapped teapot, and a polygonalized sphere with smooth shading normals. The images are simulations of a particle tracing algorithm: for each particle that strikes a surface, a dot is made at the corresponding point in the image, where the dot intensity is proportional to how much light is reflected toward the viewer. Figure 5.11(a) shows the correct result (using the adjoint BSDF), while Figure 5.11(b) shows what happens if particles are scattered just like viewing rays (i.e. if the non-symmetry caused by shading normals is not recognized). Both images use the same shading normals; the flat-shaded appearance of Figure 5.11(b) is an example of the shading artifacts described in Section 5.3.6.

5.3.8 Alternatives to shading normals

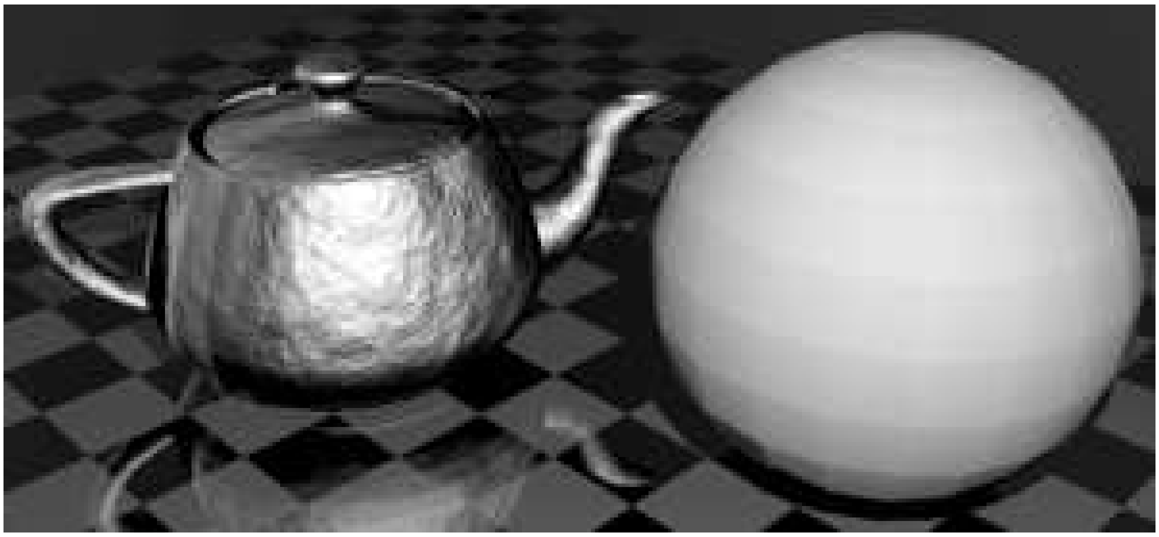
One way to avoid the problems associated with shading normals is to simply not use them. After all, they are not physically plausible. However, they are almost too useful to give up, both for approximating smooth surfaces with polygonal meshes, and for adding apparent surface detail without increasing geometric complexity.

At first, it might appear that some problems can be avoided by using the shading formula:

$$L_\mathrm{o}(\omega_\mathrm{o}) = \int_{\mathcal{S}^2} L_\mathrm{i}(\omega_\mathrm{i}) \, f_{\mathrm{s},\mathbf{N}_\mathrm{s}}(\omega_\mathrm{i} \to \omega_\mathrm{o}) \, |\omega_\mathrm{i} \cdot \mathbf{N}_\mathrm{g}| \, d\sigma(\omega_\mathrm{i}) , \tag{5.23}$$



(a)



(b)

Figure 5.11: (a) Bump mapping and Phong interpolation, reference image. Shows direct lighting as it would be computed by particle tracing. **(b)** The same model, with errors caused by assuming that shading normals do not affect the symmetry of BSDF's.

where we have used $|\omega_{\rm i} \cdot \mathbf{N}_{\rm g}|$ instead of $|\omega_{\rm i} \cdot \mathbf{N}_{\rm s}|$. This method preserves the symmetry of $f_{\rm s}$ (assuming that it was symmetric in the first place), and will often conserve energy (but not always). However, the results obtained with this formula are not very useful. For example, consider a diffuse surface. The proposed shading formula (5.23) computes the *same surface appearance* for all values of $\mathrm{N}_{\rm s}$, because the BRDF $f_{\rm r}$ is a constant. Thus, it is impossible to make a polygonal surface look smooth, or make a flat surface look bumpy with this formula.

Similarly, consider a perfect mirror. Using (5.23) and the representation (5.10) of the mirror BSDF, we see that the radiance reflected by the mirror would be

$$L_{\rm o}(\omega_{\rm o}) \;=\; \frac{|M_{\mathbf{N}_{\rm s}}(\omega_{\rm o}) \cdot \mathbf{N}_{\rm g}|}{|\omega_{\rm o} \cdot \mathbf{N}_{\rm s}|}\, L_{\rm i}(M_{\mathbf{N}_{\rm s}}(\omega_{\rm o}))\,.$$

The weighting factor in this equation causes the reflectivity of the mirror to change with $\mathrm{N}_{\rm s}$, varying in the range $0 < \rho < 2$.[8] Again, this formula does not achieve what we would expect with shading normals, since it changes the reflectivity of the surface as well as the direction of reflection.

Other BSDF's produce similarly strange effects when used with (5.23), but they do not create the appearance of a changing surface orientation (as shading normals do). Thus the usefulness of (5.23) is quite limited. It seems far better to just use traditional shading normals, and accept the fact that their use corresponds to a non-symmetric BSDF.

Another possibility is to look for new BSDF models that serve the same purpose as shading normals, and yet are symmetric and energy-conserving. This is an interesting area for future research. Perhaps it could be accomplished with a microfacet shading model [Torrance & Sparrow 1967, Glassner 1995], where the distribution of microfacets is not symmetric about the surface normal.

However, it is important to realize that this kind of approach will never replace shading normals. One of the big advantages of shading normals is that they can be applied to *any* BSDF, while a microfacet approach would obviously be limited to a particular scattering model. Second, shading normals are designed to be as effective as possible at changing the apparent surface orientation. The results achieved using any symmetric, energy-conserving

[8] Because it is possible that $\rho > 1$, this is an example of a BSDF which was originally energy conserving, but not when formula (5.23) is used.

approach will necessarily be less convincing. Note that it is impossible to duplicate the surface appearance achieved by shading normals, since if two shading formulas *always* produce the same surface appearance, then they are represented by the same BSDF.

Appendix 5.A Dirac distributions for general measures

Our goal in this section is to give a rigorous definition of the notation $\delta_\mu(\mathbf{x} - \mathbf{x}_0)$ that was used to define the mirror BSDF (Section 5.2.1.2). This concept of a Dirac distribution with respect to a general measure is apparently new (although it seems quite basic), and we have found it to be a useful tool for many problems in graphics. We show how these distributions fit into the standard framework, and we also derive several identities that give them a consistent meaning as they are manipulated during calculations.

5.A.1 Linear functionals and distributions

Although the notation $\delta(x)$ looks like a function, it is properly called a *generalized function* or *distribution*. A rigorous theory of distributions was first developed by Laurent Schwartz in the late 1940's and early 1950's [Schwartz 1966]. However, physicists had been using similar ideas well before that; for example, Dirac introduced his famous "delta function" in 1925 [Lützen 1982].

To define distributions rigorously would take us too far afield, but we will at least summarize the basic concepts. Further information can be found in [Rudin 1973, Al-Gwaiz 1992].

First, the notation

$$\int_{\mathbb{R}} f(x)\, \delta(x - x_0)\, dx \tag{5.24}$$

should be thought of as purely symbolic (there is nothing being integrated in the traditional sense). Rather, this notation defines a mapping that takes a continuous function f, and yields a real number $f(x_0)$. The mapping is called a *linear functional*, and we will denote it by Λ_{x_0}.

Formally, a linear functional is a linear operator $\Lambda : X \to \mathcal{F}$ from a vector space X onto its scalar field $\mathcal{F}$ [Taylor & Lay 1980, p. 31]. In the example (5.24), the vector space is the set $C(\mathbb{R})$ of all continuous functions $f : \mathbb{R} \to \mathbb{R}$, with the usual operations of addition and scalar multiplication, and the scalars are simply the real numbers ($\mathcal{F} = \mathbb{R}$). The functional $\Lambda_{x_0} : C(\mathbb{R}) \to \mathbb{R}$ is defined by

$$\Lambda_{x_0}(f) \;=\; f(x_0)\,.$$

Thus, the notation (5.24) is just a long way of writing $\Lambda_{x_0}(f)$.

The mapping Λ_{x_0} is actually a special kind of functional called a *distribution*. Distributions have many desirable properties: they are infinitely differentiable, they obey the usual formal rules of calculus, they are equipped with many convergence theorems, and furthermore every continuous function is a distribution [Rudin 1973, p. 135]. To achieve these wonderful properties, however certain technical restrictions must be made (see Schwartz [1966] and Rudin [1973, p. 137,141]).

5.A.2 General Dirac distributions

Returning to our notation for general Dirac distributions, the expression

$$\int_\Omega f(\mathbf{x})\,\delta_\mu(\mathbf{x}-\mathbf{x}_0)\,d\mu(\mathbf{x}) \tag{5.25}$$

represents a distribution $\Lambda_{\mathbf{x}_0}$ acting on a function $f : \Omega \to \mathbb{R}$, where $\Lambda_{\mathbf{x}_0}$ is defined by[9]

$$\Lambda_{\mathbf{x}_0}(f) \;=\; f(\mathbf{x}_0)\,. \tag{5.26}$$

At this point, the notation $\delta_\mu(\mathbf{x}-\mathbf{x}_0)$ may seem rather odd, because the measure μ does not appear anywhere in the definition of $\Lambda_{\mathbf{x}_0}$. That is, given a different measure μ' on the domain Ω, the expression

$$\int_\Omega f(\mathbf{x})\,\delta_{\mu'}(\mathbf{x}-\mathbf{x}_0)\,d\mu'(\mathbf{x})$$

denotes exactly the same distribution $\Lambda_{\mathbf{x}_0}$ that we defined above. However, the point is that we have defined the meaning of the notation

$$\int_\Omega f(\mathbf{x})\,\delta_\mu(\mathbf{x}-\mathbf{x}_0)\,d\mu'(\mathbf{x}) \tag{5.27}$$

only when the measures μ and μ' are the same. The subscript on δ_μ is a reminder of this, since it is possible to get meaningless results if an expression such as (5.27) is evaluated carelessly. For example, if we take definition (5.8) of the mirror BSDF, and substitute it in the expanded version (5.9) of the scattering equation, we obtain

$$L_\mathrm{o}(\omega_\mathrm{o}) \;=\; \int_{\mathcal{S}^2} L_\mathrm{i}(\omega_\mathrm{i})\,\delta_{\sigma^\perp}(\omega_\mathrm{i} - M_\mathbf{N}(\omega_\mathrm{o}))\,|\omega_\mathrm{i}\cdot\mathbf{N}|\,d\sigma(\omega_\mathrm{i})\,. \tag{5.28}$$

If it were not for the subscript on δ, we might apply the identity (5.7) to obtain

$$L_\mathrm{o}(\omega_\mathrm{o}) \;=\; L_\mathrm{i}(M_\mathbf{N}(\omega_\mathrm{o}))\,|\omega_\mathrm{o}\cdot\mathbf{N}|\,,$$

which is incorrect.

[9] The fact that $\Lambda_{\mathbf{x}_0}$ is a distribution, and not merely a functional, is because the formula $\Lambda_\nu(f) = \int_\Omega f\,d\nu$ defines a distribution whenever ν is a σ-finite positive measure [Rudin 1973, p. 143]. In our case, the measure ν is defined by $\nu(D) = 1$ if $\mathbf{x}_0 \in D$, and $\nu(D) = 0$ otherwise.

5.A.3 Identities for evaluating general Dirac distributions

For future reference, we give several identities that are useful for evaluating expressions containing general Dirac distributions. We can summarize these as follows:

(D1) $\delta_\mu(\mathbf{x}_0 - \mathbf{x}) \;=\; \delta_\mu(\mathbf{x} - \mathbf{x}_0)$

(D2) $\delta_\mu(\mathbf{x} - \mathbf{x}_0) \;=\; \dfrac{d\mu'}{d\mu}(\mathbf{x}_0)\, \delta_{\mu'}(\mathbf{x} - \mathbf{x}_0)$

(D3) $\delta_\mu(\beta(\mathbf{x}) - \beta(\mathbf{x}_0)) \;=\; \delta_{\mu\circ\beta}(\mathbf{x} - \mathbf{x}_0)$

For property (D3), β denotes a bijective function $\beta : \Omega \to \Omega$, and $\mu \circ \beta$ is the *composition measure* defined by

$$(\mu \circ \beta)(D) \;=\; \mu(\beta(D))\,, \tag{5.29}$$

where $\beta(D) = \{\beta(\mathbf{x}) \mid \mathbf{x} \in D\}$. Property (D3) may look more familiar when it is specialized to the case of ordinary Dirac distributions on the real line, yielding

$$\delta(f(x) - f(x_0)) \;=\; \frac{1}{|f'(x_0)|}\, \delta(x - x_0)\,, \tag{5.30}$$

where f is a bijective function that is differentiable at x_0.

Note that all three of these properties are actually *definitions*, whose purpose is to extend the notation (5.25) in a consistent way. The definitions are designed to be compatible with the usual rules of calculus, so that we may formally apply them as though Dirac distributions were ordinary functions.

Property (D1). This defines the meaning of the notation $\delta_\mu(\mathbf{x}_0 - \mathbf{x})$. Note that the expressions $\mathbf{x} - \mathbf{x}_0$ and $\mathbf{x}_0 - \mathbf{x}$ are purely symbolic; they do not imply that subtraction is defined on the domain Ω.

Property (D2). This definition gives a consistent meaning to expressions of the form

$$\int_\Omega f(\mathbf{x})\, \delta_\mu(\mathbf{x} - \mathbf{x}_0)\, d\mu'(\mathbf{x})\,, \tag{5.31}$$

where the measures μ and μ' are different. One way to evaluate an expression of this kind is to change the integration measure, a concept similar to a change of variables. To do this, we require that μ and μ' are continuous with respect to each other, i.e. they have the same sets of measure zero. This guarantees the existence of the Radon-Nikodym derivative $d\mu/d\mu'$ (see Theorem 3.2), and allows us to switch from one integration measure to the other using

$$\int_\Omega f(\mathbf{x})\, d\mu(\mathbf{x}) \;=\; \int_\Omega f(\mathbf{x})\, \frac{d\mu}{d\mu'}(\mathbf{x})\, d\mu'(\mathbf{x}) \tag{5.32}$$

[Rudin 1987, p. 23]. For example, we could use this relationship to evaluate (5.28) correctly, by first substituting

$$|\omega_{\mathrm{i}} \cdot \mathbf{N}| \;=\; \frac{d\sigma^\perp(\omega_{\mathrm{i}})}{d\sigma(\omega_{\mathrm{i}})}\,,$$

and then applying (5.32) and (5.7) to get the right answer.

Definition (D2) allows us to evaluate (5.31) in another way, by changing the *distribution* rather than the integration measure. To obtain this identity, we rewrite (5.31) as

$$\begin{aligned}\int_\Omega f(\mathbf{x})\, \delta_\mu(\mathbf{x} - \mathbf{x}_0)\, d\mu'(\mathbf{x}) \;&=\; \int_\Omega f(\mathbf{x})\, \delta_\mu(\mathbf{x} - \mathbf{x}_0) \frac{d\mu'}{d\mu}(\mathbf{x})\, d\mu(\mathbf{x}) \\ &=\; \frac{d\mu'}{d\mu}(\mathbf{x}_0) \int_\Omega f(\mathbf{x})\, \delta_\mu(\mathbf{x} - \mathbf{x}_0)\, d\mu(\mathbf{x}) \\ &=\; \frac{d\mu'}{d\mu}(\mathbf{x}_0)\, f(\mathbf{x}_0)\,.\end{aligned} \tag{5.33}$$

We now simply observe that if the substitution (D2) is made on the left-hand side of (5.33), the same result is obtained. Thus, the definition (D2) is consistent. (Note that we have not previously defined the meaning of expressions such as (5.31).)

Property (D3). The definition gives a consistent meaning to the notation

$$\int_\Omega f(\mathbf{x})\, \delta_\mu(\beta(\mathbf{x}) - \beta(\mathbf{x}_0))\, d\mu(\mathbf{x})\,, \tag{5.34}$$

where β be a bijective function $\beta : \Omega \to \Omega$. Our goal is to define this in such a way that δ_μ can be treated as an ordinary function.

To do this, we make the definition

$$\delta_\mu(\beta(\mathbf{x}) - \beta(\mathbf{x}_0))\, \frac{d(\mu \circ \beta)}{d\mu}(\mathbf{x}_0) \;=\; \delta_\mu(\mathbf{x} - \mathbf{x}_0)\,, \tag{5.35}$$

where $\mu \circ \beta$ is the composition measure (5.29). We require that $\mu \circ \beta$ is continuous with respect to μ, so that the Radon-Nikodym derivative exists in (5.35). Notice that property (D3) can be obtained from (5.35) by applying property (D2).

To show that definition (5.35) is consistent, we evaluate

$$
\begin{aligned}
&\int_\Omega f(\mathbf{x})\,\delta_\mu(\beta(\mathbf{x})-\beta(\mathbf{x}_0))\,\frac{d(\mu\circ\beta)}{d\mu}(\mathbf{x}_0)\,d\mu(\mathbf{x}) \\
&\qquad = \int_\Omega f(\mathbf{x})\,\delta_\mu(\beta(\mathbf{x})-\beta(\mathbf{x}_0))\,\frac{d\mu(\beta(\mathbf{x}))}{d\mu(\mathbf{x})}\,d\mu(\mathbf{x}) \\
&\qquad = \int_\Omega f(\mathbf{x})\,\delta_\mu(\beta(\mathbf{x})-\beta(\mathbf{x}_0))\,d\mu(\beta(\mathbf{x})) \\
&\qquad = \int_\Omega f(\beta^{-1}(\mathbf{x}'))\,\delta_\mu(\mathbf{x}'-\beta(\mathbf{x}_0))\,d\mu(\mathbf{x}') \\
&\qquad = f(\beta^{-1}(\beta(\mathbf{x}_0))) \\
&\qquad = f(\mathbf{x}_0) \\
&\qquad = \int_\Omega f(\mathbf{x})\,\delta_\mu(\mathbf{x}-\mathbf{x}_0)\,d\mu(\mathbf{x})\,,
\end{aligned}
$$

where we have defined $\mathbf{x}'=\beta(\mathbf{x})$ and used the fact the β is a bijection. Comparing the first and last lines, we get (5.35).

Another form of this identity is sometimes useful, when we are given an expression of the form $\delta_\mu(\beta(\mathbf{x})-\mathbf{x}_0')$. By letting $\mathbf{x}_0=\beta^{-1}(\mathbf{x}_0')$ in (5.35), we obtain

$$
\delta_\mu(\beta(\mathbf{x})-\mathbf{x}_0') \;=\; \left[\frac{d(\mu\circ\beta)}{d\mu}(\beta^{-1}(\mathbf{x}_0'))\right]^{-1}\,\delta_\mu(\mathbf{x}-\beta^{-1}(\mathbf{x}_0'))\,. \tag{5.36}
$$

Appendix 5.B Derivation of the adjoint BSDF for refraction

We derive the adjoint BSDF for refraction (5.15). Unlike the derivation in Section 5.2.3, this one does not depend on physical laws.

Recall from Section 5.2.3 that refraction is described by the following BSDF:

$$f_{\rm s}(\omega_{\rm i} \to \omega_{\rm t}) \;=\; \frac{\eta_{\rm t}^2}{\eta_{\rm i}^2}\, \delta_{\sigma^\perp}(\omega_{\rm i} - R(\omega_{\rm t}))\,.$$

From the definition (3.21) of the adjoint BSDF, we can immediately write it as

$$\begin{aligned} f_{\rm s}^*(\omega_{\rm i} \to \omega_{\rm t}) &= f_{\rm s}(\omega_{\rm t} \to \omega_{\rm i}) \\ &= \frac{\eta_{\rm i}^2}{\eta_{\rm t}^2}\, \delta_{\sigma^\perp}(\omega_{\rm t} - R(\omega_{\rm i}))\,, \end{aligned}$$

where $\eta_{\rm i}$ and $\eta_{\rm t}$ have been exchanged because they are functions of $\omega_{\rm i}$ and $\omega_{\rm t}$ respectively.

This is a valid expression for the adjoint BSDF, but it is certainly not obvious that it is equivalent to the expression (5.15) given in Section 5.2.3. To show that it is, we first observe that although R is a bijection, it does not preserve the measure $\sigma^\perp$, since from (5.4) we have

$$\frac{d\sigma^\perp(R(\omega_{\rm i}))}{d\sigma^\perp(\omega_{\rm i})} \;=\; \frac{d\sigma^\perp(\omega_{\rm t})}{d\sigma^\perp(\omega_{\rm i})} \;=\; \frac{\eta_{\rm i}^2}{\eta_{\rm t}^2}\,.$$

Thus, we can apply the identity (5.36) to get

$$\begin{aligned} f_{\rm s}^*(\omega_{\rm i} \to \omega_{\rm t}) &= \frac{\eta_{\rm i}^2}{\eta_{\rm t}^2}\, \delta_{\sigma^\perp}(R(\omega_{\rm i}) - \omega_{\rm t}) \\ &= \frac{\eta_{\rm i}^2}{\eta_{\rm t}^2} \left[\frac{d(\sigma^\perp \circ R)}{d\sigma^\perp}(R^{-1}(\omega_{\rm t})) \right]^{-1} \delta_{\sigma^\perp}(\omega_{\rm i} - R^{-1}(\omega_{\rm t})) \\ &= \frac{\eta_{\rm i}^2}{\eta_{\rm t}^2} \left[\frac{d\sigma^\perp(\omega_{\rm t})}{d\sigma^\perp(\omega_{\rm i})} \right]^{-1} \delta_{\sigma^\perp}(\omega_{\rm i} - R(\omega_{\rm t})) \\ &= \delta_{\sigma^\perp}(\omega_{\rm i} - R(\omega_{\rm t}))\,, \end{aligned}$$

which agrees with the expression (5.15) that we obtained before.

Appendix 5.C Angular forms of perfect specular BSDF's

Although we prefer to use general Dirac distributions to represent specular BSDF's, it may be helpful to see how reflection and refraction can be represented using ordinary δ-functions. We will use the identities from Section 3.6.3 and Appendix 5.A.

5.C.1 The BSDF for a mirror

Starting with mirrors, recall that the desired relationship between L_i and L_o is

$$L_\mathrm{o}(\theta_\mathrm{o}, \phi_\mathrm{o}) \;=\; L_\mathrm{i}(\theta_\mathrm{o}, \phi_\mathrm{o} \pm \pi)\,.$$

The corresponding BSDF is thus

$$f_\mathrm{s}(\theta_\mathrm{i}, \phi_\mathrm{i}, \theta_\mathrm{o}, \phi_\mathrm{o}) \;=\; \frac{\delta(\theta_\mathrm{i} - \theta_\mathrm{o})\,\delta(\phi_\mathrm{i} - (\phi_\mathrm{o} \pm \pi))}{|\cos\theta_\mathrm{i}|\,\sin\theta_\mathrm{i}}\,.$$

To verify this, simply substitute f_s in the scattering equation (3.17).

By using different expressions (3.16) for the projected solid angle, we can also write the mirror BSDF as

$$\begin{aligned} f_\mathrm{r}(\theta_\mathrm{i}, \phi_\mathrm{i}, \theta_\mathrm{o}, \phi_\mathrm{o}) \;&=\; \frac{\delta(\cos\theta_\mathrm{i} - \cos\theta_\mathrm{o})\,\delta(\phi_\mathrm{i} - (\phi_\mathrm{o} \pm \pi))}{\cos\theta_\mathrm{i}} \\ &=\; 2\,\delta(\sin^2\theta_\mathrm{i} - \sin^2\theta_\mathrm{o})\,\delta(\phi_\mathrm{i} - (\phi_\mathrm{o} \pm \pi))\,, \end{aligned} \tag{5.37}$$

where the last expression is valid only for one-sided mirrors (since there are two solutions for θ_i in the range $0 \le \theta_\mathrm{i} \le \pi$). These forms of the mirror BRDF were given in [Nicodemus et al. 1977, p. 44] and [Cohen & Wallace 1993, p. 31].

5.C.2 The BSDF for refraction

The BSDF for refraction can also be written in terms of the angles (θ, ϕ). This form is given by

$$f_\mathrm{s}(\theta_\mathrm{i}, \phi_\mathrm{i}, \theta_\mathrm{t}, \phi_\mathrm{t}) \;=\; \frac{\eta_\mathrm{t}^2}{\eta_\mathrm{i}^2}\,2\,\delta(\sin^2\theta_\mathrm{i} - \frac{\eta_\mathrm{t}^2}{\eta_\mathrm{i}^2}\sin^2\theta_\mathrm{t})\,\delta(\phi_\mathrm{i} - (\phi_\mathrm{t} \pm \pi))\,. \tag{5.38}$$

(compare with (5.8)). Strictly speaking, this only represents the BTDF for light flowing in one direction (rather than the full BSDF), since equation (5.38) has two solutions for θ_i in the range $0 \le \theta_\mathrm{i} \le \pi$.

The adjoint BSDF for refraction is given by

$$
\begin{aligned}
f_{\mathrm{s}}^{*}(\theta_{\mathrm{i}}, \phi_{\mathrm{i}}, \theta_{\mathrm{t}}, \phi_{\mathrm{t}}) &= \frac{\eta_{\mathrm{i}}^{2}}{\eta_{\mathrm{t}}^{2}}\, 2\, \delta(\sin^{2}\theta_{\mathrm{t}} - \frac{\eta_{\mathrm{i}}^{2}}{\eta_{\mathrm{t}}^{2}} \sin^{2}\theta_{\mathrm{i}})\, \delta(\phi_{\mathrm{t}} - (\phi_{\mathrm{i}} \pm \pi)) \\
&= 2\, \delta(\sin^{2}\theta_{\mathrm{i}} - \frac{\eta_{\mathrm{t}}^{2}}{\eta_{\mathrm{i}}^{2}} \sin^{2}\theta_{\mathrm{t}})\, \delta(\phi_{\mathrm{i}} - (\phi_{\mathrm{t}} \pm \pi))\,,
\end{aligned}
$$

where we have used (5.30).