Monte Carlo Notes

The following are definitions and derivations that are explicitly written down with the hope that they will help guide the ray tracer implementation.

1 Monte Carlo Estimator

We intend to probabilistically approximate the integral $\int_{D^*} f$ for some integrable function $f : D \to R$, where $D^* \subseteq D$. The Monte Carlo estimator will suffice: given n iid samples $\boldsymbol{X}_i \in D^*$ s.t. $\boldsymbol{X}_i \sim p \implies p(D \setminus D^*) = 0$ (with the usual restriction that $\int_D p = 1$), we define our estimator

$$\boldsymbol{M}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{X}_i)}{p(\boldsymbol{X}_i)}$$

$$\implies \begin{cases} \mathrm{E}[\boldsymbol{M}_n] & = \int_D p(\boldsymbol{x}) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{x})}{p(\boldsymbol{x})} \, d\boldsymbol{x} \\ & = \frac{1}{n} \sum_{i=1}^{n} \int_{D^*} \frac{f(\boldsymbol{x})}{p(\boldsymbol{x})} p(\boldsymbol{x}) \, d\boldsymbol{x} \\ & = \int_{D^*} f = \mu \\ \mathrm{Var}[\boldsymbol{M}_n] & = \mathrm{Var}[\frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{X}_i)}{p(\boldsymbol{X}_i)}] \\ & = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[f(\boldsymbol{X}_i)/p(\boldsymbol{X}_i)] \\ & = \frac{1}{n} \mathrm{Var}[f(\boldsymbol{X}_i)/p(\boldsymbol{X}_i)] \end{cases}$$

Thus $\lim_{n\to\infty} \mathrm{Var}[\boldsymbol{M}_n] = 0$. The definition of variance suggests that increasing the number of samples reduces squared error: $\mathrm{Var}[\boldsymbol{M}_n] = \mathrm{E}[(\boldsymbol{M}_n - \mu)^2]$, which in turn suggests that the estimator converges to the desired integral (which could perhaps be rationalized as a consequence of the law of large numbers).

2 Improving Estimator Efficiency

2.1 Importance Sampling

Suppose we pick p s.t. $p = kf$, where f is the estimated function from before. Then $\int_D p = 1 \implies k = 1/\int_{D^*} f$, in which case the estimator term $\frac{f(\boldsymbol{X}_i)}{p(\boldsymbol{X}_i)} = \int_{D^*} f = \mu$ already. Then $\mathrm{Var}[\boldsymbol{M}_n] = 0$ immediately. While this ideal p defeats the purpose of the Monte Carlo estimator, it intuitively follows that picking p that roughly conforms to the "shape" of f will decrease estimator variance. In practice, this means making p large when the contribution from f is large and vice-versa for when the contribution from f is relatively small.

2.2 Multiple Importance Sampling (MIS)

It may be desireable to utilize multiple densities p_i when estimating the rendering equation. Veach et al (1997) offers the multi-sample Monte Carlo estimator:

$$\boldsymbol{M}_n^* = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n} w_i(\boldsymbol{X}_{i,j}) \frac{f(\boldsymbol{X}_{i,j})}{p_i(\boldsymbol{X}_{i,j})}$$

given a set of densities $\{p_1, \ldots, p_n\}$ and n_i samples drawn for each p_i and $\boldsymbol{X}_{i,j} \sim p_i$.

We expect that the bias, $\beta(\boldsymbol{M}) = \mathrm{E}[\boldsymbol{M}] - \int_{D^*} f$ is still zero so long as we impose the conditions that **(W1)** $\sum_{i=1}^{n} w_i(\boldsymbol{x}) = 1$ when $f(\boldsymbol{x}) \neq 0$ and **(W2)** $w_i(\boldsymbol{x}) = 0$ when $p_i(\boldsymbol{x}) = 0$:

The multi-sample estimator is unbiased: $\beta(\boldsymbol{M}_n^*) = 0$. Each random sample $\boldsymbol{X}_{i,j}$ is not necessarily identically-distributed but they are nevertheless independent, so we can manipulate the expectation accordingly, assuming each $n_i \geq 1$:

$$\begin{aligned}
\mathrm{E}[\boldsymbol{M}_n^*] &= \int_D \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n} w_i(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{p_i(\boldsymbol{x})} \cdot p_i(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{D^*} \sum_{i=1}^{n} w_i(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{D^*} f, \text{ by } \mathbf{(W1)}
\end{aligned}$$

QED

Veach et al offers the power heuristic as a "good" weighting function: $w_i(\boldsymbol{x}) = \frac{[n_i p_i(\boldsymbol{x})]^\gamma}{\sum_k [n_k p_k(\boldsymbol{x})]^\gamma}$, where $\gamma = 1$ produces the simpler balance heuristic ($\gamma = 2$ is often sufficient). And it is clear that the power heuristic meets both weight function criteria.

2.3 Russian Roulette

Russian roulette offers a way to terminate paths while maintaining an unbiased estimate. After picking an arbitrary termination probability $q \in [0, 1]$ (usually increasing as the integrand becomes smaller), we define a new discrete estimator $\boldsymbol{R} \in \{\frac{1}{1-q}\boldsymbol{M}_n^*, \vec{0}\}$ s.t. $P(\boldsymbol{R} = \frac{1}{1-q}\boldsymbol{M}_n^*) = 1-q$ and $P(\boldsymbol{R} = \vec{0}) = q$. Then

$$\mathrm{E}[\boldsymbol{R}] = (1-q) \cdot \frac{1}{1-q} \mathrm{E}[\boldsymbol{M}_n^*] + q \cdot \vec{0} = \mathrm{E}[\boldsymbol{M}_n^*]$$

3 Light Transport

3.1 Rendering Equation

Radiance is flux per unit projected area per unit solid angle (watts/(steradian·m^2)), which is what we seek to measure. The rendering equation describes outgoing radiance from a point $\boldsymbol{x}$ in a direction $\boldsymbol{\Theta}$:

$$L(\boldsymbol{x} \to \boldsymbol{\Theta}) = L_e(\boldsymbol{x} \to \boldsymbol{\Theta}) + \int_{\Omega_{\boldsymbol{x}}} f_r(\boldsymbol{x}, \boldsymbol{\Psi} \to \boldsymbol{\Theta}) L(\boldsymbol{x} \leftarrow \boldsymbol{\Psi}) |\boldsymbol{N}_{\boldsymbol{x}} \cdot \boldsymbol{\Psi}|\, d\boldsymbol{\omega}_{\boldsymbol{\Psi}}$$

Given incoming direction(s) $\boldsymbol{\Psi}$, BRDF f_r, incoming radiance L, emitted radiance L_e, and surface normal $\boldsymbol{N}_{\boldsymbol{x}}$. Alternatively, the area formulation of the rendering equation states that

$$L(\boldsymbol{x} \to \boldsymbol{\Theta}) = L_e(\boldsymbol{x} \to \boldsymbol{\Theta}) + \int_A f_r(\boldsymbol{x}, \boldsymbol{\Psi} \to \boldsymbol{\Theta}) L(\boldsymbol{y} \to -\boldsymbol{\Psi}) V(\boldsymbol{x}, \boldsymbol{y}) \frac{|\boldsymbol{N}_{\boldsymbol{x}} \cdot \boldsymbol{\Psi}||\boldsymbol{N}_{\boldsymbol{y}} \cdot -\boldsymbol{\Psi}|}{r_{\boldsymbol{xy}}^2}\, dA_{\boldsymbol{y}}$$

because incoming radiance is equivalent to ougoing radiance from every other point $\boldsymbol{y}$ in the scene, with the visibility term V to account for obstructions. The area formulation enables importance sampling of light sources in a scene, so it is useful to use it for "direct" illumination and the preceding hemispherical formulation for "indirect" illumination:

$$L_r(\boldsymbol{x} \to \boldsymbol{\Theta}) = \int_A f_r(\boldsymbol{x}, \boldsymbol{\Theta} \to \boldsymbol{\Psi}) L_e(y \to -\boldsymbol{\Psi}) V(\boldsymbol{x}, \boldsymbol{y}) G(\boldsymbol{x}, \boldsymbol{y})\, dA_{\boldsymbol{y}} +$$

$$\int_{\omega_{\boldsymbol{x}}} f_r(\boldsymbol{x}, \boldsymbol{\Theta} \to \boldsymbol{\Psi}) L_i(x \leftarrow \boldsymbol{\Psi}) |\boldsymbol{N_x} \cdot \boldsymbol{\Psi}| \, d\omega_{\boldsymbol{\Psi}}$$

where L_i is reflected radiance from the incoming direction $\boldsymbol{\Psi}$.

4 BxDF Models

4.1 Lambertian BRDF

The Lambertian diffuse model requires just a reflectance spectrum $\boldsymbol{C}$ and is weighted by a normalization factor $1/\pi$: $f_r = \frac{\boldsymbol{C}}{\pi}$. It suffices to sample the cosine term of the rendering equation.

4.2 Specular BxDF

Specular reflection and transmission is similarly simple, operating under the assumption that $\theta_i = \theta_o$ for reflections. Transmissions abide by Snell's law, which states that given indices of refraction η_i and η_t, we have $\eta_i \sin\theta_i = \eta_t \sin\theta_t$.

The Fresnel equations describe the proportion of reflected and transmitted light at a surface and depend on the index of refraction and incident angle. Dielectrics have real-valued IORs while conductors have complex IORs. This property of dielectrics yields the following reflectance formulae for parallel polarized light and perpendicular polarized light, respectively:

$$\begin{aligned} r_{\parallel} &= \frac{\eta_t \cos\theta_i - \eta_i \cos\theta_t}{\eta_t \cos\theta_i + \eta_i \cos\theta_t} \\ r_{\perp} &= \frac{\eta_i \cos\theta_i - \eta_t \cos\theta_t}{\eta_i \cos\theta_i + \eta_t \cos\theta_t} \end{aligned}$$

For unpolarized light, the reflectance is $F_r = \left[r_{\parallel}^2 + r_{\perp}^2 \right] / 2$.

Total internal reflection (TIR) occurs when light travels into a medium with a lower index of refraction. None of the light at grazing angles will pass into the next medium, the maximum angle at which this will occur being the critical angle. The definition of sine implies TIR may be detected when the computed $\sin\theta_t \geq 1$.

For a complex IOR $\eta + ik$ we have

$$\begin{aligned} r_{\perp} &= \frac{a^2 + b^2 - 2a\cos\theta + \cos^2\theta}{a^2 + b^2 + 2a\cos\theta + \cos^2\theta} \\ r_{\parallel} &= r_{\perp} \frac{\cos^2\theta(a^2+b^2) - 2a\cos\theta\sin^2\theta + \sin^4\theta}{\cos^2\theta(a^2+b^2) + 2a\cos\theta\sin^2\theta + \sin^4\theta} \\ a^2 + b^2 &= \sqrt{(\eta^2 - k^2 - \sin^2\theta)^2 + 4\eta^2 k^2} \end{aligned}$$

where $\eta + ik = \frac{\eta_t + k_t}{\eta_i + k_i}$. Thus the specular BRDF is $f_r(\omega_o, \omega_i) = \delta(\omega_i - \omega_r) \frac{F_r(\omega_i)}{|\cos\theta_r|}$. Conversely, the specular BTDF is $f_r(\omega_o, \omega_i) = \frac{\eta_o^2}{\eta_i^2}(1 - F_r(\omega_i)) \frac{\delta(\omega_i - T(\omega_o, \boldsymbol{n}))}{|\cos\theta_i|}$. Where T denotes the specular transmission vector.

4.3 Microfacet BRDF

The microfacet BRDF model assumes that any surface consists of perfectly specular "microfacets" of differential area. Let dA denote a differential patch of macrosurface area. Let $\omega_m(p)$

denote the microfacet normal at p. We therefore expect the projection of the microsurface to cover the macrosurface: $\int_{dA}(\omega_m(p)\cdot\boldsymbol{n})\,dp = \int_{dA}\,dp$

Since we cannot expect to have an explicit ω_m, it is productive to model the distribution of microfacets via a distribution function $D(\omega_m)$ that yields the relative differential area of microfacets with normal ω_m, where surface normals exist in TBN space. The following normalization constraint follows from this concept: $\int_\Omega D(\omega_m)(\omega_m\cdot\boldsymbol{n})\,d\omega_m = 1$. The anisotrophic GGX (Trowbridge-Reitz) distribution is

$$D(\omega_m) = \frac{1}{\pi\alpha_x\alpha_y\cos^4\theta_m\left(1+\tan^2\theta_m\left(\frac{\cos^2\phi_m}{\alpha_x^2}+\frac{\sin^2\phi_m}{\alpha_y^2}\right)\right)^2}$$

When $\alpha_x = \alpha_y$, the distribution becomes isotrophic:

$$D^*(\omega_m) = \frac{1}{\pi\alpha^2\cos^4\theta_m\left(1+\frac{1}{\alpha^2}\tan^2\theta_m\right)^2}$$

The masking function $G_1(\omega,\omega_m)$ accounts for occlusion of microfacets with normal ω_m by other microfacets from a viewing angle ω. This ensures energy conservation:

$$\int_\Omega D(\omega_m)G_1(\omega,\omega_m)\max(0,\omega\cdot\omega_m)\,d\omega_m = \omega\cdot\boldsymbol{n} = \cos\theta$$

Smith's approximation operates under the assumption heights and normals of points on the surface are independently distributed (intuitively, we wouldn't expect this to be very representative of reality). Then

$$G_1(\omega) = \frac{\cos\theta}{\int_\Omega D(\omega_m)\max(0,\omega\cdot\omega_m)\,d\omega_m}$$

which has an analytic solution for the GGX microfacet distribution: $G_1(\omega) = \frac{1}{1+\Lambda(\omega)}$, where

$$\Lambda(\omega) = \frac{\sqrt{1+\alpha^2\tan^2\theta}-1}{2}$$

But since the BSDF accepts two directional arguments, occlusion may occur in two directions, resulting in masking *and* shadowing. The solution is a bidirectional generalization of the Smith approximation: $G(\omega_o,\omega_i) = \frac{1}{1+\Lambda(\omega_o)+\Lambda(\omega_i)}$.

The Torrance-Sparrow microfacet BRDF is based on the previous two models:

$$f_r(\omega_o,\omega_i) = \frac{D(\omega_h)G(\omega_o,\omega_i)F_r(\omega_o)}{4\cos\theta_0\cos\theta_i}$$

where ω_h is the half-angle vector