Mathematical Basics of Monte Carlo Rendering Algorithms

Mihai Calin Ghete*
Vienna University of Technology

Abstract

The high computational costs associated with global illumination in the field of computer graphics call for effective ways of handling complicated integrals and sums. This paper offers an overview of Monte Carlo methods, stochastic methods of approximating such constructs. Starting with the necessary mathematical basis for understanding Monte Carlo, the paper continues with a survey of sampling methods and variance reduction techniques, to conclude in an introduction to the use of Monte Carlo in global illumination.

Keywords: Monte Carlo,probability theory,sampling,variance reduction,global illumination

1 Introduction

1.1 What is Monte Carlo?

Monte Carlo integration is based on the principle that a definite integral of a certain function can be approximated using the value of that function at several randomly chosen (or *sampled*) positions from within the integration domain (see Figure 1.1). Since Monte Carlo can yield arbitrarily accurate results depending on the number of samples used, it has gained much importance in the field of computer graphics, where finding solutions to certain problems analytically is nearly impossible due to the associated computational costs.

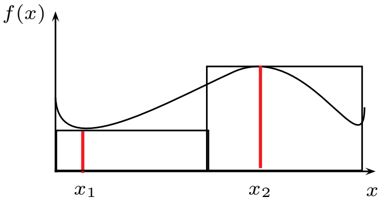

Figure 1: The principle of Monte Carlo integration. Samples x_1 and x_2 (marked in red) are chosen randomly from the domain of the integral, each sample representing half of the integration domain. The area within the two boxes is the estimate of the integral.

Descriptively, Monte Carlo integration is similar to integration by (deterministic) numerical quadrature, where samples are not chosen randomly but distributed evenly across the integration domain. Figure 1.1 shows the difference between these two numeric integration methods. Note that the "clumping" of samples in the figure results in a poor approximation of the integral; this is one of the problems that can be encountered when using Monte Carlo techniques (unless appropriate measures are taken to avoid this issue). In higher dimensions numerical quadrature becomes far less effective than Monte Carlo as the actual number of samples required grows exponentially.

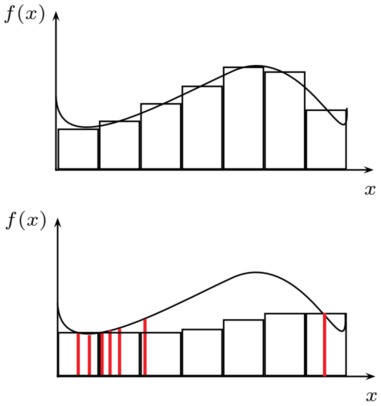

Figure 2: Comparison of deterministic numerical quadrature (top) and integration by Monte Carlo quadrature (bottom). Red lines represent randomly chosen samples.

1.2 History of Monte Carlo

The term Monte Carlo was first coined in the 1940s by scientists working on the development of nuclear weapons in Los Alamos. The term referred to games of chance whose behavior and outcome could be used to study some interesting phenomena. Since then, Monte Carlo methods have found widespread use in many fields such as operations research, radiation transport and especially statistical physics and chemistry [Kalos and Whitlock 1986].

Of course, the use of "randomness" in experiments and calculations dates much further back than the 1940s. For example, Kalos and Whitlock [1986] mention the probably earliest use of random numbers to approximate an integral, which dates back to 1777 and is attributed to Comte de Buffon.

1.3 Usage in Computer Graphics

The most prominent use of Monte Carlo in the field of computer graphics is in global illumination. Depending on the chosen approach, several highly complex, multi-dimensional integrals have to be evaluated in order to confer realistic lighting to a scene.

Basically, all approaches to the global illumination problem attempt to solve a recursive equation called the *rendering equation* or a simplification thereof. This was first pointed out by Kajiya [Kajiya 1986], who describes the rendering equation as:

$$I(x,x') = g(x,x')\left[\varepsilon(x,x') + \int_S \rho(x,x',x'')I(x',x'')dx''\right]$$

*e-mail: mihai.calin.ghete@student.tuwien.ac.at

According to the equation, the intensity of light $I(x,x')$ that passes from a point x' to another point x, is related to:

- A "geometry" term $g(x,x')$, which equals zero if the two points are not mutually visible and which otherwise encodes the distance between the two points.
- The light emitted towards x from a surface at x', denoted as $\varepsilon(x,x')$.
- The light that reaches x' from all surface points x'', modified by a scattering term $\rho(x,x',x'')$.

The rendering equation thus posesses several quantities which are hard to compute; the dependence of $I(x,x')$ on all other $I(x',x'')$ poses a particular problem. For this reason, many simplifications of the rendering equation have been considered in conjunction with algorithms that attempt to solve them; even so, the quantities involved often remain so complicated that Monte Carlo methods have to be used to approximate them. Beside what is mentioned in the rendering equation, there are also several other complex quantities, such as importance and flux, that some algorithms use; in this paper, these quantities are only briefly explained when necessary. A thorough overview of the quantities mentioned and their properties can be found in literature [Dutré et al. 2003].

2 Foundation of Monte Carlo

In order to achieve a better understanding of what Monte Carlo is, an overview of its mathematical foundation is given. The foundation lies in probability theory; this section will describe, among other things, the concepts of random events, random variables, expectation, variance, estimators, bias and consistency, concluding in an examination of the Monte Carlo estimator. A more in-depth review of probability theory with respect to Monte Carlo methods is given by Kalos and Whitlock [Kalos and Whitlock 1986].

2.1 Random Events

As explained by Kalos and Whitlock [1986], a Monte Carlo calculation is basically a sequence of random events. Random events – which can be, for example, the result of flipping a coin or throwing a dice – have a set of possible outcomes. Associated with each possible outcome E_k is a number between 0 and 1 called the *probability of* E_k – or p_k – which is a measure of how likely it is for E_k to occur. If E_k never occurs, then $p_k = 0$ and if E_k is bound to occur, then $p_k = 1$.

When the occurence of an event E_i implies that another event E_j does not occur and vice versa, these two events are called *mutually exclusive*. The probability of these two events occuring together is 0, and the probability that either one of these events will occur is the sum of the two events' probabilities:

$$P\{E_i \text{ and } E_j\} = 0$$
$$P\{E_i \text{ or } E_j\} = p_i + p_j$$

Furthermore, considering a class of events which are all mutually exclusive to each other, and where all possible events are enumerated, these events' probabilities sum up to one:

$$\sum_i p_i = 1$$

2.2 Composite Events

Composite events can be created from two or more elementary events. Considering a composite event consisting of only two elementary events E and F, the probability of a specific outcome (E_i, F_j) is called the *joint probability* for E_i and F_j. One can say that E_i and F_j are *independent* if:

$$p_{ij} = p_{1i}p_{2j}$$

If E_i and F_j are not independent, the joint probability p_{ij} can be written as:

$$p_{ij} = \left(\sum_k p_{ik}\right)\left(\frac{p_{ij}}{\sum_k p_{ik}}\right) \qquad (1)$$

Each of the two events has a *marginal probability* that can be extracted from the formulation of joint probability above. The marginal probability for event E_i can be described as the probability that E_i will occur, regardless of the outcome of F:

$$p(i) = \sum_k p_{ik}$$

The other part of Equation 1 above is called the *conditional probability* $p(j|i)$. This is the probability of event F_j occuring, considering that event E_i has occured:

$$p(j|i) = \frac{p_{ij}}{\sum_k p_{ik}}$$

Clearly, Equation 1 can be written so that $p(j)$ and $p(i|j)$ are generated, the results are analogous.

Furthermore, these definitions can be generalized in order to apply to more than two events.

2.3 Random Variables

Random variables are numerical values that can be attributed to the outcomes of events.

2.3.1 Expectation

The stochastic mean, or *expectation* of a random variable, is defined as:

$$E(x) = \sum_i p_i x_i$$

When applying a function g to a random variable x, the result $g(x)$ is a random variable as well. The expectation of this new random variable is defined as:

$$E(g(x)) = \sum_i p_i g(x_i)$$

The following properties can be shown to hold:

- The expectation of a constant is the constant itself.
- For any constants λ_1, λ_2 and two functions g_1, g_2:

$$E(\lambda_1 g_1(x) + \lambda_2 g_2(x)) = \lambda_1 E(g_1(x)) + \lambda_2 E(g_2(x)).$$

Independent random variables have the property that the expectation of their product is the product of their expectations. This is implied by the definition of independence:

$$\begin{aligned} E(xy) &= \sum_{i,j} x_i y_j p_{ij} = \sum_{i,j} x_i y_j p_{1i} p_{2j} \\ &= \sum_i x_i p_{1i} \sum_j y_i p_{2j} \\ &= E(x)E(y) \end{aligned} \tag{2}$$

This property becomes important when calculating the variance of independent variables.

2.3.2 Variance

Variance is another important characteristic of a random variable; it is defined as:

$$var\{x\} = E(x - E(x))^2$$

and can be analogously written as:

$$var\{x\} = E(x^2) - E(x)^2$$

The square root of the variance, called the *standard deviation*, is an often used measure for the dispersion of the random variable.

The variance of a function $g(x)$ can be written as:

$$var\{g(x)\} = E(g(x)^2) - E(g(x))^2$$

Due to the linearity of the expectation operator, the variance of a linear combination of two functions $g_1(x)$ and $g_2(x)$ becomes [Kalos and Whitlock 1986]:

$$\begin{aligned} var\{\lambda_1 g_1(x) + \lambda_2 g_2(x)\} =& \lambda_1^2 var\{g_1(x)\} + \lambda_2^2 var\{g_2(x)\} \\ & + 2[\lambda_1 \lambda_2 E(g_1(x) g_2(x)) \\ & - \lambda_1 \lambda_2 E(g_1(x)) E(g_2(x))] \end{aligned}$$

Considering Equation 2, we see that when $g_1(x)$ and $g_2(x)$ are independent, the last term of the sum becomes zero. *Covariance* is derived from this term and measures the degree of independence between two random variables[1]:

$$cov\{x,y\} = E(xy) - E(x)E(y)$$

In case covariance is zero, we are left with a shorter formula for the variance of a linear combination of two functions:

$$var\{\lambda_1 g_1(x) + \lambda_2 g_2(x)\} = \lambda_1^2 var\{g_1(x)\} + \lambda_2^2 var\{g_2(x)\} \tag{3}$$

2.3.3 Continuous Random Variables

Random variables do not have to be confined to a set of discrete values, they can also be continuous. In this case however, expectation is defined with the aid of a *probability density function* (PDF)[2]. For a continuous real-valued random variable x, the probability that the variable will have the value x is given by the value of $p(x)dx$, where $p(x)$ is the probability density function [Dutré et al. 2003]. The PDF is similar to discrete probability, as the (definite) integral of the PDF over all possible values of x is 1:

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

The *cumulative distribution function* (CDF) is related to the PDF:

$$F(x) = \int_{-\infty}^{x} p(y)dy$$

This function yields the probability that the random variable will be lower or equal to x; the CDF has the important property that it is a nondecreasing function with values between 0 and 1.

The expectation of x is thus:

$$E(x) = \int_{-\infty}^{\infty} xp(x)dx = \int_{-\infty}^{\infty} xdF(x)$$

and has the same properties as noted in the discrete case. The expectation of a function $g(x)$ is, analogously:

$$E(g(x)) = \int_{-\infty}^{\infty} g(x)p(x)dx \tag{4}$$

The variance of x is defined and can be calculated just as in the discrete case.

2.3.4 The Uniform Probability Distribution

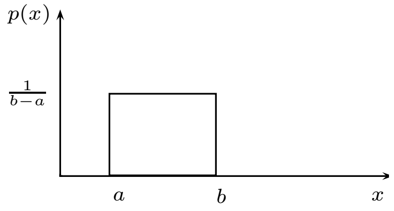

Figure 3: The uniform probability distribution.

The uniform probability distribution (Figure 2.3.4) is quite common in probability theory and will serve as an example of a continuous distribution. The PDF of this distribution is constant within a specific interval $[a,b]$ and zero everywhere else. Since the integral over the whole domain of the PDF has to be 1, we find that:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

For convenience, we will analyze the uniform distribution in the interval $[0,1]$, for which $p(x)$ is 1 within the given domain $[0,1]$ and 0 outside of it. More on the uniform distribution can be found in literature [Kalos and Whitlock 1986; Viertl 2003]. In this case, the CDF is:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \le x \le 1 \\ 1 & \text{otherwise} \end{cases}$$

The expectation is, intuitively, 0.5:

$$E(x) = \int_0^1 xdx = \frac{1}{2}(1-0) = 0.5$$

[1] Zero covariance does not imply independence, for the details see Kalos and Whitlock [1986].

[2] The reason for this lies in the fact that if there are infinitely many possible outcomes of a stochastic event, each possible outcome will have a probability of zero [Viertl 2003].

The variance can be calculated easily as well:

$$\begin{aligned} var\{x\} &= E(x^2) - E(x)^2 = \int_0^1 x^2 dx - \frac{1}{4} \\ &= \frac{1}{3} - \frac{1}{4} \\ &= \frac{1}{12} \end{aligned}$$

2.4 Estimators

2.4.1 A Monte Carlo Estimator

Examining Equation 4 above, we see that we can express the integral of a function $f(x)$ as the expectation of a function $g(x) = \frac{f(x)}{p(x)}$, where $p(x)$ is an arbitrary PDF which is nonzero within the integration range:

$$g(x) = \frac{f(x)}{p(x)} \Rightarrow E(g(x)) = \int_{-\infty}^{\infty} \frac{f(x)}{p(x)} p(x) dx = \int_{-\infty}^{\infty} f(x) dx \qquad (5)$$

This method of describing integrals is crucial to Monte Carlo integration because expectation can be estimated by the **arithmetic mean** of several random variables drawn from $g(x)$, as follows:

If random variables $x_1, x_2, \ldots, x_N$ are drawn independently from $p(x)$, we can build the sum:

$$G = \frac{1}{N} \sum_{i=1}^{N} g(x_i) \qquad (6)$$

which is also a random variable. We can now calculate the expectation of G, which is:

$$\begin{aligned} E(G) &= E\left(\frac{1}{N}\sum_{i=1}^{N} g(x_i)\right) = \frac{1}{N}\sum_{i=1}^{N} E(g(x_i)) \\ &= \frac{1}{N}\sum_{i=1}^{N} E(g(x)) \\ &= E(g(x)) \end{aligned}$$

Thus, the **expectation** of G is equal to the expectation of a single random variable $g(x)$. This fact alone is not satisfactory enough, as an estimator has to be a useful approximation of a given quantity [Kalos and Whitlock 1986]. Thus, we also have to observe the variance of G.

2.4.2 Variance of the Estimator

Using the formula for the variance of a sum of independent random variables illustrated in Equation 3, we obtain:

$$\begin{aligned} var\{G\} &= var\left\{\frac{1}{N}\sum_{i=1}^{N} g(x_i)\right\} = \sum_{i=1}^{N} \frac{1}{N^2} g(x_i) \\ &= \frac{1}{N} var\{g(x)\} \qquad (7) \end{aligned}$$

Assuming the number of samples drawn from $g(x)$ is one, the variance is equal to the actual variance of $g(x)$. By increasing the number of samples however, variance diminishes; for an infinite number of samples, the variance becomes zero and with it the probability that G equals $E(g(x))$ becomes one. Thus, G can be considered an estimator for the expectation of $g(x)$.

The actual error of the Monte Carlo estimator still has to be determined; two theorems can be used for this purpose. The first is *Chebyshev's inequality* [1986; 2003], which states that the probability that the estimator deviates from its expected value by more than $\sqrt{var\{G\}/\delta}$ is smaller than δ, where δ is an arbitrary positive number:

$$P\left\{|G - E(G)| \geq \sqrt{\frac{var\{G\}}{\delta}}\right\} \leq \delta$$

By substituting the variance of the Monte Carlo estimator, we obtain:

$$P\left\{(G - E(G))^2 \geq \frac{1}{N\delta} var\{g\}\right\} \leq \delta$$

meaning that by knowing the variance $var\{g\}$ and fixating a certain probability δ, we can always calculate a number of samples N such that the probability that our estimator deviates from the expected value is lower than δ.

A second theorem for determining the error of Monte Carlo calculations is the *central limit theorem* of probability. The central limit theorem states that the Monte Carlo estimator will asymptotically (for $N \to \infty$) have a normal distribution [1986; 2003]. When N is large enough, the standard deviation will vary as much as $1/\sqrt{N}$. This is the normally encountered error in Monte Carlo calculations.

2.5 Bias and Consistency

Clearly, not all estimators are necessarily equally good. Some may converge slower toward the desired value, some faster; likewise, the expectation of some estimators may not always be equal to the desired value, either.

Assuming that the quantity Q is to be approximated by an estimator θ, the *bias* of θ is defined as $E(\theta) - Q$. If the bias is zero, θ is called an *unbiased estimator* for Q. While it is more desirable to have an unbiased estimator, one should also consider the variance of the estimator as a factor of interest. A biased estimator with low variance is usually preferred to an unbiased one with high variance.

Most of the time, an unbiased estimator cannot be found, so a weaker constraint is imposed: An estimator θ is *consistent* if θ converges to Q with probability 1 as the number of samples N approaches infinity [1986], or:

$$P\{\lim_{N\to\infty} \theta(x_1, x_2, \ldots, x_N) = Q\} = 1$$

2.6 Summary

As we have seen, an estimator for the expectation of a function $g(x) = \frac{f(x)}{p(x)}$ is:

$$G_N = \frac{1}{N}\sum_{i=1}^{N} g(x_i)$$

Using this estimator, we can approximate the definite integral of $f(x)$:

$$E(G_N) = E(g(x)) = \int \frac{f(x)}{p(x)} p(x) dx = \int f(x) dx$$

The expectation of G_N has been shown to equal the expectation of $g(x)$, regardless of the number of samples N or of the sampling distribution $p(x)$; G_N is therefore **unbiased** (and implicitly consistent).

The variance of G_N decreases with the number of samples, meaning that the probability of finding a G_N which is a fixed distance away from $E(G_N)$ becomes smaller, reaching zero as N approaches infinity.

As a side note, Monte Carlo can also be used to approximate sums instead of integrals, especially when the number of summands is very high [1986]; the principle remains the same, with the major difference being that values are sampled from a discrete distribution instead of a continous one.

3 Sampling and Variance Reduction

Although Monte Carlo can calculate an asymptotically correct result, the method used often converges badly, meaning that variance is quite high given a low number of samples. For this reason, variance reduction methods have been developed, which aim to counter this problem. Variance is reduced either by introducing bias or by incorporating additional information about the integral.

Also, some calculations require a special integration domain or – more generally – values sampled from a certain distribution. Unfortunately, one usually only has access to random (or pseudo-random) variables that follow a single distribution, most commonly a uniform distribution between 0 and 1. This calls for ways to transform one distribution into another without introducing bias.

Finally, some algorithms require that the termination criteria be random, this criteria also has to be chosen wisely with respect to the entire calculation as to not introduce bias.

3.1 ICDF Sampling

ICDF sampling [Kalos and Whitlock 1986; Dutré et al. 2003] is an analytical method of transforming a sample taken from a uniform distribution over the interval $[0,1)$ into one that follows a given distribution $p(x)$. This is done by applying the *inverse cumulative distribution function* of $p(x)$ to the uniformly generated sample. Assuming $F(x)$ is the CDF based on $p(x)$ as defined in Section 2.3.3, it is easy to calculate the inverse CDF by, for example, isolating x from $F(x) = y$, knowing that $x = F^{-1}(y)$.

Let u be the uniformly generated sample from $[0,1)$ and $y = F^{-1}(u)$. In order to validate this approach, we need to prove that p really is the PDF of y. We will do so by showing that:

$$P\{y \leq Y\} = \int_{-\infty}^{Y} p(x)dx = F(Y)$$

First of all, two things have to be noted. One of them is the value of the CDF for a uniform distribution in $[0,1)$, as explained in Section 2.3.4, which is:

$$P\{u \leq X\} = X$$

The other is the fact that the CDF is a monotonically nondecreasing function by definition and thus that applying it to both sides of an inequality of the form $a \leq b$ yields $F(a) \leq F(b)$. We proceed by applying $F(x)$ to both sides of the inequality contained in $P\{y \leq Y\}$. As u is uniformly distributed, the result proves our assumption:

$$P\{F^{-1}(u) \leq Y\} = P\{u \leq F(Y)\} = F(Y)$$

It is thus possible to use samples taken from a uniform distribution, which is commonly available through a computer function, to gain samples distributed according to any desired PDF. The downside of this approach is that the CDF needs to be computable at all and invertable analytically. Rejection sampling techniques, as will be discussed in succession, do not have this downside.

3.1.1 Sampling the Cosine Lobe

Many global illumination algorithms have to solve an integral equation that includes a cosine term; in these cases, sampling according to the cosine term over the hemisphere can make Monte Carlo integration easier [Dutré et al. 2003].

Sampling can be done using the ICDF technique; we define the PDF according to which we want to generate samples as:

$$p(\theta,\phi) = \frac{\cos\theta}{\pi}$$

The CDF can be computed as:

$$F(\theta,\phi) = \frac{\phi}{2\pi}(1-\cos^2\theta)$$

Since the CDF is a product of a function of ϕ and a function of θ, we can use two independent random variables ξ_1 and ξ_2 uniformly distributed in the domain $[0,1)$ for each of the two values:

$$F_\phi = \frac{\phi}{2\pi} \Rightarrow \phi_i = 2\pi\xi_1$$
$$F_\theta = 1-\cos^2\theta \Rightarrow \theta_i = \cos^{-1}\sqrt{\xi_2}$$

In the second equation, $1-\xi$ was replaced by ξ_2 for simplicity (since $1-\xi$ is a uniform random variable in $[0,1)$, ξ will be uniform and in the same domain). ϕ_i and θ_i are now distributed according to the cosine PDF.

3.2 Rejection Sampling

Rejection sampling is another general method of sampling an arbitrary probability distribution [Kalos and Whitlock 1986; Dutré et al. 2003]. A sample is first proposed, then tested for acceptance. If the sample does not pass the acceptance test, it is said to be rejected, meaning that it is simply discarded and the process has to be repeated for another sample. Looking at the procedure, one can see that its most obvious disadvantage is the low efficiency that may arise from rejecting a lot of samples. The main difficulty thus lies in finding adequately efficient rejection tests for certain distributions.

We will now have a closer look at how samples can be proposed and tested; a straightforward method is to sample uniformly from the box that encloses the entire PDF [2003] (see Figure 3.2).

For a one-dimensional PDF $p(x)$, which is to be sampled over the domain $[a,b]$, the value of the PDF can lie between 0 and a maximum M. Thus, sampling will be done uniformly from a two-dimensional region $[a,b] \times [0,M]$, producing a sample (x,y). This sample will be accepted if $y \leq p(x)$, or rejected otherwise. The distribution of the accepted samples thus follows $p(x)$ (intuitively, the higher the value of $p(x)$ at a certain position, the more samples taken at this position will be accepted).

In practice, sampling from the two-dimensional uniform distribution can also be done by sampling x from the given region and y from $[0,1)$ as opposed to $[0,M)$. The test is then adjusted to reflect this change, by accepting the samples only when $y \leq \frac{p(x)}{M}$.

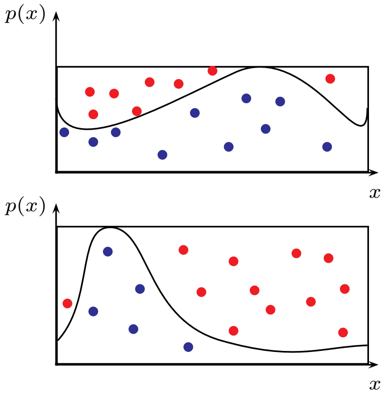

Figure 4: Rejection sampling. Red samples are rejected, blue samples (below the function curve). Note that when sampling uniformly and the maximum of a function is high in relation to the rest of the function (bottom), many samples are rejected.

As the efficiency of this technique is proportional to the probability that a sample will be accepted, we can see that the higher the area of the PDF within the bounds of the sampling rectangle, the higher the efficiency of the technique will be (see Figure 3.2, down).

A more general approach to rejection techniques is discussed by Kalos and Whitlock [Kalos and Whitlock 1986]. Here a random variable Z with a PDF $g(Z)$ is considered. The random variable is accepted if a uniformly distributed sample ξ_2 is lower than the value of a function of Z, $h(Z)$: $\xi_2 \leq h(Z) < 1$. By looking at the test as a composite event, the probability of success, that is, the probability that a sample will be accepted is found to be:

$$P\{\text{success}\} = \varepsilon = \int_{-\infty}^{\infty} h(z)g(z)dz$$

Here, ε is also called the *efficiency* of the method. Furthermore, the probability distribution of Zs that results from the rejection technique is:

$$f(z) = \frac{h(z)g(z)}{\int_{-\infty}^{\infty} h(t)g(t)dt}$$

The number of trials until a sample is accepted follows a geometrical distribution [Viertl 2003; Kalos and Whitlock 1986], the expected number of trials is thus $\frac{1}{\varepsilon}$. All these quantities offer quite an effective framework for analyzing rejection techniques.

3.3 Importance Sampling

The methods described until here primarily dealt with transformations of random variables from one PDF into another, not considering variance. The following methods attempt to lower the variance of a given PDF by optimizing the Monte Carlo estimator.

A very widespread and widely discussed variance reduction technique is *importance sampling* [Kalos and Whitlock 1986; Dutré et al. 2003; László 1999]. When computing an integral using Monte Carlo (see Equation 5), the variance of the estimator (Equation 7) depends on the number of samples, $f(x)$ and $p(x)$. Since $f(x)$ cannot be influenced, this leaves us with two ways of modifying the variance:

- Increasing or decreasing the number of samples.
- Choosing a certain sampling probability density function $p(x)$.

Out of these two options, importance sampling pursues the latter.

We want to find a $p(x)$ such that the variance is as low as possible (ideally, the variance should be zero, yielding a *perfect* estimator). This can be done by minimizing the equation of the variance using variational techniques and Lagrange multipliers [2003]. The optimal $p(x)$ is given by:

$$p(x) = \frac{|f(x)|}{\int_D f(x)dx}$$

where D is the integration domain. If $f(x)$ does not change sign, this $p(x)$ will yield a variance of zero when used. An interesting explanation for the fact that variance can be minimized by choosing a $p(x)$ proportional to $f(x)$ is given in [László 1999]:

The ratio $\frac{f(x)}{p(x)}$ can be expressed as:

$$\frac{f(x)}{g(x)} = \alpha + \beta \cdot \delta(x)$$

where $\alpha = E\left[\frac{f(x)}{p(x)}\right]$ and $\int_D (\delta(x))^2 \cdot p(x)dx = 1$. The variance of the Monte Carlo estimator can be written as:

$$\begin{aligned} var\left\{\frac{f(x)}{p(x)}\right\} &= E[(\alpha + \beta \cdot \delta(x) - E[\alpha + \beta \cdot \delta(x)])^2] \\ &= E[(\beta \cdot \delta(x))^2] \\ &= \beta^2 \cdot E[(\delta(x))^2] \\ &= \beta^2 \end{aligned}$$

We can see that the minimum variance is achieved when $\beta = 0$ and thus $\frac{f(x)}{p(x)}$ is a constant (meaning $f(x)$ and $p(x)$ are proportional).

Both results mentioned lead however to a single conclusion: in order to achieve the lowest possible variance, we have to know the value of $\int_D f(x)dx$, which is exactly the quantity we are trying to approximate. This might not be directly visible for the second result, but considering that $p(x)$ is a PDF and its integral over the whole domain is 1, the only way we can compute a $p(x)$ that is also proportional to $f(x)$ is by dividing $f(x)$ by the value of its integral.

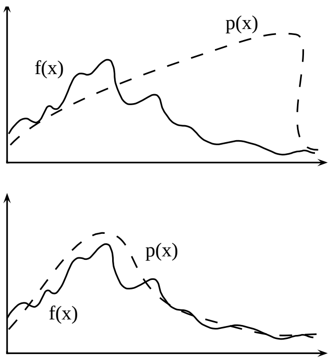

Figure 5: Importance sampling. The continuous function is the function we want to integrate, while the dotted function is the PDF used. The PDF above will yield a higher variance than the PDF below.

Although the usage of the perfect estimator can be ruled out, variance can be reduced considerably by using PDFs that "look like" the function we are trying to integrate (see Figure 3.3). Several examples are given by Kalos and Whitlock [1986], one of which we will analyze here for clarification:

Suppose the following integral should be approximated by Monte Carlo:

$$G = \int_0^1 \cos\left(\frac{\pi x}{2}\right) dx$$

When approximating the integral using uniformly distributed samples from $[0,1)$ (that is, $p(x) = 1$), the function $g_1(x)$ used by the estimator (as g in Equation 6) has the variance:

$$\begin{aligned} var\{g_1\} &= E\left(\cos^2\left(\frac{\pi x}{2}\right)\right) - E\left(\cos\frac{\pi x}{2}\right)^2 \\ &\approx 0.0947 \end{aligned}$$

By expanding $cos(\pi x/2)$ in a power series, an approximation of the function for low values of x can be easily found:

$$\cos\left(\frac{\pi x}{2}\right) = 1 - \frac{\pi^2 x^2}{8} + \frac{\pi^4 x^4}{384} - \ldots$$

From this result, an even more convenient approximation is made. By choosing $\tilde{p}(x) = a(1-x^2)$ and calculating the factor $a = \frac{3}{2}$ such that the integral of $\tilde{p}(x)$ over $[0,1)$ is 1, $\tilde{p}(x)$ thus becoming a valid PDF, we get:

$$\tilde{g}(x) = \frac{g_1(x)}{\tilde{p}(x)} = \frac{2}{3}\frac{\cos(\pi x/2)}{1-x^2}$$

The variance of $\tilde{g}(x)$ is approximately:

$$var\{\tilde{g}(x)\} \approx 0.000990$$

yielding a variance decrease by a factor of 100.

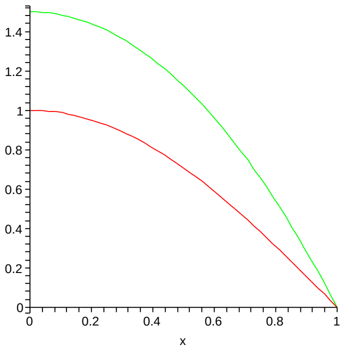

Figure 6: The function $\cos(\pi x/2)$ (bottom, red) and its approximation function (top, green).

As we have seen from this example, sampling according to even a highly approximated PDF proportional to $f(x)$ can lower the variance greatly.

More importance sampling based methods, which serve special problems in rendering, will be discussed later.

3.4 Use of Expected Values

This technique can be used to reduce the dimension of the approximated integral when one dimension can be integrated analytically [Kalos and Whitlock 1986; Veach 1998]. Consider that the integral to be approximated is:

$$G = \int g(x,y)p(x,y)dxdy$$

and let $m(x)$ be the marginal distribution for x:

$$m(x) = \int f(x,y)dy$$

The integral can be written as:

$$\begin{aligned} G &= \int m(x)\frac{1}{m(x)}\int g(x,y)p(x,y)dydx \\ &= \int m(x)h(x)dx \end{aligned}$$

If $m(x)$ and the inner integral $h(x)$ can be evaluated analytically, we can remove y completely from the calculation.

The difference between the variance of $g(x,y)$ and the variance of $h(x)$ can be shown to equal $E(var\{g|x\})$, meaning that it is always positive [1986]. Thus, calculating part of the integral analytically will result in lower variance.

3.5 Correlation Methods

When two random variables are correlated, that is, one depends on the other, this correlation can be used to decrease variance in a Monte Carlo calculation [Kalos and Whitlock 1986].

In the method of *control variates*, the integral to be estimated is written as:

$$G = \int (g(x) - h(x))p(x)dx + \int h(x)p(x)dx$$

where $\int h(x)p(x)$ can be calculated analytically. The estimator for G becomes:

$$\langle G \rangle = \int h(x)p(x)\ dx + \frac{1}{N}\sum_{i=1} N[g(x_i) - h(x_i)]$$

This method is useful when the variance of $g(x) - h(x)$ is much lower than the variance of $g(x)$, that is, when $h(x)$ is very similar to $g(x)$. This technique appears similar to importance sampling; in fact, when $|g(x) - h(x)|$ is approximately constant, the method is more efficient than importance sampling, but when $|g(x) - h(x)|$ is approximately proportional to $|h(x)|$, importance sampling is more efficient [1986].

The method of *antithetic varieties* [1986] exploits negative correlation and is useful when $g(x)$ is linear; however, it has been shown that this is not a very good method for variance reduction in many dimensions.

3.6 Stratified Sampling

Stratified sampling [Kalos and Whitlock 1986; Dutré et al. 2003] is another important variance reduction method; it tries to address the problem that samples are often chosen unevenly from the integration domain, thus resulting in "clumping". The solution is to split the integration domain into m disjoint subdomains (called *strata*) and approximate each resulting integral separately. Note that here, for simplicity, samples are taken from a uniform distribution:

$$\int_0^1 f(x)dx = \int_0^{\alpha_1} f(x)dx + \int_{\alpha_1}^{\alpha_2} f(x)dx + \ldots + \int_{\alpha_{m-1}}^1 f(x)dx$$

This method does indeed reduce variance; considering that each stratum receives a number of uniformly distributed samples n_j, the variance of the corresponding estimator is [2003]:

$$\begin{aligned}\sigma^2 = & \sum_{j=1}^{m} \frac{\alpha_j - \alpha_{j-1}}{n_j} \int_{\alpha_{j-1}}^{\alpha_j} f(x)^2 dx \\ & - \sum_{j=1}^{m} \frac{1}{n_j} \left(\int_{\alpha_{j-1}}^{\alpha_j} f(x) dx \right)^2\end{aligned}$$

Furthermore, if all strata are of equal size and each stratum only contains one uniformly generated sample, the variance becomes:

$$\begin{aligned}\sigma^2 &= \sum_{j=1}^{m} \frac{1}{N} \int_{\alpha_{j-1}}^{\alpha_j} f(x)^2 dx - \sum_{j=1}^{m} \left(\int_{\alpha_{j-1}}^{\alpha_j} f(x) dx \right)^2 \\ &= \frac{1}{N} \int_0^1 f(x)^2 dx - \sum_{j=1}^{m} \left(\int_{\alpha_{j-1}}^{\alpha_j} f(x) dx \right)^2\end{aligned}$$

By comparing this to the variance of the naive Monte Carlo estimator we can see that the variance obtained using stratified sampling is always lower than the one of the naive estimator. Thus, instead of using two or more samples to approximate the integral of a stratum, performing a subdivision on the stratum such that only one sample is attributed to each stratum will yield a lower variance.

Achieving the smallest possible variance with stratified sampling is not very easy; for this, the size of the strata relative to each other and the number of samples per stratum have to be adjusted. It can be shown that the optimal number of samples per stratum is proportional to the variance of the function with regard to its mean within that stratum [2003]. If only one sample per stratum is used, the boundaries of strata have to be chosen such that the function variance within all strata is equal, which requires detailed knowledge of the function.

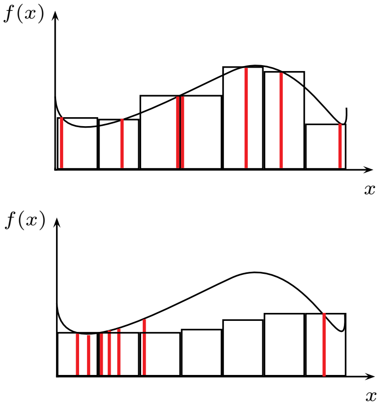

Figure 7: The advantage of stratified sampling (above) - in one dimension. With stratified sampling and one sample per stratum, each sample is confined within the bounds of its stratum, below samples are taken randomly. Clearly, stratified sampling yields a better result in this case.

In order to further lower variance, stratified sampling can be combined with importance sampling, retaining the advantages of both techniques. In practice, this can be done by applying stratified sampling on a uniform distribution and then using ICDF sampling to distribute the values according to a PDF of choice [2003].

An interesting sampling method was introduced by Arvo [Arvo 1995] which makes use of stratification to more efficiently sample spherical triangles. Techniques for sampling spherical polygons, especially useful when requiring integration over the unit hemisphere, are also described by Arvo [Dutré et al. 2004].

Stratified sampling is slightly reminescent of deterministic numerical quadrature as it shares one of its downsides: the number of samples required by the technique grows exponentially with the dimension of the function (that is, when approximating the integral of a d-dimensional function with one sample per stratum, N^d samples are required). Considering that Monte Carlo is primarily used to estimate highly-dimensional integrals, we see that this shortcoming has to be addressed. Fortunately, Quasi Monte Carlo or techniques such as the N-rooks algorithm, which are both described below, can be used for this purpose.

3.6.1 N-Rooks

The *N-Rooks*, or *latin hypercube algorithm* [Dutré et al. 2003; Kalos and Whitlock 1986; Owen 1998], offers a way to distribute N samples evenly across multiple dimensions, thus allowing a constant number of samples with stratified sampling.

Consider that we want to sample N times from a d-dimensional hypercube. This is done by first dividing each dimension of the hypercube into N segments. Then we form $d-1$ independent random permutations of $1, 2, \ldots, N$. Assuming that the i'th member of the j'th permutation is $n(i, j)$, we can sample from the regions:

$$\begin{aligned}&[1, n(1,1), n(1,2), \ldots, n(1, d-1)], \\ &[2, n(2,1), n(2,2), \ldots, n(2, d-1)], \ldots, \\ &[N, n(N,1), n(N,2), \ldots, n(N, d-1)], \ldots,\end{aligned}$$

In two dimensions, each combination of a row and a column will contain exactly one sample (see Figure 3.6.1).

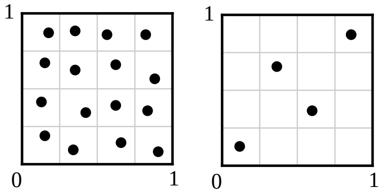

Figure 8: Stratified sampling (left) and N-Rooks sampling in two dimensions (right). With 4 subdivisions, stratified sampling already needs 16 samples.

A generalization of the N-Rooks algorithm, called *orthogonal array sampling* was introduced by Owen [Owen 1992]. This method increases the rate of convergence in some cases compared to N-Rooks.

3.7 Quasi Monte Carlo

Quasi-Monte Carlo (QMC) [Dutré et al. 2003; László 1999] describes a class of methods that use non-random numbers to perform Monte Carlo calculations. In order for this approach to work, the numbers chosen need to have low *discrepancy*, that is, they must be distributed as uniformly as possible.

The key principle of quasi-Monte Carlo is thus the same as that of Monte Carlo: the integral of an s-dimensional function $f(\mathbf{z})$ (for simplicity, we choose the domain of the integral as $[0,1]$ in each dimension) can be approximated using a sequence $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ such that [1999]:

$$\int_{\mathbf{z}\in[0,1]^s} f(\mathbf{z})d\mathbf{z} \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{z}_i)$$

A property that needs to be fulfilled in this case, and which can be seen as analogous to choosing a consistent estimator in Monte Carlo, is the *stability* of the sequence. This means that the error in approximation becomes zero asymptotically:

$$\int_{\mathbf{z}\in[0,1]^s} f(\mathbf{z})d\mathbf{z} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{z}_i)$$

Sequences that meet this criterium are called *uniform* [1999].

Discrepancy is most often quantified using a measure called *star discrepancy*, which can be explained as follows: Consider an s-dimensional box with one corner positioned at the origin, and the outmost corner at coordinates $A = (v_1, v_2, \ldots, v_s)$. The volume of this box is:

$$V(A) = \int_{\mathbf{z}\in[0,A]} d\mathbf{z} = \prod_{j=1}^{s} v_j$$

The volume can also be approximated by QMC using N samples:

$$\begin{aligned} V(A) &\approx \frac{1}{N}\sum_{i=1} Nf(\mathbf{z}_i) \\ &\approx \frac{m(A)}{N} \end{aligned}$$

where $m(A)$ is the number of samples in the box. The star discrepancy measure shows how much the distribution of samples deviates from the ideal situation:

$$D^*(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N) = \sup_A \left| \frac{m(A)}{N} - V(A) \right|$$

The *Koksma-Hlawka* inequality is uses the star discrepancy measure to describe the error bounds of quasi-Monte Carlo:

$$\left| \int_{\mathbf{z}\in[0,1]^s} f(\mathbf{z})d\mathbf{z} - \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{z}_i) \right| \leq V_{HK} \cdot D^*((\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N)$$

where V_{HK} is the *Hardy-Krause variation* [1999], which is basically a measure of how fast the function can change. As we can see, the error of quasi-Monte Carlo becomes lower with discrepancy. A lot of research has been done in the area of minimizing discrepancy, and several low discrepancy sequences have been found. Examples include Halton, Hammersley, Sobol and Niederreiter sequences [2003].

Halton sequences are very popular in QMC. They are based on the radical inverse function, which basically mirrors the representation of a number i expressed in base b about the decimal point in that base. For example, the representation of $i = 4$ in base $b = 2$ is 100_2. This representation is mirrored about the decimal point, resulting in 0.001_2, or 0.125. A multidimensional Halton sequence uses a radical inverse sequence with a different base for each dimension, with all bases being relatively prime to each other. It can be shown that the discrepancy of an s-dimensional Halton sequence is of order $O((logN)^s/N)$ [2003].

Considering that pure Monte Carlo has an error bound of $1/\sqrt{N}$, QMC appears very advantageous. Indeed, low-discrepancy sequences work best for low dimensions (10-20), while at higher dimensions, their performance does not offer a direct advantage. Low-discrepancy sequences are also highly correlated, thus trading randomness for uniformity when sampling from them [2003].

3.8 Combining Estimators of Different Distributions

Often, more than one estimator for an integral is available. It is also possible that each estimator is good at approximating one particular feature of the integral, such as one of the terms in the rendering equation. Since the estimators may thus have different relevance depending on the parameters of the integrated function, it would be best to calculate all estimators and weigh them appropriately, depending on these parameters [Dutré et al. 2003].

3.8.1 Weighting by Variance

Mathematically speaking, it can be shown that if $I_1, I_2, \ldots I_m$ are all estimators for a certain integral, a linear combination of these estimators:

$$I = \sum_{i=1}^{m} w_i I_i$$

will also be a valid estimator of the integral, provided that the sum of all w_i is 1. For simplicity, let us analyze the variance when $m = 2$:

$$\begin{aligned} var\{I\} &= var\{w_1 I_1 + w_2 I_2\} \\ &= w_1^2 var\{I_1\} + w_2^2 var\{I_2\} + 2w_1 w_2 cov(I_1, I_2) \end{aligned}$$

By minimizing the variance, we obtain a formula for the optimal ratio of the two weights [2003]:

$$\frac{w_1}{w_2} = \frac{var\{I_2\} - cov(I_1, I_2)}{var\{I_1\} - cov(I_1, I_2)}$$

Knowing that the covariance of independent variables is zero (as explained in Section 2.3.2), we can see that, for independent estimators, the weights should be chosen inversely proportional to the variance. This is quite straightforward, as the function with the higher variance is weighted less and the function with the lower variance is weighted more.

3.8.2 Multiple Importance Sampling

Multiple importance sampling was introduced by Veach [1998]. Like the previously discussed method, it allows estimators to be combined using a weighting scheme, only the weight can also depend on the sample, not only on the variance of the estimator.

Let n be the number of estimating techniques used, n_i the total number of samples generated using the i'th technique, $X_{i,j}$ the j'th sample generated using the technique and finally $w_i(X)$ a weighting function for the technique. The estimator used by multiple importance sampling (called the *multi-sample estimator*) is:

$$F = \sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i} w_i(X_{i,j})\frac{f(X_{i,j})}{p_i(X_{i,j})}$$

If the weighting functions w_i satisfy two simple conditions:

- whenever $f(x) \neq 0$, the weights of all estimators sum up to one:
$$\sum_{i=0}^{n} w_i(x) = 1$$
- and whenever the probability density $p_i(x) = 0$, the weight must be zero as well,

the estimator can be shown to be unbiased [1998].

The approach makes use of specialized weighting functions. One of them is the so-called *balance heuristic* for which it is proven [1998] that no other combination strategy is much better. The balance heuristic implies weighting functions of the form:

$$\hat{w}_i(x) = \frac{n_i}{p_i(x)} \sum_k n_k p_k(x)$$

which result in an estimator:

$$F = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{f(X_{i,j})}{\sum_k c_k p_k(X_{i,j})}$$

where $N = \sum_i n_i$ is the total number of samples taken and $c_k = n_k/N$ is the fraction of samples originating from technique k.

Other weighting functions for the multi-sample estimator, designed especially for low-variance problems, have also been proposed by Veach [1998].

3.9 Adaptive sampling

Adaptive sampling refers to a class of sampling techniques that attempt to reduce variance by sampling more values at positions where the variation of the integrand is high [Veach 1998]. This is usually done dynamically by analyzing the samples already taken.

The downside of adaptive sampling techniques is that they can introduce bias, as information about the number of samples taken is usually calcuated on-the-fly; furthermore, given a high number of dimensions, it is usually quite expensive to sample in all dimensions that require it, especially if only one or two dimensions are responsible for the variance.

3.10 Other Variance Reduction Methods

As variance reduction is a necessity for Monte Carlo algorithms, this field has received much attention. Many other specialized variance reduction methods have been developed, most of them extending importance or stratified sampling. Some examples are listed below, while other more specialized ones can be found in literature [Kajiya 1986; Agarwal et al. 2003; David Burke 2004; Ghosh and Heidrich 2005].

3.10.1 Weighted Importance Sampling

Weighted importance sampling is an extension to importance sampling and was introduced by Bekært et al. [2000]. This technique uses two PDFs, a so-called “source” PDF $q(x)$ and a “target” PDF $p(x)$. It should be possible to sample from the source PDF, but not necessarily from the target PDF.

The integral of the function $f(x)$ to be approximated over the domain D is written in such a way that both PDFs are contained within the integral:

$$\int_D f(x)dx = \int_D \frac{f(x)}{p(x)} \frac{p(x)}{q(x)} q(x)dx$$

Let the weighting function $w(x) = p(x)/q(x)$. The integral can now be written as:

$$\int_D f(x)dx = \int_D \frac{f(x)}{p(x)} w(x)q(x)dx$$

In order to estimate the integral, samples are taken from the source PDF $q(x)$ and weighted according to the ratio $w(x)$. The proposed estimator is:

$$\begin{aligned} I &= \sum_{i=1}^{N} \frac{f(x_i)}{p(x_i)} \frac{w(x_i)}{\sum_{j=1}^{N} w(x_j)} \\ &= \frac{\sum_{i=1}^{N} f(x_i)/q(x_i)}{\sum_{i=1}^{N} p(x_i)/q(x_i)} \end{aligned}$$

Although this technique is biased, it is proven to be consistent (the bias disappears by $1/N$) [2000].

3.10.2 Resampled Importance Sampling

Resampled importance sampling is another more robust technique based on *importance resampling* [Talbot et al. 2005]. The principle of importance resampling is as follows:

1. A number M of samples $x_1, \ldots, x_M$ is taken from the distribution $p(x)$.
2. Each acquired sample is given a weight w_j.
3. A random sample y from $x_1, \ldots, x_M$ is chosen according to the defined weights $w_1, \ldots, w_M$.

If the weights are chosen such that:

$$w_j = \frac{g(x_j)}{p(x_j)}$$

then the x_j will be distributed approximately according to $g(x)$.

Thus, resampled importance sampling offers a Monte Carlo estimator based on this principle:

$$I_{ris} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{f(y_i)}{g(y_i)} \cdot \frac{1}{M} \sum_{j=1}^{M} \frac{g(x_{ij})}{p(x_{ij})} \right)$$

This technique can be proven to be unbiased, the only source of error coming from variance.

3.11 Russian Roulette

Sometimes, an algorithm requires a certain more or less arbitrary termination condition, as it would otherwise run indefinitely (an example of such an algorithm is stochastic ray tracing discussed further below). By automatically terminating the algorithm after a fixed number of iterations, bias may be introduced. *Russian Roulette* is a stochastic, unbiased method for determining termination [Dutré et al. 2003].

Suppose we have a recursive algorithm and a random variable determines the weighting of the next recursion. Thus, if the random variable is 0, the algorithm will terminate. With Russian Roulette, the PDF of the random variable can be adjusted to return 0 with a certain additional probability $\alpha = 1 - P$, called the *absorbtion probability*:

The integral

$$I = \int_0^1 f(x)dx$$

can be “compressed” horizontally, more precisely scaled by P horizontally and by $1/P$ vertically, resulting in:

$$I = \int_0^P \frac{1}{P} f(\frac{x}{P}) dx$$

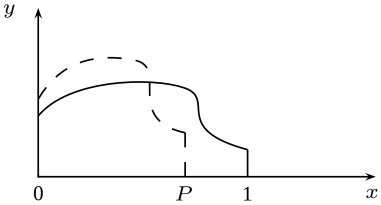

Figure 9: Russian Roulette. The dotted function is the PDF after applying Russian Roulette with abosrbtion probability $1-P$.

If x is a sample from the uniform distribution between 0 and 1, an estimator for the integral can be written as:

$$\langle I \rangle = \begin{cases} \frac{1}{P} f(\frac{x}{P}) & \text{if } x \leq P \\ 0 & \text{otherwise} \end{cases}$$

Clearly, the expectation of $\langle I \rangle$ remains I and, since the $1/P$ weighting factor compensates for the fact that a value of 0 is chosen between P and 1, the estimator remains unbiased.

Variance will clearly be influenced by α (the higher α is, the higher the variance), however this is to be expected as the average recursion length will decrease as α grows, making the algorithm generate less accurate results.

3.12 Metropolis Sampling

As a final sampling technique, Metropolis sampling [Kalos and Whitlock 1986] is presented. First introduced by Metropolis et al. in 1953 in the field of computational physics, this sampling technique has been adapted by Veach and Guibas [1997] for use in global illumination, resulting in the *Metropolis Light Transport* algorithm.

The principle of Metropolis sampling is quite simple and was inspired by the behavior of systems in statistichal mechanics that approach an equilibrium regardless of the kinetics of the system [1986]. Metropolis sampling can be used to generate a random walk, that is, a series of random variables $X_0, X_1, X_2, \ldots$. The desired property of this random walk is that starting from a certain X_i, the random variables are distributed proportionally to a specific function $f(x)$, regardless of what X_0 was; this is achieved by conducting a random “change” between X_{i-1} and X_i.

The “change” between X_{i-1} and X_i can be described by a probability function. The probability of transitioning from a state x to a state y is given by $K(y|x)$, the probability that y will be the next state, given that x is the current state.

In order for the random walk to eventually reach a distribution proportional to $f(x)$, it has to be just as likely to transition from a state X to a state Y as the other way around. More precisely:

$$K(X|Y)f(Y) = K(Y|X)f(x)$$

This property is known as *detailed balance*.

We thus have to find a K such that the above equation is fulfilled; in practice, this is done by splitting K into two parts: a *tentative transition function T* which can be chosen arbitrarily and an *acceptance probability a*. A rejection technique is then applied:

Suppose that we want to transition from a state X_{i-1} to a state X_i:

1. We choose a *tentative sample* $X_i^{'}$ according to the tentative transition function T.
2. We accept the sample $X_i^{'}$ with probability $a(X_i^{'}|X_{i-1})$ or we set $X_i = X_{i-1}$ otherwise.

The *acceptance probability a* is chosen such that detailed balance is fulfilled. A commonly used acceptance probability function (as in Metropolis Light Transport) is:

$$a(X|Y) = \min\left(1, \frac{T(Y|X)f(X)}{T(X|Y)f(Y)}\right)$$

The key strength of Metropolis sampling is that the only requirement for the function $f(x)$ to be sampled is that it can be evaluated at a given position. However, Metropolis sampling is biased at the beginning, so commonly the first few elements of the random walk are discarded [1986]; Veach and Guibas use a weighting scheme to eliminate this bias [1997]. Another problem is that consecutive samples are also highly correlated [1986].

4 Applications

This section will explain how the theoretical framework outlined above can be used in actual Monte Carlo rendering algorithms. For this purpose, the two main approaches to realistic rendering, stochastic ray tracing and stochastic radiosity, are briefly explained.

4.1 The Rendering Equation

4.1.1 Radiometric Quantities

The rendering equation introduced by Kajiya [1986] and actually describes, just like most of its variations, a radiometric quantity called *radiance*. This quantity is fundamental in global illumination, as it describes the “appearance” of an object.

Physically speaking, radiance is the total energy flow per unit projected area per unit solid angle [Dutré et al. 2003]. Radiance varies with position and direction and is thus a five-dimensional quantity. Radiance shall be denoted as $L(x, \Theta)$, where x is the position and Θ the direction, or alternatively as $L(x \rightarrow \Theta)$ to symbolize the radiance leaving from x in direction Θ and $L(x \leftarrow \Theta)$ to express the radiance arriving at x from direction Θ, as in Dutré et al. [2003].

Radiance is invariant along straight paths, meaning that the radiance leaving point x in direction of point y equals the radiance leaving point y and heading for point x. This property implies the absence of a participating medium (such as fog or water) between the two points. Radiance itself also does not diminish with distance, this effect is indirectly encoded in the definition of the solid angle [2003].

Furthermore, radiance is the quantity observed by sensors such as cameras and the human eye. This explains why brightness or color does not diminish with distance and why radiance is so important in rendering.

Other important quantities used in global illumination can be derived from radiance [2003]:

- *Irradiance* (noted E)[3] is the incident radiant power on a surface, per unit surface area. It can be calculated by integrating the radiance over the hemisphere:

$$E(x) = \int_{\Omega} L(x \leftarrow \Theta) \cos\theta d\omega_{\Theta}$$

- *Radiant Exitance* or *Radiosity* (noted B) is the exitant radiant power per unit surface area:

$$B(x) = \int_{\Omega} L(x \rightarrow \Theta) \cos\theta d\omega_{\Theta}$$

Importance is also often used in global illumination algorithms; although it does not exactly belong to radiometry, it can be mentioned that importance is a quantity analogous to radiance, only it “flows” in the opposite direction.

4.1.2 BRDF

The *bidirectional reflectance distribution function* (BRDF) describes the (approximate) interaction of light with a material. It assumes that light always exits at the same point it enters an object and thus not accounting for subsurface scattering. The BRDF is a four-dimensional function defined at every surface point, with the incident direction of light Ψ and the exitant direction Θ as parameters.

The BRDF can thus be expressed as the ratio of the differential radiance reflected in direction Θ and the differential irradiance coming from Ψ [2003]:

$$f_r(x, \Psi \rightarrow \Theta) = \frac{dL(x \rightarrow \Theta)}{dE(x \leftarrow \Psi)}$$

The BRDF is also reciprocal, meaning that if Θ and Ψ are reversed, the result is the same.

Several material types exist and can be characterized by the BRDF, including diffuse (the BRDF is constant), specular (for an incident angle, the BRDF is nonzero only for a special exitant angle) and glossy materials.

4.1.3 Formulations of the Rendering Equation

The rendering equation can be written in many ways to allow evaluation by a specific type of algorithm. Basically, the rendering equation can be described for a pair of two points, or for a point and a direction. Either the incoming or the outgoing radiance at a point can be described, with contributions being made by (or to) either all surfaces in the scene, or all surfaces reached through the hemisphere [2003].

For example, a formulation of the rendering equation based on exitant radiance (that is, we describe the radiance leaving x for direction Θ) and on integration over the hemisphere can be written as:

$$\begin{aligned} L(x \rightarrow \Theta) &= L_e(x \rightarrow \Theta) \\ &\quad + \int_{\Omega_x} f_r(x, \Psi \leftrightarrow \Theta) L(y \rightarrow -\Psi) \cos(N_x, \Psi) d\omega_{\Psi} \end{aligned}$$

where L_e is the emitted light and N_x is the surface normal of x.

When integrating over all surfaces in the scene, the equation becomes:

$$\begin{aligned} L(x \rightarrow \Theta) &= L_e(x \rightarrow \Theta) \\ &\quad + \int_{A} f_r(x, \Psi \leftrightarrow \Theta) L(y \rightarrow \overrightarrow{xy}) V(x,y) G(x,y) dA_y \qquad (8) \end{aligned}$$

where:

$$G(x,y) = \frac{\cos(N_x, \Psi)\cos(N_y, -\Psi)}{r_{xy}^2}$$

Here, $V(x,y)$ is a visibility term that is 1 if x and y are mutually visible and $G(x,y)$ is a geometry term basically making the transition between solid angle and distance.

4.2 Stochastic Path Tracing

Stochastic path tracing is based on classical path tracing (or ray tracing) but uses probabilities to determine the destination of rays.

4.2.1 Principle

The principle of stochastic path tracing is the same as that of ray tracing: for each pixel on the image plane, one or more rays are shot which are used to determine the radiance value of that pixel. A ray then hits a certain point in the scene, and in order to calculate the contribution of that point to the original ray, another ray can be shot from this point.

Rays are shot from point x in a direction Θ according to the radiance at (x, Θ) [2003]:

$$\begin{aligned} L(x \rightarrow \Theta) &= L_e(x \rightarrow \Theta) + L_r(x \rightarrow \Theta) \\ &= L_e(x \rightarrow \Theta) + \int_{\Omega_x} L(x \leftarrow \Psi) f_r(x, \Psi \leftrightarrow \Theta) \cos(\Psi, N_x) d\omega_{\Psi} \end{aligned}$$

The integral can be approximated using Monte Carlo integration.

N random directions over the hemisphere, $\Psi_1, \ldots, \Psi_N$, are sampled according to a given PDF $p(\Psi)$. The classic Monte Carlo estimator for the integral is:

$$\langle L_r(x \rightarrow \Theta) \rangle = \frac{1}{N} \sum_{i=1}^{N} \frac{L(x \leftarrow \Psi_i) f_r(x, \Omega \leftrightarrow \Psi_i) \cos(\Psi_i, N_x)}{p(\Psi_i)}$$

However, we do not yet know the value of $L(x \leftarrow \Psi_i)$. For this reason, a ray is shot in the direction of Ψ_i and the resulting radiance is added to the calculation.

Termination of this recursion can be done best with a Russian Roulette technique, since this will result in an unbiased image. Terminating after a fixed number of iterations would possibly discard more important paths.

The estimator can also be enhanced using importance sampling, one option being to sample according to the cosine of the direction and the normal.

One problem that exists with this simple approach is that light sources are rarely hit, especially when they are small. That is why the radiance calculated at a certain point is usually split into two parts: direct illumination, which quantifies the radiance coming from light sources visible at that point and indirect illumination, which contains an approximation of interreflected light coming from more distant sources.

[3] Not to be confused with expectation in probability theory, noted with E as well.

4.2.2 Direct Illumination

For efficiency and better results, the light at each point is thus split into a direct illumination component and an indirect illumination component.

Direct illumination includes the light that is explicitly emitted by light sources in the direction of the current point. It can be written as [2003]:

$$L_{direct}(x \rightarrow \Theta) = \int_{\Omega_x} L_e(r(x,\Psi) \rightarrow -\Psi) \\ \cdot f_r(x, \Theta \leftrightarrow \Psi) \cos(\Psi, N_x) d\omega_\Psi$$

where $r(x,\Psi)$ denotes the point closest to x in direction Ψ.

Since the emitted light will only be nonzero at a low number of points, it might be more advantageous to integrate over the areas of the light sources, requiring the calculation of the same terms (visibility and geometry) as in Equation 8.

The visibility terms can be calculated by tracing so called *shadow rays* between the point and the light sources. With the aid of these shadow rays, it is simple to calculate the direct contribution of light sources.

An estimator can be created using the formulation of L_{direct} to approximate the contribution made by light sources to x; this estimator uses several points y_i on the light sources and depends on:

- The incident light from those points on x;
- The value of the light sources' BRDFs at the given y_i;
- The distance between x and y_i and the cosine terms residing in $G(x, y_i)$
  and finally on:
- The visibility predicate.

Knowing this, the sampling PDF $p(y_i)$ used by the estimator can be chosen accordingly. Dutre et al. [2003] illustrate a few options, such as uniform sampling over the light source area or sampling according to the cosine terms. Furthermore, if the scene contains more light sources which all have different power, meaning there are weaker and stronger light sources, it might be a good idea to include the power of the light source in the estimator, too.

4.2.3 Indirect Illumination

The indirect illumination of a given point in a given direction is illustrated by the following equation [2003]:

$$L_{indirect}(x \rightarrow \Theta) = \int_{\Omega_x} L_r(r(x,\Psi) \rightarrow -\Psi) \\ \cdot f_r(x, \Theta \leftrightarrow \Psi) \cos(\Psi, N_x) d\omega_\Psi$$

Indirect illumination is thus determined by the radiance not emitted directly from the visible points across the hemisphere, but only by the light that these points themselves receive.

Since indirect illumination remains quite general and cannot be sampled better using special methods as is the case with direct illumination, Monte Carlo variance reduction techniques gain a much greater importance here.

The general Monte Carlo estimator for indirect illumination is [2003]:

$$\langle L_{indirect}(x \rightarrow \Theta) \rangle = \frac{1}{N} \sum_{i=1}^{N} L_r(r(x,\Psi_i) \rightarrow -\Psi_i) \\ \cdot \frac{f_r(x, \Theta \leftrightarrow \Psi_i) \cos(\Psi_i, N_x)}{p(\Psi_i)}$$

To estimate the integral, we thus simply choose N directions $\Psi_1, \ldots, \Psi_N$ and evaluate all components above for the given direction. Since the reflected radiance $L_r(r(x,\Psi_i) \rightarrow -\Psi_i)$ will depend on the value of indirect lighting at $r(x,\Psi_i)$, a similar approach as in simple stochastic path tracing has to be taken, by evaluating $L_{indirect}(r(x,\Psi_i) \rightarrow -\Psi_i)$ recursively.

As outlined above, variance reduction techniques, especially importance sampling, play a more important role in indirect illumination. The PDF $p(\Psi)$ can be made proportional to an arbitrary combination of functions used in the estimator, for example:

- The PDF can be made proportional to the cosine factor $\cos(\Psi_i, N_x)$ (a technique for sampling according to the cosine factor is discussed in Section 3.1.1). In this case, the PDF will be:

  $$p(\Psi) = \frac{\cos(\Psi, N_x)}{\pi}$$

  If the BRDF f_r is diffuse (that is, the value of the BRDF is a constant), a more compact estimator results [2003]:

  $$\langle L_{indirect}(x \rightarrow \Theta) \rangle = \frac{\pi f_r}{N} \sum_{i=1}^{N} L_r(r(x,\Psi_i) \rightarrow -\Psi_i)$$

- BRDF sampling can be employed, meaning that regions with a high value of the BRDF are more likely to be sampled. BRDF sampling offers good results when the surface is glossy or specular, however, only a few BRDF models allow proper sampling (although a technique such as resampled importance sampling, Section 3.10.2 can still be used for this purpose).
- The recursive term $L_r(r(x,\Psi_i) \rightarrow -\Psi_i)$ can also be used in importance sampling; but as this quantity is often unknown, adaptive methods or approximations (using a photon-map algorithm, for example) can be used.

4.2.4 Light tracing

It should be noted that methods analogous to path tracing exist which do not start at the camera, but at the light sources [2003]. Such methods are dubbed *light tracing* methods and function basically the same way as path tracing, with the only difference being that they try to solve the importance equation, which is the dual of the rendering equation:

$$W(x \rightarrow \Theta) = W_e(x \rightarrow \Theta) \\ + \int_{\Omega_x} W(x \leftarrow \Psi) f_r(x, \Theta \leftrightarrow \Psi) \cos(\Psi, N_x) d\omega_\Psi$$

Light tracing is seldom used as a primary image rendering technique as it does not necessarily want to compute all pixels of an image; some pixels will most likely not be reached unless the number of light rays shot is very high.

4.3 Stochastic Radiosity

Another approach to solving the rendering equation lies in radiosity; this technique requires the scene to be split into a large number of so called *surface patches*, polygons that act as single entities in regard to light transport in the scene. Only diffuse light transport is modeled in radiosity.

The radiosity integral equation is a specialization of the rendering equation. The radiance at point x only regarding diffuse surfaces is:

$$L(x) = L_e(x) + \int_{\Omega_x} f_r(x) L(x \leftarrow \Theta') \cos(\Theta', N_x) d\omega'_{\Theta}$$

This equation can be transformed to integrate over all surfaces in the scene:

$$L(x) = L_e(x) + \rho(x) \int_S K(x,y) L(y) dA_y$$

As, in a diffuse environment, radiance and radiosity are proportional by a factor of π [2003], multiplying the equation above by this factor will yield the *radiosity integral equation*:

$$B(x) = B_e(x) + \rho(x) \int_S K(x,y) B(y) dA_y$$

where $K(x,y)$ is $G(x,y)V(x,y)$ as in the area formulation of the rendering equation.

By regarding the average radiosity emitted by a surface patch i with the area A_i [2003] and by assuming that reflectivity is constant over each patch, we obtain the classical radiosity system of equations:

$$B'_i = B_e i + \rho_i \sum_j F_{ij} B'_j$$

where F_{ij} are so called *form factors*:

$$F_{ij} = \frac{1}{A_i} \int_{S_i} \int_{S_j} K(x,y) dA_y dA_x$$

4.3.1 Computing Form Factors

It soon becomes clear that calculating the form factors is the most difficult part in solving the radiosity system of equations. Therefore, several methods have been developed to make form factor calculations efficient. Two methods that are interesting with regard to Monte Carlo are sampling using local lines and sampling using global lines [2003].

Form factor sampling using *local lines* can be employed as follows: From a certain patch i, N_i virtual particles that behave like photons are shot throughout the scene. The number of particles that land on another patch j, N_{ij} is an estimate of the form factor F_{ij}:

$$\frac{N_{ij}}{N_i} \approx F_{ij}$$

The variance of this estimator is shown to be $\frac{F_{ij}(1-F_{ij})}{N_i}$ [2003].

Form factor sampling using *global lines* is done using a similar technique; this time, "global lines" are shot uniformly throughout the scene, thereby connecting pairs of patches i and j. It can be shown that form factors can again be estimated by the number of lines that passed though a patch:

$$\frac{N_{ij}}{N_i} \approx F_{ij}$$

While global lines can usually be generated more efficiently by exploiting the coherence of the scene, the methods used cannot be easily adapted to increase the number of lines passing through a certain patch, also, the variance of the estimator depends on the area of the examined patch A_i; the lower the area, the higher the variance.

4.3.2 Solutions of the Radiosity System of Equations

Many solutions for the radiosity system of equations exist, but they will not be described in detail here. Instead, an extensive overview can be found in Dutre et al. [2003].

Basically, two types of stochastic solutions can be distinguished:

- *Stochastic relaxation methods* attempt to solve the radiosity system using a mathematical iterative solution; each iteration of such a solution is basically a sum that can be approximated using Monte Carlo techniques.
- *Discrete random walk* radiosity methods try to reach a solution of the radiosity system by regarding random walks in a discrete state space. A particle will be "born" on a light-emitting patch with a certain probability, it will transition to another patch according to a transition probability and will again be absorbed with a certain probability, thus ending the random walk. The estimated radiosity of a patch then results from how often a patch is "visited" by such particles.

Random walk approaches used for radiosity have lead to *photon density estimation* methods, which regard a continuous state space and can thus be used to approximate integrals instead of sums. This also leads to the technique of photon mapping.

4.4 Combined Approaches

Monte Carlo is also used as an integral technique in approaches that combine path tracing and radiosity, such as final gathering (which uses a final path tracing pass to enhance a radiosity solution), bidirectional tracing (which uses a combination of path tracing and light tracing) [Lafortune and Willems 1994; Veach and Guibas 1994] and general multipass methods.

Also, the Metropolis Light Transport (MLT) algorithm [Veach and Guibas 1997] makes use of Metropolis sampling as discussed in Section 3.12 to compute a random walk that converges to the actual solution. MLT starts by generating a path with bidirectional path tracing and then subjecting this path to certain mutations; each mutation is designed to optimize a specific feature such as caustics.

5 Conclusions

This paper has given an overview of the mathematical basics behind Monte Carlo methods, starting with elements of probability theory, continuing with general sampling methods and methods for reducing variance in Monte Carlo calculations.

We have seen that the integral of a given function f can be approximated by evaluating f at several randomly chosen positions and calculating the weighted sum thereof; in fact, we have seen that the expected value of this technique is the value of the integral itself. We have analyzed the error bounds that arise when performing such an approximation; in probability theory, the error bounds are determined by variance and bias. In the standard case, variance was shown to be inversely proportional to the square root of N, the number of samples. By carefully choosing the distribution of the samples used when approximating, we observed that the error can be drastically reduced, which was the principle of importance sampling.

We have also addressed the problem of random samples "clumping together"; the solution was to either confine samples to a subdomain of the integrand, much in the spirit of deterministic numerical quadrature (stratified sampling) or use special sequences of non-random numbers with low discrepancy instead of random numbers; the according technique being called quasi-Monte Carlo.

We also took a short turn at analyzing various methods that could combine different estimators, techniques such as multiple importance sampling and weighted importance sampling were noted. We also saw how, using Metropolis sampling, we can sample from complicated functions without even needing to know what they look like. Additionally, we observed the downsides of some of these mentioned methods, be it the limited usage of inverse CDF sampling, the complexity of finding a good distribution for importance sampling or the correlation introduced by low discrepancy sequences in QMC.

Finally, a few applications of Monte Carlo in rendering algorithms were discussed, most importantly the basics of stochastic ray tracing and stochastic radiosity. We have seen that the entire lighting of a scene can be calculated using integrals, which can in turn be approximated using Monte Carlo methods.

References

AGARWAL, S., RAMAMOORTHI, R., BELONGIE, S., AND JENSEN, H. W. 2003. Structured importance sampling of environment maps. *ACM Trans. Graph. 22*, 3, 605–612.

ARVO, J. 1995. Stratified sampling of spherical triangles. In *Computer Graphics* Proceedings, Annual Conference Series, ACM SIGGRAPH, 437–438.

BEKAERT, P., SBERT, M., AND WILLEMS, Y. D. 2000. Weighted importance sampling techniques for monte carlo radiosity. In *Proceedings of the Eurographics Workshop on Rendering Techniques 2000*, Springer-Verlag, London, UK, 35–46.

COOK, R. L., PORTER, T., AND CARPENTER, L. 1984. Distributed ray tracing. In *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, ACM Press, New York, NY, USA, 137–145.

DAVID BURKE, M. 2004. *Bidirectional Importance Sampling for Illumination from Environment Maps*. Master's thesis, University of British Columbia.

DUTRÉ, P., BALA, K., AND BEKAERT, P. 2003. *Advanced Global Illumination*. A. K. Peters, Ltd., Natick, MA, USA.

DUTRÉ, P., JENSEN, H. W., ARVO, J., BALA, K., BEKAERT, P., MARSCHNER, S., AND PHARR, M. 2004. State of the art in monte carlo global illumination. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Course Notes*, ACM Press, New York, NY, USA, 5.

GHOSH, A., AND HEIDRICH, W. 2005. Correlated visibility sampling for direct illumination. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Sketches*, ACM Press, New York, NY, USA, 107.

KAJIYA, J. T. 1986. The rendering equation. In *SIGGRAPH '86: Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, ACM Press, New York, NY, USA, 143–150.

KALOS, M. H., AND WHITLOCK, P. A. 1986. *Monte Carlo methods. Vol. 1: basics*. Wiley-Interscience, New York, NY, USA.

LAFORTUNE, E. P., AND WILLEMS, Y. D. 1994. A theoretical framework for physically based rendering. In *Computer Graphics Forum*, vol. 13, 97–107.

LAFORTUNE, E. 1996. *Mathematical Models and Monte Carlo Algorithms for Physically Based Rendering*. PhD thesis, Katholieke University, Leuven, Belgium.

LÁSZLÓ, S.-K. 1999. *Monte-Carlo Methods in Global Illumination*. Institute of Computer Graphics, Vienna University of Technology. Course notes.

OWEN, A. B. 1992. Orthogonal arrays for computer experiments, integration and visualization. In *Statistica Sinica 2*, 439–452.

OWEN, A. B. 1998. Monto carlo extension of quasi-monte carlo. In *WSC '98: Proceedings of the 30th conference on Winter simulation*, IEEE Computer Society Press, Los Alamitos, CA, USA, 571–578.

TALBOT, J., CLINE, D., AND EGBERT, P. 2005. Importance resampling for global illumination. In *Rendering Techniques 2005: 16th Eurographics Workshop on Rendering*, 139–146.

VEACH, E., AND GUIBAS, L. 1994. Bidirectional estimators for light transport. In *Proceedings of the 5th Eurographics Workshop on Rendering*, 147–162.

VEACH, E., AND GUIBAS, L. J. 1995. Optimally combining sampling techniques for monte carlo rendering. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, ACM Press, New York, NY, USA, 419–428.

VEACH, E., AND GUIBAS, L. J. 1997. Metropolis light transport. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 65–76.

VEACH, E. 1998. *Robust monte carlo methods for light transport simulation*. PhD thesis. Adviser-Leonidas J. Guibas.

VIERTL, R. 2003. *Einführung in die Stochastik. Mit Elementen der Bayes-Statistik und der Analyse unscharfer Information.*, 3rd ed. Springer-Verlag Wien New York. In German.