A Crash Course in Monte Carlo Methods for the Busy Mathematician

A. B. Marnie

Abstract

We introduce Monte Carlo quadrature and its applications. We begin by reviewing some basic essential notions from probability theory. We then describe how integration can be viewed as a probabilistic problem so that Monte Carlo integration can be applied. We discuss accuracy of the Monte Carlo method, and we briefly discuss sampling. We end with by introduction a few variance reduction techniques in order to increase the rate of convergence.

1 What is a Monte Carlo method?

Monte Carlo, in essence, is the solution of non-probabilistic problems by probabilistic means. The name "Monte Carlo" was given to a class of mathematical methods first used by scientists working on the development of nuclear weapons in Los Alamos in the 1940s [3]. Monte Carlo methods involve inventing "games of chance" whose solution produces useful answers to questions of interest. These games of chance, which most often involve some form of "random" sampling, lead to answers which are statistical in nature. However, theory has been developed to determine how accurate the answer is (e.g., the correct solution will likely lie within a specified "error range"), so the statistical nature of the solution is often not a fatal flaw [3]. Furthermore, mathematical theory has been developed that demonstrates the rate of convergence (in terms of the number of samples taken) of the Monte Carlo solution to the correct solution. In general, Monte Carlo methods are straightforward to implement on a computer (even parallel), and they are very robust since their accuracy depends on easily controlled parameters (irrespective of the dimension of the problem).

2 Probability theory background

The information here is commonly included in textbooks on mathematical probability theory. The brief review given here is taken from [5] and [3], which treats them in full form.

2.1 Basic definitions

Definition 2.1 (Probability space)**.** Let Ω be any set. A *σ-algebra* in Ω is a collection of subsets of Ω which

- Contains $\varnothing$ and Ω.
- Is closed under countable unions and countable intersections.

- Is closed under complements.

A *measurable space* is a pair $\Omega = (\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is a $\sigma-$algebra in Ω. Elements of $\mathcal{F}$ are called *measurable sets* in Ω.

A function $\phi : \Omega \to R$ between two measurable space $(\Omega, \mathcal{F})$, $(R, \mathcal{B})$ is said to be *measurable* if $\phi^{-1}(S) \in \mathcal{F}$ for all $S \in \mathcal{B}$.

A *probability measure* on a measurable space $(\Omega, \mathcal{F})$ is a map $\mu : \mathcal{F} \to [0, +\infty]$ which obeys the following axioms:

- $\mu(\varnothing) = 0$.
- If $\{E_n\}_{n=1}^{\infty}$ are a countable sequence of disjoint sets in $\mathcal{F}$, then $\mu(\cup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mu(E_n)$.
- $\mu(\Omega) = 1$.

A *probability space* is a triplet $(\Omega, \mathcal{F}, \mu)$, where $(\Omega, \mathcal{F})$ is a measurable space equipped with a probability measure μ.

Remark 2.1. σ-algebras are the sets for which you can actually properly define a measure for; in particular, they ensure that we can get all the intuitive properties of measures (like additivity) that are needed. The non-measurable subsets of X (that is, those which do not lie in any $\sigma-$algebra) are usually extremely pathological (when talking about measures generated from the standard topology, see next note); their existence relies on the axiom of choice.

Remark 2.2. There may be more than one σ-algebra for Ω (and so, there may be more than one valid notion of "measure" on Ω). The most common σ-algebra one sees is called the Borel algebra. It consists of the closure of a topological space X under the "σ-operations (countable unions/intersections, complements). The Borel algebra can be thought of as being generated by a given topology, and so, if one has a "standard topology" (e.g., one induced from an inner product / norm / metric), then the Borel algebra is the "standard σ-algebra".

Remark 2.3. A measurable function is one which preserves the "measure structure". When using non-standard σ-algebras $\mathcal{F}, \mathcal{B}$, it is common to write $\phi : (\Omega, \mathcal{F}) \to (R, \mathcal{B})$ to emphasize the sigma-algebras. The definition should be reminiscent of that of continuous functions between topological spaces.

Remark 2.4. For integration, we will take Ω to be some subset of $\mathbb{R}^d$, usually the unit cube I^d, and $\mathcal{F}_\Omega$ will be the standard Borel σ-algebra on I^d. Since $\mathcal{F}$ is a Borel measure, it induces a unique probably measure on $\mathcal{F}$.

Definition 2.2 (Basic probabilistic terminology)**.** Let $\Omega = (\Omega, \mathcal{F}_\Omega, \mu)$ be a probability space. The set Ω (without structure) is called the *sample space*, and an element of Ω is called an *outcome*. An *event* is a measurable set $E_\Omega \in \mathcal{F}_\Omega$. The *probability* of an event

E_Ω is defined by its probability measure,

$$P[E_\Omega] = \mu(E_\Omega).$$

An event E_Ω is *surely true* if $E_\Omega = \Omega$, and is surely false if $E_\Omega = \varnothing$. An event E_Ω is *almost surely true* if $P(E_\Omega) = 1$, and is *almost surely false* if $P(E_\Omega) = 0$.

A (real) *random variable* X_Ω is a measurable function $X_\Omega : \Omega \to R$ from a sample space Ω to the range $R \subset \mathbb{R}$.

Remark 2.5. Random variables can be manipulated with respect to measurable operations. That is, if X is a random variable taking values in a measurable space $(R, \mathcal{F}_R)$, and $f : R \to S$ is a measurable map, then $f(X)$ is a random variable taking values in S. In general, measurable functions preserve probabilistic properties of their input.

Remark 2.6. Given a measurable relation $F : R_1 \times R_2 \to \{\text{true}, \text{false}\}$ on ranges R_1, R_2, and given random variables X_1, X_2, one can define an event

$$F(X_1, X_2) := \{\omega \in \Omega : F(X_1(\omega), X_2(\omega)) = \text{true}\}.$$

Thus, for instance, given random variables X, Y, we can define the event $(X > Y)$, or $(X = Y)$. Furthermore, this can easily be extended to relations on more than two events.

Definition 2.3 (More probabilistic and statistical terminology)**.** Let X be a real random variable taking values in a measurable space $(\Omega, \mathcal{F})$. The *probability distribution* of X is the probability measure μ_X on Ω defined by the formula

$$\mu_X(E) := \boldsymbol{P}(X \in E), \qquad \forall E \in \mathcal{F}.$$

We say that two random variables X, Y *agree in distribution* and write $X =^d Y$ if we have $\mu_X = \mu_Y$. A *cumulative distribution function* (or cdf for short) of a real random variable X is a function $F : \boldsymbol{R} \to [0, 1]$ given by $F(t) := \boldsymbol{P}(X \leq t)$. The cdf of X completely characterizes the probability distribution μ_X, so we often only concern ourselves with cdfs. Thus, two real random variables (potentially on different sample spaces) *agree in distribution* if they have the same cdf. When the cdf $F(t) = \boldsymbol{P}(X \leq x)$ for X has the special form

$$F(t) = \int_{-\infty}^{t} f(x)dx,$$

we say that f is a *probability density function* (or pdf for short) for X. The *expectation* or *mean* of a random variable X is just the (Lebesgue) integral

$$E[X] := \int_\Omega X(\omega)d\mu(\omega).$$

The *variance* of X is defined as

$$\mathrm{Var}(X) := E[(X - E[X])^2].$$

Two random variables X, Y taking values in R_X, R_Y are said to be *independent* if

$$E[F(X), G(Y)] = E[F(X)]E[G(Y)]$$

for any real absolutely integrable (or unsigned) functions F, G on R_X, R_Y, respectively. A sequence of random variables which are all pairwise-independent are called *jointly independent*. A sequence of jointly independent random variables which have the same distribution are said to be *independent identically distributed* (or iid for short). If X, Y are square-integrable real random variables, then the *covariance* between them is defined to be

$$\mathrm{Cov}(X, Y) := E[(X - E[X])(Y - E[Y])].$$

Remark 2.7. Note that $E[X] < \infty$ if and only if $E[|X|] < \infty$. Furthermore, the expectation operator inherits all the properties of the Lebesgue integral. Most notably, the expectation operator is linear, identifies functions up to almost sure equivalence, obeys monotonicity, obeys the triangle inequality, obeys Markov's inequality, obeys the change of variables formula, and we also have the standard Lebesgue convergence theorems (Fatou's lemma, monotone convergence theorem, dominated convergence theorem...).

Remark 2.8. By the Cauchy-schwarz inequality for square-integrable random variables, we have

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y],$$
$$\mathrm{Var}(X) = \mathrm{Cov}(X, X).$$

Note that if X, Y are independent square-integrable random variables, then

$$\mathrm{Cov}(X, Y) = 0.$$

Definition 2.4 (Notions of convergence)**.**

- Let X_n be a sequence of random variables in a metric space. Let X be a random variable which takes values in R. We say that X_n *converges almost surely* to X if

  $$P(\lim_{n\to\infty} X_n = X) = 1,$$

  that is, if $X_n \to X$ pointwise almost surely.

- Let X_n be a sequence of random variables taking values in a separable metric space $R = (R, d)$, e.g., X_n could be real random variables. Let X be a random variable which takes values in R. We say that X_n *converges in probability* to X if, for every radius $\epsilon > 0$, one has $P(d(X_n, X) > \epsilon) \to 0$ as $n \to \infty$.

- Let R be a locally compact Hausdorff topological space with the Borel σ-agebra. A sequence of finite measures μ_n on R is said to *converge vaguely* to another finite measure μ if one has

$$\int_R G(x)d\mu_n(x) \to \int_R G(x)d\mu(x)$$

as $n \to \infty$ for all continuously compactly supported functions $G : R \to \mathbb{R}$. More importantly, a sequence of random variables X_n taking values in R is said to *converge in distribution* to another random variable X if the distributions μ_{X_n} converge vaguely to the distribution u_X, or equivalently if

$$E[G(X_n)] \to E[G(X)]$$

as $n \to \infty$ for all continuous compactly supported functions $G : R \to \mathbb{R}$.

Remark 2.9. Almost sure convergence (strong convergence) implies convergence in probability. Convergence in probability implies convergence in distribution (weak convergence).

2.2 Main results

We can finally state the two main theorems we need.

Theorem 2.1 (Law of large numbers (iid version))**.** *Let $X_1, X_2, \ldots$ be a sequence of iid random variables, and let X be an absolutely integrable random variable which also is also identically distributed. Set the mean $\mu := E[X]$, and for each natural number n, set the empirical/sample sum $S_n := X_1 + \ldots X_n$, so the empirical/sample average $\frac{S_n}{n}$ is a random variable. Then,*

- *(Strong law) $\frac{S_n}{n}$ converge almost surely to μ.*
- *(Weak law) $\frac{S_n}{n}$ converge in probability to μ.*

Theorem 2.2 (Central limit theorem (iid version))**.** *Let $X_1, X_2, \ldots$ be a sequence of iid random variables, and let X be an absolutely integrable random variable which also is also identically distributed. Set the mean $\mu := E[X]$, set the (nonzero) variance $\sigma^2 := Var(X)$, and for each natural number n, set the empirical/sample sum $S_n := X_1 + \ldots X_n$. Then the random variables $\frac{\sqrt{n}}{\sigma}(\frac{S_n}{n} - \mu)$ converges in distribution to a random variable ν with the standard normal distribution $N(0,1)$ (e.g., a random variable with pdf $x \mapsto \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Thus we have*

$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right) \to^d \nu.$$

In the normalized case when X has mean zero and unit variance, this simplifies to

$$\frac{S_n}{\sqrt{n}} \to^d \nu.$$

3 Integration

We finally turn to how the problem of computing a definite integral can be solved by probabilistic means.

3.1 Probabilistic view of integration

Suppose we would like to compute the definite integral of a real-valued function $f(x)$, and, for simplicity, we consider $x \in [0,1]$. In other words, we would like to find the quantity

$$\text{Int}[f] := \int_0^1 f(x)dx = \int_{\mathbb{R}} 1_{[0,1]}(x)f(x)dm(x),$$

where m is the standard measure. Note that the restriction of m to the unit interval is a probability measure on $(\Omega, \mathcal{F})$, where $\Omega = [0,1]$, and $\mathcal{F}$ contains all measurable subsets of Ω. Relabelling $m := m|_\Omega$, we see that $(\Omega, \mathcal{F}, m)$ is a probability space. Now let x be a random variable that is uniformly distributed on Ω. If f is measurable, then $X := f(x)$ is itself a random variable. With this set up, we can view the original quantity of interest $\text{Int}[f]$ as the average of a random variable:

$$\mu := E[X] = \text{Int}[f].$$

Note, that all of this works if we consider real-valued measurable functions on $\mathbb{R}^d$. Thus we can compute definite integrals on the unit cube I^d in d by finding the average of a random variable $X := f(\boldsymbol{x})$ (where $\boldsymbol{x}$ is a random variable uniformly distributed on I^d) in exactly the same way:

$$\mu := \int_{\mathbb{R}^d} f(\boldsymbol{x})1_{I^d}(\boldsymbol{x})dm(\boldsymbol{x}).$$

Since X is a random variable with a cdf, we can then perhaps build a sequence of N empirical measurements (or samples) $\{X_n\}$ from the underlying distribution. In fact, we can just uniformly sample points in I^d to build an iid sequence $\boldsymbol{x}_n$ and set $X_n := f(\boldsymbol{x}_n)$ (once again using the fact that measurable functions preserve probabilistic properties of random variables to ensure that the X_n are iid). Then, setting $S_N = X_1 + \ldots X_N$, we have the following empirical mean approximation to the true mean:

$$\frac{S_N}{N} \approx \mu.$$

The theoretical justification for using this empirical mean approximation is Theorem 2.1 (the strong law of large numbers) [1]. By the strong law of large numbers, we have the pointwise limit

$$\lim_{N\to\infty} \frac{S_N}{N} = \mu$$

with probability 1. Furthermore, this also implies that the empirical approximation is *unbiased* in the sense that $E[\frac{S_N}{N}] = \mu$ for all N [1]. In general, we define the Monte Carlo integration error to be

$$\varepsilon_N[X] = \frac{S_N}{N} - \mu,$$

so that the *root mean square error* (or RMSE for short) is [1]

$$\sqrt{E[(\varepsilon_N[X])^2]}.$$

The most basic Monte Carlo procedure can thus be described by the following algorithm:

Pick N random uniformly distributed points $\boldsymbol{x}_n \in I^d$.
$X_n \leftarrow f(x_n), \quad \forall n.$
$S_N \leftarrow X_1 + \ldots + X_N.$
return $\frac{S_N}{N}$.

3.2 Accuracy of Monte Carlo integration

In order to precisely describe the size and statistical properties of Monte Carlo integration, we turn to Theorem 2.2 (the central limit theorem). By the central limit theorem, we have the distributional limit

$$\lim_{N\to\infty} \frac{\sqrt{N}}{\sigma}\left(\frac{S_N}{N} - \mu\right) = \nu,$$

where ν is the random variable with standard normal distribution $N(0,1)$, and σ is the variance. Performing a truncation after the N-th measurement, we have the Monte Carlo integration error

$$\varepsilon_N[X] \approx \sigma N^{-1/2}\nu. \tag{1}$$

Hence, the error in Monte Carlo integration is of size $O(N^{-1/2})$, with a constant that is exactly the square root of the variance of the integrand. Furthermore, the statistical distribution of the error is approximately a normal random variable. In contrast with the standard upper-bound (worst-case) results of numerical analysis, the result obtained here is a probabilistic result that says that the error is of a specific size with some probability [1].

Since this error bound is probabilistic, the precision of the Monte Carlo method can only be established within some confidence level [1]. To ensure an error of size at most ε with confidence level c, the number of sample points N needs to be (from Eq (1)):

$$N = \varepsilon^{-2}\sigma^2 s, \tag{2}$$

where $s = s(c)$ is the confidence function for a normal variable [2], meaning

$$\begin{aligned} c &= \frac{1}{\sqrt{2\pi}}\int_{-s(c)}^{s(c)} e^{-x^2/2}dx, \\ &= \mathrm{erf}(s(c)/\sqrt{2}). \end{aligned}$$

For example, for a 95 percent confidence in the error size, we should approximately take $s = 1.96$. This number is often taken from

Another important thing to consider is that Eq (2) can only be used if the exact value of the variance is known, which is usually not the case [1]. In order to overcome this, one instead must use the empirical error and empirical variance [1]. To do this, for $j = 1, \ldots, M$ perform M computations using independent uniformly distributed points $\boldsymbol{x_i}$ for $1 \leq i \leq MN$ to obtain the j Monte Carlo outputs $\frac{S_N^{(j)}}{N}$. One can then compute the empirical RMSE to be [2]

$$\tilde{\varepsilon_N} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\frac{S_N^{(j)}}{N} - \bar{I_N}\right)^2},$$

where

$$\bar{I_N} = \frac{1}{M}\sum_{j=1}^{M}\frac{S_N^{(j)}}{N}.$$

Also, the empirical variance can be computed by [1]

$$\tilde{\sigma} = N^{1/2}\tilde{\varepsilon_N}.$$

These empirical estimates for σ and ε_N can be used in Eq (2) to determine the number of samples needed for a given precision level ϵ and a given confidence level c.

3.3 Monte Carlo integration compared to grid-based integration

Compared to standard methods analyzed in numerical analysis, $O(N^{-1/2})$ is not a very fast order of convergence. The upside is that this rate of convergence is completely independent of the dimension of the problem, thus avoiding the "curse of dimensionality" that one encounters when performing numerical integration in high dimensions. For example, the rate of convergence for grid-based quadrature is $O(N^{-k/d})$ for a k-th order method in dimension d [2]. Hence, Monte Carlo integration beats grid-based methods in dimension d whenever one has $k/d < 1/2$. Most importantly perhaps, is that it is extremely difficult to lay down grids in high dimensions. For example, the simplest uniform grid in d dimensions requires at least 2^d points [2]. In addition to this, refining a grid in d dimension requires increasing the number of points by another factor of 2^d [2].

4 Random number generation and sampling

4.1 Pseudo-random numbers

In order to use Monte Carlo integration, one needs a way to uniformly choose points $\boldsymbol{x}_n \in I^d$. This process may be thought of as generating a sequence of *random numbers*. In practice,

these numbers are not truly random, but that does not matter long as these numbers are approximately correctly distributed [2]. There are many statistical tests that can determine whether an infinite sequence of numbers is correctly distributed, but in practice we can only generate finitely many numbers before repeating, and we do not care if our sequence is passes all of these tests [2]. Requiring our sequence to only have a few key properties of true random number sequences means that we are no longer generating random numbers, but instead are generating *pseudo-random numbers* that are more suitable for quadrature. There are implementations of pseudo-random number generators for quadrature in pretty much all programming languages [1]. One of the more popular pseudo-random number generators is called Mersenne Twister [4]; it is the default algorithm used for generating pseudo-random numbers in MATLAB. For an in-depth discussion of pseudo-random number generators, see chapter 9 of [3].

4.2 Inverse transform sampling

For a non-uniform random variable $X := f(x)$ which admits a pdf $p(x)$, the expectation of X is given by

$$E_p[X] = \int_{[0,1]} f(x)p(x)dx.$$

Standard pseudo-random number generators such as Mersenne Twister produce uniformly distributed variables, but one can sample non-uniform random variables by means of a transformation of uniform random variables [1]. Suppose we seek a random variable x with a particular pdf $p(x)$. Let y be a uniform random variable. Then what we seek is a function T so that $x = T(y)$ has the desired density $p(x)$ [1]. Using the pdf $p(x)$ we can define the cdf

$$P(x) := \int_{-\infty}^{x} p(t)dt.$$

Now, for any function f we have

$$E_p[f(x)] = E_{\text{unif}}[f(T(y))],$$

so by a change of variables followed by implicit differentiation, we have

$$\begin{aligned} E_{\text{unif}}[f(T(y))] &:= \int_{[0,1]} f(T(y))dy, \\ &= \int f(x)\frac{dy}{dx}dx, \\ &= \int f(x)\frac{1}{T'(y)}dx. \end{aligned}$$

This implies $p(x) = \frac{dy}{dx} = \frac{1}{T'(y)}$, so that we have

$$y = \int_{-\infty}^{T(y)} p(x)dx = P(T(y)).$$

Hence, we have $T(y) = P^{-1}(y)$. This means we can obtain a random variable with a desired pdf by applying the associated inverse cdf to a uniform distributed random variable.

For an example of this, consider the problem of sampling from a Gaussian (normal) distributed random variable x, which has pdf $p(x)$ and cdf $P(x)$ given by

$$\begin{aligned} p(x) &= \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \\ P(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2dt}, \\ &= \frac{1}{2}(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right). \end{aligned}$$

To do this, we first sample a uniformly distributed random variable y, and then apply the transformation $y \mapsto x$ given by

$$x = T(y) = P^{-1}(y) = \sqrt{2}\text{erf}^{-1}(2y - 1).$$

This demonstrates how easy it is to implement the transformation. The main downside is that it may be difficult to compute P^{-1} [1]. For the Gaussian distribution, as well as many other distributions, there are special transformations one can perform instead of computing P^{-1} [1]. A simple method for generating two Gaussian normal distributed variables x_1, x_2 starting with two uniformly distributed random variables y_1, y_2 proceeds by the formula

$$\begin{aligned} x_1 &= \sqrt{-2\log(y_1)}\cos(2\pi y_2), \\ x_2 &= \sqrt{-2\log(y_1)}\sin(2\pi y_2). \end{aligned}$$

This is called the Box-Muller method, and its main advantage is that one does not need to invert the error function [1].

There other methods that can be used when inverting the cdf is too costly or difficult. Two popular methods are the adaptive-rejection method, and its extension the Metropolis-Hasting method. More information about sampling random variables, including discussions on adaptive-rejection and Metropolis-Hastings can be found in chapter 3 of [3].

5 Variance reduction

Recall that we had the following Monte Carlo integration error

$$\begin{aligned} &\varepsilon_N[X] \approx \sigma N^{-1/2}\nu, \\ \implies\ &N \approx \sigma^2/\epsilon^2, \end{aligned}$$

where N is the number of samples, σ is the square root of the variance, and ν is the random variable with standard normal distribution $N(0, 1)$. There are two ways to reduce this error. First, one can attempt to reduce the constant σ by performing a suitable transformation on the integrand; this is called *variance reduction*. Second, one could attempt to replace the

random variables with an alternative sequence which improves the exponent 1/2; one way to do this is by using *low-discrepancy sequences* (or *quasi-random numbers*), and doing so would be called using a *quasi-Monte Carlo method*. These acceleration methods tend to require additional computations which must be balanced against the savings gained by reducing the number of samples taken, but in many cases, the savings to be gained are significant [1].

5.1 Antithetic variates

We first talk about a very simple and widely used variance reduction method known as using *antithetic variates*. For each sample value $\boldsymbol{x} \in I^d$, one also uses the value $1 - \boldsymbol{x}$. Setting $X_n := f(\boldsymbol{x}_n)$ and $X_{-n} := f(1 - \boldsymbol{x}_n)$, we then have the following empirical mean approximation

$$\frac{S_n + S_{-n}}{2N} = \frac{1}{2N} \sum_{n=1}^{N} X_n + X_{-n}.$$

To see why this works, suppose that we would like to estimate $E[X]$ where $X = f(\boldsymbol{x})$ and $\boldsymbol{x}$ is an $N(0, \sigma^2)$ random variable. Given samples X_1, X_2 we then have unbiased approximation $E[X] \approx \frac{X_1+X_2}{2}$. Then expanding the definition of variance and covariance, we get

$$\mathrm{Var}(\frac{X_1 + X_2}{2}) + \frac{1}{4}\mathrm{Var}(X_1) + \frac{1}{4}\mathrm{Var}(X_2) + \frac{1}{2}\mathrm{Cov}(X_1, X_2),$$

so that the variance can be reduced by choosing X_1, X_2 such that $\mathrm{Cov}(X_1, X_2) < 0$ [2]. Furthermore, the number of pseudo-random numbers needed to be generated using antithetic variates is halved.

5.2 Control variates

The next method is to the called using *control variates*. The basic idea is to replace the integrand f with $f - \lambda g + \lambda g$, where g is easy to integrate by some other means, and $\sigma_{f-\lambda g} << \sigma_f$. Then we can write

$$\int_{I^d} f(\boldsymbol{x}) d\boldsymbol{x} = \int_{I^d} (f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})) d\boldsymbol{x} + \lambda \int_{I^d} g(\boldsymbol{x}) d\boldsymbol{x}.$$

Setting $X_n := f(\boldsymbol{x}_n)$ and $Y_n := g(\boldsymbol{x}_n)$, we then have the empirical mean approximation

$$\frac{S_N}{N} = \frac{1}{N} \sum_{n=1}^{N} X_n + \lambda Y_n + \lambda \int_{I^d} g(x) dx.$$

One can find the optimal value of λ by minimizing the variance $\sigma^2_{f-\lambda g}$ to obtain [1]

$$\lambda = \left(\int_{I^d} \overline{f}\overline{g} dx \right) / \left(\int_{I^d} \overline{g}^2 dx \right),$$

where $\overline{f}, \overline{g}$ are the mean-centered versions of f, g respectively. In order to have $\sigma_{f-\lambda g} < \sigma_f$, we should pick g to be highly correlated with f.

This method reduces the variance in a manner differently than antithetic variates, so the two methods can be combined together.

5.3 Stratification

The next method is called *stratification*. To perform stratification, subdivide the integration domain Ω into M disjoint (up to a set of measure zero) pieces Ω_k with $\Omega = \cup_{k=1}^{M} \Omega_k$. We then take N_k random variables in each piece Ω_k with $\sum_{k=1}^{M} N_k = N$. In each piece Ω_k we choose points $x_n^{(k)}$ distributed with density $p^{(k)}(x)$ in which

$$\begin{aligned} \bar{p}_k &:= \int_{\Omega_k} p(\boldsymbol{x}) d\boldsymbol{x} \\ p^{(k)}(\boldsymbol{x}) &:= p(x)/\bar{p}_k. \end{aligned}$$

Setting $X_n^{(k)} := f(\boldsymbol{x}_n^{(k)})$, we then have the following stratified empirical mean approximation,

$$\frac{S_N}{N} = \sum_{k=1}^{M} \frac{\bar{p}_k}{N_k} \sum_{n=1}^{N} {}_k X_n^{(k)},$$

where the stratified error is just the sum of the local errors

$$\begin{aligned} \varepsilon_N[X] &= \sum_{k=1}^{M} \varepsilon_N^{(k)}[X], \\ \varepsilon_N^{(k)}[X] &= \sqrt{\frac{\bar{p}_k}{N_k}} \sigma^{(k)}, \end{aligned}$$

where $\sigma^{(k)}$ is the local variance defined on Ω_k. If the number of points in Ω_k is proportional to the weighted size $\bar{p}_k$, meaning for all k we have

$$\bar{p}_k / N_k = 1/N,$$

then the resulting stratified quadrature error is

$$\varepsilon_N[X] \approx N^{-1/2} \sigma_s, \qquad \sigma_s^2 = \sum_{k=1}^{M} \left(\sigma^{(k)}\right)^2,$$

which implies $\sigma_s \leq \sigma$. We can lower the variance even further by placing more sub-intervals where the variance is highest.

5.4 Importance sampling

The last variance reduction method we cover is called *importance sampling*. Consider introducing a density function $p(x)$ by writing

$$\int_{I^d} f(x)dx = \int_{I^d} \frac{f(\boldsymbol{x})}{p(\boldsymbol{x})} p(\boldsymbol{x})d\boldsymbol{x}.$$

Setting $g(\boldsymbol{x}) = f(x)/p(x)$, we can sample points $\boldsymbol{x}_n$ from the distribution with density $p(x)$ and set $Y_n := g(\boldsymbol{x}_n)$ to form the empirical mean approximation

$$\frac{S_n}{N} = \frac{1}{N} \sum_{n=1}^{N} Y_n.$$

The resulting error is the same as usual, but with variance $\sigma_{f/p}$. This is effective with f/p is nearly constant, so that $\sigma_{f/p}$ is small [1]. The main difficulty of this method is that one choose a suitable density function, and furthermore, one needs a way to efficiently sample from it (e.g., by the inverse transform method, or by rejection-based methods) [1]. The Metropolis-Hastings algorithm is one of the most commonly used methods for importance sampling [3].

References

[1] R. E. Caflisch, *Monte carlo and quasi-monte carlo methods*, Acta numerica, 7 (1998), pp. 1–49.

[2] J. M. Hammersley and D. C. Handscomb, *General principles of the monte carlo method*, in Monte Carlo Methods, Springer, 1964, pp. 50–75.

[3] M. H. Kalos and P. A. Whitlock, *Monte carlo methods*, John Wiley & Sons, 2008.

[4] M. Matsumoto and T. Nishimura, *Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator*, ACM Transactions on Modeling and Computer Simulation (TOMACS), 8 (1998), pp. 3–30.

[5] T. Tao, *Math 275a probability theory notes*. https://terrytao.wordpress.com/category/teaching/275a-probability-theory/, 2015.