3 Monte Carlo Methods

Broadly speaking, a Monte Carlo method is any technique that employs randomness as a tool to calculate, estimate, or simply investigate a quantity of interest. Monte Carlo methods can be used in applications as diverse as estimating definite integrals, performing sensitivity analysis on parameters fit to noisy data, and solving differential equations. This section sketches the probability background necessary to understand certain core Monte Carlo principles, and sketches some concrete applications in the setting of integration.

3.1 Probability Basics

3.1.1 Random Variables

A random variable is a quantity whose value is determined by the outcome of some random process. Random variables can be discrete (meaning that the possible values they can assume can be enumerated in a finite or countably infinite set) or continuous (meaning that the possible values they can assume lie in an interval, or a collection of intervals.) The set of values a random variable can take on is called the sample space.

Example 3.1. (Roll of a 6-sided die) Let $X$ be the value of a single roll of a fair six-sided die. Then $X$ is a discrete random variable whose sample space is the set $\{1, 2, 3, 4, 5, 6\}$, and the random process that generates $X$ is the rolling of the die.

Example 3.2. (Waiting times) Let $X$ be the amount of time a person who arrives at a certain bus stop must wait until her bus arrives. Then $X$ is a continuous random variable whose sample space is the set $[0, \infty)$, and the random process that generates $X$ is whatever combination of traffic conditions, driver disposition, and mechanical capacity that leads to the arrival of the bus.

Note that it could be argued that the “random” processes in these two cases are not really random at all, but the product of deterministic forces. The question of whether or not true randomness exists is a philosophical point that shall not concern us here.

3.1.2 Distributions

Informally speaking, a probability distribution is a specification of what values a random variable can assume, and how often it assumes them. More formally, let $\Omega$ denote the sample space of a random variable $X$, and define an “event” $E$ as any subset of $\Omega$. A probability distribution is a mapping $P$ from the set of all events of $\Omega$ to the interval $[a, b]$ such that the following hold:

1. $P(\Omega) = 1$ (i.e. the probability that the value of $X$ lies somewhere in $\Omega$ is 1.)
2. If $E$ and $F$ are events with $E \cap F = \emptyset$, then $P(E \cup F) = P(E) + P(F)$.

Probability distributions are generally specified by density functions. A density function is a map $\rho : \Omega \to \mathbb{R}$ that is used to define the distribution $P$. Density functions take different forms depending on whether the random variable is discrete or continuous.

For a discrete random variable, a density function $\rho$ is a map that associates each element $x_i \in \Omega$ to a probability $p_i$. This map yields $P$ under the convention that if $E$ is a subset of $\Omega$,

$$P(E) = \sum_{i: x_i \in E} p_i.$$ 

Note that $P$ defined this way automatically satisfies property (2) above, and that property (1) requires $\sum_{i=1}^{\infty} p_i = 1$.

Example 3.3. For the fair 6-sided example given above, the probability that $X$ assumes any particular value is 1/6. Let $E$ be the event that $X$ is odd, i.e. $E = \{1, 3, 5\}$. Then the probability of $E$ is $1/6 + 1/6 + 1/6 = 1/2$. 

17
For a continuous random variable, a density function is a map \( \rho : \Omega \rightarrow [0, \infty) \) such that if \( E \) is a subset of \( \Omega \), then the probability of \( E \) is given by

\[
P(E) = \int_E \rho(x)dx.
\]

Note \( P \) defined this way automatically satisfies property (2) above, and that property (1) requires that \( \int_\Omega \rho(x)dx = 1 \).

**Example 3.4.** Let \( X \) be a random variable with sample space \( [0, \infty) \) and density function \( \rho(x) = e^{-x} \). If \( E \) is the event that \( X < 2 \), then \( P(E) \) is given by

\[
P(E) = \int_0^2 e^{-x}dx = 1 - e^{-2}.
\]

### 3.1.3 Expected Value

Informally, the **expected value** (also called the **mean**) of a random variable \( X \) is its average, were you to run the underlying random process an infinite number of times. The expectation of a random variable is generally denoted either by \( E(X) \) or by \( \mu_X \). Formally, the expectation of a discrete random variable is defined as

\[
\mu_X = E(X) = \sum_{x_i \in \Omega} x_ip_i, \quad \text{(Expected Value of Discrete R.V.)}
\]

while the expectation of a continuous random variable is

\[
\mu_X = E(X) = \int_\Omega xp(x)dx. \quad \text{(Expected Value of Continuous R.V.)}
\]

### 3.1.4 Variance and Standard Deviation

Informally, the **variance** of a random variable \( X \) is a measure of its spread, i.e. how far \( X \) tends to be from its average value. The variance of \( X \) is generally denoted by \( \sigma^2_X \), and is defined formally as

\[
\sigma^2_X = E([X - \mu_X]^2).
\]

Note that since this definition is given abstractly (i.e. in terms of the expectation) it applies to both discrete and continuous random variables. Note too that the quantity \( [X - \mu_X]^2 \) is a random variable in its own right: it will be discrete if \( X \) is discrete and continuous if \( X \) is continuous, but in neither case will it be equal to \( X \).

The **standard deviation** of a random variable \( X \) is defined as the square root of the variance. The standard deviation is denoted \( \sigma_X \).

### 3.1.5 Special Distributions

Here is a short list of probability distributions that show up a lot in applications. Note that these distributions are all characterized by **parameters**, i.e. certain constants that need to be specified. Changing the values of these parameters will change the values of the density functions, but not their basic shapes.

- **Uniform:** \( X \) takes values in an interval \([a, b]\), and has density function
  \[
  f_X(x) = \frac{1}{b-a}
  \]
  The mean is \( \mu_X = (a + b)/2 \), and the variance is \( \sigma^2 = (b - a)^2/12 \).

- **Exponential:** \( X \) takes values in \([0, \infty)\), and has density function
  \[
  f_X(x) = \lambda e^{-\lambda x}
  \]
  for some positive \( \lambda \). The mean is \( \mu_X = 1/\lambda \), and the variance is \( \sigma^2_X = 1/\lambda^2 \).
• **Normal (Gaussian):** $X$ takes values in $(-\infty, \infty)$, and has density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$  

The mean is $\mu$ and the variance is $\sigma^2$.

### 3.1.6 Independence

Informally speaking, two random variables $X$ and $Y$ will be **independent** if the outcome of one has no influence on the outcome of the other.

**Example 3.5. (Independence)** Let $X$ be the outcome of one toss of a die, $Y$ the outcome of a second toss. Since the result of the first throw has no influence on the result of the second, $X$ and $Y$ are independent.

The above example is paradigmatic: if the value of $X$ is the result of some random process, and then the same random process is run again to generate $Y$, then $X$ and $Y$ will be independent and identically distributed (i.i.d.). This notion can be extended (in an obvious way) to more than two random variables.

A very useful fact about independent random variables $X$ and $Y$ is the following:

$$E(XY) = E(X)E(Y).$$

This simply states that if the two random quantities $X$ and $Y$ don’t influence one another, then the average of their product is the product of their averages.

**Example 3.6.** Suppose you roll a die twice and let $Z$ be the product of the outcomes. Since the rolls are independent, the expected value of $Z$ is just the product of the means of each individual roll, i.e. $E(Z) = 3.5^2$.

Note that a consequence of the above fact is that if $X$ and $Y$ are independent and both have mean 0, then $E(XY) = 0$.

### 3.1.7 Functions of Random Variables

Suppose $X$ is a continuous random variable with sample space $\Omega \subset \mathbb{R}$ and density function $f_X(x)$. Suppose $g : \mathbb{R} \rightarrow \mathbb{R}$ is a function. Then $Y = g(X)$ is a random variable with density function

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}.$$  

**Example 3.7.** Suppose $X$ is uniformly distributed on $[0, 2]$ and $g(x) = x^2$. Then if $Y = g(X)$, the density function of $Y$ is given by

$$f_Y(y) = \frac{1}{2\sqrt{y}}$$

since $f_X(x) = 1/2$, $g^{-1}(y) = \sqrt{y}$, and $g'(x) = 1/2x$.

### 3.2 Application: Estimating Definite Integrals

Suppose we wish to estimate the quantity

$$\int_a^b g(x)dx$$

for some function $g : \mathbb{R} \rightarrow \mathbb{R}^+$. In other words, we wish to calculate the area between the graph of $g$ and the $x$–axis.

One approach is first calculate the maximum value of $g$ in the interval $[a, b]$, and then fill the two dimensional box $[a, b] \times [0, g_{\text{max}}]$ with $n$ uniformly spaced random points. Intuitively, the proportion of these points that fall in the area we’re trying to estimate should be pretty close to the ratio of that area and the area of the box. In other words,

$$\left(\text{area beneath curve}\right) \approx \frac{\hat{n}}{n} \times \left(\text{area of box}\right),$$
where \( \hat{n} \) is the number of sample points that lie below the graph of \( g \). This suggests our first algorithm for estimating \( S_{bT} \):

### Algorithm 1

- Calculate \( g_{\text{max}} \), the maximum of \( g \) in the interval \([a,b]\).
- Fill the 2-dimensional box \([a, b] \times [0, g_{\text{max}}]\) with \( n \) uniformly spaced sample points.
- Count how many of these fall in the region below the graph of \( g \).
- Estimate the integral as
  \[
  \int_a^b g(x) \, dx \approx \frac{\hat{n}}{n} \times (b - a) \times g_{\text{max}}.
  \]

An alternative formulation takes advantage of the mean-value theorem for integrals, which states that the average value of a function \( y = g(x) \) on an interval \([a, b]\) is

\[
y_{\text{avg}} = \frac{1}{b - a} \int_a^b g(x) \, dx,
\]
or in other words,

\[
\int_a^b g(x) \, dx = y_{\text{avg}} \times (b - a).
\]

This latter formulation is useful because we can estimate \( y_{\text{avg}} \) via random sampling. The simplest way to do this is take \( n \) random sample points \( x_i \) in the domain \([a, b]\), evaluate \( g \) at each of these points, and form an average of the results, i.e., form

\[
\overline{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \text{where } y_i = g(x_i).
\]

The quantity \( \overline{y} \) is an approximation of \( y_{\text{avg}} \). The approximation of the integral is then \( \overline{y} \times (b - a) \). This leads to our second algorithm:

### Algorithm 2

- Choose \( n \) random sample points \( x_i \) distributed uniformly in \([a, b]\).
- Calculate \( y_i = g(x_i) \) for each \( i \).
- Form \( \overline{y} = (1/n)\left( \sum_{i=1}^n y_i \right) \).
- Estimate the integral as
  \[
  \int_a^b g(x) \, dx \approx \frac{1}{n} \sum_{i=1}^n y_i \cdot (b - a) = \overline{y} \cdot (b - a).
  \]

Note that Algorithm 2 is better than Algorithm 1 in a number of ways: in particular, it requires fewer sample points and it doesn’t require calculating \( g_{\text{max}} \).
3.3 Error Analysis

Estimates achieved through the use of random quantities will necessarily contain errors. On a practical level, it is useful to be able to quantify these errors. For example, if we use 100 sample points to estimate the integral via Algorithm 2, it would be nice to be able to say something like “This algorithm produces an answer that is within 0.7% of the true answer 99% of the time. Therefore, there is a 99% change that this particular answer is within 0.7% of the true answer.” Elementary probability theory allows us to make claims like this. The following explains how.

3.3.1 Sampling

Suppose $Y$ is a random variable. As an abstract quantity, $Y$ can take on any of multiple values. A sample of $Y$ is a concrete outcome under one run of the underlying random process. A set of samples $y_1, \cdots, y_n$ is a set of outcomes under $n$ runs of the underlying random process. If the underlying process is the same for each $i$, then the $y_i$ are independent and identically distributed.

Given $n$ samples of $Y$, the sample mean is the quantity

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

and the sample variance is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \mu_Y)^2.$$

The sample mean is just the actual average of the samples. The sample variance is (basically) the mean square distance between the samples and the sample mean. The reasons for dividing by $n-1$ and not $n$ are arcane. Our main interest will be with $\overline{y}$.

Note that $\overline{y}$ is itself a random variable. It thus has a mean and a variance. The mean is easy to calculate, since the expectation operator is linear:

$$\mu_{\overline{y}} = E\left( \frac{1}{n} \sum_{i=1}^{n} y_i \right) = \frac{1}{n} \sum_{i=1}^{n} E(y_i) = \frac{1}{n} \cdot n \cdot \mu_Y = \mu_Y.$$

Note that we use the fact that the $y_i$ are identically distributed in setting $E(y_i) = \mu_Y$ for all $i$.

To calculate the variance, we note

$$\sigma_{\overline{y}}^2 = E((\overline{y} - \mu_Y)^2) = E(\overline{y}^2) - 2E(\overline{y} \cdot \mu_Y) + E(\mu_Y^2).$$

It is not hard to show (and is left as an exercise) that if the $y_i$ are independent, this reduces to

$$\sigma_{\overline{y}}^2 = \frac{\sigma_Y^2}{n}.$$  \hspace{1cm} (15)

3.3.2 Central Limit Theorem

We have seen that $\overline{y}$ is a good estimator for the mean of a random variable $Y$, since its mean is the correct one and its variance reduces as the number of samples increases. It turns out that as $n$ gets larger, the distribution of $\overline{y}$ approaches a Gaussian distribution with mean $\mu_Y$ and variance $\sigma_{\overline{y}}^2/n$. What is amazing about this statement is that it doesn’t depend on the distribution of $Y$: it holds for any random variable $Y$. This is the content of the central limit theorem:

**Theorem 3.1** (Central Limit Theorem). Let $Y$ be any random variable with mean $\mu_Y$ and variance $\sigma_Y^2$. If for $i = 1, \cdots, n$ the variables $y_i$ represent independent random samples of $Y$, then as $n \to \infty$, the sample mean $\overline{y} = (\sum y_i)/n$ approaches a Gaussian random variable with mean $\mu_Y$ and variance $\sigma_{\overline{y}}^2/n$.

The significance of the central limit theorem (for our purposes) is that it allows us to make claims like the one sketched in the beginning of this section. The reason for this is that Gaussian random variables happen to satisfy the following property:
Theorem 3.2 (64-95-99.7 Rule for Gaussians). Let $X$ be a Gaussian random variable with mean $\mu$ and variance $\sigma^2$. Then

- 64\% of the time, random samples of $X$ will lie within distance $\sigma$ of the mean $\mu$.
- 95\% of the time, random samples of $X$ will lie within distance $2\sigma$ of the mean $\mu$.
- 99.7\% of the time, random samples of $X$ will lie within distance $3\sigma$ of the mean $\mu$.

In other words, if $X$ is Gaussian, there is a very high probability that a random sample will lie within $3\sigma$ of $\mu$. If we have an estimate which we know is approximately Gaussian, and we can approximate how big $\sigma$ is, we can thus make some pretty good claims about how accurate the estimate is.

### 3.3.3 Errors in Monte-Carlo Estimate

Recall from Algorithm 2 that we chose sample points $x_i$ uniformly in $[a, b]$, set $y_i = g(x_i)$, and formed

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$ 

The $y_i$ represent samples of the random variable $Y = g(X)$. By the Mean Value Theorem for Integrals, the true mean value of $Y$ is just

$$\mu_Y = \frac{1}{b - a} \int_{a}^{b} g(x) \, dx,$$

and thus the mean value of $\bar{y} \cdot (b - a)$ is just

$$\mu_{\bar{y}(b-a)} = (b - a) \mu_Y = \int_{a}^{b} g(x) \, dx.$$

This means that the expected value of our Monte Carlo estimate is the true value of the integral. This is good.

To calculate the variance of $\bar{y} \cdot (b - a)$, we note that the variance of $Y$ is just

$$\sigma_Y^2 = E((Y - \mu_Y)^2) = \frac{1}{b - a} \int_{a}^{b} (g(x) - \mu_Y)^2 \, dx,$$

whence

$$\sigma_{\bar{y}(b-a)}^2 = \sigma_Y^2 (b - a)^2 = \frac{\sigma_Y^2}{n} (b - a)^2 = \frac{b - a}{n} \int_{a}^{b} (g(x) - \mu_Y)^2 \, dx. \quad (16)$$

In other words, the variance of the Monte Carlo estimate decreases in inverse proportion to the number of samples used, and is proportional to the average square distance between $g(x)$ and its mean.

**Example 3.8.** Consider the integral

$$\int_{0}^{2} x^2 \, dx.$$ 

Using Algorithm 2 with $n$ sample points yields an estimate $\bar{y} \cdot (b - a)$ whose variance is

$$\frac{\sigma_{\bar{y}(b-a)}^2}{n} = \frac{2}{n} \int_{0}^{2} (x^2 - \frac{4}{3})^2 \, dx = \frac{5.689}{n}.$$ 

With $n = 100$, this means that the variance of the estimate should be 0.05689, and the standard deviation is $\sqrt{0.05689} = 0.24$. Since $n = 100$ is relatively large, the central limit theorem states that the estimate $\bar{y}(b-a)$ is roughly Gaussian, whence, for this value of $n$, our answer should like within $3 \times 0.24 \approx 0.72$ of the true answer 99.7\% of the time.