Notes by Dave Edwards

Monte Carlo Integration

Monte Carlo integration is a powerful method for computing the value of complex integrals using probabilistic techniques. This document explains the math involved in Monte Carlo integration. First I give an overview of discrete random variables. Then I show how concepts from discrete random variables can be combined with calculus to reason about continuous random variables. Finally, with a knowledge of continuous random variables, I discuss the concept of Monte Carlo integration.

To get the most of our this explanation, the reader should have a knowledge of one-dimensional calculus. A background in probability should also be helpful, although I have made an attempt to explain all necessary probability as intuitively as possible.

Discrete Probability

If we roll a fair six-sided die, it has an equal probability of landing on any of its six faces. Consider an experiment in which we roll the die N times and record the result of each roll. Let n_i be the number of times an i is rolled in our experiment; for example, n_1 is the number of times we roll a 1, n_2 is the number of times we roll a 2, and so on.

Since we must roll a number between 1 and 6, inclusive, we know that

$$\sum_{i=1}^{6} n_i = N$$

that is, if we sum the number of rolls of each possible result, we will get N, the total number of rolls.

We can use a histogram to graphically explain the results of our experiment. For reasons we will explain shortly, we draw our histrogram subject to two constraints: 1) the total width of the histogram should be 1, and 2) the total area of all bars in the histogram should sum to 1.

Since the total width of our histogram is 1, we will draw the histogram on the x-interval $[0, 1)$. Therefore, the bar representing the trials where we rolled a 1 will span the interval $[0, \frac{1}{6})$; the bar representing the trials where we rolled a 2 will span the interval $[\frac{1}{6}, \frac{2}{6})$, and so on. Note that the width of each bar is $\frac{1}{6}$.

The second constraint says that the total area of the histogram bars should sum to 1. Let A_i be the area of the ith bar, this means that

$$\sum_{i=1}^{6} A_i = 1$$

Since each histogram bar is a rectangle, we know that $A_i = w_i h_i$, where w_i and h_i are the width and height of bar i, respectively. From the previous

paragraph, we know that $w_i = \frac{1}{6}$ for all bars. This leads to the equation

$$\sum_{i=1}^{6} A_i = \sum_{i=1}^{6} w_i h_i = \sum_{i=1}^{6} \frac{h_i}{6} = 1$$

by factoring the $\frac{1}{6}$ out of the sum, we find that

$$\sum_{i=1}^{6} h_i = 6$$

Finally, we know that the height h_i of each bar should be proportional to n_i, the number of times the corresponding number i was rolled. In mathematical terms, this is:

$$h_i = kn_i$$

Combining this information with some of the previous equations, we can now calculate a value for k:

$$\sum_{i=1}^{6} h_i = 6 = k \sum_{i=1}^{6} n_i = kN$$

so $k = \frac{6}{N}$. Given k, we find that the height h_i of each bar is $\frac{6n_i}{N}$, and the area of each bar is $\frac{n_i}{N}$.

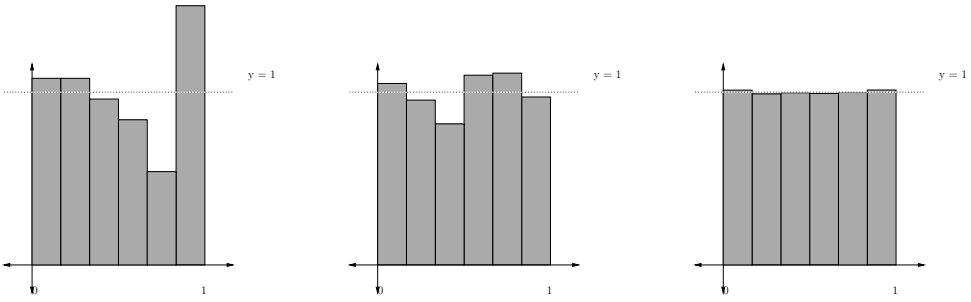

Figure 1: Histograms for the six-sided die experiment with $N = 100$, $N = 1000$, and $N = 100000$, respectively

We now explain the reasoning behind the constraints we mentioned earlier. We found that the total area of bar i is $\frac{n_i}{N}$; this is the empirical probability of rolling an i. If we increase N (thus rolling the die more times), we expect this probability to approach $\frac{1}{6}$ for all i. In the limit (as N approaches infinity), this means that the area of all six bars should be the same. Because the width of the bars is the same, this means that the heights of the bars should also be the same. We can use the equations above to see that the height of each bar will

approach 1 in the limit. Therefore, if we roll the die enough times, we should expect our histogram to look like a 1-unit-by-1-unit square (see figure 1).

Why is this shape significant? Consider performing the same experiment, this time using a 10-sided die. If we draw the histogram subject to the same two constraints, we now find that the width of each bar is $\frac{1}{10}$. After enough rolls, we also expect the area of each bar to be $\frac{1}{10}$. This means that the height of each bar should be 1, and once again, our histogram should look like a 1-unit-by-1-unit square. Further reasoning will show that, given these constraints, the shape of the final histogram is the same for dice with any number of sides!

Finally, if we consider the piecewise constant function formed by the tops of the histogram bars, we find that in the limit, the entire histogram can be described by the function $f(x) = 1$ on the interval $[0, 1)$. Once again, this is true regardless of the number of sides on the die.

Discrete Random Variables

If the die has 6 sides, we saw that the probability of rolling each side is $\frac{1}{6}$. We can represent the possible outcomes by a *discrete random variable* that we will call X. The value of X is 1 with probability $\frac{1}{6}$, 2 with probability $\frac{1}{6}$, and so on. We call a single roll of the die a *sample* of the random variable X. It is important to note that a sample has a well-defined value (in this case a sample can be 1, 2, 3, 4, 5, or 6), whereas the random variable X does not have a single value, but a distribution of values that are defined probabilistically. An intuitive way to reason about this is that a sample represents one roll of the die, but the random variable represents the results of the entire experiment (all N rolls).

We now explain a common bit of notation: $P(X = i)$ means, "the probability that a sample of the random variable X is equal to i." For example, if X is the random variable representing the die-rolling experiment, $P(X = 3)$ is the probability that we rolled a 3, which is $\frac{1}{6}$.

We can also evaluate the probability of X falling within a certain range of values. For example, if X represents a die roll, $P(2 \leq X \leq 4)$ means "the probability that we rolled a number between 2 and 4, inclusive." We can calculate this by summing the associated probabilities:

$$P(2 \leq X \leq 4) = \sum_{i=2}^{4} P(X = i) = \sum_{i=2}^{4} \frac{1}{6} = \frac{1}{2}$$

Summing probabilities over a range of possible values is a key part in continuous probability, as we shall see shortly.

Expected Values

We define the *expected value* (written $E[X]$) of a random variable X to be the average value of a sample over all samples. If X is the result of a die roll, we

can calculate the expected value using the formula

$$E[X] = \frac{1}{N} \sum_{i=1}^{6} i \cdot n_i$$

We multiply each possible outcome by the number of times we rolled it, sum the results, and then divide by the total number of rolls. This gives us the average, or expected, outcome.

Once again, we can examine the behavior of the experiment as we increase the number of rolls N. As we do this, we expect $\frac{n_i}{N}$ to approach $\frac{1}{6}$ for all i. So in the limit, the above equation becomes

$$E[X] = \frac{1}{6} \sum_{i=1}^{6} i$$

and by calculating the sum, we find that $E[X] = 3.5$. This makes sense, since 3.5 is the midpoint of 1 and 6, the smallest and largest values we can roll on the die.

Continuous Probability

Now, we make the jump from the discrete case to the continuous case. After reading the previous section, the mathematically-minded reader may be asking, "If this information is true regardless of the number of sides on the die, what happens when we roll an infinite-sided die?"

There are an infinite number of real numbers on the interval $[0, 1)$. One way to construct an infinite-sided die would be to label each side of the die with a real number on this interval. We can now think of the die-rolling problem as the problem of randomly choosing a real number on the interval $[0, 1)$.

In the last section, we saw that a histogram of a die-rolling experiment drawn according to two certain constraints looks the same in the limit, regardless of the number of sides on the die. For the infinite-sided die, the histogram should look the same, but each bar will have an infinitesimal width (and therefore, an infinitesimal area). Fortunately, we can use the methods of calculus to reason about this histogram for our infinite-sided die.

In the inifinite case, the bars in the histogram have no width or area. Remember that the area of each bar represents the probability of rolling the corresponding value. Since the bars now have zero area, this means that the probability of rolling any individual real number must be 0! This might seem counterintuitive, but we can think of the problem in terms of limits: as the number of sides on the die grows without bound, the probability of rolling each individual side must go to 0.

However, the height of each histogram bar is still 1. In fact, we have seen that after enough rolls, the histogram can be described by the function $f(x) = 1$ over the interval $[0, 1)$. In some sense, the fact that the heights of the bars are the same represents the fact that it is equally likely to roll any real number on

the interval. Because of this fact, the function $f(x) = 1$ is known as the *uniform density function* over the interval $[0, 1)$, because it describes this uniform likelihood of rolling any random number.

We have seen that in the continuous case, the probability of rolling an individual number is 0. If X is the random variable representing the experiment, we can write this fact as $P(X = x) = 0$ for all $x \in [0, 1)$. When we discussed discrete probability, we saw how we could sum probabilities over a range of possible outcomes. We can extend this idea to continuous probability. Even though the area of our histogram bars are individually 0, we can still sum the areas of many bars using integration to acheive a non-zero result. For example, we write $P(0 \leq X \leq \frac{1}{2})$ to mean "the probability that we roll a real number between 0 and $\frac{1}{2}$, inclusive." To calculate this probability, we integrate the histogram function $f(x) = 1$ over the interval $[0, \frac{1}{2})$:

$$P(0 \leq X \leq \frac{1}{2}) = \int_0^{\frac{1}{2}} f(x)dx = \int_0^{\frac{1}{2}} dx = \frac{1}{2}$$

So we can expect that half the numbers we roll will lie between 0 and $\frac{1}{2}$.

Probability Density Functions

What does the value of $f(x)$ represent? The value of $f(x)$ comes from the height of the histogram bars. We can think of the height of each bar as area divided by width; since the area of each bar is the probability of rolling the corresponding number, this means that the height of the bar at a given point x is the probability *per unit length* of rolling an x on the die. For this reason, we call $f(x)$ the *probability density function* (or PDF for short).

The uniform density function $f(x) = 1$ is just one example of a PDF. In fact, any function p over a domain D is a PDF as long as the following two properties are true:

1. $p(\vec{x}) \geq 0$ for all $\vec{x} \in D$
2. $\displaystyle\int_D p(\vec{x})d\mu(\vec{x}) = 1$

The first property corresponds to the fact that "negative probabilities" do not exist, the second ensures that the sum of probabilities of all possible outcomes is 1. Although the examples we have seen so far are one-dimensional functions, we use the notation $p(\vec{x})$ here to denote the fact that we can define a multidimensional PDF. For example, after studying a lifetime's worth of dart games, we could define a PDF over the dart board representing the probability per unit area of a dart hitting certain points on the board.

If we let $D = [0, 1)$, we see that the uniform density function $f(x) = 1$ does indeed satisfy these two properties.

Cumulative Distribution Functions

The PDF represents probability per unit length. Therefore, by integrating the PDF over an interval, we can find the probability that a random sample lies within that interval. We have already done this to show that the probability that a uniform sample on $[0, 1)$ is less than $\frac{1}{2}$ is $\frac{1}{2}$.

For a one-dimensional PDF $p(x)$, we can define another function F called the *cumulative distribution function* (or CDF) as follows:

$$F(x) = \int_{-\infty}^{x} p(t)dt$$

If we consider the meaning of the PDF, we find that the value of the CDF at a point x is the probability that a random sample has a value less than x. Written mathematically, this is:

$$F(x) = P(-\infty \leq X \leq x)$$

where X is a random variable. Note that the lower bound on the integral should actually be the lower bound of the domain over which p is defined. We have used $-\infty$ for generality. For example, the CDF of the uniform density function $f(x) = 1$ is

$$F(x) = \int_{0}^{x} f(t)dt = \int_{0}^{x} dt = x$$

since the function is defined over the interval $[0, 1)$. Therefore, we find that for a random variable X representing an infinite die roll,

$$P(0 \leq X \leq x) = x$$

It is possible to define CDFs for multi-dimensional PDFs. However, this requires the ability to divide the multi-dimensional space into dimensions in a consistent way. A "left endpoint" (corresponding to $-\infty$ in the one-dimensional case) must also be consistently defined for each dimension. For example, assume we have a two-dimensional PDF $p(x, y)$ defined over the Cartesian plane. Then we can define the corresponding CDF P as

$$P(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p(s, t)dtds$$

Continuous Random Variables

We have already referred to continuous random variables in this section by appealing to the reader's intuition and knowledge of discrete random variables. More formally, a *continuous random variable* is a variable that probabilistically takes any real-number value over a certain domain. Just as discrete random variables can be described by a discrete set of possible values and their associated probabilities, a continuous random variable can be fully described by its

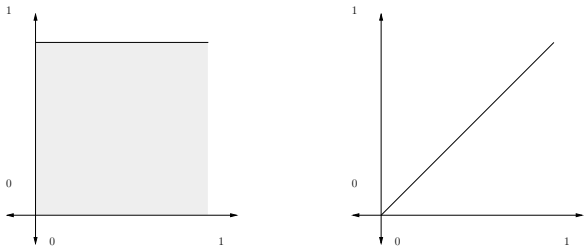

Figure 2: PDF $f(x) = 1$ and CDF $F(x) = x$ for the uniform distribution. The region under $f(x)$ is shaded to denote that $f(x)$ is a density function.

PDF (or CDF). We use the notation $X \sim p$ to mean, "X is a continuous random variable with PDF p." We also write "X is *distributed* according to p." Random variables are sometimes called distributions. For example, the random variable distributed according to the uniform density function is usually called the uniform distribution.

As with the discrete case, we say that a sample of a continuous random variable is a single value generated randomly according to the PDF of that variable. Samples from the infinite die would simply be the numbers that we rolled. From what we know about the uniform distribution, we would expect the samples to be evenly scattered across the interval $[0, 1)$.

Random samples from other distributions would be scattered differently. For example, the well-known "bell curve" is actually the PDF for the *normal distribution*. If we took samples from this distribution, we would find that most of them were clustered around the peak of the bell curve, with fewer samples farther away from the center.

If we have a random variable X and an arbitrary function g, note that $g(X)$ is also a continuous random variable. If we take a set of random samples of X and apply the function g to them, we will get another set of random samples distributed according to a different PDF (unless of course g is the identity function).

Expected Values

As in the discrete case, the expected value (or mean) of a continuous random variable X is the average value of all samples taken from X. To find the expected value, we multiply the infinitesimal probability of each real value in X's domain by the value itself, and sum the result. We can easily represent this using an integral. If we have a random variable $X \sim p$ on a domain D, then

$$E[X] = \int_D \vec{x} p(\vec{x}) d\mu(\vec{x})$$

For example, the expected value of a uniformly distributed random variable

on $[0, 1)$ is

$$E[X] = \int_0^1 x dx = \frac{1}{2}$$

which makes sense: if we take the average of all rolls on our infinite die, we should get $\frac{1}{2}$, which is half way between 0 and 1.

There are some useful rules for expected values of continuous random variables. If we have two random variables X and Y and a constant k, the following equations are true:

1. E[X + k] = E[X] + k
2. E[k X] = k E[X]
3. E[X + Y] = E[X] + E[Y], even if X and Y are dependent random variables

Finally, we can also compute the expected value of a function of a continuous random variable. Once again, if we have a random variable $X \sim p$ on domain D, then

$$E[g(X)] = \int_D g(\vec{x}) p(\vec{x}) d\mu(\vec{x})$$

Transforming Samples Between Distributions

Assume that we have two one-dimensional distributions $X \sim f$ and $Y \sim q$ and a random sample x taken from X. It is possible to transform this sample into a sample y that is distributed according to q.

Let F be the CDF for X. So we have

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

Remember also that the value of F has a clear probabilistic meaning:

$$F(x) = P(X \leq x)$$

Of course, we can also calculate the CDF Q for Y such that

$$Q(x) = P(Y \leq x)$$

Finally, we need to ask: what does it mean for a sample from X to be equivalent to a sample from Y? Assume we have a sample x from X and a sample y from Y such that $F(x) = Q(y)$. This means that $P(X \leq x) = P(Y \leq y)$. Since these probabilities are equivalent, we know that x and y divide their corresponding distributions in exactly the same way: random samples from X and Y are equally likely to be less than x and y, respectively. Therefore, in a probabilistic sense, the two samples are equivalent.

From this information, we find that, given the sample x, we need to find a y such that $F(x) = Q(y)$. Since Q is a CDF, it is monotonically increasing, and hence invertible. Therefore, we find that

$$y = Q^{-1}(F(x))$$

Following this logic, we can apply the function $Q^{-1} \circ F$ to any sample (or set of samples) from X to find the corresponding sample (or set of samples) in Y.

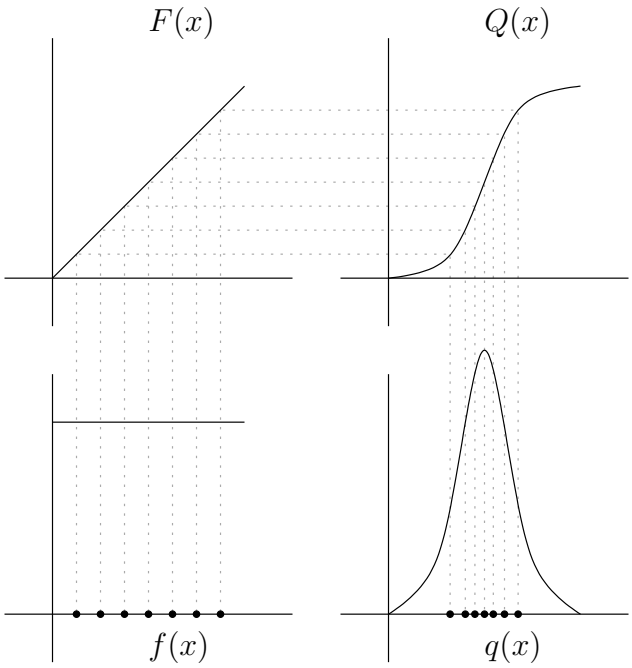

Figure 3: A graphical description of transforming samples between two distributions. We begin with the samples in the lower left, distributed according to $f(x)$. For each sample x_i, we follow the dotted lines to compute $F(x_i)$, then $Q^{-1}(F(x_i))$, which gives us the set of samples distributed according to $q(x)$. Whereas the samples from f are uniform, the samples from q occur near the peak of q. Note that we could follow the dotted lines in the other direction to transform the samples back to the original distribution f.

We commonly use this transformation in computation. Most programming languages have a facility for generating uniformly-distributed random numbers. If we would like a set of random samples with a different distribution, we can get it by transforming a set of uniform random samples using the method above. Conveniently, the CDF for the uniform distribution is the identity function, so in this case, we only have to compute

$$y = Q^{-1}(x)$$

to generate our target samples. Unfortunately, it is not always possible to invert Q analytically; in these cases we resort to approximation or to numerical methods to handle this problem.

Monte Carlo Integration

Monte Carlo Integration is a simple and powerful technique for approximating complicated integrals. Assume we are trying to estimate the integral of a function f over some domain D:

$$F = \int_D f(\vec{x}) d\mu(\vec{x})$$

Once again, we use vector notation to indicate that f need not be one-dimensional. In fact, Monte Carlo techniques are used mostly for higher-dimensional integrals, or integrals that cannot be evaulated analytically.

Assume that we have a PDF p defined over a domain D. Then the above integral is equivalent to

$$F = \int_D \frac{f(\vec{x})}{p(\vec{x})} p(\vec{x}) d\mu(\vec{x})$$

When discussing continuous probability eariler, we saw that the above interval is equal to

$$E\left[\frac{f(\vec{x})}{p(\vec{x})}\right]$$

the expected value of

$$\frac{f(\vec{x})}{p(\vec{x})}$$

with respect to a random variable distributed according to $p(\vec{x})$. This equality is true for any PDF on D, as long as $p(\vec{x}) \neq 0$ whenever $f(\vec{x}) \neq 0$.

We can also estimate the value of $E\left[\frac{f(\vec{x})}{p(\vec{x})}\right]$ by generating a number of random samples according to p, computing f/p for each sample, and finding the average of these values. As more and more samples are taken, this average is guaranteed to converge to the expected value, which is also the value of the integral. This process of averaging the value of $\frac{f(\vec{x})}{p(\vec{x})}$ for multiple random samples to estimate the value of an integral is called *Monte Carlo integration*.

It might seem like we should be wary of samples where $p(\vec{x}) = 0$ (to avoid dividing by 0), but conveniently, the probability of generating samples where $p = 0$ is 0, so we know that none of our random samples will cause this problem.

In summary, to evaluate the integral of a function f using Monte Carlo integration, we generate a number of random samples according to a PDF p, and compute the average of $\frac{f}{p}$ over all samples. As we have seen, this sample average will approximate the value of the integral. Futhermore, Monte Carlo integration is a consistent method: as the number of samples increases to ∞, the estimate is guaranteed to converge to the value of the integral.

Practical Issues

So far our discussion has been mostly theoretical. However, Monte Carlo integration is commonly implemented in computer programs; as a result, there are

a number of practical issues that we should address about the implementation of Monte Carlo integration.

Variance

Before we begin our discussion of practical implementation, we need to discuss the concept of *variance*. In our discussion of Monte Carlo integration, we glossed over an important practical issue: how accurate is the estimate of the integral that we get from Monte Carlo methods? We have seen that taking an infinite number of samples yields an infinitely accurate result, but infinite loops are prohibitive in practice. It would be useful to have some measure of the accuracy of our Monte Carlo estimate.

There is a quantity known as the *variance* of a random variable that measures the average squared distance between a sample and the mean of the distribution. For a random variable $X \sim p$ with domain D, we denote the variance of X with $V[X]$ and calculate the variance using the formula:

$$V[X] = E[(X - E[X])^2] = \int_D (x - E[X])^2 p(x) dx$$

Referring to the section on expected values, we can determine that

$$\begin{aligned} V[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X^2] - E[2XE[X]] + E[E[X]^2] \\ &= E[X^2] - 2E[X]^2 + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

This form of the variance is convenient to use in computations.

As with expected value, there are a number of rules for adding and multplying the variance of different distributions. The only rule we refer to later is the following: for independent random variables X and Y,

$$V[X + Y] = V[X] + V[Y]$$

We refer to this rule later when we discuss a method called stratified sampling.

The formulas above are analytical; we can also compute the variance for a set of random samples. Assume that we have a set of samples $x_1, x_2, \ldots, x_n$ of a random variable X. Then we could estimate $E[X]$ by finding the average of the samples (we call our estimate $\tilde{\mu}$):

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \approx E[X]$$

and we can use the sample mean $\tilde{\mu}$ to estimate the sample variance:

$$V[X] \approx \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \tilde{\mu})^2$$

In our estimate of the variance, we divide by $n - 1$ so that the estimate is unbiased. We must do this because $\tilde{\mu}$ is itself an estimate. See [1] for more information about this.

In general, we find that as the sample variance decreases, the sample mean becomes a more accurate estimate of the true mean of the underlying distribution. However, there is no definite rule relating the sample variance to the accuracy of the sample mean. For example, the variance of a single sample is always 0! Despite this, we say that sample sets with lower variance are more *efficient*: that is, we need to take fewer samples before the sample mean is a good approximation of the true mean.

There are a number of practical methods that use this principle to attempt to achieve better estimates of the mean with fewer random samples. We call these methods *variance reduction techniques*. In the rest of this document, we explain several of these techniques and the logic behind them.

Adaptive sampling

Given a set of samples, lower variance usually means a better estimate of the mean. We saw in the last section how the mean and variance of a set of samples are actually estimates of the true mean and variance of the underlying distribution. If we fix an upper bound on the sample variance, we can iteratively generate a new sample, recalculate the sample mean, and recalculate the sample variance until it falls below the bound we have set. When this happens, we can use the sample mean as a good estimate of the true mean, which in this case is also the value of the integral.

There are some difficulties that arise from the method. To fix an upper bound on the variance we need to have a good knowledge of the underlying problem; there is no universal bound that will give good estimates in all cases. We must also take care not to take too few samples for our estimate (remember that a single-sample set has 0 variance, which will be below any bound we can set!). Most importantly, even if the variance of our samples is low, there is no guarantee that the sample mean is a good estimate of the integral value because we are using non-deterministic methods. However, in practice adaptive sampling is effective in allowing us to take only as many samples as we need to accurately estimate the integral value.

Stratified sampling

Consider what happens when we reduce the size of the domain of a random variable. Any random samples we generate will be closer and closer together, and therefore the variance of those samples will decrease.

Therefore, another way to reduce variance is to divide the initial domain of integration D into smaller non-overlapping subsets. Then we generate the random samples in such a way that we guarantee that at least one sample lies in each subset.

As we saw earlier, for independent variables (such as our random samples), the total variance of the resulting distribution is the sum of the variances of the distribution over each smaller subset. In general, the smaller a region, the more constant a function over that region will be, and so our hope is that the variance in some of these regions will be decreased significantly. Therefore, when we sum the individual variances to recover the total variance of our sample set, the result should be smaller than the variance of a set of samples taken over the entire domain.

This technique of ensuring that smaller subregions of the entire domain each have a certain number of samples is called stratified sampling. A major benefit of stratified sampling is that there is no potential penalty: integral estimates from stratified samples can be no worse probabilistically than estimates from non-stratified samples.

One practical issue with stratification is that each subset of the original domain should have at least one sample. So although a large number of subsets can better reduce the overall variance of the samples, it also requires more samples to be taken. The ideal number of subsets depends on the problem; there is no global value that works well for every case.

Importance sampling

We mentioned earlier that we do not need to worry about generating random samples where the PDF p is 0. However, there is a related issue that can cause problems. What happens when we generate a random sample where the value of p is very small?

Remember that in Monte Carlo integration, we are finding the average of $\frac{f}{p}$ for a number of samples. Hence, if p is very small for a given sample, $\frac{f}{p}$ will be arbitrarily large. This large sample will greatly skew the sample mean away from the true mean, and the sample variance will also increase greatly. We will now need to take many more samples to cancel out these effects caused by this "bad sample."

One way to avoid such cases is to ensure that values of p are not small except where f is small; that is, by bounding the maximum value of $\frac{f}{p}$. In practice this can be difficult, but one general rule of thumb to follow is that p should "look like" f. The peaks and valleys of p should correspond to peaks and valleys of f. If we can design a PDF p that satisfies this rule, we can avoid troublesome samples of the kind we studied earlier.

This method of wisely choosing a PDF that corresponds to the integrand f is called importance sampling. In effect, we should get more random samples in areas where the value of f is high, and fewer samples where f is smaller. Note that this is the first method where we have discussed choosing the PDF function (as opposed to using a PDF that has already been selected). As we discussed earlier, we can use any PDF in our Monte Carlo method, subject only to the constraint that the PDF must be nonzero whereever the integrand f is nonzero.

Combined sampling

As we saw in the last section, we are free to choose a PDF for sampling, and some choices yield more efficient sampling methods than others. Ideally, the PDF p should be proportional to the integrand f.

However, designing such a PDF can be difficult. For example, if the integrand has two separate peaks, we cannot use a PDF with one peak and expect good sampling properties. However, in [2], Veach showed how different sampling techniques could be combined, in effect forming a new sampling distribution with low variance. Veach's method is really an extension of importance sampling that allows us to combine simpler PDFs into a resulting PDF that more accurately mimics the integrand function.

Assume we have a number of PDFs $p_1, p_2, \ldots, p_n$ over the same domain as the integrand. We will generate N random samples, with an equal number of samples (namely, $\frac{N}{n}$) coming from each p_i. This leads to the following equation:

$$E\left[\frac{f}{p}\right] = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{\frac{N}{n}}\frac{f(x_{ij})}{p(x_{ij})}$$

where x_{ij} is the jth sample taken from the PDF p_i.

This is equivalent to combining the PDFs p_i into a single combined distribution $\tilde{p}$, defined as follows:

$$\tilde{p}(x) = \frac{1}{n}\sum_{i=1}^{n} p_i(x)$$

Now we can simply generate random samples according to $\tilde{p}$ and use these samples in our Monte Carlo estimate. We can use this technique, which we call combined sampling, to build a relatively complicated PDF out of simpler PDFs.

Summary

In this document, we have studied discrete and continuous probability. Using the principles of continuous probability, we introduced the idea of Monte Carlo integration. Finally, we discussed some of the practical aspects of implementing Monte Carlo methods in computer software.

Monte Carlo integration in software is a very common method for estimating complicated integrals. To implement Monte Carlo integration, we only need a way to generate a set of random samples according to some probability distribution. If we can do this, we can implement basic Monte Carlo integration, which we can make more efficient using a number of practical techniques including adaptive sampling, stratification, importance sampling, and combined sampling.

References

[1] J.F. Kenney and E.S. Keeping. *Mathematics of Statistics, Part 2*. Van Nostrand, Princeton, NJ, second edition, 1951.

[2] Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 419–428. ACM Press, 1995.