Chapter 2

Basics of direct Monte Carlo

2.1 The probabilistic basis for direct MC

We start our study of Monte Carlo methods with what is usually called direct or simple Monte Carlo. We will refer to it as direct Monte Carlo.

We assume that our original problem can be put in the following form. There is a probability space (Ω, P) and a random variable X on it. The quantity we want to compute is the the mean of X which we denote by $\mu = E[X]$. (The sample space Ω is the set of possible outcomes. Subsets of Ω are called events, and the probability measure P a function that assigns a number between 0 and 1 to each event. Usually the probability measure is only defined on a σ-field $\mathcal{F}$, which is a sub-collection of the subsets of Ω, but we will not worry about this.) We emphasize that the original problem need not involve any randomness, even if it does the probability space we use for the Monte Carlo may not have been part of the original problem.

Let X_n be an independent, identically distributed (iid) sequence which has the same distribution as X. Recall that saying X_n and X are indentically distributed means that for all Borel sets B, $P(X_n \in B) = P(X \in B)$. A standard result in probability says that if they are equal for all B whic are intervals, then that is sufficient to insure they are equal for all Borel sets. The key theorem that underlies direct Monte Carlo is the Strong Law of Large Numbers.

Theorem 1 *Let X_n be an iid sequence such that $E[|X_n|] < \infty$. Let $\mu = E[X_n]$. Then*

$$P(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mu) = 1 \tag{2.1}$$

The conclusion of the theorem is often written as $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \mu$ a.s. Here a.s. stands for almost surely.

Suppose we want to compute the probability of an event rather than the mean of a random variable. If E is the event, we can think of its probability $P(E)$ as the expected value of the indicator function of E, i.e., $P(E) = E[1_E]$. In this case the sample is bunch of 0's and $1's$ indicating whether or not the outcome was in E. The sample mean is the fraction of outcomes that were in E is usually denoted $\hat{p}_n$. In this situation the strong law says that $\hat{p}_n$ converges to $P(E)$ almost surely.

We have seen several examples of direct Monte Carlo in the introduction. The integration examples and the network examples with independent edges were all examples of direct Monte Carlo. Note that in the network example where we were concerned with connectivity, we were computing a probability. The network example in which the edges are not independent and the self-avoiding walk example cannot be studied with direct Monte Carlo. Unless the network or walk is really small, there is no practical way to generate samples from the probability distribution .

2.2 Error estimation

The strong law says that we can approximate μ by generating a sample $X_1, X_2, \cdots, X_n$ and then computing the sample mean

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \tag{2.2}$$

Note that μ is a constant while $\hat{\mu}_n$ is a random variable. In the language of statistics, μ is a parameter and $\hat{\mu}_n$ is a statistic that estimates μ. We follow the notational convention of using a hat to denote the statistic that estimates the corresponding parameter.

The strong law tells us that $\hat{\mu}_n$ converges to μ but it does not tell us anything about how close $\hat{\mu}_n$ is to μ for a given value of μ. In any Monte Carlo simulation we do not actually let n go to infinity, we only use a (hopefully) large value of n. So it is crucial to address this question of how close our approximation is. $\hat{\mu}_n$ is a random variable. Since all the random variables X_i in our sample have mean μ, the mean of the sample mean is

$$E\hat{\mu}_n = \mu \tag{2.3}$$

In the language of statistics, μ_n is said to be an unbiased estimator of μ. We assume that the X_i have finite variance. Since they are identically distributed, they have the same variance

and we denote this common variance by σ. Since the X_i are independent, we have

$$var(\sum_{i=1}^{n} X_i) = n\sigma^2 \tag{2.4}$$

and so the variance of the sample mean is

$$var(\hat{\mu}_n) = \frac{1}{n^2} var(\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n} \tag{2.5}$$

Thus the difference of μ_n from μ should be of order $\sigma/\sqrt{n}$.

The $1/\sqrt{n}$ rate of convergence is rather slow. If you compute a one dimensional integral with Simpon's rule the rate of convergence is $1/n^4$. However, this rate of convergence requires some smoothness assumptions on the integrand. By contrast, we get the $1/\sqrt{n}$ rate of convergence in Monte Carlo without any assumptions other than a finite variance. While Simpson's rule is fourth order in one dimension, as one goes to higher dimensional integrals the rate of convergence gets worse as the dimension increases. In low dimensions with a smooth integrand, Monte Carlo is probably not the best method to use, but to compute high dimensional integrals or integrals with non-smooth integrands, Monte Carlo with it slow $1/\sqrt{n}$ convergence may be the best you can do.

The central limit theorem gives a more precise statemnt of how close $\hat{\mu_n}$ is to μ. Note that since $\hat{\mu}_n$ is random, even if n is very large, there is always some probability that $\hat{\mu}_n$ is not close to μ. The central limit theorem says

Theorem 2 *Let X_n be an iid sequence such that $E[|X_n|^2] < \infty$. Let $\mu = E[X_n]$ and let σ^2 be the variance of X_n, i.e., $\sigma^2 = E[X_n^2] - E[X_n]^2$. Then*

$$\frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu) \tag{2.6}$$

converges in distribution to a standard normal random variable. This means that

$$\lim_{n\to\infty} P(a \le \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu) \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \tag{2.7}$$

In terms of the sample mean, the central limit theorem says that $(\hat{\mu}_n - \mu)\sqrt{n}/\sigma$ converges in distrbution to a standard normal distribution.

The statement of the central limit theorem involves σ^2. It is unlikely that we know σ^2 if we don't even know μ. So we must also use our sample to estimate σ^2. The usual estimator of

the variance σ^2 is the sample variance. It is typically denoted by s^2, but we will denote it by s_n^2 to emphasize that it depends on the sample size. It is defined to be

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 \tag{2.8}$$

(s_n is defined to be $\sqrt{s_n^2}$.) A straightforward calculation show that $Es_n^2 = \sigma^2$, so s_n is an unbiased estimator of σ^2. This is the reason for the choice of $1/(n-1)$ as the normalization. It makes the estimator unbiased. An application of the strong law of large numbers shows that $s_n^2 \to \sigma^2$ a.s. Since $(\hat{\mu}_n - \mu)\sqrt{n}/\sigma$ converges in distribution to a standard normal distribution and that $s_n^2 \to \sigma^2$ converges almost surely to σ^2, a standard theorem in probability implies that $(\hat{\mu}_n - \mu)\sqrt{n}/s_n$ converges in distribution to a standard normal. (In statistics the theorem being used here is usually called Slutsky's theorem.) So we have the following variation on the central limit theorem.

Theorem 3 *Let X_n be an iid sequence such that $E[|X_n|^2] < \infty$. Let $\mu = E[X_n]$ and let σ^2 be the variance of X_n, i.e., $\sigma^2 = E[X_n^2] - E[X_n]^2$. Then*

$$\frac{(\mu_n - \mu)\sqrt{n}}{s_n} \tag{2.9}$$

converges in distribution to a standard normal random variable. This means that

$$\lim_{n\to\infty} P(a \le \frac{(\mu_n - \mu)\sqrt{n}}{s_n} \le b) = \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx \tag{2.10}$$

The central limit theorem can be used to construct confidence intervals for our estimate $\hat{\mu}_n$ for μ. We want to construct an interval of the form $[\hat{\mu}_n - \epsilon, \hat{\mu}_n + \epsilon]$ such that the probability μ is in this interval is $1 - \alpha$ where the confidence level $1 - \alpha$ is some number close to 1, e.g., 95%. Let Z be a random variable with the standard normal distribution. Note that μ belongs to $[\hat{\mu}_n - \epsilon, \hat{\mu}_n + \epsilon]$ if and only if $\hat{\mu}_n$ belongs to $[\mu - \epsilon, \mu + \epsilon]$. The central limit theorem says that

$$\begin{aligned} & P(\mu - \epsilon \le \hat{\mu}_n \le \mu + \epsilon) = P(-\epsilon \le \hat{\mu}_n - \mu \le \epsilon) \\ = \;& P(-\frac{\epsilon\sqrt{n}}{s_n} \le \frac{(\hat{\mu}_n - \mu)\sqrt{n}}{s_n} \le \frac{\epsilon\sqrt{n}}{s_n}) \approx P(-\frac{\epsilon\sqrt{n}}{s_n} \le Z \le \frac{\epsilon\sqrt{n}}{s_n}) \end{aligned} \tag{2.11}$$

Let z_c be the number such that $P(-z_c \le Z \le z_c) = 1 - \alpha$. Then we have $\frac{\epsilon\sqrt{n}}{s_n} = z_c$, i.e., $\epsilon = z_c s_n/\sqrt{n}$. Thus our confidence interval for μ is

$$\hat{\mu}_n \pm \frac{z_c s_n}{\sqrt{n}} \tag{2.12}$$

Stop - Wed, 1/20

Common choices for $1-\alpha$ are 95% and 99%, for which $z_c \approx 1.96$ and $z_c \approx 2.58$, respectively. The central limit theorem is only a limit statement about what happens as the sample size n goes to infinity. How fast the distribution in question converges to a normal distribution depends on the distribution of the original random variable X.

Suppose we are using direct Monte Carlo to compute the probability $p = P(E)$ of some event E. As discussed above we can think of this as computing the expected value of 1_E. The central limit theorem still applies, so we can construct confidence intervals for our estimate. The variance of 1_E is easily found to be $p(1-p)$. So we can use our estimate $\hat{p}_n$ of p to estimate the variance by $\hat{p}_n(1-\hat{p}_n)$ rather than s_n. Thus the confidence interval is

$$\hat{p}_n \pm \frac{z_c\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} \tag{2.13}$$

This is one possible problem with the above. Suppose p is very small, so small that np is of order one. There is some chance that none of the outcomes in our sample will lie in the event E and so $\hat{p}_n = 0$. In this case the above confidence interval would be $[0,0]$. This is clearly nonsense. The problem with the above derivation of the confidence interval in this case is that if np is not reasonable large, the central limit theorem is not a good approximation. When $\hat{p}_n = 0$ a reasonable confidence interval can be obtained as follows. Obviously the confidence interval should be of the form $[0, p_0]$. Note that $P(\hat{p}_n = 0) = (1-p)^n$. If p is not small enough this is a small probability. But $\hat{p}_n$ did in fact equal 0. So we will choose p_0 so that $(1-p_0)^n = \alpha$, where $1-\alpha$ is the confidence level, e.g., 95%. This is the same as $n\ln(1-p_0) = \ln(\alpha)$. Since p_0 is small, $\ln(1-p_0) \approx -p_0$. So we let

$$p_0 = \frac{-\ln(\alpha)}{n} \tag{2.14}$$

With $\alpha = 5\%$, $-\ln(\alpha)$ is approximately 3, so the confidence interval is $[0, 3/n]$.

Another approach to treating $\hat{p}_n$ is the Agresti confidence interval. See Owen for a discussion.

2.3 Accuracy vs. computation time

We have seen that the error in a Monte Carlo computation is proportional to $\sigma/\sqrt{n}$. Obviously we can reduce the error by increasing the number of samples. Note, however, that

to reduce the error by half we must increase the number of samples by a factor of four. Another way to improve the accuracy is to reduce the variance σ^2. We will study this topic in detail in a later chapters.

It is important to keep in mind that from a practical point of view, what is important is not how many samples are needed to achieve a given level of accuracy, but rather how much CPU time is need to achieve that accuracy. Suppose we have two Monte Carlo methods that compute the same thing. They have variances σ_1^2 and σ_2^2. Let τ_1 and τ_2 be the CPU time needed by the methods to produce a single sample. Then with a fixed amount T of CPU time we can produce $N_i = T/\tau_i$ samples for the two methods. So the errors of our two methods with be

$$\frac{\sigma_i}{\sqrt{N_i}} = \frac{\sigma_i\sqrt{\tau_i}}{\sqrt{T}} \tag{2.15}$$

Thus the method with the smaller $\sigma_i^2\tau_i$ is the better method.

It is also important to keep in mind that the time needed to compute a sample typically consists of two parts. First we have to generate a sample ω from the probability space. Then we have to evaluate the random variable X on ω. In many applications this second step can be as time consuming (or more so) than the first step. As an illustration, consider the network reliability example from the introduction. To generate a sample, all we have to do is generate a uniformly distributed random number from $[0, 1]$ for each edge and compare that random number with p_e to decide if the edge is included in the network or not. This take a time proportional to the number of possible edges. Finding the shortest path in the resulting network can take much longer.

Many problems involve a size or scale. In the network examples there is the number of edges. In the self-avoiding walk we have the number of steps. In integration problems there is the dimension. One should pay attention to how the times required for different parts of the Monte Carlo simulation depend on n. In particular one should keep in mind that while one part of the computation may be the most time consuming for moderate values of n, for larger values of n another part of the computation may start to dominate.

2.4 How good is the confidence interval

Our confidence interval was chosen so that the probability that the confidence interval contains the mean μ is the given confidence level $1 - \alpha$. In working this out we used the central limit theorem which is only an approximation which gets better as n increases. If the distribution of X is normal, then for any n the distribution of μ_n is a well-studied distribution known as student's t distribution. One can use this distribution (with $n - 1$ degrees of

freedom) in place of the standard normal. Of course this is only a reasonble thing to do if the distribution of X is approximately normal. Unless n is pretty small, the effect of using student's t instead of normal is negligible. With a confidence level of 95%, the critical z from the normal distribution is 1.960 while the critical t value for $n = 100$ is 1.984 and for $n = 20$ is 2.086. So unless you are in the very usual situation of doing a Monte Carlo with a very small number of samples, there is no need to use students t.

Even if n is not small, one can still worry about how much error the central limit theorem approximation introduces, i.e., how close is $P(\hat{\mu}_n - \frac{z_c s_n}{\sqrt{n}} \leq \mu \leq \hat{\mu}_n + \frac{z_c s_n}{\sqrt{n}})$ to $1 - \alpha$? Typically it is off by something of order $1/n$, so this is not really an issue for large values of n.

If we want to be really paranoid and we have an a priori upper bound on σ^2, then we can use Chebyshev's inequality. It says that for any $\epsilon > 0$,

$$P(|\hat{\mu}_n - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2} E[(\hat{\mu}_n - \mu)^2] = \frac{1}{\epsilon^2} var(\hat{\mu}_n) = \frac{1}{n\epsilon^2}\sigma^2 \tag{2.16}$$

Suppose we know that $\sigma^2 \leq M$. Then the above is bounded by $M/(n\epsilon^2)$. If we set this equal to α, we get $\epsilon = \sqrt{\frac{M}{n\alpha}}$. So if we take the confidence interval to be

$$\hat{\mu}_n \pm \sqrt{\frac{M}{n\alpha}} \tag{2.17}$$

then Chebyshev insures that the probability μ is not in this confidence interval is at most α. Comparing with our central limit theorem confidence interval we see that $z_c s_n$ has been replaced by $\sqrt{M/\alpha}$. If M is close to σ^2, then s_n is close to $\sqrt{M}$ and so the effect is to replace z_c by $1/\sqrt{\alpha}$. For $\alpha = 5\%$ this amounts to replacing 1.96 by 4.47. So we can get a confidence interval for which we are certain that the probability the interval does not capture μ is at most 5%, but at the expense of a much wider confidence interval.

Finally, if one does not have a bound on σ^2, but the random variable X is known to always be in the interval $[a, b]$, then one can use Hoeffding's inequality in place of Chebyshev. See Owen for more on this.

2.5 Estimating a function of several means

Sometimes the quantity we want to compute is the quotient of two means

$$\theta = \frac{E[X]}{E[Y]} \tag{2.18}$$

Suppose we generate an independent samples $\omega_1, \cdots, \omega_n$ from our probability space, distributed according to P. We then let $X_i = X(\omega_i)$ and $Y_i = Y(\omega_i)$. We want to use the

sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ to estimate θ. Note that in this approach X_i and Y_i are not independent. The natural estimator for θ is

$$\hat{\theta} = \frac{\overline{X}_n}{\overline{Y}_n} \tag{2.19}$$

Here we use the notation

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2.20}$$

for the two sample means. The nontrivial thing here is to find a confidence interval for our estimate. We do this using the Delta method.

There is nothing special about ratios. More generally we can consider a function of several means. So we assume we have a random vector $(X_1, X_2, \cdots, X_d)$ and we want to estimate a function of their means

$$\theta = f(E[X_1], \cdots, E[X_d]) \tag{2.21}$$

for some function f on R^d. Since we are using subscripts to indicate the components, we will now use superscripts to label our samples. We suppose we have n i.i.d. samples of our random vector. We denote them by $(X_1^i, \cdots, X_d^i)$ where $i = 1, 2, \cdots, n$. We let $\hat{\mu}_j^n$ be the sample mean of the jth component

$$\hat{\mu}_j^n = \frac{1}{n}\sum_{i=1}^{n} X_j^i \tag{2.22}$$

The natural estimator for θ is

$$\hat{\theta} = f(\hat{\mu}_1, \cdots, \hat{\mu_d}) \tag{2.23}$$

To get a confidence interval we need a multivariate version of the central limit theorem. We first recall a couple of definitions. The covariance of X and Y is
$cov(X, Y) = E[XY] - E[X]E[Y]$. Letting $\mu_x = E[X]$ and $\mu_y = E[Y]$, the covariance can also be written as

$$cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] \tag{2.24}$$

The correlation of X and Y is

$$\rho = \frac{cov(X, Y)}{\sigma_x \sigma_y} \tag{2.25}$$

where σ_x^2, σ_y^2 are the variance of X and Y. We let σ_j^2 be the variance of X^j, and let $\rho_{j,k}$ be the correlation of X_j and X_k. So the covariance of X_j and X_k is $\rho_{j,k}\sigma_j\sigma_k$. As $j, k = 1, 2, \cdots, d$ this gives a $d \times d$ matrix which we denote by Σ. It is called the covariance matrix.

Theorem 4 *Let $(X_1^n, \cdots, X_d^n)$ be an i.i.d. sequence of random vectors with finite variances. Let $\mu_j = E[X_j^n]$, let σ_j^2 be the variance of X_j^n and $\rho_{j,k}$ be their correlations. Then*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (X_1^k - \mu_1, \cdots, X_d^k - \mu_d) \tag{2.26}$$

converges in distribution to a multivariate normal random variable with zero means and covariance matrix Σ.

The crucial fact we will need about a multivariate normal distribution is the following. Let $(Z_1, Z_2, \cdots, Z_d)$ be a multivariate normal distribution with zero mean and covariance matrix Σ. Then the linear combination $\sum_{j=1}^{d} c_j Z_j$ is a normal random variable with mean zero and variance equal to

$$\sum_{j,k=1}^{d} c_j c_k \Sigma_{j,k} \tag{2.27}$$

Now we turn to the delta method. The idea is simple. We just use a first order Taylor expansion of f about $(\mu_1, \cdots, \mu_d)$ and use the central limit theorem.

$$\hat{\theta} = f(\hat{\mu}_1, \cdots, \hat{\mu_d}) \approx f(\mu_1, \cdots, \mu_d) + \sum_{j=1}^{d} f_j(\mu_1, \cdots, \mu_d)(\hat{\mu_j} - \mu_j) \tag{2.28}$$

where f_j denotes the jth partial derivative of f. The mean of the right side is $f(\mu_1, \cdots, \mu_d)$. This says that to first order the estimator $\hat{\theta}$ is an unbiased estimator. It is not exactly unbiased. By looking at the second order Taylor expansion one can see that the bias is of order $1/n$. The variance of $\hat{\theta}$ is

$$var(\hat{\theta}) = \sum_{j,k=1}^{d} f_j(\mu_1, \cdots, \mu_d) f_k(\mu_1, \cdots, \mu_d) \Sigma_{j,k} \tag{2.29}$$

This can be written more succintly as $(\nabla f, \Sigma \nabla f)$

Before our application of the central limit theorem involved σ which we do not know. So we had to replace σ by s_n. Here there are several things in the above that we do not know: μ_i and Σ. We approximate $f_j(\mu_1, \cdots, \mu_d)$ by $f_j(\hat{\mu_1}, \cdots, \hat{\mu_d})$. We denote the resulting approximation of ∇f by $\hat{\nabla} f$. We approximate Σ by $\hat{\Sigma}$ where the entries are

$$\hat{\Sigma}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} (X_j^i - \hat{\mu}_j)(X_k^i - \hat{\mu}_k) \tag{2.30}$$

So our estimate for the variance of $\hat{\theta}$ is $(\hat{\nabla} f, \hat{\Sigma}\hat{\nabla} f)$. Thus the confidence interval is

$$\hat{\theta} \pm \frac{z_c}{\sqrt{n}} \sqrt{(\hat{\nabla} f, \hat{\Sigma}\hat{\nabla} f)} \tag{2.31}$$

We can now return to the problem of estimating the ratio $\theta = E[X]/E[Y]$. So $n = 2$ and $f(x, y) = x/y$. Some computation leads to the follows. The variance of $\hat{\theta}$ is approximately

$$(\hat{\nabla} f, \hat{\Sigma}\hat{\nabla} f) = \frac{1}{n} \frac{\sum_{i=1}^{n} (Y_i - \hat{\theta} X_i)^2}{n\overline{X}^2} \tag{2.32}$$

where $\overline{X}$ and $\overline{Y}$ are the sample means for X and Y and $\hat{\theta} = \overline{X}/\overline{Y}$.

2.6 References

Most books on Monte Carlo include the topics in this section. Our treatment follows Owen closely. Fishman's *A first course in Monte Carlo* has some nice examples, one of which we have used (networks).