The Monte Carlo Method

Introduction

Let X denote a set of numbers or vectors, and let real-valued function $g(x)$ be defined over X. There are essentially three fundamental computational problems associated with X and g.

Integral Evaluation Evaluate $\int_X g(x)dx$. Note that if X is discrete, then integration can be replaced by summation.

Optimization Find $x_0 \in X$ that minimizes g over X. In other words, find $x_0 \in X$ such that, for all $x \in X$, $g(x_0) \leq g(x)$.

Mixed Optimization and Integration If g is defined over the space $X \times \Theta$, where Θ is a **parameter space**, then the problem is to determine $\theta \in \Theta$ for which $\int_X g(x, \theta)dx$ is a minimum.

The mathematical disciplines of real analysis (e.g. calculus), functional analysis, and combinatorial optimization offer deterministic algorithms for solving these problems (by *solving* we mean finding an exact solution or an approximate solution) for special instances of X and g. For example, if X is a closed interval of numbers $[a, b]$ and g is a polynomial, then

$$\int_X g(x)dx = \int_a^b g(x)dx = G(b) - G(a),$$

where G is an antiderivative of g.

As another example, if $G = (V, E, w)$ is a connected weighted simple graph, X is the set of all spanning trees of G, and $g(x)$ denotes the sum of all the weights on the edges of spanning tree x, then Prim's algorithm may be used to find a spanning tree of G that minimizes g. Such a tree is called a **minimum spanning tree**.

However, many problems in science, engineering, and business do not have known deterministic solutions, and **Monte Carlo simulation** refers to the general technique of finding approximate solutions to these fundamental problems by randomly sampling from X, evaluating $g(x)$ at the sample points, and using the resulting values to find an approximate solution. In this lecture we introduce **Independent Monte Carlo (IMC)** simulation, where samples from X are drawn independently. In a later lecture we study **Markov-Chain Monte Carlo (MCMC)** simulation, where samples are drawn with the help of a Markov chain.

Independent Monte Carlo Simulation

The first step of IMC is to represent X has a random variable, by assigning it a density function $\omega(x)$. Note that for simplicity, we are assuming X is continuous. But in the case X is discrete, then $\omega(x)$ becomes a probability distribution, and the statistics of X are calculated with sums instead of integrals. We may also assume that there is a function $h(x)$ defined over X, and for which we desire to compute $\lambda = E[h(X)]$.

Example 1. Suppose the goal is approximate the integral $\int_0^\infty g(x)dx$, for some integrable function $g(x)$. Then by defining $X = [0, \infty)$ with density function $\omega(x) = e^{-x}$, and $h(x) = g(x)/\omega(x)$, we have

$$\lambda = E[h(X)] = \int_0^\infty \frac{g(x)}{\omega(x)}\omega(x)dx = \int_0^\infty g(x)dx.$$

Therefore, we may obtain a good approximation of $\int_0^\infty g(x)dx$ by obtaining a good approximation of the statistic $\lambda = E[h(X)]$. Often, the choice of $\omega(x)$ will affect the approximation accuracy.

Posterior statistics

Once X, $\omega(x)$, and $h(x)$ have been identified, the next step is to draw n independent random samples $x_1, \ldots, x_n$ from X that are distributed according to $\omega(x)$. These samples may then be used to compute **posterior statistics** for $h(X)$; i.e. statistics that are obtained *a posteriori*, meaning after having observed the data. These statistics are defined as follows.

Sample Mean used to approximate $E[h(X)]$

$$\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

Sample Variance used to approximate $\text{Var}(h(X))$

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} h^2(x_i) - n\hat{\lambda}_n^2\right]$$

Standard Error If Z is defined as $Z = \frac{1}{n}(h(x_1) + \cdots + h(x_n))$, then the standard error represents an approximation of the standard deviation of Z. In other words, since $E[Z] = E[h(X)]$, if we were to repeat the IMC experiment several times, then it would represent the average observed deviation from $E[h(X)]$.

$$\text{s.e.} = \frac{\hat{\sigma}_n}{\sqrt{n}}$$

Relative Error The ratio of the standard error to the sample mean

$$\text{r.e} = \frac{\hat{\sigma}_n}{\sqrt{n}\hat{\lambda}_n}$$

To simplify notation, when the value of n is understood or not important, we write $\hat{\lambda}$ instead of $\hat{\lambda}_n$. Similarly, $\hat{\sigma}^2$ denotes $\hat{\sigma}_n^2$. The justification of the posterior statistics representing good approximations of the prior statistics for increasingly large values of n is justified by the following theorem(s) from probability.

Theorem 1. Asymptotic Properties of Posterior Statistics. Assume that $x_1, \ldots, x_n$ are drawn independently from X, and are distributed according to $\omega(x)$. Moreover, suppose that $E[h^4(X)] < \infty$. Define $\lambda = E[h(X)]$ and $\sigma^2 = \text{Var}(h(X))$. Then

Law of Large Numbers For every $\epsilon > 0$ and $\delta > 0$, n can be chosen sufficiently large so that

$$P(|\hat{\lambda} - \lambda| > \epsilon) < \delta.$$

Convergence of Sample Variance For every $\epsilon > 0$ and $\delta > 0$, n can be chosen sufficiently large so that

$$P(|\hat{\sigma}^2 - \sigma^2| > \epsilon) < \delta.$$

Central Limit Theorem As n increases, the CDF of

$$Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\sigma^2/n}}$$

converges to the CDF $\Phi(z)$ of the standard normal distribution.

Central Limit Theorem with Sample Variance Replacing Prior Variance As n increases, the CDF of

$$Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\sigma}^2/n}}$$

converges to the CDF $\Phi(z)$ of the standard normal distribution.

The Central Limit Theorem using sample variance can be used to construct a **δ-confidence interval** for prior statistic λ, i.e. an intervaly I centered about $\hat{\lambda}$, such that $P(\lambda \in I) \geq 1-\delta$. Here probability measure P pertains to the entire IMC simulation that involves sampling n values from X. In words it is saying, "if the just-completed Monte Carlo simulation were repeated several times, then in a fraction $1 - \delta$ of the experiments would one find λ inside I.

Example 2. Suppose that for $n = 10000$, a Monte Carlo simulation yields $\hat{\lambda} = 4.6$, and a standard error s.e. $= 0.93$. We compute the $\delta = 0.95$ confidence interval for λ. By Theorem 1, the random variable

$$Z = \frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\sigma}^2/n}} = \frac{\hat{\lambda} - \lambda}{\text{s.e.}}$$

has approximately the distribution of $N(0, 1)$. Moreover, since $N(0, 1)$ is two-tailed and symmetric about the y-axis, it suffices to find a z value for which $\Phi(z) \leq 0.975$. Here $z \approx 1.96$. Thus the 0.95-confidence interval for Z is $I = (-1.96, 1.96)$. In other words,

$$-1.96 \leq \frac{\hat{\lambda} - \lambda}{\text{s.e.}} \leq 1.96 \Longrightarrow$$

$$(-1.96)(0.93) + 4.6 \leq \lambda \leq 4.6 + (1.96)(0.93).$$

Therefore, the 0.95-confidence interval for λ is $I = (2.77, 6.43)$.

Using relative error as a stopping criterion

Quite often the goal is to approximate λ so that, with high probability, $\hat{\lambda}$ is within a desired percentage of the actual value. In other words, for some $\delta > 0$ and $\theta > 0$, the goal is for

$$P(|\hat{\lambda} - \lambda| \geq \theta\lambda) \leq \delta$$

to be true. For n sufficiently large, we claim that a good indicator for the truth of the above statement is for

$$\text{r.e.} = \frac{\text{s.e.}}{\hat{\lambda}} \leq \frac{\theta}{\Phi^{-1}(1 - \frac{\delta}{2})}.$$

To see this, note that

$$P(|\hat{\lambda} - \lambda| \geq \theta\lambda) = P(|\frac{\hat{\lambda} - \lambda}{\text{s.e.}}| \geq \frac{\theta\lambda}{\text{s.e.}}).$$

But for large vaues of n, Statement 4 of Theorem 1 implies that

$$Z = \frac{\hat{\lambda} - \lambda}{\text{s.e.}}$$

is approximately standard normal. This implies that, a sufficient conditon for

$$P(|\hat{\lambda} - \lambda| \geq \theta\lambda) \leq \delta$$

to be true, is for

$$\Phi^{-1}(1 - \frac{\delta}{2}) \leq \frac{\theta\lambda}{\text{s.e.}},$$

which is true iff

$$\frac{\text{s.e.}}{\lambda} \leq \frac{\theta}{\Phi^{-1}(1 - \frac{\delta}{2})}.$$

Finally, the left side can be estimated using relative error s.e./$\hat{\lambda}$.

Now suppose that, upon using n samples, we find that

$$\text{r.e.} = \frac{\text{s.e.}}{\hat{\lambda}} > \frac{\theta}{\Phi^{-1}(1-\frac{\delta}{2})}.$$

As a next step, we compute how many additional samples m might be sufficient in order to obtain

$$\text{r.e.} = \frac{\text{s.e.}}{\hat{\lambda}} \leq \frac{\theta}{\Phi^{-1}(1-\frac{\delta}{2})}.$$

First, we assume that, after acquiring m additional samples and re-computing $\hat{\lambda}$ and $\hat{\sigma}$, that these values do not significantly change. If this is the case, then the relative error will change by a factor of $\sqrt{\frac{n}{m+n}}$. Moreover we want,

$$\sqrt{\frac{n}{m+n}}\frac{\hat{\sigma}_n}{\hat{\lambda}_n} \leq \frac{\theta}{\Phi^{-1}(1-\frac{\delta}{2})}$$

which is true provided

$$m \geq n\left[\frac{\Phi^{-1}(1-\frac{\delta}{2})\hat{\sigma}_n}{\theta\hat{\lambda}_n}\right]^2 - n.$$

Example 3. Suppose the goal is approximate λ to within 10% of its true value with 95% certainty. Suppose that for $n = 10000$, a Monte Carlo simulation yields $\hat{\lambda} = 4.6$, and a standard error s.e. $= 0.93$. How many additional samples should be taken as a next step towards achieving this goal?

Example 3 Solution.

Applying IMC to Integral Approximation

Returning to the problem of integration evaluation, Theorem 1 can be applied to $\int_X g(x)dx$ by choosing a distribution $\omega(x)$ and letting $h(x) = \frac{g(x)}{\omega(x)}$. Then, assuming $\int_X (\frac{g(x)}{\omega(x)})^4 \omega(x)dx < \infty$, Theorem 1 implies that IMC can be used to approximate

$$\int_X g(x)dx = \int_X \frac{g(x)}{\omega(x)} \omega(x)dx.$$

Example 4. Approximate $\int_0^1 \cos x e^{\sin x} dx$ by performing IMC sampling with

$$n = 10, 100, 500, 10^3, 10^4, 10^5, 10^6, 10^7$$

and for each experiment, compute the sample mean, standard error, and relative error. Use $\omega(x) = 1$ for all $x \in [0, 1]$. Note: by direct calculation the above integral evaluates to approximately 1.31977682472.

```
Number of samples: 10
Sample mean: 1.31629
Sample variance: 3.860359
Standard error: 0.6213179
Relative error: 0.4720222
0.95-confidence interval: [0.09852892,2.53405]

Number of samples: 100
Sample mean: 1.313812
Sample variance: 3.504397
Standard error: 0.1872004
Relative error: 0.1424865
0.95-confidence interval: [0.9469056,1.680717]

Number of samples: 1000
Sample mean: 1.320727
Sample variance: 3.509703
Standard error: 0.05924274
Relative error: 0.04485617
0.95-confidence interval: [1.204613,1.436841]

Number of samples: 1e+06
Sample mean: 1.319877
Sample variance: 3.50137
Standard error: 0.001871195
Relative error: 0.001417704
0.95-confidence interval: [1.316209,1.323544]
Processing time: 17.44
```

Cost and Variance Ratios

Notice that the standard error provides an estimate of the amount that an IMC experiment using n samples is expected to deviate from the exact value λ. It would thus seem wise to design the IMC experiment in such a way that reduces this standard error. There are essentially two approaches for accomplishing this.

1. increase the value of n
2. decrease the value of σ

Although the first approach has the advantage of seeming straightforward, it has the disadvantage of increasing the computational cost in proportion to the increase in samples. The second approach is called **variance reduction**, and has the disadvantage of requiring more insight into the specific problem that is being solved. On the other hand, it offers a true computational savings if a sampling technique can be devised that reduces the sample variance, yet does not significantly increase the computational cost of sampling from X. We refer to this cost as **sampling cost**. For example, if a sampling technique reduced the variance by one half, yet only incurs a ten percent increase in sampling cost, then this plan is preferred over the original one. We use sampling cost and variance ratios as measures for comparing two different sampling techniques.

Given two MC sampling techniques T_1 and T_2, let c_i, $i = 1, 2$, denote the sampling cost (in terms of computing time or number of elementary operations) that is required to generate a sample using T_i. Similarly, let σ_i^2 denote the variance of the samples that is induced by T_i. Then the **cost ratio (CR)** and **variance ratio (VR)** of T_1 and T_2 are respectively defined as $\mathrm{CR} = \frac{c_1}{c_2}$ and $\mathrm{VR} = \frac{\sigma_1^2}{\sigma_2^2}$.

Theorem 2. Given sampling techniques T_1 and T_2, along with their cost and variance ratios CR and VR,

1. If $\mathrm{CR} \cdot \mathrm{VR} < 1$, then, $(1 - \sqrt{\mathrm{CR} \cdot \mathrm{VR}}) \times 100$ gives the percent decrease in standard error when using T_1 instead of T_2 for an IMC experiment having a fixed computational cost.
2. If $\mathrm{CR} \cdot \mathrm{VR} > 1$, then $(1 - \frac{1}{\sqrt{\mathrm{CR} \cdot \mathrm{VR}}}) \times 100$ gives the percent decrease in standard error when using T_2 instead of T_1 for an IMC experiment having a fixed computational cost.
3. If $\mathrm{CR} \cdot \mathrm{VR} = 1$, then T_1 and T_2 induce the same standard error for an IMC experiment having some fixed computational cost.

Proof of Theorem 2. It suffices to prove the first statement (using ≤ 1 instead of < 1). The second statement follows by i) inverting the ratios and ii) invoking Statement 1 for the favorable technique T_2.

Let C denote the cost limit that is afforded either technique. Let c_i and σ_i^2, $i = 1, 2$ denote the respective sampling cost and variance for T_1 and T_2. Then T_1 can afford C/c_1 samples for a standard error of

$$\sigma_1/\sqrt{C/c_1} = \frac{\sqrt{c_1\sigma_1^2}}{\sqrt{C}}.$$

Similarly T_2 provides a standard error of $\frac{\sqrt{c_2\sigma_2^2}}{\sqrt{C}}$. Next, notice that the ratio of these standard errors is equal to

$$\frac{\sqrt{c_1\sigma_1^2}}{\sqrt{c_2\sigma_2^2}} = \sqrt{\mathrm{CR}\cdot\mathrm{VR}} \leq 1.$$

Finally, for any ratio $x/y < 1$, the fractional reduction that x represents in terms of y is $(1 - x/y)$. For example, if $x/y = 0.7$, then x represents a fractional reduction of 0.3, or 30%. Therefore, T_1 provides fractional reduction in standard error equal to $1 - \sqrt{\mathrm{CR}\cdot\mathrm{VR}}$.

Example 5. A new sampling technique has been proven to reduce variance by 40%, while only incurring a 5% increase in computational cost, relative to some existing technique. Provide the percentage decrease in standard error (relative to the current technique) that results from using this new technique.

Example 5 Solution.

Variance Reduction

Variance reduction refers to the development of sampling techniques that on average reduce the sample variance, and hence the standard error. One such technique is **importance sampling** which is an attempt at reducing variance by using distributions that better fit the integrand $g((x)$. Recall that Theorem 1 implies that IMC can be used to approximate

$$\lambda = \int_X g(x)dx = \int_X \frac{g(x)}{\omega(x)}\omega(x)dx,$$

assuming that appropriate conditions are placed on probability distribution $\omega(x)$ (e.g. $(g/\omega)^4$ is integrable). So one might ask, "what is the best choice of ω in terms of reducing the variance that occurs in a Monte Carlo sampling plan?". To answer this, consider $\omega(x) = g(x)/\lambda$. Assuming that g is nonnegative, this represents a probability density function over X, since

$$\int_X g(x)/\lambda dx = \lambda/\lambda = 1.$$

Moreover, we see that, by substituting ω for g/λ, the function g/ω equals the constant λ, and hence Monte Carlo sampling will yield $\hat{\lambda}_n = \lambda$, with zero variance. Unfortunately, however, this ideal distribution is not known, since it depends on λ.

So given that the "ideal distribution" is not known, the next best thing is to first approximate this distribution as well as possible, and then proceed with Monte Carlo sampling. We give an example of this for integration of a single real variable over a closed interval $[a, b]$. The method used to approximate the ideal distribution is known as the **partitioning method**, which requires the following steps.

1. Partition the interval $[a, b]$ into k sub-intervals of equal length.
2. For each sub-interval $[a_i, b_i]$, $i = 1, \ldots, k$, peform a short (for example, using 100 samples) Monte Carlo experiment to approximate

   $$\lambda_i = \int_{a_i}^{b_i} g(x)dx.$$

   Call the approximation $\hat{\lambda}_i$, and let $\hat{\lambda}$ denote the sum of these approximations.
3. Then the density function $\hat{\omega}(x)$ provides an approximation to the ideal distribution, where, for x in sub-interval i, $\hat{\omega}(x) = \frac{\hat{\lambda}_i}{(b_i - a_i)\hat{\lambda}}$.

One can check that $\hat{\omega}(x)$ is indeed a density function byintegrating it over the interval $[a, b]$. Why would it make for a good approximation to the ideal? To see why, notice that, for each sub-interval

i, $\hat{\omega}(x)$ is really just a constant multiplied with $\hat{\lambda}_i$. Moreover, $\hat{\lambda}_i/(b_i - a_i)$ gives an approximation of the mean value of g in the i th sub-interval. Now if k is chosen large enough, for well-behaved g, $\hat{\lambda}_i/(b_i - a_i)$ will make for a very good approximation of the mean-value, in that the variance will be small, since g does not change much in the small sub-interval. Thus, the ratio, $g(x)/\hat{\omega}(x)$ will tend to have very little variance, and hence a better approximation of λ should be obtained.

Example 6. Apply the partitioning method with a $k = 10$ partition size to approximate the integral from $[0, 10]$ of the function $g(x) = \cos x e^{\sin x} dx$ from Example 1.

Example 6 Solution. We set $n = 10^6 - S$, where S is the total number of samples needed to approximate $\hat{lambda}_1, \ldots, \hat{\lambda}_{10}$. Here we use $S = 10 \times 100 = 1000$, so $n = 999,000$.

Use the values $k = 10, 100, 1000, 5000$. Report on the mean, variance, standard error, and 90% confidence intervals for each k, and compare with the same statistics that are obtained when using the standard IMC algorithm. Use 10^6 samples. Note: the exact value for λ is close to -0.41959. Furthermore, the cost ratio between the standard IMC sampling plan versus the sampling plan required for partitioning has been approximated between 0.60 and 0.70. Hence, the standard plan can have a cost savings of near 40%.

Example 6 Solution.

```
Standard Monte Carlo
mean: -0.41287
variance: 95.35378
standard error: 0.00976
90% confidence interval: [-0.42888,-0.39685]

Partitioning (k=10)
mean: -0.31065
variance: 0.01245
standard error: 0.11159
90% confidence interval: [-0.31083,-0.31047]

Partitioning (k=100)
mean: -0.42238
variance: 0.00022
standard error: 0.00001
90% confidence interval: [-0.42240,-0.42236]

Partitioning (k=1000)
mean: -0.41955
variance: 0.0
standard error: 0.0
90% confidence interval: [-0.41955,-0.41954]

Partitioning (k=5000)
mean: -0.41959
variance: 0.0
standard error: 0.0
90% confidence interval: [-0.41959,-0.41959]
```

Approximating the value of π

The most common way of approximating π via Monte Carlo simulation is to define the function $g(x, y)$ over the square $\mathcal{S}$ of length 2 that is centered about the origin. Notice that the unit circle centered about the origin is inscribed within the square. Then define $g(x, y)$ as

$$g(x,y) = \begin{cases} 1 & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\omega(x, y) = \omega(x)\omega(y) = (1/2)(1/2) = 1/4.$$

Then, letting $h(x, y) = \frac{g(x,y)}{\omega(x,y)}$, we have

$$h(x,y) = \begin{cases} 4 & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Example 7. Compute $E[h(X, Y)]$ and $\text{Var}(h(X, Y))$.

Example 7 Solution. $h(X, Y) = 4I(X, Y)$, where $I(X, Y)$ equals 1 if (X, Y) lies in the unit circle, and 0 otherwise. Thus

$$E[h(X, Y)] = 4E[I(X, Y)] = 4P(X^2 + Y^2 \leq 1) = 4(\pi/4) = \pi,$$

where the 2nd-to-last equality is due to ω being uniformly distributed over the square, and the circle area comprising a fraction of $\pi/4$ of the square's area.

Finally,

$$E[h^2(X, Y)] = 16E[I^2(X, Y)] = 16E[I(X, Y)] = 16P(X^2 + Y^2 \leq 1) = 16(\pi/4) = 4\pi.$$

Thereofore, $\text{Var}(h(X, Y)) = 4\pi - \pi^2 = \pi(4 - \pi)$.

Now suppose that, instead of circumscribing the unit circile with a square, we use an octagon instead. Moreover, a sample (X, Y) is now drawn from a triangle T whose bottom side is determned by one side of the octagon, and has a unit height that is represented by a radius of the unit circle that bisects the angle opposite the bottom side. Now $g(x, y)$ is defined as

$$g(x,y) = \begin{cases} 8 & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

We use 8 since $\int_T g(x, y)dxdy = (8)(\pi/8) = \pi$. Moreover, since the area of T is

$$(1/2)\text{base} \times \text{height} = (1/2)(2\tan(22.5°)(1) = \tan(22.5°) = 0.414,$$

and the area of the circle that lies in T is $\pi/8 = 0.393$, notice how over 94% of the area of T is occupied by the circle. This will lead to a significant reduction in variance for $h(X, Y)$. Moreover, in the next example, we show that (X, Y) can be sampled using $\omega(x, y) = 1/\tan(22.5°)$.

Now, letting $g(x,y) = 8$ if $x^2 + y^2 \leq 1$, and 0 otherwise, then $h(x,y) = \frac{g(x,y)}{\omega(x,y)}$ has rule

$$h(x,y) = \begin{cases} 8\tan(22.5^\circ) & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Moreover,

$$E[h(X,Y)] = 4E[I(X,Y)] = 8\tan(22.5^\circ)P(X^2 + Y^2 \leq 1) = 8\tan(22.5^\circ)(\pi/8)/\tan(22.5^\circ) = \pi,$$

and

$$E[h^2(X,Y)] = 64\tan^2(22.5^\circ)E[I^2(X,Y)] = 64\tan^2(22.5^\circ)E[I(X,Y)] =$$
$$64\tan^2(22.5^\circ)P(X^2 + Y^2 \leq 1) = 8\tan(22.5^\circ)\pi.$$

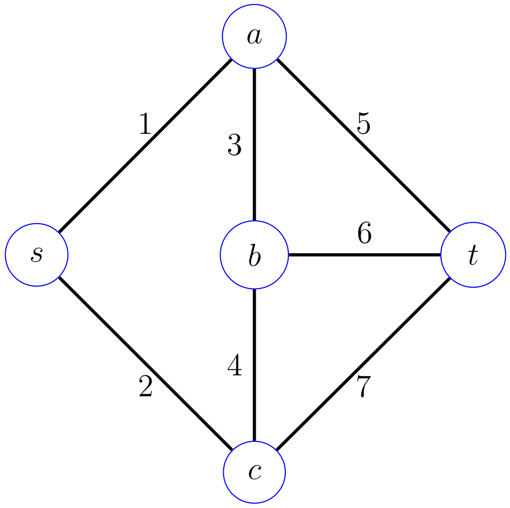
Thereofore, $\text{Var}(h(X,Y)) = 8\tan(22.5^\circ)\pi - \pi^2 = \pi(8\tan(22.5^\circ) - \pi)$. This yields a variance ratio of

$$\text{VR} = \frac{8\tan(22.5^\circ) - \pi}{4 - \pi} = 0.20.$$

Thus, the new sampling method provides an 80% reduction in variance.

Example 8. Provide an efficient method for uniformly sampling points (x,y) from triangle T described above.

Estimating the reliability of a network



It is often essential to estimate the likelihood that a communcation network will become disconnected at some point in time. As an example, consider a simple graph similar to the one above, where each edge is labeled 1 through m, and suppose we are interested in determining the probability that node s will become disconnected from node t. We assume that each edge i of the graph has a probability p_i of being removed, and that the event of edge i being removed is independent of the event that edge j is removed for all $1 \leq i < j \leq m$. Then the state of the network at any given time can be represented by the vector $\overline{x} = (x_1, \ldots, x_m)$, where $x_i \in \{0, 1\}$ indicates whether or not edge i is working (i.e. has not been removed), for all $i = 1, \ldots, m$. Let S denote the set of all possible network states, while $D \subseteq S$ represents the event that s is disconnected from t. Let $h(\overline{x}) = 1$ iff $\overline{x} \in D$, and

$$\pi(\overline{x}, \overline{p}) = \prod_{i=1}^{m} \pi_i(x_i, p_i)$$

denote the probability of $\overline{x}$, where

$$\pi_i(x_i, p_i) = \begin{cases} p_i & \text{if } x_i = 1 \\ 1 - p_i & \text{otherwise} \end{cases}$$

The goal then is to estimate

$$\lambda(\overline{p}) = E[D] = P(\overline{x} \in D) = \sum_{\overline{x} \in S} h(\overline{x}) \pi(\overline{x}, \overline{p}) = \sum_{\overline{x} \in D} \pi(\overline{x}, \overline{p}).$$

Then we define $g(\overline{x})$ as $g(\overline{x}) = h(\overline{x}) \pi(\overline{x}, \overline{p})$.

One problem with using IMC to estimate $\lambda(\overline{p})$ is in the number of samples needed to within a desired percentage of the true value. This is due to the fact that $\lambda(\overline{p})$ will be very small for most practical networks. Notice also that, for this problem, $\lambda(\overline{p})$ represents a Bernoulli probability (i.e. the probability of s being disconnected from t). In general, estimating a very small Bernoulli probability

p may require a large number of samples in order to within θp of p. Moreover, since the variance of a Bernoulli random variable is $\sqrt{p(1-p)}$, we know from above that approximately

$$\left[\frac{\Phi^{-1}(1-\frac{\delta}{2})\sqrt{p(1-p)}}{\theta p}\right]^2$$

samples are needed to get within θp with probability $1-\delta$. For example, if $\delta = 0.01$, and $p = 0.0001$, and $\theta = 0.1$, then

$$\left[\frac{\Phi^{-1}(1-\frac{\delta}{2})\sqrt{p(1-p)}}{\theta p}\right]^2 = \left[\frac{\Phi^{-1}(0.995)\sqrt{0.0001(0.9999)}}{0.00001}\right]^2 = 6,625,476.$$

This number may seem reasonable for some simulations, but, in the case of network reliability, one sample requires checking a network for connectivity which takes $\mathrm{O}(m)$ time, and performing this task millions of times can seem prohibitive for large values of m. Therefore, we desire to find a different distribution $\pi(\overline{x}, \overline{q})$ that can reduce the variance (notice that the variance is the only parameter that we are able to change, assuming that someone is demanding an estimate that lies within $\theta\lambda$ with probability at least $1-\delta$).

To simplify the analysis, assume that $p_i = p > 1/2$ for all $i = 1, \ldots, m$. Then we look to sample with $\pi(\overline{x}, q)$ instead of $\pi(\overline{x}, p)$. Intuitiely, we should $q < p$, since $q \geq p$ produces an increase in variance (event D becomes even less probabile). But how small should we make q?

To help answer the above question, recall the primary lesson of importance sampling: when choosing an $\omega(\overline{x})$ for approximating

$$\sum_{\overline{x} \in X} (g(\overline{x})/\omega(\overline{x}))\omega(\overline{x}),$$

one should choose an $\omega(\overline{x})$ that is proportional to $g(\overline{x})$. Notice that this is far from the case when using $\pi(\overline{x}, p)$, since this distribution places most of the weight at points $\overline{x}$ for which $g(\overline{x}) = 0$. Thus, ideally q would be chosen so that i) there is little to no weight assigned to connected graphs, and ii) the weight assigned to a disconnected graph is proportional to the likelihood of that graph occurring in accordance with $\pi(\overline{x}, p)$. Now, in practice, p is usually very large, and so a network will fail with small probability. Moreover, when it does fail, there will likely be as few destroyed links as possible (since p is so high). Therefore, q should be made just small enough so that there are a sufficient number of destroyed links to create a disconnected graph. Note that a **minimal cut set** for a graph G.is defined as a minimum set of edges S which, if removed from G, disconnect G. Also $r = |S|$ is defined as the **min-cut number** of G. Finally, since a disconnected graph is more likely to have exactly r edges destroyed than having any other number destroyed, it follows that the ideal value for q appears to satisfy $qm = m - r$, or $q = 1 - r/m$. In what follows, we set out to quantitatively verify the above qualitative analysis.

To begin, consider the variance of $h(\overline{x})$ with respect to $\pi(\overline{x}, q)$. We have

$$\mathrm{Var}\left(\frac{g(\overline{x}, p)}{\pi(\overline{x}, q)}\right)$$

with respect to probability distribution $\pi(\overline{x}, q)$ can be expressed as

$$\operatorname{Var}\left(\frac{g(\overline{x}, p)}{\pi(\overline{x}, q)}\right) = \sum_{\overline{x}\in S}\left(\frac{h(\overline{x})\pi(\overline{x}, p)}{\pi(\overline{x}, q)}\right)^2 \pi(\overline{x}, q) - \lambda^2(p) =$$

$$\sum_{\overline{x}\in D}\left(\frac{\pi(\overline{x}, p)}{\pi(\overline{x}, q)}\right)\pi(\overline{x}, p) - \lambda^2(p).$$

Thus, a low variance can be obtained by ensuring that

$$\frac{\pi(\overline{x}, p)}{\pi(\overline{x}, q)} = \prod_{i=1}^{m}\left(\frac{p}{q}\right)^{x_i}\left(\frac{1-p}{1-q}\right)^{1-x_i} < 1$$

for each $\overline{x} \in D$. Thus, ideally we desire a q for which

$$\max_{\overline{x}\in D}\frac{\pi(\overline{x}, p)}{\pi(\overline{x}, q)}$$

is minimized. Unfortunately, this can seem very difficult to compute. However, as a consolation, we can minimize the ratio for the special case when the state $\overline{x} \in D$ is most probable; namely, when it has exactly r destroyed edges. States with more than r removed edges are significantly less probable, and hence contribute less to the variance. Thus, we seek q for which

$$\frac{\pi(\overline{x}, p)}{\pi(\overline{x}, q)} = \left(\frac{p}{q}\right)^{m-r}\left(\frac{1-p}{1-q}\right)^{r}$$

is minimized.

Example 9. Show that

$$\frac{\pi(\overline{x}, p)}{\pi(\overline{x}, q)} = \left(\frac{p}{q}\right)^{m-r}\left(\frac{1-p}{1-q}\right)^{r}$$

is iminimized when $q = 1 - r/m$.

Example 9 Solution.

As an example, using $p = 0.99$, $\lambda(p) = 1.032 \times 10^{-4}$, $m = 7$, $r = 2$, $q = 1 - 2/7 = 5/7$, and $n = 10^6$ we obtain a variance raio of VR = 173.6, bringing the number of required samples down from millions to tens of thousands.

Exercises

1. For $k = 2, 3, 4, 5, 6$, use IMC simulation with $n = 10^k$ samples to approximate each of the following integrals. Use an appropriate uniform distribution $\omega(x)$. Compare your approximations with the exact answer.

   a. $\int_0^1 x(1-x^2)^{3/2}dx$

   b. $\int_{-2}^2 e^x dx$

2. For $k = 2, 3, 4, 5, 6$, use IMC simulation with $n = 10^k$ samples to approximate each of the following double integrals. Use an appropriate uniform distribution $\omega(x, y)$. Compare your approximations with the exact answer.

   a. $\int_0^1 \int_0^1 2xy\ dxdy$

   b. $\int_{-1}^2 \int_1^4 (2x + 6x^2y)dxdy$

3. For $k = 2, 3, 4, 5, 6$, use IMC simulation with $n = 10^k$ samples to approximate the double integral
$$\int_0^4 \int_0^x e^{-(x+y)}dydx.$$
   Hint: first sample $U_1 \sim \mathcal{U}(0, 4)$, followed by $U_2 \sim \mathcal{U}(0, x)$. What density function $f(x, y)$ does this represent?

4. For $k = 2, 3, 4, 5, 6$, use IMC simulation with $n = 10^k$ samples to approximate $E[N]$, where
$$N = \min\{n : \sum_{i=1}^{n} U_i > 1\},$$
   where $U_i \sim \mathcal{U}(0, 1)$, $i = 1, 2, \ldots$. In other words, N counts the minimum number of $\mathcal{U}(0, 1)$ that are needed to sum to at least 1. Based on the results, give a likely approximation of $E[N]$ to two decimal places.

5. Verify that $\hat{\lambda}_n = \frac{X_1+\cdots+X_n}{n}$ can be recursively computed by defining $\hat{\lambda}_0 = 0$, and using the following recurrence.
$$\hat{\lambda}_{n+1} = \hat{\lambda}_n + \frac{X_{n+1} - \hat{\lambda}_n}{n+1}.$$

6. Verify that
$$\hat{\sigma}_n^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\hat{\lambda}_n^2\right]$$
   can be recursively computed by defining $\hat{\sigma}_1^2 = 0$, and using the following recurrence.
$$\hat{\sigma}_{n+1}^2 = (1 - \frac{1}{n})\hat{\sigma}_n^2 + \hat{\lambda}_n^2 + \frac{X_{n+1}^2}{n} - (1 + \frac{1}{n})\hat{\lambda}_{n+1}^2.$$

7. One million samples of random variable X produed a sample mean of $\hat{\lambda} = 4.27$, and sample variance of $\hat{\sigma}^2 = 13.21$. Determine the 0.90-confidence interval for λ.

8. Prove that a Bernoulli random variable $X \sim \text{Be}(p)$ has variance equal to $p(1-p)$, and that this variance attains it maximum when $p = 0.5$.

9. Let $X_1, \ldots, X_n$ represent samples of a Bernoulli random variable $X \sim \text{Be}(p)$. Let $\hat{p} = \hat{\lambda} = m/n$, where m is the number of samples that produced a value of 1. Prove that

$$\hat{\sigma}^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}).$$

10. Suppose a model exists for simulating the outcome of a soccer game between two teams A and B. The winner of the game varies from simulation to simulation. Let p denote the prior probability that A wins when the model is simulated. The goal is to obtain a 0.95-confidence interval $[a, b]$ for p, with the property that $b - a \leq 0.001$, so that, with 0.95-confidence, p is known to within two decimal places. Use the previous exercise to determine a minimum number n of game simulations that will produce the desired accuracy.

11. For the sampling data from Exercise 7, are we 90% certain that $\hat{\lambda}$ is within 10% of λ?

12. For a Monte Carlo simulation if computer A can generate four times as many samples as computer B in the same time period, then what will be the reduction in standard error if replacing computer B with computer A?

13. A new sampling is able to reduce variance by 50%, while only incurring a 5% increase in sampling cost. Determine the percent reduction in standard error that it will produce in comparison with the existing plan.

14. Recall the Buffon Needle problem where a needle of unit length is tossed onto a plane surface with parallel lines that are unit distance apart. Describe an IMC procedure for estimating the probability that the needle will intersect one of the lines. Hint: assume the needle center point falls in the lower half of a region bounded by two parallel lines, and that the right half of the needle makes a counterclockwise angle between 0^{deg} and 90^{deg} with a vertical line.

15. Suppose that a rectangular region S, whose area is known and is denoted as $A(S)$, encloses another region R whose area $A(R)$ is unknown, but for which there is a procedure for checking if any point in S is also a point in R. Suppose an IMC simulation is conducted to approximate $A(R)$. The procedure approximates the integral

$$\int_S g(s)ds,$$

where $g(s) = 1$ if $s \in R$, and zero otherwise. It does this by using $\omega(s)$ which is uniformly distributed over S. Provide an equation for both $\omega(s)$ and $h(s) = g(s)/\omega(s)$. Show that $\text{Var}(h(S))$ is direcctly proportional to both $A(R)$ and $A(S) - A(R)$.

Exercise Solutions

1. Exact answers are a) 0.2 and b) $e^2 - e^{-2} \approx 7.2537208$.

2. Exact answers are a) 0.5 and b) 234.

3. To sample (x, y) points, use density function $f(x, y) = 1/4x$, and the sampling method described in the hint. Exact answer is $1/2 + e^{-4} + e^{-8} \approx 0.5186511$.

4. $E[N] \approx 2.71$.

5. We have

$$\hat{\lambda}_n + \frac{X_{n+1} - \hat{\lambda}_n}{n+1} = \frac{(n+1)\hat{\lambda}_n + X_{n+1} - \hat{\lambda}_n}{n+1} =$$

$$\frac{n\hat{\lambda}_n + X_{n+1}}{n+1} = \frac{(X_1 + \cdots + X_n) + X_{n+1}}{n+1} = \hat{\lambda}_{n+1}.$$

6. We have

$$(1 - \frac{1}{n})\hat{\sigma}_n^2 + \hat{\lambda}_n^2 + \frac{X_{n+1}^2}{n} - (1 + \frac{1}{n})\hat{\lambda}_{n+1}^2 =$$

$$(\frac{n-1}{n})\frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\hat{\lambda}_n^2) + \hat{\lambda}_n^2 + \frac{X_{n+1}^2}{n} - (1 + \frac{1}{n})\hat{\lambda}_{n+1}^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\lambda}_n^2 + \hat{\lambda}_n^2 + \frac{X_{n+1}^2}{n} - (1 + \frac{1}{n})\hat{\lambda}_{n+1}^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n+1} X_i^2 - \frac{n+1}{n}\hat{\lambda}_{n+1}^2 =$$

$$\frac{1}{n}(\sum_{i=1}^{n+1} X_i^2 - (n+1)\hat{\lambda}_{n+1}^2) = \hat{\sigma}_{n+1}^2.$$

7. With 0.90 confidence,

$$\lambda = 4.27 \pm \Phi^{-1}(0.95)(\hat{\sigma}/\sqrt{10^6}) \Rightarrow$$

$$\lambda = 4.27 \pm (1.64)(\sqrt{13.21}/1000) = 4.27 \pm 0.0036 \Rightarrow$$

$\lambda \in [4.2664, 4.2736]$.

8. We have

$$\mathrm{Var}(X) = E[X^2] - E^2[X] = p - p^2 = p(1-p).$$

Finally, $p(1-p) = p - p^2$ is a concave parabolic function in term of p, and thus reaches its maximum value of $1/4$ at $p = -b/2a = 1/2$.

9. We have

$$\hat{\sigma}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\hat{p}^2\right] = \frac{1}{n-1}\left[n\hat{p} - n\hat{p}^2\right] = \frac{n}{n-1}\hat{p}(1-\hat{p}).$$

10. The length L of a 0.95-confidence interval is given as

$$L = 2(1.96)\frac{\hat{\sigma}}{\sqrt{n}}.$$

Moreover, since

$$\hat{\sigma} = \frac{\sqrt{n\hat{p}(1-\hat{p})}}{\sqrt{n-1}} \leq \frac{\sqrt{n/4}}{\sqrt{n-1}} = \frac{\sqrt{n}}{2\sqrt{n-1}},$$

it follows that

$$L \leq \frac{1.96}{\sqrt{n-1}} \leq 0.001 \Longleftrightarrow n \geq \left(\frac{1.96}{0.001}\right)^2 = 3,841,600.$$

11. We have r.e. $= \sqrt{13.21}/(4.27)1000 = 0.000851$, On the other hand, $0.1/\Phi^{-1}(0.95) = 0.0608$. Therefore, since r.e. < 0.0608, there is justification to believe that, with 0.90 certainty, $\hat{\lambda}$ is within 10% of λ.

12. Computer A produces a

$$(1 - \sqrt{0.25}) \times 100 = (1/2) \times 100 = 50\%$$

reduction in standard error.

13. The new plan produces a

$$(1 - \sqrt{(1.05)(0.5)}) \times 100 = 27.5\%$$

reduction in standard error.

14. Create n independent samples $X_1, \ldots, X_n$, where each sample X_i is obtained as follows. First sample a random variable $Y \sim U(0, 1/2)$. Next, sample $\Theta \sim U(0, \pi/2)$. If $\Theta \leq \cos^{-1}(2Y)$, then assign $X_i = 1$ (such an angle implies that the needle has intersected the lower line). Otherwise, assign $X_i = 0$.

15. $\omega(s) = 1/A(S)$, and hence

$$h(s) = \begin{cases} A(S) & \text{if } s \in R \\ 0 & \text{otherwise} \end{cases}$$

Finally,

$$\mathrm{Var}(h(S)) = A^2(S)\left(A(R)/A(S)\right) - A^2(R) = A(R)\left(A(S) - A(R)\right),$$

which is directly proportional to both $A(R)$ and $A(S) - A(R)$.