Monte Carlo Simulation: IEOR E4703 Fall 2004

Variance Reduction Methods II

1 Importance Sampling

Suppose we wish to estimate $\theta := \mathbf{P}(X \geq 20) = \mathrm{E}[I_{\{X \geq 20\}}]$ where $X \sim \mathsf{N}(0,1)$. The raw estimator given by the "usual approach" is obtained as follows:

1. Generate $X_1, \ldots, X_n$ IID $\mathsf{N}(0,1)$
2. Set $I_j = I_{\{X_j \geq 20\}}$ for $j = 1, \ldots, n$
3. Set $\widehat{\theta}_n = \sum_{j=1}^n I_j/n$
4. Compute approximate CI's

For this problem, however, the usual approach would be completely inadequate since approximating θ to any reasonable degree of accuracy would require n to be inordinately large. For example, on average we would have to set $n \approx 2.7014 \times 10^{89}$ in order to obtain just *one* non-zero value of I.[1] Clearly this is impractical and a much smaller value of n would have to be used. Using a much smaller value of n, however, would almost inevitably result in an estimate, $\widehat{\theta}_n = 0$, and an approximate confidence interval $[L, U] = [0, 0]$! So the naive approach does not work. We could try to use the variance reduction techniques we have seen in the course so far, but they would provide little, if any, help.[2]

Before proceeding any further, it is not unreasonable to ask why such a problem would be important. After all, if you want to estimate $\theta = \mathbf{P}(X \geq 20)$, isn't it enough to know that θ is very close to 0? Put another way, who cares whether $\theta = 10^{-10}$ or $\theta = 10^{-20}$? For many problems, this is a valid objection, and indeed for such problems the answer is that we don't care. However, for many other other problems it is very important to know θ to a much greater level of accuracy.

For example, suppose we are designing a nuclear power plant and we want to estimate the probability, θ, of a meltdown occurring sometime in the next 100 years. Then to start with, we would expect θ to be very small, even for a *poorly* designed power plant. However, this is not enough. Should a meltdown occur, then clearly the consequences would be extremely significant and we care enough about those consequences that we want to know θ to a high degree of accuracy.

As another example, suppose we want to price a deep-out-of-the-money option using simulation. Then the price of the option will be very small, perhaps lying between 10 cents and $.1$ cents. Clearly a bank is not going to suffer if it misprices an option and sells it for $.1$ cents when the correct value is 10 cents. But what if the bank sells 1 million of these options? And what if the bank makes similar trades several times a week? Then it becomes very important to price the option correctly.

Note that both of these examples involve estimating the probability of a *rare event*. Even though the events are rare, they are very important because when they do occur, their impact can be very significant. We will study *importance sampling*, a variance reduction technique that can be invaluable when estimating rare event probabilities and expectations.[3]

[1] We will see where this figure comes from later.

[2] Why is this?

[3] It can of course also be used for estimating non rare event probabilities and expectations but it is not as useful in such circumstances. An exception to this statement is when importance sampling is easier than the regular simulation method.

1.1 The Importance Sampling Estimator

Suppose we wish to estimate $\theta = \mathrm{E}_f[h(X)]$ where X has PDF[4] $f(\cdot)$. Let $g(\cdot)$ be another PDF with the property that $g(x) \neq 0$ whenever $f(x) \neq 0$. That is, g has the same *support* as f. Then

$$\begin{aligned}\theta &= \mathrm{E}_f[h(X)] = \int h(x)f(x)\; dx \\ &= \int h(x)\frac{f(x)}{g(x)}g(x)\; dx.\end{aligned}$$

Since g is a PDF, it is therefore also the case that

$$\theta = \mathrm{E}_g\left[\frac{h(X)f(X)}{g(X)}\right]$$

where $\mathrm{E}_g[\,\cdot\,]$ denotes an expectation with respect to the density $g(\cdot)$. This has very important implications for estimating θ. The original simulation method is to generate n samples of X from the density, $f(\cdot)$, and set $\widehat{\theta}_n = \sum h(X_j)/n$. An alternative method, however, is to generate n values of X from the density, $g(\cdot)$, and set

$$\widehat{\theta}_{n,is} = \sum_{j=1}^{n} \frac{h(X_j)f(X_j)}{ng(X_j)}.$$

$\widehat{\theta}_{n,is}$ is then[5] an *importance sampling* estimator of θ.

Notation

Suppose we wish to estimate $\theta = \mathrm{E}_f[h(X)]$ where X has PDF, $f(\cdot)$. We saw above that it is also the case that

$$\theta = \mathrm{E}_g\left[\frac{h(X)f(X)}{g(X)}\right]$$

where g is another PDF with the same support as f. Then we define

$$h^*(X) := \frac{h(X)f(X)}{g(X)}$$

so that $\theta = \mathrm{E}_g[h^*(X)]$. We will refer to f as the *original* density, and call g the *importance sampling* density[6] or simply, the *sampling* density.

Example 1 (Estimating $\mathbf{P}(X \geq 20)$)

Consider again our original problem where we want to estimate

$$\theta = \mathbf{P}(X \geq 20) = \mathrm{E}[I_{\{X\geq 20\}}]$$

when $X \sim \mathsf{N}(0,1)$. We may then write

$$\begin{aligned}\theta = \mathrm{E}[I_{\{X\geq 20\}}] &= \int_{-\infty}^{\infty} I_{\{X\geq 20\}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\; dx \\ &= \int_{-\infty}^{\infty} I_{\{X\geq 20\}} \left(\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}}} \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}\; dx\end{aligned}$$

[4] Or PMF, if X is a discrete random variable.

[5] We will see later where the name "importance sampling" comes from.

[6] Obviously if X was discrete we would use mass function in place of density function.

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} I_{\{X\geq 20\}} e^{-\mu x+\mu^2/2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}\, dx \\
&= \mathrm{E}_\mu \left[I_{\{X\geq 20\}} e^{-\mu X+\mu^2/2} \right]
\end{aligned}
$$

where now $X \sim \mathsf{N}(\mu, 1)$. (This is clear from our notation, $\mathrm{E}_\mu[\ .\ ]$.) Let us now estimate θ by simulating X from the $\mathsf{N}(\mu, 1)$ distribution, so that

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

If we set $\mu = 20$, for example, then we have the following **Matlab** code.

Matlab Code for Estimating $\mathbf{P}(X \geq 20)$

```
> n=1000000;
> mu=20;
> x=randn(n,1) + mu;
> hprime = [x >= 20] .*exp(-mu*x + mu^2/2);
> theta_est = mean(hprime)

   theta_est =    2.7514e-089

> se = std(hprime)

    se =    1.3526e-088

> CI95 = [theta_est - 1.96*se/sqrt(n), theta_est + 1.96*se/sqrt(n)]

   CI95 =  1.0e-088 *

                  0.2725    0.2778
```

∎

We can of course also estimate expectations using importance sampling.

Example 2

Suppose we wish to estimate $\theta = \mathrm{E}[X^4 e^{X^2/4} I_{\{X\geq 2\}}]$ where $X \sim \mathsf{N}(0, 1)$. Then the same argument as before implies that we may also write

$$\theta = \mathrm{E}_\mu[X^4 e^{X^2/4} e^{-\mu X+\mu^2/2} I_{\{X\geq 2\}}]$$

where now $X \sim \mathsf{N}(\mu, 1)$. In our importance sampling notation we would write

$$\theta = \mathrm{E}_g[X^4 e^{X^2/4} e^{-\mu X+\mu^2/2} I_{\{X\geq 2\}}]$$

where g refers to the $\mathsf{N}(\mu, 1)$ distribution. ∎

In general, we have the following importance sampling algorithm for estimating $\theta = \mathrm{E}_f[h(X)]$ where we simulate with respect to the sampling density, $g(\cdot)$.

Importance Sampling Algorithm for Estimating $\theta = \mathrm{E}_f[h(X)]$

```
for $j = 1$ to $n$
    generate $X_j$ from density $g(\cdot)$
    set $h_j^* = h(X_j)f(X_j)/g(X_j)$
end for
set $\widehat{\theta}_{n,is} = \sum_{j=1}^n h_j^*/n$
set $\widehat{\sigma}^2_{n,is} = \sum_{j=1}^n (h_j^* - \widehat{\theta}_{n,is})^2/(n-1)$
set approx. $100(1-\alpha)$ % CI $= \widehat{\theta}_{n,is} \pm z_{1-\alpha/2}\frac{\widehat{\sigma}_{n,is}}{\sqrt{n}}$
```

1.2 The General Formulation

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with joint PDF $f(x_1, \ldots, x_n)$ and suppose we wish to estimate $\theta = \mathrm{E}_f[h(\mathbf{X})]$. Let $g(x_1, \ldots, x_n)$ be another PDF such that $g(\mathbf{x}) \neq 0$ whenever $f(\mathbf{x}) \neq 0$. Then

$$\begin{aligned}
\theta = \mathrm{E}_f[h(\mathbf{X})] &= \int_{x_1} \cdots \int_{x_n} h(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\ dx_1 \ldots dx_n \\
&= \int_{x_1} \cdots \int_{x_n} h(x_1, \ldots, x_n) \frac{f(x_1, \ldots, x_n)}{g(x_1, \ldots, x_n)} g_(x_1, \ldots, x_n)\ dx_1 \ldots dx_n \\
&= \mathrm{E}_g[h^*(\mathbf{X})]
\end{aligned}$$

where

$$h^*(\mathbf{X}) := \frac{h(\mathbf{X})f(\mathbf{X})}{g(\mathbf{X})}.$$

So now, we again have two methods for estimating θ: the original method where we simulate with respect to the density function f, and the importance sampling method where we simulate with respect to the density function, g.

Example 3 (Estimating $\theta = \mathbf{P}\left(\sum_{i=1}^n X_i^2 \geq 50\right)$)

Suppose we wish to estimate $\theta = \mathbf{P}\left(\sum_{i=1}^n X_i^2 \geq 50\right)$ where the X_i's are IID $\mathsf{N}(0,1)$. Then $\theta = \mathrm{E}[h(\mathbf{X})]$ where $h(\mathbf{X}) := I_{\{\sum X_i^2 \geq 50\}}$ and $\mathbf{X} := (X_1, \ldots, X_n)$. We could estimate θ using importance sampling as follows.

$$\begin{aligned}
\theta &= \mathrm{E}[h(\mathbf{X})] = \int_{x_1} \cdots \int_{x_n} \frac{e^{-x_1^2/2}}{\sqrt{2\pi}} \cdots \frac{e^{-x_n^2/2}}{\sqrt{2\pi}}\ I_{\{\sum X_i^2 \geq 50\}}\ dx_1 \ldots dx_n \\
&= \sigma^n \int_{x_1} \cdots \int_{x_n} \left(\frac{e^{-x_1^2/2}}{e^{-x_1^2/2\sigma^2}} \cdots \frac{e^{-x_n^2/2}}{e^{-x_n^2/2\sigma^2}} \right) \frac{e^{-x_1^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \cdots \frac{e^{-x_n^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\ I_{\{\sum X_i^2 \geq 50\}}\ dx_1 \ldots dx_n \\
&= \sigma^n \int_{x_1} \cdots \int_{x_n} \left(e^{-\frac{x_1^2}{2}(1-1/\sigma^2)} \cdots e^{-\frac{x_n^2}{2}(1-1/\sigma^2)} \right) \frac{e^{-x_1^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \cdots \frac{e^{-x_n^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\ I_{\{\sum X_i^2 \geq 50\}}\ dx_1 \ldots dx_n \\
&= \mathrm{E}_g\left[\sigma^n \left(e^{-\frac{X_1^2}{2}(1-\frac{1}{\sigma^2})} \cdots e^{-\frac{X_n^2}{2}(1-\frac{1}{\sigma^2})} \right) I_{\{\sum X_i^2 \geq 50\}} \right] \\
&= \mathrm{E}_g[h^*(\mathbf{X})]
\end{aligned}$$

where $\mathrm{E}_g[.]$ denotes expectation under a multivariate normal distribution where the X_i's are IID $\mathsf{N}(0, \sigma^2)$. So to estimate θ using importance sampling we could use the following algorithm.

```
for $j = 1$ to $m$
        generate $\mathbf{X_j} = (X_{1,j}, \ldots, X_{n,j})$ where $X_{i,j} \sim$ IID N$(0, \sigma^2)$
        set $Y_j = \sigma^n \left( e^{-\frac{X_{1,j}^2}{2}(1-\frac{1}{\sigma^2})} \ \ldots \ e^{-\frac{X_{n,j}^2}{2}(1-\frac{1}{\sigma^2})} \right) I_{\{\sum_i X_{i,j}^2 \geq 100\}}$
end for
set $\widehat{\theta}_{m,is} = \sum_{j=1}^{m} Y_j/m$
set $\widehat{\sigma}^2_{m,is} = \sum_{j=1}^{m} (Y_j - \widehat{\theta}_{m,is})^2/(m-1)$
set approx. $100(1-\alpha)$ % CI $= \widehat{\theta}_{m,is} \pm z_{1-\alpha/2}\frac{\widehat{\sigma}_{m,is}}{\sqrt{n}}$
```

∎

So far we have not addressed the issue of how to choose a good sampling density, g, so that we obtain a variance reduction when we sample from g instead of f. We will now address this question in the next two sections, as well as explaining the term, "importance sampling".

1.3 Obtaining a Variance Reduction

As before, suppose we wish to estimate $\theta = \mathrm{E}_f[h(\mathbf{X})]$ where $\mathbf{X}$ is a random vector with joint PDF, f. We will assume that $h(\mathbf{X}) \geq 0$. Now let g be another density with support equal to that of f. Then we know

$$\theta = \mathrm{E}_f[h(\mathbf{X})] = \mathrm{E}_g[h^*(\mathbf{X})]$$

and this gives rise to two estimators:

1. $h(\mathbf{X})$ where $\mathbf{X} \sim f(\cdot)$
2. $h^*(\mathbf{X})$ where $\mathbf{X} \sim g(\cdot)$

The variance of the importance sampling estimator is given by

$$\begin{aligned}\mathrm{Var}_g(h^*(\mathbf{X})) &= \int h^*(\mathbf{x})^2 g(\mathbf{x})\ d\mathbf{x} \ - \ \theta^2 \\ &= \int \frac{h(\mathbf{x})^2 f(\mathbf{x})}{g(\mathbf{x})} f(\mathbf{x})\ d\mathbf{x} \ - \ \theta^2\end{aligned}$$

while the variance of the original estimator is given by

$$\mathrm{Var}_f(h(\mathbf{X})) = \int h(\mathbf{x})^2 f(\mathbf{x})\ d\mathbf{x} \ - \ \theta^2.$$

So the reduction in variance[7] is then given by

$$\mathrm{Var}_f(h(\mathbf{X})) - \mathrm{Var}_g(h^*(\mathbf{X})) \ = \ \int h(\mathbf{x})^2 \left(1 - \frac{f(\mathbf{x})}{g(\mathbf{x})}\right) f(\mathbf{x})\ d\mathbf{x}. \tag{1}$$

In order to achieve a variance reduction, the integral in (1) should be positive. For this to happen, we would like

1. $f(\mathbf{x})/g(\mathbf{x}) > 1$ when $h(\mathbf{x})f(\mathbf{x})$ is small and

[7] A negative reduction means a variance increase.

2. $f(\mathbf{x})/g(\mathbf{x}) < 1$ when $h(\mathbf{x})f(\mathbf{x})$ is large.

Now the *important* part of the density, f, could plausibly be defined to be that region, A say, in the support of f where $h(\mathbf{x})f(\mathbf{x})$ is large. But, by the above observation, we would like to choose g so that $f(\mathbf{x})/g(\mathbf{x})$ is small whenever $\mathbf{x}$ is in A. That is, we would like a density, g, that puts more weight on A: hence the name *importance sampling*. Note that when h involves a *rare event* so that $h(\mathbf{x}) = 0$ over "most" of the state space, it can then be particularly valuable to choose g so that we sample often from that part of the state space where $h(\mathbf{x}) \neq 0$. This is why importance sampling is most useful for simulating rare events. Further guidance on how to choose g is obtained from the following observation.

As we are free to choose g, let's suppose we choose[8] $g(\mathbf{x}) = h(\mathbf{x})f(\mathbf{x})/\theta$. Then it is easy to see that

$$\mathrm{Var}_g(h^*(\mathbf{X})) = \theta^2 - \theta^2 = 0$$

so that we have a zero variance estimator! This means that if we sample with respect to this particular choice of g, then we would only need one sample and this sample would equal θ with probability one.[9] Of course this is not feasible in practice. After all, since it is θ that we are trying to *estimate*, it does not seem likely that we could simulate a random variable whose density is given by $g(\mathbf{x}) = h(\mathbf{x})f(\mathbf{x})/\theta$. However, all is not lost and this observation can often guide us towards excellent choices of g that lead to extremely large variance reductions.

1.4 How to Choose a Good Sampling Distribution

We saw above that if we could choose $g(\mathbf{x}) = h(\mathbf{x})f(\mathbf{x})/\theta$, then we would obtain the best possible estimator of θ, that is, one that has zero variance. In general, we cannot do this, but it does suggest that if we could choose $g(\cdot)$ so that it is *similar* to $h(\cdot)f(\cdot)$, then we might reasonably expect to obtain a large variance reduction.

What does the phrase "similar" mean? One obvious thing to do would be to choose $g(\cdot)$ so that it has a similar *shape* to $h(\cdot)f(\cdot)$. In particular, we could try to choose g so that $g(\mathbf{x})$ and $h(\mathbf{x})f(\mathbf{x})$ both take on their maximum values at the same value, $\mathbf{x}^*$, say. When we choose g this way, we say that we are using the **maximum principle**. Of course this only partially defines g since there are infinitely many density functions that could take their maximum value at $\mathbf{x}^*$. Nevertheless, this is often enough to obtain a significant variance reduction and in practice, we often take g to be from the same family of distributions as f. For example, if f is multivariate normal, then we might also take g to be multivariate normal but with a different mean vector and / or variance-covariance matrix.[10]

Example 4 (Example 2 Continued)

Recall that we wish to estimate $\theta = \mathrm{E}[h(X)] = \mathrm{E}[X^4 e^{X^2/4} I_{\{X \geq 2\}}]$ where $X \sim \mathsf{N}(0,1)$. Then if we sample from a PDF, g, that is also normal with variance 1, but mean μ, we know that $g(\cdot)$ takes it maximum value at $x = \mu$. Therefore, a good choice of μ might be

$$\begin{aligned}
\mu &= \text{arg} \max_x \; h(x)f(x) \\
&= \text{arg} \max_x \; x^4 e^{x^2/4} I_{\{x \geq 2\}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \\
&= \text{arg} \max_{x \geq 2} \; x^4 e^{-x^2/4} \\
&= \sqrt{8}.
\end{aligned}$$

[8] Note that this choice of g is valid since $\int g(\mathbf{x})\, d\mathbf{x} = 1$ and we have assumed h is non-negative.

[9] Recall that with this choice of g, $h^*(\mathbf{x}) = h(\mathbf{x})f(\mathbf{x})/g(\mathbf{x}) = \theta$.

[10] We note that it is not necessary that f and g come from the same family of distributions. In fact sometimes it is necessary to choose g from a *different* family of distributions. This might occur, for example, if it is difficult or inefficient to simulate from the family of distributions to which f belongs. In that case, our reason for using importance sampling in the first place is so that we can simulate from an 'easier' distribution, g.

Then, as we saw before, $\theta = \mathrm{E}_g[h^*(X)] = \mathrm{E}_g[X^4 e^{X^2/4} e^{-\mu X+\mu^2/2} I_{\{X \geq 2\}}]$ where $g(\cdot)$ denotes the $\mathsf{N}(\mu, 1)$ PDF. So to estimate θ, we might use the following algorithm where we have set $\mu = \sqrt{8}$.

```
for $j = 1$ to $m$
    generate $X_j \sim \mathsf{N}(\sqrt{8}, 1)$
    set $h_j^* = X_j^4 e^{X_j^2/4} e^{-\sqrt{8}X_j+4} I_{\{X_j \geq 2\}}$
end for
set $\widehat{\theta}_{m,is} = \sum_{j=1}^m h_j^*/m$
set $\widehat{\sigma}^2_{m,is} = \sum_{j=1}^m (h_j^* - \widehat{\theta}_{m,is})^2/(m-1)$
set approx. $100(1-\alpha)$ % CI $= \widehat{\theta}_{m,is} \pm z_{1-\alpha/2}\frac{\widehat{\sigma}_{m,is}}{\sqrt{n}}$
```

∎

Example 5 (Pricing an Asian call option)

For the purpose of option pricing, we assume as usual that $S_t \sim GBM(r, \sigma^2)$, where S_t is the price of the stock at time t and r is the risk-free interest rate. Suppose now that we want to price an Asian call option whose payoff at time t is given by

$$h(\mathbf{S}) := \max\left(0, \frac{\sum_{i=1}^m S_{iT/m}}{m} - K\right) \tag{2}$$

where $\mathbf{S} := \{S_{iT/m} : i = 1, \ldots, m\}$, T is the expiration date and K is the strike price. The price of this option is then given by

$$C_a = \mathrm{E}[e^{-rT} h(\mathbf{S})].$$

Now we can write

$$S_{iT/m} = S_0 e^{(r-\sigma^2/2)\frac{iT}{m} + \sigma\sqrt{\frac{T}{m}}(X_1+\ldots+X_i)}$$

where the X_i's are IID $\mathsf{N}(0, 1)$. This means that if f is the joint PDF of $\mathbf{X} = (X_1, \ldots, X_m)$, then (with a mild abuse of notation) we may write

$$C_a = \mathrm{E}_f[h(X_1, \ldots, X_n)].$$

Now if K is large relative to S_0 then the option is said to be *out-of-the-money* which means that most of the time, the option expires worthless, or unexercised. Pricing such options using simulation amounts to doing a rare event simulation. This is particularly true when K is very large relative to S_0, and the option is deep out-of-the-money. As a result, estimating C_a using importance sampling will often result in a very large variance reduction. In order to apply importance sampling, we need to choose the sampling density, $g(\cdot)$. For this, we could take $g(\cdot)$ to be multivariate normal with variance-covariance matrix equal to the identity, I_m, and mean vector, μ^*. That is we shift $f(\mathbf{x})$ by μ^*. As before, a good possible value of μ^* might be

$$\mu^* = \text{arg } \max_{\mathbf{x}} \ h(\mathbf{x})f(\mathbf{x})$$

which can be found using numerical methods. ∎

1.4.1 Potential Problems with the Maximum Principle

Sometimes applying the maximum principle to choose $g(\cdot)$ will be difficult. For example, it may be the case that there are multiple or even infinitely many solutions to $\mu^* = \text{arg } \max_{\mathbf{x}} \ h(\mathbf{x})f(\mathbf{x})$. Even when there is a unique solution, it may be the case that finding it is very difficult. In such circumstances, an alternative method for choosing g is to **scale** f. We will demonstrate this by example.

Example 6 (Using Scaling to Select $g(\cdot)$)

Assume in Example 3 that $n = 2$. Then $\theta = \mathbf{P}\left(X_1^2 + X_2^2 \geq 50\right) = \mathrm{E}[I_{\{X_1^2+X_2^2 \geq 50\}}]$ where X_1, X_2 are IID N$(0, 1)$. Then

$$h(\mathbf{x})f(\mathbf{x}) = I_{\{x_1^2+x_2^2 \geq 50\}} \frac{e^{-(x_1^2+x_2^2)/2}}{2\pi}.$$

Therefore, $h(\mathbf{x})f(\mathbf{x}) = 0$ inside the circle $x_1^2 + x_2^2 \leq 50$ and it takes on its maximum value at every point *on* the circle $x_1^2 + x_2^2 = 50$. As a result, it is not possible to apply the maximum principle.

Before choosing a sampling density, g, recall that we would like g to put more weight on those parts of the sample space where $h(\mathbf{x})f_x(\mathbf{x})$ is large. One way to achieve this is by *scaling* the density of $\mathbf{X} = (X_1, X_2)$ so that $\mathbf{X}$ is more "dispersed". For example, we could take g to be multivariate normal with mean vector $\mathbf{0}$ and variance-covariance matrix

$$\Sigma_g = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

where $\sigma^2 > 1$. Note that this simply means that under g, X_1 and X_2 are IID N$(0, \sigma^2)$. Furthermore, when $\sigma^2 > 1$, then more probability mass is given to the region $X_1^2 + X_2^2 \geq 50$ as desired.

We should choose the value of σ using heuristic methods. One method would be to choose σ so that $\mathrm{E}_g[X_1^2 + X_2^2] = 50$ which in this case would imply that $\sigma = 5$. Why? Then[11]

$$\theta = \mathrm{E}[I_{\{X_1^2+X_2^2 \geq 50\}}] = \mathrm{E}_g\left[\sigma^2 \exp\left(-\frac{X_1^2}{2}(1 - 1/\sigma^2) - \frac{X_2^2}{2}(1 - 1/\sigma^2)\right) I_{\{X_1^2+X_2^2 \geq 50\}}\right].$$

For the more general case where $n > 2$, we could proceed by again choosing σ so that $\mathrm{E}_g[\sum_{i=1}^n X_i^2] = 50$. ▮

1.5 Tilted Densities

A common way of generating the sampling density, g, from the original density, f, is to use the moment generating function (MGF) of X. We use $M_x(t)$ to denote the MGF and it is defined by

$$M_x(t) = \mathrm{E}_f[e^{tX}].$$

Then a *tilted* density of f is given by

$$f_t(x) = \frac{e^{tx} f(x)}{M_x(t)}$$

for $-\infty < t < \infty$. The tilted densities are useful since a random variable with density $f_t(\cdot)$ tends to be larger than one with density f when $t > 0$, and smaller when $t < 0$. This means, for example, that if we want to sample more often from the region where X tends to be large, we might want to use a tilted density with $t > 0$ as our sampling density g. Similarly, if we want to sample more often from the region where X tends to be small, then we might use a tilted density with $t < 0$.

Example 7

Suppose X is an exponential random variable with mean $1/\lambda$. Then $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and it is easy to see that $f_t(x) = Ce^{-(\lambda - t)x}$ where C is the constant that makes the density integrate to 1. ▮

Example 8

Suppose $X_1, \ldots, X_n$ are independent random variables, where X_i has density $f_i(\cdot)$. Let $S_n := \sum_{i=1}^n X_i$ and suppose we want to estimate $\theta := \mathbf{P}(S_n \geq a)$ for some constant, a. If a is large so that we are dealing with a

[11] This is the same expression we had earlier in Example 3 except n =2.

rare event we should use importance sampling to estimate θ. Since S_n is large when the X_i's are large it makes sense to sample each X_i from its tilted density function, $f_{i,t}(\cdot)$ for some value of $t > 0$. Then we may write

$$\begin{aligned}\theta &= \mathrm{E}[I_{\{S_n \geq a\}}] \\ &= \mathrm{E}_t\left[I_{\{S_n \geq a\}} \prod_{i=1}^{n} \frac{f_i(X_i)}{f_{i,t}(X_i)}\right] \\ &= \mathrm{E}_t\left[I_{\{S_n \geq a\}} \left(\prod_{i=1}^{n} M_i(t)\right) e^{-tS_n}\right]\end{aligned}$$

where $\mathrm{E}_t[.]$ denotes expectation with respect to the X_i's under the tilted densities, $f_{i,t}(\cdot)$, and $M_i(t)$ is the moment generating function of X_i. If we write $M(t) := \prod_{i=1}^{n} M_i(t)$, then it is easy to see that the importance sampling estimator, $\widehat{\theta}_{n,i}$, satisfies

$$\widehat{\theta}_{n,i} \leq M(t)e^{-ta}. \tag{3}$$

Therefore a good choice of t would be that value that minimizes the bound in (3). Why is this? See Ross (2002) for further details. ▮

1.6 Estimating Conditional Expectations

Importance sampling can also be very useful for computing conditional expectations when the event being conditioned upon is a rare event. For example, suppose we wish to estimate

$$\theta = \mathrm{E}[h(\mathbf{X})|\mathbf{X} \in A]$$

where A is a rare event and $\mathbf{X}$ is a random vector with PDF f. Then the density of $\mathbf{X}$, given that $\mathbf{X} \in A$, is

$$f(\mathbf{x}|\mathbf{x} \in A) = \frac{f(\mathbf{x})}{\mathbf{P}(\mathbf{X} \in A)} \qquad \text{for } \mathbf{x} \in A$$

so

$$\theta = \frac{\mathrm{E}[h(\mathbf{X})I_{\{\mathbf{X} \in A\}}]}{\mathrm{E}[I_{\{\mathbf{X} \in A\}}]}.$$

Now since A is a rare event we would be much better off if we could simulate using a sampling density, g, that makes A more likely to occur. Then, as usual, we would have

$$\theta = \frac{\mathrm{E}_g[h(\mathbf{X})I_{\{\mathbf{X} \in A\}} f(\mathbf{X})/g(\mathbf{X})]}{\mathrm{E}_g[I_{\{\mathbf{X} \in A\}} f(\mathbf{X})/g(\mathbf{X})]}.$$

So to estimate θ using importance sampling, we would generate $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with density $g(\cdot)$, and set

$$\widehat{\theta}_{n,i} = \frac{\sum_{i=1}^{n} h(\mathbf{X_i})I_{\{\mathbf{X_i} \in A\}} f(\mathbf{X_i})/g(\mathbf{X_i})}{\sum_{i=1}^{n} I_{\{\mathbf{X_i} \in A\}} f(\mathbf{X_i})/g(\mathbf{X_i})}.$$

In contrast to our usual estimators, $\widehat{\theta}_{n,i}$ is no longer an average of n IID random variables but instead, it is the *ratio* of two such averages. This has implications for computing approximate confidence intervals for θ. In particular, confidence intervals should now be estimated using the *bootstrapping* technique.[12]

An important application of this methodology in financial engineering is the estimation of *conditional value-at-risk*. For example, suppose a bank has estimated that its 10 day 99% VAR is 1 million dollars. That is, the probability of losing more than 1 million dollars over the next 10 days is approximately $1 - .99 = .01$. In this case, the conditional value-at-risk is the expected loss over the next 10 days *given* that the loss is greater than 1 million dollars.[13]

[12] See Ross, Chapter 7, for details. We will not study the bootstrapping technique in this course.

[13] Sometimes banks like to reduce their VAR by adopting various hedging strategies but in so doing, it is quite possible that they are actually *increasing* their conditional VAR. This is hardly desirable.

1.7 Difficulties with Importance Sampling

The most difficult aspect to importance sampling is in choosing a good sampling density, g. In general, one needs to be very careful for it is possible to choose g according to some good heuristic such as the maximum principle, but to then find that g results in a variance *increase*. In fact it is possible to choose a g that results in an importance sampling estimator that has an infinite variance! This situation would typically occur when g puts too little weight relative to f on the tails of the distribution.

2 Stratified Sampling

Consider a game show where contestants first pick a ball at random from an urn and then receive a payoff, Y. The payoff is random and depends on the color of the selected ball so that if the color is c then Y is drawn from the PDF, $f_c(\cdot)$. The urn contains red, green, blue and yellow balls, and each of the four colors is equally likely to be chosen. The producer of the game show would like to know how much a contestant will win on average when he plays the game. To answer this question, she decides to simulate the payoffs of n contestants and take their average payoff as her estimate. The payoff, Y, of each contestant is simulated as follows:

1. Simulate a random variable, I, where I is equally likely to take any of the four values r, g, b and y
2. Simulate Y from the density $f_I(y)$.

The average payoff, $\theta := \mathrm{E}[Y]$, is then estimated by

$$\widehat{\theta}_n := \frac{\sum_{j=1}^n Y_j}{n}.$$

Now suppose $n = 1000$, and that a red ball was chosen 246 times, a green ball 270 times, a blue ball 226 times and a yellow ball 258 times.

Question: Would this influence your confidence in $\widehat{\theta}_n$? What if f_g tended to produce very high payoffs and f_b tended to produce very low payoffs?

Is there anything that we could have done to avoid this type of problem occurring? The answer is yes. We know that each ball color should be selected $1/4$ of the time so we could force this to be true by conducting four separate simulations, one each to estimate $\mathrm{E}[X|I = c]$ for $c = r, g, b, y$. Note that

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y|I]] = \frac{1}{4}\mathrm{E}[Y|I = r] + \frac{1}{4}\mathrm{E}[Y|I = g] + \frac{1}{4}\mathrm{E}[Y|I = b] + \frac{1}{4}\mathrm{E}[Y|I = y]$$

so that an unbiased estimator of θ is obtained by setting

$$\widehat{\theta}_{st,n} := \frac{1}{4}\widehat{\theta}_{r,n_r} + \frac{1}{4}\widehat{\theta}_{g,n_g} + \frac{1}{4}\widehat{\theta}_{b,n_b} + \frac{1}{4}\widehat{\theta}_{y,n_y} \tag{4}$$

where $\theta_c := \mathrm{E}[Y|I = c]$ for $c = r, g, b, y$.[14]

How does the variance of $\widehat{\theta}_{st,n}$ compare with the variance of $\widehat{\theta}_n$, the original raw simulation estimator? To answer this question, assume for now that $n_c = n/4$ for each c, and that Y_c is a sample from the density, $f_c(\cdot)$.

Then a fair[15] comparison of $\mathrm{Var}(\widehat{\theta}_n)$ with $\mathrm{Var}(\widehat{\theta}_{st,n})$ should compare

$$\mathrm{Var}(Y_1 + Y_2 + Y_3 + Y_4) \quad \text{with} \quad \mathrm{Var}(Y_r + Y_g + Y_b + Y_y) \tag{5}$$

[14] $\widehat{\theta}_{c,n_c}$ is an estimate of θ_c using n_c samples. $\widehat{\theta}_{st,n}$ is an estimate of θ using n samples, so it is implicitly assumed in (4) that $n_r + n_g + n_b + n_y = n$.

[15] Fair here means that each estimator is based on the same total number of samples.

where Y_1, Y_2, Y_3 and Y_4 are IID samples from the original simulation algorithm (i.e., where we first select the ball randomly and then receive the payoff), and the Y_c's are independent with density $f_c(\cdot)$, for $c = r, g, b, y$. Now recall the conditional variance formula which states

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|I)] + \mathrm{Var}(\mathrm{E}[Y|I]). \tag{6}$$

Each term in the right-hand-side of (6) is non-negative so this implies

$$\begin{aligned}
\mathrm{Var}(Y) &\geq \mathrm{E}[\mathrm{Var}(Y|I)] \\
&= \frac{1}{4}\mathrm{Var}(Y|I = r) + \frac{1}{4}\mathrm{Var}(Y|I = g) + \frac{1}{4}\mathrm{Var}(Y|I = b) + \frac{1}{4}\mathrm{Var}(Y|I = y) \\
&= \frac{\mathrm{Var}(Y_r + Y_g + Y_b + Y_y)}{4}
\end{aligned}$$

which implies

$$\mathrm{Var}(Y_1 + Y_2 + Y_3 + Y_4) = 4\mathrm{Var}(Y) \geq \mathrm{Var}(Y_r + Y_g + Y_b + Y_y). \tag{7}$$

As a result, we may conclude that using $\widehat{\theta}_{st,n}$ instead of $\widehat{\theta}_n$ leads to a variance reduction. This variance reduction will be substantial if I accounts for a large fraction of the variance of Y. Note also that the computational requirements for computing $\widehat{\theta}_{st,n}$ are similar[16]to those required for computing $\widehat{\theta}_n$.

We call $\widehat{\theta}_{st,n}$ a *stratified sampling* estimator of θ, and we say that I is the *stratification* variable.

2.1 Stratified Sampling Algorithm

We will now formally describe the stratified sampling algorithm. Suppose as usual that we wish to estimate $\theta := \mathrm{E}[Y]$ where Y is a random variable. Let W be another random[17] variable that satisfies the following two conditions:

Condition 1: For any $\Delta \subseteq \mathbf{R}$, $\mathbf{P}(W \in \Delta)$ can be easily computed.

Condition 2: It is easy to generate $(Y|W \in \Delta)$, i.e., Y given $W \in \Delta$.

Now divide $\mathbf{R}$ into m non-overlapping subintervals, $\Delta_1, \ldots, \Delta_m$, such that $\sum_{j=1}^m p_j = 1$ where $p_j := \mathbf{P}(W \in \Delta_j) > 0$.

Note that if W can take any value in $\mathbf{R}$, then the first interval should be $[-\infty, b]$, while the final interval should be $[a, \infty]$ for some finite a and b.

Notation

1. Let $\theta_j := \mathrm{E}[Y|W \in \Delta_j]$ and $\sigma_j^2 := \mathrm{Var}(Y|W \in \Delta_j)$.
2. We define the random variable I by setting $I := j$ if $W \in \Delta_j$.
3. Let $Y^{(j)}$ denote a random variable with the same distribution as $(Y|W \in \Delta_j) \equiv (Y|I = j)$.

Our notation then implies $\theta_j = \mathrm{E}[Y|I = j] = \mathrm{E}[Y^{(j)}]$ and $\sigma_j^2 = \mathrm{Var}(Y|I = j) = \mathrm{Var}(Y^{(j)})$. In particular we have

$$\begin{aligned}
\theta \;=\; \mathrm{E}[Y] \;=\; \mathrm{E}[\mathrm{E}[Y|I]] &= p_1\mathrm{E}[Y|I = 1] \;+\; \ldots \;+\; p_m\mathrm{E}[Y|I = m] \\
&= p_1\theta_1 \;+\; \ldots \;+\; p_m\theta_m.
\end{aligned}$$

[16] For this example, the stratified estimator will actually require less work, but it is also possible in general for it to require more work.

[17] Y and W must be dependent to achieve a variance reduction.

Note that to estimate θ we only need to estimate the θ_i's since by condition 1 above, the p_i's are easily computed. Furthermore, we know how to estimate the θ_i's by condition 2. If we use n_i samples to estimate θ_i, then an estimate of θ is given by

$$\widehat{\theta}_{st,n} = p_1\widehat{\theta}_{1,n_1} + \ldots + p_m\widehat{\theta}_{m,n_m}.$$

It is clear that $\widehat{\theta}_{st,n}$ will be unbiased if for each i, $\widehat{\theta}_{i,n_i}$ is an unbiased estimate of θ_i.

2.2 Obtaining a Variance Reduction

How does the stratification estimator compare with the usual raw simulation estimator? As was the case with the game show example, to answer this question we would like to compare $\mathrm{Var}(\widehat{\theta}_n)$ with $\mathrm{Var}(\widehat{\theta}_{st,n})$. First we need to choose $n_1, \ldots, n_m$ such that $n_1 + \ldots + n_m = n$. That is, we need to determine the number of samples, n_i, that will be used to estimate each θ_i, but in such a way that the total number of samples is equal to n. Clearly, the optimal approach would be to choose the n_i's so as to minimize $\mathrm{Var}(\widehat{\theta}_{st,n})$.

Consider for now, however, the *sub-optimal* allocation where we set $n_j := np_j$ for $j = 1, \ldots, m$. Then

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}_{st,n}) &= \mathrm{Var}(p_1\widehat{\theta}_{1,n_1} + \ldots + p_m\widehat{\theta}_{m,n_m}) \\ &= p_1^2\frac{\sigma_1^2}{n_1} + \ldots + p_m^2\frac{\sigma_m^2}{n_m} \\ &= \frac{\sum_{j=1}^m p_j\sigma_j^2}{n}.\end{aligned}$$

On the other hand, the usual simulation estimator has variance σ^2/n where $\sigma^2 := \mathrm{Var}(Y)$. Therefore, we need only show that $\sum_{j=1}^m p_j\sigma_j^2 < \sigma^2$ to prove that the non-optimized[18] stratification estimator has a lower variance than the usual raw estimator.[19]

The proof that $\sum_{j=1}^m p_j\sigma_j^2 < \sigma^2$ is precisely the same as that used for the game show example. In particular, equation (6) implies

$$\sigma^2 = \mathrm{Var}(Y) \geq \mathrm{E}[\mathrm{Var}(Y|I)] = \sum_{j=1}^m p_j\sigma_j^2$$

and the proof is complete!

2.3 Optimizing the Stratified Estimator

We know

$$\begin{aligned}\widehat{\theta}_{st,n} &= p_1\widehat{\theta}_{1,n_1} + \ \ldots \ + p_m\widehat{\theta}_{m,n_m} \\ &= p_1\frac{\sum_{i=1}^{n_1} Y_i^{(1)}}{n_1} + \ \ldots \ + p_m\frac{\sum_{i=1}^{n_m} Y_i^{(m)}}{n_m}\end{aligned}$$

where, for a fixed j, the $Y_i^{(j)}$'s are IID $\sim Y^{(j)}$. Then this implies

$$\mathrm{Var}(\widehat{\theta}_{st,n}) = p_1^2\frac{\sigma_1^2}{n_1} + \ \ldots \ + p_m^2\frac{\sigma_m^2}{n_m} = \sum_{j=1}^m \frac{p_j^2\sigma_j^2}{n_j}. \tag{8}$$

[18] The optimized stratification estimator refers to the estimator where we choose the n_i's to minimize the variance of $\widehat{\theta}_{st,n}$.

[19] The optimized stratification estimator would then of course achieve an even greater variance reduction.

Therefore, to minimize $\mathrm{Var}(\widehat{\theta}_{st,n})$ we must solve the following constrained optimization problem:

$$\min_{n_j} \sum_{j=1}^{m} \frac{p_j^2 \sigma_j^2}{n_j}$$

$$\text{subject to} \qquad n_1 + \ldots + n_m = n.$$

We can easily solve this problem using a Lagrange multiplier and the optimal solution is given by

$$n_j^* = \left(\frac{p_j \sigma_j}{\sum_{j=1}^m p_j \sigma_j} \right) n \tag{9}$$

with the minimized variance given by

$$\mathrm{Var}(\widehat{\theta}_{st,n^*}) = \frac{\left(\sum_{j=1}^m p_j \sigma_j\right)^2}{n}. \tag{10}$$

Note that the solution in (9) makes intuitive sense: if p_j is large, then other things being equal, it makes sense to expend more effort simulating from stratum j, i.e., the region where $W_j \in \Delta_j$. Similarly, if σ_j^2 is large then, other things again being equal, it makes sense to simulate more often from stratum j so as to get a more accurate estimate of θ_j.

Remark 1 *It is interesting at this point to note how stratified sampling is related to importance sampling. We saw in the last lecture that when we importance sample, we would like to sample more often from the* important *region. The choice of n_j in (9) also means that we simulate more often from the important region when we use optimized stratified sampling.*

Remark 2 *Note also the connection of stratified sampling to the method of conditional expectations. Both methods rely on the conditional variance formula to prove that they lead to a variance reduction. The difference between the two methods can best be explained as follows. Suppose we wish to estimate $\theta := \mathrm{E}[Y]$ using simulation and we do this by first generating random variable, W, and then generating Y given W. In the conditional expectation method, we simulate W first, but then compute $\mathrm{E}[Y|W]$ analytically. In the stratified sampling method, we effectively generate W analytically, and then simulate Y given W.*

2.4 Advantages and Disadvantages of Stratified Sampling

The obvious advantage of stratified sampling is that it leads to a variance reduction which can be very substantial if the stratification variable, W, accounts for a large fraction of the variance of Y. The main disadvantage of stratified sampling is that typically we do not know the σ_j^2's so it is impossible to compute the optimal n_j's exactly. Of course we can overcome this problem by first doing m pilot simulations to estimate each σ_j. If we let N_p denote the total number of pilot simulations, then a good heuristic is to use N_p/m runs for each individual pilot simulation. In order to obtain a reasonably good estimate of σ_j^2, a useful rule-of-thumb is that N_p/m should be greater than 30. If m is large however, and each simulation run is computationally expensive, then it may be the case that a lot of effort is expended in trying to estimate the optimal n_j's.

One method of overcoming this problem is to abandon the pilot simulations and simply use the sub-optimal allocation where $n_j = np_j$. We saw earlier that this allocation still results in a variance reduction which sometimes can be substantial. In practice, both methods are used. The decision to conduct pilot simulations should depend on the problem at hand. For example, if you have reason to believe that the σ_j's will not vary too much then it should be the case that the optimal allocation and the sub-optimal allocation will be very similar. In this case, it is probably not worth doing the pilot simulations. On the other hand, if the σ_j's vary considerably, then conducting the pilot runs may be worthwhile. Of course, a combination of the two is also possible where a only a subset of the pilot simulations is conducted.

The stratified simulation algorithm is given below. We assume that the pilot simulations have already been completed, or it has been decided not to conduct them at all; either way, the n_j's have been computed. We also show how the estimate, $\widehat{\theta}_{n,st}$, and the estimated variance, $\widehat{\sigma}^2_{n,st}$, can be computed without having to store all the generated samples. That is, we simply keep track of $\sum {Y_i^{(j)}}^2$ and $\sum Y_i^{(j)}$ for each j since these quantities are all that is required[20] to compute $\widehat{\theta}_{n,st}$ and $\widehat{\sigma}^2_{n,st}$.

Stratification Simulation Algorithm for Estimating θ

```
set $\widehat{\theta}_{n,st} = 0$;   $\widehat{\sigma}^2_{n,st} = 0$;
for $j = 1$ to $m$
    set $sum_j = 0$;   $sum\_squares_j = 0$;
    for $i = 1$ to $n_j$
        generate $Y_i^{(j)}$
        set $sum_j = sum_j + Y_i^{(j)}$
        set $sum\_squares_j = sum\_squares_j + {Y_i^{(j)}}^2$
    end for
    set $\theta_j = sum_j / n_j$
    set $\widehat{\sigma}^2_j = \left(sum\_squares_j - sum_j^2/n_j\right)/(n_j - 1)$
    set $\widehat{\theta}_{n,st} = \widehat{\theta}_{n,st} + p_j\theta_j$
    set $\widehat{\sigma}^2_{n,st} = \widehat{\sigma}^2_{n,st} + \widehat{\sigma}^2_j p_j^2 / n_j$
end for
set approx. $100(1-\alpha)$ % CI $= \widehat{\theta}_{n,st} \pm z_{1-\alpha/2}\, \widehat{\sigma}_{n,st}$
```

2.5 Applications

Example 9

Suppose we want to estimate $\theta := \mathrm{E}[\sqrt{1-U^2}]$ where $U \sim U(0,1)$. We set $Y = \sqrt{1-U^2}$ and we can choose $W = U$ as our stratification variable. We can do this since

1: $\mathbf{P}(W \in \Delta)$ can easily be computed.

2: $Y^{(j)} := (Y | W \in \Delta_j)$ can easily be generated.

To see that $Y^{(j)}$ can easily be generated, suppose $\Delta_j = [a,b]$. Then $Y^{(j)} = (\sqrt{1-U^2} \mid U \in [a,b])$. Now it is easy to see that

$$(U \mid U \in [a,b]) \sim U(a,b),$$

[20] Recall that $\widehat{\theta}_{n,st} = \sum_{j=1}^m \left(\frac{\sum_{i=1}^{n_j} Y_i^{(j)}}{n_j}\right) p_j$, $\mathrm{Var}(\widehat{\theta}_{n,st}) = \sum_{j=1}^m \mathrm{Var}\left(\frac{\sum_{i=1}^{n_j} Y_i^{(j)}}{n_j}\right) p_j^2$ and $\mathrm{Var}\left(\frac{\sum_{i=1}^{n_j} Y_i^{(j)}}{n_j}\right)$ can be estimated knowing just $\sum_{i=1}^{n_j} Y_i^{(j)}$ and $\sum_{i=1}^{n_j} {Y_i^{(j)}}^2$. As stated previously, any simulation study that requires a large number of samples should only keep track of these quantities, thereby avoiding the need to store every sample.

so to generate $Y^{(j)}$ we first generate $U \sim U(a,b)$, and then set $Y^{(j)} = \sqrt{1-U^2}$.

Let's choose m equi-probable strata so that

$$\Delta_1 = \left[0, \frac{1}{m}\right], \ \Delta_2 = \left[\frac{1}{m}, \frac{2}{m}\right], \ \ldots, \ \Delta_m = \left[\frac{m-1}{m}, 1\right]$$

and $p_j = 1/m$ for all j. To avoid conducting pilot runs, we set $n_j = np_j = n/m$. We then have the following code for solving this problem.

Matlab Code for Estimating $\mathrm{E}[\sqrt{1-U^2}]$ using Stratified Sampling

```
function[theta,CI] = strat(N,m);

p=1/m;
n=N/m;    % This is n_j
theta=0; var=0;

for j=1:m
   U = (j-1)/m + rand(n,1)/m;
   X=sqrt(1-U.^2);
   theta = theta + p*mean(X);
   Sum = sum(X);
   Sum_squares =  sum(X.^2);
   Sig_square_j = (Sum_squares - (Sum^2)/n )/(n-1);
   var = var + Sig_square_j * p^2/n;
end;

CI = [theta - 1.96*sqrt(var), theta + 1.96*sqrt(var)]
% This is an approx 95% CI
```

At the Matlab prompt we can then execute *strat.m* with $N = 10000$ and $m = 100$:

```
>> [theta, CI] = strat(10000,100)

theta =   0.7854

CI =    0.7853    0.7855
```

∎

Example 10 (Pricing a European Call Option)

Suppose that we wish to price a European call option where we assume as usual that $S_t \sim GBM(r, \sigma^2)$. Then

$$C_0 = \mathrm{E}\left[e^{-rT} \max(0, S_T - K)\right] \ = \ \mathrm{E}[Y]$$

where

$$Y = h(X) = e^{-rT} \max\left(0, \ S_0 e^{(r-\sigma^2/2)T + \sigma\sqrt{T}X} - K\right)$$

for $X \sim \mathsf{N}(0,1)$. While we know how to compute C_0 analytically, it is worthwhile seeing how we could estimate it using stratified simulation. Let $W = X$ be our stratification variable. To see that we can stratify using this choice of W note that:

(1) Computing $\mathbf{P}(W \in \Delta)$

For $\Delta \subseteq \mathbf{R}$, $\mathbf{P}(W \in \Delta)$ can easily be computed. Indeed, if $\Delta = [a,b]$, then $\mathbf{P}(W \in \Delta) = \Phi(b) - \Phi(a)$, where $\Phi(.)$ is the CDF of a standard normal random variable.

(2) Generating $(Y|W \in \Delta)$

$(h(X)|X \in \Delta)$ can easily be generated. We do this by first generating $\tilde{X} := (X|X \in \Delta)$ and then take $h(\tilde{X})$. We generate $\tilde{X}$ as follows.

First note that if $X \sim \mathsf{N}(0,1)$, then we can generate an X using the inverse transform method by setting $X = \Phi^{-1}(U)$. The problem with such an X is that it may not lie in $\Delta = [a,b]$. However, we can overcome this problem by simply generating $\tilde{U} \sim U(\Phi(a), \Phi(b))$ and then setting $\tilde{X} = \Phi^{-1}(\tilde{U})$. It is then straightforward to check that $\tilde{X} \sim (X|X \in [a,b])$.

It is therefore clear that we can estimate C_0 using X as a stratification variable. ∎

Exercise 1 *Are there other ways to generate* $(X|X \in [a,b])$ *in Example 10 ?*

Example 11 (Pricing an Asian Call Option)

Recall that the *discounted* payoff of an Asian call option is given by

$$Y := e^{-rT} \max\left(0, \frac{\sum_{i=1}^{m} S_{iT/m}}{m} - K\right) \tag{11}$$

and that it's price is given by $C_a = \mathrm{E}[Y]$ where we assume $S_t \sim GBM(r, \sigma^2)$. Now each $S_{iT/m}$ may be expressed as

$$S_{iT/m} = S_0 \ \exp\left((r - \sigma^2/2)\frac{iT}{m} + \sigma\sqrt{\frac{T}{m}}(X_1 + \ \ldots \ + X_i)\right) \tag{12}$$

where the X_i's are IID $\mathsf{N}(0,1)$. This means that we may then write $C_a = \mathrm{E}\left[h(X_1, \ldots, X_m)\right]$ where the function $h(.)$ is given implicitly by equations (11) and (12). So to estimate C_a using our standard simulation algorithm, we would simply generate sample values of $h(X_1, \ldots, X_m)$ and take their average as our estimate. We can also, however, estimate C_a using stratified sampling.[21]

To do so, we must first choose a stratification variable, W. One possible choice would be to set $W = X_j$ for some j. However, this is unlikely to capture much of the variability of $h(X_1, \ldots, X_m)$. A much better choice would be to set $W = \sum_{j=1}^{m} X_j$. Of course, we need to show that such a choice is possible. That is, we need to show that $\mathbf{P}(W \in \Delta)$ is easily computed, and that $(Y|W \in \Delta)$ is easily generated.

(1) Computing $\mathbf{P}(W \in \Delta)$

Since $X_1, \ldots, X_m$ are IID $\mathsf{N}(0,1)$, we immediately have that $W \sim \mathsf{N}(0,m)$. If $\Delta = [a,b]$ then

$$\begin{aligned}\mathbf{P}(W \in \Delta) = \mathbf{P}\left(\mathsf{N}(0,m) \in \Delta\right) &= \mathbf{P}\left(a \le \mathsf{N}(0,m) \le b\right) \\ &= \mathbf{P}\left(\frac{a}{\sqrt{m}} \le \mathsf{N}(0,1) \le \frac{b}{\sqrt{m}}\right) \\ &= \Phi\left(\frac{b}{\sqrt{m}}\right) - \Phi\left(\frac{a}{\sqrt{m}}\right).\end{aligned}$$

[21] The method we now describe is also useful for pricing other path dependent options. See Glasserman, Heidelberger and Shahabuddin (1998) for further details.

Similarly, if $\Delta = [b, \infty)$, then $\mathbf{P}(W \in \Delta) = 1 - \Phi\left(\frac{b}{\sqrt{m}}\right)$, and if $\Delta = (-\infty, a]$, then $\mathbf{P}(W \in \Delta) = \Phi\left(\frac{a}{\sqrt{m}}\right)$.

(2) Generating $(Y|W \in \Delta)$

We need two results from the theory of multivariate normal random variables. The first result was studied earlier in the course.

Result 1

Suppose $\mathbf{X} = (X_1, \ldots, X_m) \sim \mathsf{MVN}(\mathbf{0}, \mathbf{\Sigma})$. If we wish to generate a sample vector $\mathbf{X}$, we first generate $\mathbf{Z} \sim \mathsf{MVN}(\mathbf{0}, \mathbf{I_m})$[22] and then set

$$\mathbf{X} = \mathbf{C}^T\mathbf{Z} \tag{13}$$

where $\mathbf{C}^T\mathbf{C} = \mathbf{\Sigma}$. One possibility of course is to let $\mathbf{C}$ be the Cholesky decomposition of $\mathbf{\Sigma}$, but in fact any matrix $\mathbf{C}$ that satisfies $\mathbf{C}^T\mathbf{C} = \mathbf{\Sigma}$ will do.

Result 2

Let $\mathbf{a} = (a_1\ a_2\ \ldots\ a_m)$ satisfy $||a|| = 1$, i.e. $\sqrt{a_1^2 + \ldots + a_m^2} = 1$, and let $\mathbf{Z} = (Z_1, \ldots, Z_m) \sim \mathsf{MVN}(\mathbf{0}, \mathbf{I_m})$. Then

$$\left\{(Z_1, \ldots, Z_m) \;\Big|\; \sum_{i=1}^m a_i Z_i = w\right\} \sim \mathsf{MVN}(w\mathbf{a}^T,\ \mathbf{I_m} - \mathbf{a^T a}).$$

Therefore, to generate $\{(Z_1, \ldots, Z_m)| \sum_{i=1}^m a_i Z_i = w\}$ we just need to generate a vector, $\mathbf{V}$, where

$$\mathbf{V} \sim \mathsf{MVN}(w\mathbf{a}^T,\ \mathbf{I_m} - \mathbf{a^T a}) = w\mathbf{a}^T + \mathsf{MVN}(\mathbf{0},\ \mathbf{I_m} - \mathbf{a^T a}).$$

Generating such a $\mathbf{V}$ is very easy since

$$(\mathbf{I_m} - \mathbf{a^T a})^T(\mathbf{I_m} - \mathbf{a^T a}) = \mathbf{I_m} - \mathbf{a^T a}.$$

That is, $\mathbf{\Sigma}^T\mathbf{\Sigma} = \mathbf{\Sigma}$ where $\mathbf{\Sigma} = \mathbf{I_m} - \mathbf{a^T a}$, so we can take $\mathbf{C} = \mathbf{\Sigma}$ in (13).

Now, we can return to the problem of generating $(Y \mid W \in \Delta)$. Since $Y = h(X_1, \ldots, X_m)$, we can clearly generate $(Y \mid W \in \Delta)$ if we can generate $[(X_1, \ldots, X_m) \mid \sum_{i=1}^m X_i \in \Delta]$. To do this, suppose again that $\Delta = [a, b]$. Then

$$\left[(X_1, \ldots, X_m) \;\Big|\; \sum_{i=1}^m X_i \in [a, b]\right] \;\equiv\; \left[(X_1, \ldots, X_m) \;\Big|\; \frac{1}{\sqrt{m}}\sum_{i=1}^m X_i \in \left[\frac{a}{\sqrt{m}}, \frac{b}{\sqrt{m}}\right]\right].$$

Now we can generate $[(X_1, \ldots, X_m) \mid \sum_{i=1}^m X_i \in \Delta]$ in two steps.

Step 1

Generate $\left[\frac{1}{\sqrt{m}}\sum_{i=1}^m X_i \;\Big|\; \frac{1}{\sqrt{m}}\sum_{i=1}^m X_i \in \left[\frac{a}{\sqrt{m}}, \frac{b}{\sqrt{m}}\right]\right]$.

This is easy to do since $\frac{1}{\sqrt{m}}\sum_{i=1}^m X_i \sim \mathsf{N}(0, 1)$ so we just need to generate

$$\left(\mathsf{N}(0,1) \;\Big|\; \mathsf{N}(0,1) \in \left[\frac{a}{\sqrt{m}}, \frac{b}{\sqrt{m}}\right]\right)$$

[22] That is, we generate m IID $\mathrm{N}(0, 1)$ random variables.

which we can do using the method described in Example 10. Let w be the generated value.

Step 2

Now generate $\left[(X_1, \ldots, X_m) \,\middle|\, \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i = w\right]$ which we can do by the second result above and the comments that follow it.

▮

Example 12 (Pricing a Barrier Option)

Recall again the problem of pricing an option that has payoff

$$h(X) = \begin{cases} \max(0, S_T - K_1) & \text{if } S_{T/2} \leq L, \\ \max(0, S_T - K_2) & \text{otherwise.} \end{cases}$$

where $X = (S_{T/2}, S_T)$. We can write the price of the option as

$$C_0 = \mathrm{E}\left[e^{-rT}\left(\max(0, S_T - K_1) I_{\{S_{T/2} \leq L\}} + \max(0, S_T - K_2) I_{\{S_{T/2} > L\}}\right)\right]$$

where as usual, we assume that $S_t \sim GBM(r, \sigma^2)$. Using conditional expectations, we saw earlier that we could write $C_0 = \mathrm{E}[Y]$ where

$$Y := e^{-rT/2}\left(c(S_{T/2}, T/2, K_1, r, \sigma) I_{\{S_{T/2} \leq L\}} + c(S_{T/2}, T/2, K_2, r, \sigma) I_{\{S_{T/2} \geq L\}}\right) \tag{14}$$

and where $c(x, t, k, r, \sigma)$ is the price of a European call option with strike k, interest rate r, volatility σ, time to maturity t, and initial stock price x.

Question 1: Having conditioned on $S_{T/2}$, could we now also use stratified sampling?

Question 2: Could we use importance sampling?

Question 3: What about using importance sampling *before* doing the conditioning?

▮

3 Low Discrepancy Sequences and Quasi Monte Carlo Methods

Consider the problem of computing an integral over the d-dimensional unit cube. One of the principle advantages of using Monte Carlo simulation to do this is that the convergence rate has order $1/\sqrt{n}$ which is independent of d, and where n is the number of simulated points. In contrast, standard numerical integration schemes based on a rectangular grid of points converge as $n^{-2/d}$. Since many interesting problems in financial engineering are high-dimensional, either due to multiple state variables or path-dependence, it is clear that Monte Carlo simulation can provide a significant computational advantage. On the other hand, a sample of uniformly distributed points in the d-dimensional unit cube covers the cube *inefficiently*. This is clear, for example, in Figure 1, where uniform samples from $[0, 1] \times [0, 1]$ are plotted.

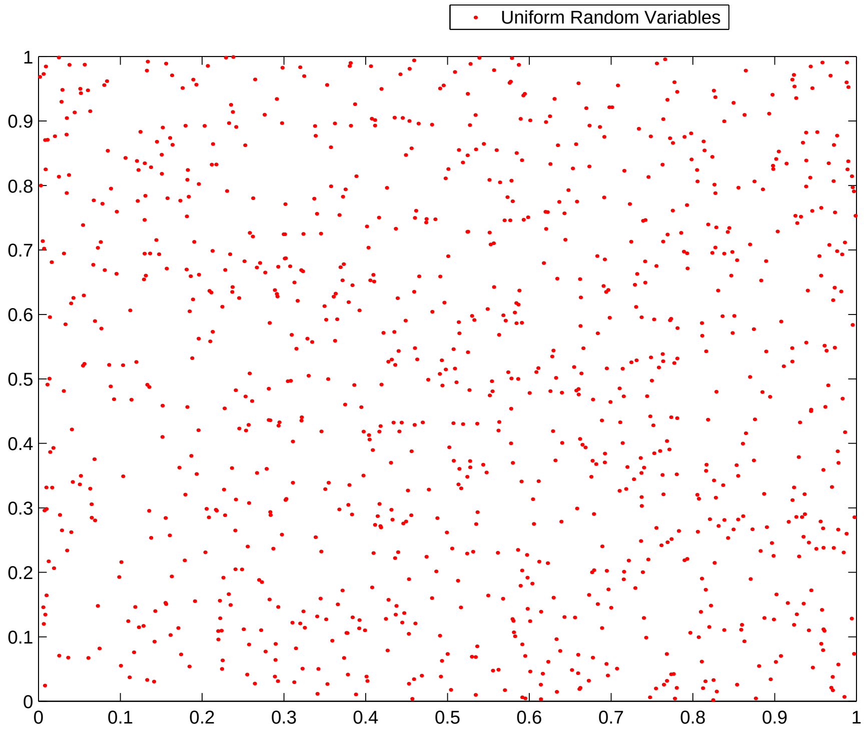

Figure 1: Uniform Random Variables

A d-dimensional *low discrepancy sequence*[23] is a deterministic sequence of points in the d-dimensional unit cube that fills the cube efficiently, i.e. it has a low *discrepancy*. This low discrepancy property results in a convergence rate of $(\log n)^d/n$, implying in particular that they can often be much more effective than Monte Carlo methods. An example of a 2-dimensional low discrepancy sequence is plotted in Figure 2 where it is clear that there is nothing random about these points whatsoever. Despite this, the term *"Quasi Monte Carlo methods"* is often used to refer to approaches that use low discrepancy sequences as an alternative to standard Monte Carlo methods.

Exercise 2 *How might you evaluate an expectation, $\theta = \mathrm{E}[f(\mathbf{X})]$, where $\mathbf{X}$ is a d-dimensional multivariate normal random vector? Consider first the case where the d normal random variables are independent.*

[23] This topic does not rightly belong to lecture notes titled *"Variance Reduction Methods"*. Nonetheless, we place it here as variance reduction methods and the use of low discrepancy sequences are both employed with a view towards reducing estimator errors.

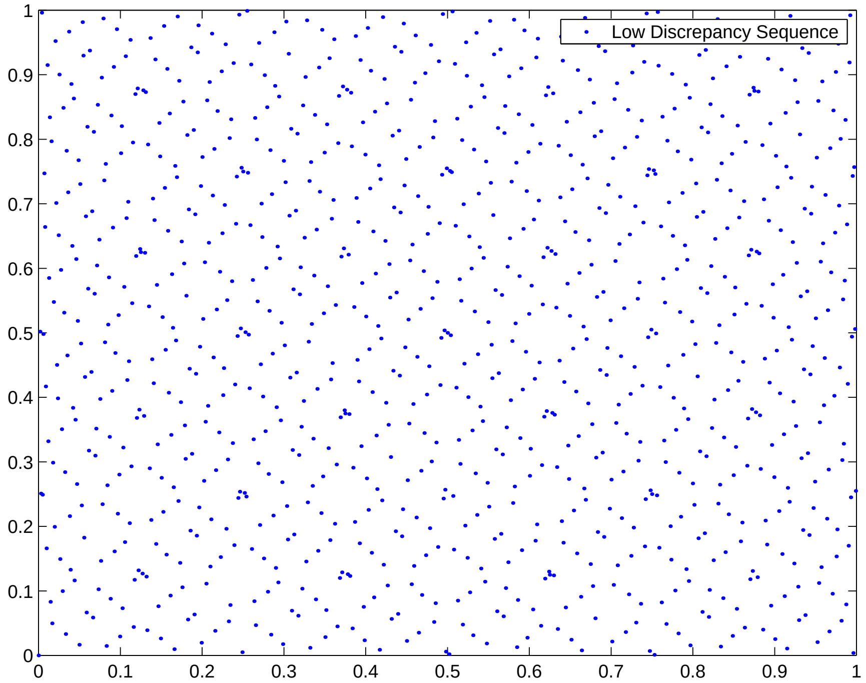

Figure 2: Low Discrepancy Points

The use of low discrepancy sequences have a number of advantages and disadvantages, some of which we outline below:

Advantages

1. Their asymptotic convergence properties are superior to those of Monte Carlo simulation.
2. The number of points, n, need not be known in advance. This is a property shared with Monte Carlo but not with numerical integration techniques that are based on regular grids.

Disadvantages

1. For a fixed sample size, n, there is no guarantee that low discrepancy sequences will outperform Monte Carlo simulation.
2. Since they are deterministic, confidence intervals are not available and so it is difficult to tell whether or not an estimate is sufficiently accurate. (There have been attempts to randomize low discrepancy sequences, motivated in part by the desire to overcome this problem.)
3. The sample size, n, may be too small relative to the dimension, d. For example, many popular low discrepancy sequences cover the initial coordinates, (x_1, x_2), more or less uniformly, but do not cover the final coordinates, (x_{d-1}, x_d), in a sufficiently uniform manner. In such circumstances, it might be necessary to raise n to an unsatisfactorily high level.
4. In general, more care is needed when applying low-discrepancy sequences than when applying Monte Carlo methods. However, they often produce significantly better estimates.

For further information on low discrepancy sequences, see Glasserman (2003).

References

Glasserman, P. 2003. *Monte Carlo Methods in Financial Engineering.* Springer-Verlag, New York.

Ross, S.M. 2002. *Simulation.* Academic Press, New York.