Further Variance Reduction Methods

In these lecture notes we discuss more advanced variance reduction techniques, in particular importance sampling and stratified sampling. Importance sampling is a vital technique when estimating quantities associated with *rare events*. When combined with stratified sampling it can be a particularly powerful variance reduction technique. We consider several applications of these techniques to financial engineering and insurance.

1 Variance Reduction and Importance Sampling

We begin with a seemingly contrived example but one that nonetheless gained some notoriety at the beginning of the financial crisis in August 2007 when Goldman Sachs Asset Management's (GSAM) *Global Alpha* fund incurred steep losses. In explaining these losses, the CFO of Goldman Sachs claimed they had seen 25 standard deviation moves several days in a row. Not surprisingly this claim resulted in considerable criticism.

Example 1 (Just How Unlucky is a 25 Standard Deviation (Negative) Return?)
Suppose we wish to estimate $\theta := \mathbf{P}(X \geq 25) = \mathrm{E}[I_{\{X \geq 25\}}]$ where $X \sim \mathsf{N}(0,1)$. The usual Monte-Carlo approach to this problem proceeds as follows:

1. Generate $X_1, \ldots, X_n$ IID $\mathsf{N}(0,1)$
2. Set $I_j = I_{\{X_j \geq 25\}}$ for $j = 1, \ldots, n$
3. Set $\widehat{\theta}_n = \sum_{j=1}^n I_j/n$
4. Compute an approximate 95% CI as $\widehat{\theta}_n \pm 1.96 \times \widehat{\sigma}_n/\sqrt{n}$ where $\widehat{\sigma}_n$ is the standard deviation of the I_j'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, we will soon see that on average we would have to set $n \approx 3.26 \times 10^{137}$ in order to obtain just *one* non-zero value of I. 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 Monte-Carlo approach does not work here. ∎

Before proceeding further, it is not unreasonable to ask why such a problem would be important. After all, if you want to estimate $\theta = \mathbf{P}(X \geq 25)$, isn't it enough to know that θ is very close to 0? Put another way, do we care whether $\theta = 10^{-10}$ or $\theta = 10^{-20}$? For many problems, this is a valid objection, as we may care just how small θ as long as we know that it is indeed "small". 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. 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 could be catastrophic and so we would like to know θ to a very high degree of accuracy.

For 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 .1 cents and 10 cents. Clearly a bank is not going to suffer if it misprices this 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. A particularly rich source of examples can be found in risk-management where we seek to estimate risk measures such as the VaR_α or ES_α of a given portfolio.

1.1 Introduction and Main Results

Note that these examples require 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 now consider importance sampling, a variance reduction technique that can be invaluable when estimating rare event probabilities and expectations.

Suppose we wish to estimate $\theta = \mathrm{E}_f[h(X)]$ where X has PDF f (or PMF, if X is a discrete random variable). Let g 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

$$\theta \;=\; \mathrm{E}_f[h(X)] \;=\; \int h(x)\frac{f(x)}{g(x)}g(x)\,dx \;=\; \mathrm{E}_g\left[\frac{h(X)f(X)}{g(X)}\right]$$

where $\mathrm{E}_g[\,\cdot\,]$ denotes an expectation with respect to the density g. This has very important implications for estimating θ. The original simulation method is to generate n samples of X from the density, f, 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, 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 an *importance sampling* estimator of θ. We often define $h^*(X) := h(X)f(X)/g(X)$ so that $\theta = \mathrm{E}_g[h^*(X)]$. We refer to f and g as the *original* and *importance sampling* densities, respectively. We also refer to f/g as the *likelihood ratio*.

Example 2 (Revisiting the 25 Standard Deviation Example)
Consider again the problem where we want to estimate $\theta = \mathbf{P}(X \geq 25) = \mathrm{E}[I_{\{X\geq 25\}}]$ when $X \sim \mathsf{N}(0,1)$. We may then write

$$\begin{aligned}
\theta = \mathrm{E}[I_{\{X\geq 25\}}] &= \int_{-\infty}^{\infty} I_{\{X\geq 25\}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx \\
&= \int_{-\infty}^{\infty} I_{\{X\geq 25\}} \left(\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}}}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}\,dx \\
&= \mathrm{E}_\mu\left[I_{\{X\geq 25\}} e^{-\mu X+\mu^2/2}\right]
\end{aligned}$$

where now $X \sim \mathsf{N}(\mu,1)$ and $e^{-\mu X+\mu^2/2}$ is the likelihood ratio. If we set $\mu = 25$, for example, and use $n = 10$ million samples, then we find an approximate 95% confidence interval for θ is given by $[3.053,\ 3.074] \times 10^{-138}$. ∎

We can of course also estimate expectations using importance sampling.

Example 3 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 $X \sim \mathsf{N}(\mu,1)$. ∎

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 we easily obtain

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

where $h^*(\mathbf{X}) := h(\mathbf{X})f(\mathbf{X})/g(\mathbf{X})$. Again we 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, g.

Example 4 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 $\mathrm{N}(0,1)$. 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.

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

where $\mathrm{E}_g[.]$ denotes expectation under a multivariate normal distribution where the $\mathbf{X} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\mathbf{I}_n)$. ∎

Thus 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.

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. Without loss of generality 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 $\mathbf{X} \sim f$ and
2. $h^*(\mathbf{X})$ where $\mathbf{X} \sim g$

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 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 $h(\mathbf{x})^2 f(\mathbf{x})$ is small and
2. $f(\mathbf{x})/g(\mathbf{x}) < 1$ when $h(\mathbf{x})^2 f(\mathbf{x})$ 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 $g(\mathbf{x}) = h(\mathbf{x})f(\mathbf{x})/\theta$. 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[1] one. Of course this is not feasible in practice since we don't know θ and therefore don't know g either. However, all is not lost and this observation can often guide us towards excellent choices of g that lead to extremely large variance reductions.

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.[2]

Example 5 Returning to Example 3, recall that we wished to estimate $\theta = \mathrm{E}[h(X)] = \mathrm{E}[X^4 e^{X^2/4} I_{\{X \geq 2\}}]$ where $X \sim \mathsf{N}(0,1)$. If we sample from a PDF, g, that is also normal with variance 1 but mean μ, then we know that g takes it maximum value at $x = \mu$. Therefore, a good choice of μ might be

$$\mu \;=\; \arg\max_x \; h(x)f(x) \;=\; \arg\max_{x \geq 2} \; x^4 e^{-x^2/4} \;=\; \sqrt{8}.$$

Then $\theta = \mathrm{E}_g[h^*(X)] = \mathrm{E}_g[X^4 e^{X^2/4} e^{-\sqrt{8}X+4} I_{\{X \geq 2\}}]$ where $g(\cdot)$ denotes the $\mathsf{N}(\sqrt{8}, 1)$ PDF. ∎

Example 6 (Pricing an Asian Call Option)
For the purpose of option pricing, we assume 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

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

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

[2] 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.

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 very large relative to S_0 so that the option is deep out-of-the-money then pricing the option using simulation amounts to performing a rare event simulation. 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. For this, we could take g to be the multivariate normal PDF with variance-covariance matrix equal to the identity, I_m, and mean vector, μ^*. As before, a good possible value of μ^* might be $\mu^* = \arg\max_{\mathbf{x}} \; h(\mathbf{x}) f(\mathbf{x})$ which can be found using numerical methods. ∎

Potential Problems with the Maximum Principle

Sometimes applying the maximum principle to choose g will be difficult. For example, it may be the case that there are multiple or even infinitely many solutions to $\mu^* = \arg\max_{\mathbf{x}} \; h(\mathbf{x}) f(\mathbf{x})$. 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 7 (Using Scaling to Select g)
Assume in Example 4 that $n = 2$ so that $\theta = \mathbf{P}\left(X_1^2 + X_2^2 \geq 50\right) = \mathrm{E}[I_{\{X_1^2 + X_2^2 \geq 50\}}]$ where X_1, X_2 are IID $\mathsf{N}(0,1)$. Then

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

so that $h(\mathbf{x}) f(\mathbf{x}) = 0$ inside the circle $x_1^2 + x_2^2 \leq 50$ and $h(\mathbf{x}) f(\mathbf{x})$ 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 $h(\mathbf{x}) f(\mathbf{x})$ 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 $\mathsf{N}(0, \sigma^2)$. 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 could 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? We then have

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

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

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. In more sophisticated applications of importance sampling it is desirable to have (or prove) some guarantee that the importance sampling variance will be finite. As an example of such a guarantee see Example 9 below.

1.2 Using Tilted Densities to Obtain a Good Sampling Distribution

Suppose f is *light-tailed* so that it has a moment generating function (MGF). Then a very useful way of generating the sampling density, g, from the original density, f, is to use the MGF of f. 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. 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 8 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 9 (The Probability that a Sum of Random Variables Will Exceed Some Value) 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 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 $t > 0$. We may then write

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

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

$$0 \leq \widehat{\theta}_{n,i} \leq M(t)e^{-ta}. \tag{3}$$

Therefore a good choice of t would be that value that minimizes the bound in (3). We can minimize this by minimizing $\log(M(t)e^{-ta}) \;=\; \log(M(t)) - ta$. It is straightforward to check that the minimizing value of t satisfies $\mu_t = a$ where $\mu_t := \mathrm{E}_t[S_n]$. This can easily be found numerically. ∎

Applications From Insurance: Estimating Ruin Probabilities

Continuing on from Example 9, if we define the stopping time $\tau_a := \min\{n \geq 0 : S_n \geq a\}$, then $P(\tau_a < \infty)$ is the probability that S_n ever exceeds a. If $\mathrm{E}[X_1] > 0$ and the X_i's are IID with MGF, $M_X(t)$, then this probability equals one. The case of interest is then when $\mathrm{E}[X_1] \leq 0$. A similar argument to that of Example 9 yields

$$\begin{aligned}
\theta \;=\; \mathrm{E}[I_{\{\tau_a < \infty\}}] \;=\; \mathrm{E}\left[\sum_{n=1}^{\infty} 1_{\{\tau_a = n\}}\right] &= \sum_{n=1}^{\infty} \mathrm{E}\left[1_{\{\tau_a = n\}}\right] \\
&= \sum_{n=1}^{\infty} \mathrm{E}_t\left[1_{\{\tau_a = n\}} \left(M_X(t)\right)^n e^{-tS_n}\right] \\
&= \sum_{n=1}^{\infty} \mathrm{E}_t\left[1_{\{\tau_a = n\}} \left(M_X(t)\right)^{\tau_a} e^{-tS_{\tau_a}}\right] \\
&= \mathrm{E}_t\left[I_{\{\tau_a < \infty\}} e^{-tS_{\tau_a} + \tau_a \psi(t)}\right]
\end{aligned}$$

where $\psi(t) := \log(M_X(t))$ is the *cumulant generating function*. Note that if $\mathrm{E}_t[X_1] > 0$ then $\tau_a < \infty$ almost surely and so we obtain $\theta = \mathrm{E}_t\left[e^{-tS_{\tau_a}+\tau_a\psi(t)}\right]$. In fact, this is an example where importance sampling can be used to ensure that the simulation stops almost surely. It is possible to choose a good value of t based on the cumulant generating function.

Note that this example has direct applications to the estimation of ruin probabilities in the context of insurance risk. For example, suppose $X_i := Y_i - cT_i$ where Y_i is the *size* of the i^{th} claim, T_i is the *inter-arrival* time between claims, c is the premium received per unit time and a is the *initial reserve*. Then θ is the probability that the insurance company ever goes bankrupt. Only in very simple models is it possible to calculate θ analytically. In general, Monte-Carlo approaches are required.

1.3 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

$$\theta = \frac{\mathrm{E}[h(\mathbf{X})I_{\{\mathbf{X}\in A\}}]}{\mathrm{E}[I_{\{\mathbf{X}\in A\}}]}. \tag{4}$$

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 *bootstrapping* techniques. An obvious application of this methodology in risk management is in the estimation of quantities *similar* to ES or CVaR.

Exercise 1 *How does the practical problem of estimating the α-CVaR of a loss distribution, i.e. estimating $\theta := \mathrm{E}[L \,|\, L \geq \mathit{VaR}_\alpha]$, generally differ from the problem of estimating θ in (4)?*

---

2 An Application of Importance Sampling to Credit Risk

We consider[3] a portfolio loss of the form $L = \sum_{i=1}^m e_i Y_i$ where e_i is the deterministic and positive exposure to the i^{th} credit and Y_i is the default indicator with corresponding default probability, p_i. We assume that $\mathbf{Y}$ follows a Bernoulli mixture model which we now define.

Definition 1 *Let $p < m$ and let $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_p)^\top$ be a p-dimensional random vector. Then we say the random vector $\mathbf{Y} = (Y_1, \ldots, Y_m)^\top$ follows a* **Bernoulli mixture model** *with* **factor vector** $\mathbf{\Psi}$ *if there are functions $p_i : \mathbb{R}^p \to [0,1],\ 1 \leq i \leq m$, such that conditional on $\mathbf{\Psi}$ the components of $\mathbf{Y}$ are independent Bernoulli random variables satisfying $P(Y_i = 1 \mid \mathbf{\Psi} = \psi) = p_i(\psi)$.*

[3]This application is described in Section 8.5 of *MFE*, but is based on the original paper of Glasserman and Li (2005).

We are interested in the problem of estimating $\theta := P(L \geq c)$ where c is substantially larger than $\mathrm{E}[L]$. Note that a good importance sampling distribution for θ should also yield a good importance sampling distribution for computing risk measures associated with the α-tail of the loss distribution where $q_\alpha(L) \approx c$. We begin with the case where the default indicators are independent.

2.1 Independent Default Indicators

We define Ω to be the state space of $\mathbf{Y}$ so that $\Omega = \{0,1\}^m$. Then

$$P(\{\mathbf{y}\}) \;=\; \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{1-y_i}, \quad \mathbf{y} \in \Omega$$

so that

$$M_L(t) \;=\; \mathrm{E}[e^{tL}] \;=\; \prod_{i=1}^{m} \mathrm{E}[e^{te_iY_i}] \;=\; \prod_{i=1}^{m} \left(p_i e^{te_i} + 1 - p_i\right).$$

Let Q_t be the corresponding tilted probability measure so that

$$\begin{aligned} Q_t(\{\mathbf{y}\}) \;=\; \frac{e^{t\sum_{i=1}^m e_i y_i}}{M_L(t)}\, P(\{\mathbf{y}\}) \quad &= \quad \prod_{i=1}^{m} \frac{e^{te_iy_i}}{(p_i e^{te_i} + 1 - p_i)}\, p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \quad \prod_{i=1}^{m} q_{t,i}^{y_i}(1-q_{t,i})^{1-y_i} \end{aligned}$$

where $q_{t,i} := p_i e^{te_i}/(p_i e^{te_i} + 1 - p_i)$ is the Q_t probability of the i^{th} credit defaulting. Note that the default indicators remain independent Bernoulli random variables under Q_t. Since $q_{t,i} \to 1$ as $t \to \infty$ and $q_{t,i} \to 0$ as $t \to -\infty$ it is clear that we can shift the mean of L to any value in $(0, \sum_{i=1}^m e_i)$. The same argument that was used at the end of Example 9 suggests that we take t equal to that value that solves $\mathrm{E}_t[L] = \sum_{i=1}^m q_{i,t} e_i = c$. This value can be found easily using numerical methods.

2.2 Dependent Default Indicators

Suppose now that there is a p-dimensional **factor vector**, $\mathbf{\Psi}$, such that the default indicators are independent with default probabilities $p_i(\psi)$ *conditional* on $\mathbf{\Psi} = \psi$. Suppose also that $\mathbf{\Psi} \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{\Sigma})$. The Monte-Carlo scheme for estimating θ is to first simulate $\mathbf{\Psi}$ and to then simulate $\mathbf{Y}$ conditional on $\mathbf{\Psi}$. We can apply importance sampling to the second step using our discussion of independent default indicators above. However, we can also apply importance sampling to the first step. A natural way to do this is to simulate $\mathbf{\Psi}$ from the $\mathrm{MVN}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution for some $\boldsymbol{\mu} \in \mathbb{R}^p$. The corresponding likelihood ratio, $r_{\boldsymbol{\mu}}(\mathbf{\Psi})$ say, (f/g in our earlier notation) is given by the ratio of the two multivariate normal densities and satisfies

$$r_{\boldsymbol{\mu}}(\mathbf{\Psi}) \;=\; \frac{\exp\left(-\frac{1}{2}\mathbf{\Psi}^\top \mathbf{\Sigma}^{-1}\mathbf{\Psi}\right)}{\exp\left(-\frac{1}{2}(\mathbf{\Psi}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{\Psi}-\boldsymbol{\mu})\right)} \;=\; \exp(-\boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1}\mathbf{\Psi} + \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1}\boldsymbol{\mu}).$$

How Do We Choose μ?

Recall that the quantity of interest is $\theta := P(L \geq c) = \mathrm{E}[P(L \geq c \mid \mathbf{\Psi})]$. We know from our earlier discussion of importance sampling that we would like to choose the importance sampling density, $g^*(\mathbf{\Psi})$ say, so that

$$g^*(\mathbf{\Psi}) \;\propto\; P(L \geq c \mid \mathbf{\Psi})\, \exp(-\frac{1}{2}\mathbf{\Psi}^\top \mathbf{\Sigma}^{-1}\mathbf{\Psi}). \tag{5}$$

Of course this is not possible since we do not know $P(L \geq c \mid \mathbf{\Psi})$, the very quantity that we wish to estimate. The maximum principle applied to the $\mathrm{MVN}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution would then suggest taking $\boldsymbol{\mu}$ equal to the value of $\mathbf{\Psi}$ which maximizes the right-hand-side of (5). Again it is not possible to solve this problem exactly as we do not know $P(L \geq c \mid \mathbf{\Psi})$ but numerical methods can be used to find good approximate solutions. See Glasserman and Li (2005) for further details.

The Importance Sampling Algorithm for Estimating $\theta = P(L \geq c)$

1. Generate $\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_n$ independently from the $\text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution
2. For each $\mathbf{\Psi}_i$ estimate $P(L \geq c \mid \mathbf{\Psi} = \mathbf{\Psi}_i)$ using the importance sampling distribution that we described in our discussion of independent default indicators. Let $\widehat{\theta}^{IS}_{n_1}(\mathbf{\Psi}_i)$ be the corresponding estimator based on n_1 samples.
3. The full importance sampling estimator is then given by

$$\widehat{\theta}^{IS}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} r_{\boldsymbol{\mu}}(\mathbf{\Psi}_i)\, \widehat{\theta}^{IS}_{n_1}(\mathbf{\Psi}_i).$$

---

3 Variance Reduction and 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. 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

$$\widehat{\theta}_{st,n} := \frac{1}{4}\widehat{\theta}_{r,n_r} + \frac{1}{4}\widehat{\theta}_{g,n_g} + \frac{1}{4}\widehat{\theta}_{b,n_b} + \frac{1}{4}\widehat{\theta}_{y,n_y} \tag{6}$$

where $\theta_c := \mathrm{E}[Y|I = c]$ for $c = r, g, b, y$.[4] 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. Then a fair comparison of $\mathrm{Var}(\widehat{\theta}_n)$ with $\mathrm{Var}(\widehat{\theta}_{st,n})$ should compare

$$\mathrm{Var}(Y_1 + Y_2 + Y_3 + Y_4) \quad \text{with} \quad \mathrm{Var}(Y_r + Y_g + Y_b + Y_y) \tag{7}$$

[4] $\widehat{\theta}_{c,n_c}$ 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 (6) that $n_r + n_g + n_b + n_y = n$.

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

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|I)] + \mathrm{Var}(\mathrm{E}[Y|I]). \tag{8}$$

Each term in the right-hand-side of (8) 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

$$\mathrm{Var}(Y_1 + Y_2 + Y_3 + Y_4) = 4\mathrm{Var}(Y) \geq \mathrm{Var}(Y_r + Y_g + Y_b + Y_y). \tag{9}$$

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[5] 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.

3.1 The 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 variable that satisfies the following two conditions:

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

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

In order to achieve a variance reduction it must also be the case that Y and W are *dependent*. Now divide $\mathbb{R}$ into m non-overlapping subintervals, $\Delta_1, \ldots, \Delta_m$, such that $p_j := \mathbf{P}(W \in \Delta_j) > 0$ and $\sum_{j=1}^m p_j = 1$. Note that if W can take any value in $\mathbb{R}$, then the first interval should be $[-\infty, a]$, while the final interval should be $[b, \infty]$ for some finite a and b.

We will use the following 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

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

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

$$\widehat{\theta}_{st,n} = p_1\widehat{\theta}_{1,n_1} + \ldots + p_m\widehat{\theta}_{m,n_m}. \tag{10}$$

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.

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

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 stratification estimator has a lower[6] variance than the usual raw estimator. 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 (8) 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!

Optimizing the Stratified Estimator

We know from (10) that

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

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

$$\mathrm{Var}(\widehat{\theta}_{st,n}) = p_1^2\frac{\sigma_1^2}{n_1} + \ \ldots\ + p_m^2\frac{\sigma_m^2}{n_m} = \sum_{j=1}^m \frac{p_j^2\sigma_j^2}{n_j}. \tag{11}$$

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

$$\min_{n_j} \sum_{j=1}^m \frac{p_j^2\sigma_j^2}{n_j} \quad \text{subject to} \quad n_1 + \ldots + n_m = n. \tag{12}$$

We can easily solve (12) using a Lagrange multiplier. The optimal solution is given by

$$n_j^* = \left(\frac{p_j\sigma_j}{\sum_{j=1}^m p_j\sigma_j}\right) n \tag{13}$$

and the minimized variance is given by $\mathrm{Var}(\widehat{\theta}_{st,n^*}) = \left(\sum_{j=1}^m p_j\sigma_j\right)^2/n$. Note that the solution in (13) 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 to note the following connection between stratified sampling and importance sampling. We know that when we importance sample, we like to sample more often from the* important *region. The choice of n_j in (13) also means that we simulate more often from the important region (in this case the regions with large σ_j's) when we use optimized stratified sampling.*

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

Remark 2 *Note also the connection between stratified sampling and conditional Monte-Carlo. 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*.*

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 pre-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[7] to compute $\widehat{\theta}_{n,st}$ and $\widehat{\sigma}^2_{n,st}$.

[7]Note that $\widehat{\theta}_{n,st} = \sum_{j=1}^m \left(\frac{\sum_{i=1}^{n_j} Y_i^{(j)}}{n_j} \right) p_j$ and $\mathrm{Var}(\widehat{\theta}_{n,st}) = \sum_{j=1}^m \mathrm{Var}\left(\frac{\sum_{i=1}^{n_j} Y_i^{(j)}}{n_j} \right) p_j^2$. Both quantities can be estimated knowing just $\sum_{i=1}^{n_j} Y_i^{(j)}$ and $\sum_{i=1}^{n_j} {Y_i^{(j)}}^2$. 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.

Stratification Simulation Algorithm for Estimating θ

set $\widehat{\theta}_{n,st} = 0; \quad \widehat{\sigma}^2_{n,st} = 0;$
for $j = 1$ to m
  set $sum_j = 0; \quad sum\_squares_j = 0;$
  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 $= \widehat{\theta}_{n,st} \pm z_{1-\alpha/2}\,\widehat{\sigma}_{n,st}$

3.2 Some Applications of Stratified Sampling to Option Pricing

Our examples relate to option pricing in a geometric Brownian motion (GBM) setting. Since GBM is a very poor model of asset price dynamics, however, these examples are not directly applicable in practice. However, the specific techniques and results that we discuss are certainly applicable. For example, we know a multivariate t random vector can be generated using a multivariate normal random vector together with an independent chi-squared random variable. Therefore the techniques we discuss below for multivariate normal random vectors can also be used[8] for multivariate t random vectors. They are also applicable more generally to normal-mixture distributions, Gaussian and t copulas, as well as the simulation of SDE's. As stated earlier, Chapter 9 of Glasserman (2004) should be sufficient to convince any reader of the general applicability of stratified sampling (as well as importance sampling) to risk management.

Example 10 (Pricing a European Call Option)
Suppose we wish to price a European call option where $S_t \sim GBM(r, \sigma^2)$. 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 $\Delta \subseteq \mathbb{R}$, $\mathbf{P}(W \in \Delta)$ 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

[8] See Chapter 9 of Glasserman (2004) for a risk management application in a multivariate t setting.

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. ∎

Example 11 (Pricing an Asian Call Option)
Recall that the *discounted* payoff of an Asian call option is given by

$$Y := e^{-rT} \max\left(0, \frac{\sum_{i=1}^{m} S_{iT/m}}{m} - K\right) \tag{14}$$

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

$$S_{iT/m} = S_0 \ \exp\left((r - \sigma^2/2)\frac{iT}{m} + \sigma\sqrt{\frac{T}{m}}(X_1 + \ \ldots \ + X_i)\right) \tag{15}$$

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 (14) and (15). 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[9] sampling.

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

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

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 is well known to us:

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 $\mathbf{Z} \sim \mathsf{MVN}(\mathbf{0}, \mathbf{I_m})$ and then set
$$\mathbf{X} = \mathbf{C}^T\mathbf{Z} \tag{16}$$
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.

[9] The method we now describe is also useful for pricing other path dependent options. See Glasserman (2004) for further details.

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

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

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

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

Generating such a $\mathbf{V}$ is very easy since $(\mathbf{I_m} - \mathbf{a}^\top\mathbf{a})^\top(\mathbf{I_m} - \mathbf{a}^\top\mathbf{a}) = \mathbf{I_m} - \mathbf{a}^\top\mathbf{a}$. That is, $\mathbf{\Sigma}^\top\mathbf{\Sigma} = \mathbf{\Sigma}$ where $\mathbf{\Sigma} = \mathbf{I_m} - \mathbf{a}^\top\mathbf{a}$, so we can take $\mathbf{C} = \mathbf{\Sigma}$ in (16).

We can now 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

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

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

Step 1: Generate $\left[\frac{1}{\sqrt{m}} \sum_{i=1}^m X_i \;\middle|\; \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i \in \left[\frac{a}{\sqrt{m}}, \frac{b}{\sqrt{m}} \right] \right]$. 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 $\left(\mathsf{N}(0,1) \;\middle|\; \mathsf{N}(0,1) \in \left[\frac{a}{\sqrt{m}}, \frac{b}{\sqrt{m}} \right] \right)$ 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

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

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

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

where we again assume that $S_t \sim GBM(r, \sigma^2)$. Using conditional Monte-Carlo, we can write (why?) $C_0 = \mathrm{E}[Y]$ where

$$Y := e^{-rT/2} \left(c(S_{T/2}, T/2, K_1, r, \sigma) I_{\{S_{T/2} \le L\}} \;+\; c(S_{T/2}, T/2, K_2, r, \sigma) I_{\{S_{T/2} \ge L\}} \right) \tag{17}$$

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? ∎