IEOR E4703: Monte-Carlo Simulation

Further Variance Reduction Methods

Martin Haugh

Department of Industrial Engineering and Operations Research
Columbia University
Email: martin.b.haugh@gmail.com

Outline

Importance Sampling
- Introduction and Main Results
- Tilted Densities
- Estimating Conditional Expectations

An Application to Portfolio Credit Risk
- Independent Default Indicators
- Dependent Default Indicators

Stratified Sampling
- The Stratified Sampling Algorithm
- Some Applications to Option Pricing

Just How Unlucky is a 25 Standard Deviation Return?

Suppose we wish to estimate $\theta := P(X \geq 25) = \mathsf{E}[I_{\{X \geq 25\}}]$ where $X \sim \mathsf{N}(0,1)$.

Standard Monte-Carlo approach 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 $\hat{\theta}_n = \sum_{j=1}^{n} I_j / n$
4. Compute approximate 95% CI as

$$\hat{\theta}_n \pm 1.96 \times \hat{\sigma}_n / \sqrt{n}.$$

Question: Why is this a bad idea?

Question: Beyond knowing that θ is very small, do we even care about estimating θ accurately?

The Importance Sampling Estimator

Suppose we wish to estimate $\theta = \mathsf{E}_f[h(X)]$ where X has PDF f.

Let g be another PDF with the property that $g(x) \neq 0$ whenever $f(x) \neq 0$. Then

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

- has very important implications for estimating θ.

Original simulation method generates n samples of X from f and sets $\hat{\theta}_n = \sum h(X_j)/n$.

Alternative method is to generate n values of X from g and set

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

The Importance Sampling Estimator

$\hat{\theta}_{n,is}$ is then an unbiased estimator of θ.

We often define

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

- so that $\theta = \mathsf{E}_g[h^*(X)]$.

We refer to f and g as the original and importance sampling densities, respectively.

Also refer to f/g as the likelihood ratio.

Just How Unlucky is a 25 Standard Deviation Return?

Recall we want to estimate $\theta = P(X \geq 25) = \mathsf{E}[I_{\{X \geq 25\}}]$ when $X \sim \mathsf{N}(0, 1)$.

We write

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

and where now $X \sim \mathsf{N}(\mu, 1)$.

Leads to a much more efficient estimator if say we take $\mu \approx 25$.

Find an approx. 95% CI for θ is given by $[3.053, \; 3.074] \times 10^{-138}$.

The General Formulation

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with joint PDF $f(x_1, \ldots, x_n)$.

Suppose we wish to estimate $\theta = \mathsf{E}_f[h(\mathbf{X})]$.

Let $g(x_1, \ldots, x_n)$ be another PDF such that $g(\mathbf{x}) \neq 0$ whenever $f(\mathbf{x}) \neq 0$. Then

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

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

Obtaining a Variance Reduction

We wish to estimate $\theta = \mathsf{E}_f[h(\mathbf{X})]$ where $\mathbf{X}$ is a random vector with joint PDF, f.

We assume wlog (why?) that $h(\mathbf{X}) \geq 0$.

Now let g be another density with support equal to that of f.

Then we know

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

and this gives rise to two estimators:

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

Obtaining a Variance Reduction

The variance of importance sampling estimator is given by

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

Variance of original estimator is given by

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

So **reduction** in variance is

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

– would like this reduction to be **positive**.

Obtaining a Variance Reduction

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

Could define important part of f to be that region, A say, in the support of f where $h(\mathbf{x})^2 f(\mathbf{x})$ is large.

But by the above observation, 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 term importance sampling.

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 $h(\mathbf{x}) \neq 0$.

Obtaining a Variance Reduction

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:

- Suppose we choose $g(\mathbf{x}) = h(\mathbf{x})f(\mathbf{x})/\theta$.
- Then easy to see that

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

  so that we have a zero variance estimator!
- Would only need one sample with this choice of g.

Of course this is not feasible in practice. Why?

But this observation can often guide us towards excellent choices of g that lead to extremely large variance reductions.

The Maximum Principle

Saw 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, i.e. a zero-variance estimator.

This suggests that if we could choose $g \approx hf$, then might reasonably expect to obtain a large variance reduction.

One possibility is to choose g so that it has a similar shape to hf.

In particular, could 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 are applying the maximum principle.

Of course this only partially defines g as there are infinitely many density functions that could take their maximum value at $\mathbf{x}^*$.

Nevertheless, often enough to obtain a significant variance reduction.

In practice, often take g to be from the same family of distributions as f.

The Maximum Principle

e.g. If f is multivariate normal, then might also take g to be multivariate normal but with a different mean and / or variance-covariance matrix.

We wish to estimate $\theta = \mathsf{E}[h(X)] = \mathsf{E}[X^4 e^{X^2/4} I_{\{X \geq 2\}}]$ 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

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

Then

$$\theta = \mathsf{E}_g[h^*(X)] = \mathsf{E}_g[X^4 e^{X^2/4} e^{-\sqrt{8}X+4} I_{\{X \geq 2\}}]$$

where $g(\cdot)$ denotes the $\mathsf{N}(\sqrt{8},1)$ PDF.

Pricing an Asian Option

e.g. $S_t \sim GBM(r, \sigma^2)$, where S_t is the stock price at time t.

Want to price an Asian call option whose payoff at time T is given by

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

where $\mathbf{S} := \{S_{iT/m} : i = 1, \ldots, m\}$ and K is the strike price.

The price of this option is then given by $C_a = \mathsf{E}_0^Q[e^{-rT} h(\mathbf{S})]$.

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 $\mathsf{N}(0,1)$.

If f is the risk-neutral PDF of $\mathbf{X} = (X_1, \ldots, X_m)$, then (with mild abuse of notation) may write

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

Pricing an Asian Option

If K very large relative to S_0 then the option is deep out-of-the-money and using simulation amounts to performing a rare event simulation.

As a result, estimating C_a using importance sampling will often result in a large variance reduction.

To apply importance sampling, we need to choose the sampling density, g.

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

- can be found using numerical methods.

Potential Problems with the Maximum Principle

Sometimes applying the maximum principle to choose g is 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.

Difficulties with Importance Sampling

Most difficult aspect to importance sampling is in choosing a good sampling density, g.

In general, need 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**.

Possible in factto 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.

Tilted Densities

Suppose f is light-tailed so that it has a moment generating function (MGF).

Then a common way of generating the sampling density, g, from the original density, f, is to use the MGF of f.

Let $M_x(t) := \mathsf{E}[e^{tX}]$ denote the MGF.

Then for $-\infty < t < \infty$, a tilted density of f is given by

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

If we want to sample more often from region where X tends to be large (and positive), then could use f_t with $t > 0$ as our sampling density g.

Similarly, if we want to sample more often from the region where X tends to be large (and negative), then could use f_t with $t < 0$.

An Example: Sums of Independent Random Variables

Suppose $X_1, \ldots, X_n$ are independent r. vars, where X_i has density $f_i(\cdot)$.

Let $S_n := \sum_{i=1}^n X_i$ and want to estimate $\theta := P(S_n \geq a)$ for some constant, a.

If a is large then can use importance sampling.

Since S_n is large when 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 $t > 0$.

May then write

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

where $\mathsf{E}_t[.]$ 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.

An Example: Sums of Independent Random Variables

If we write $M(t) := \prod_{i=1}^{n} M_i(t)$, then easy to see the importance sampling estimator, $\hat{\theta}_{n,i}$, satisfies

$$\hat{\theta}_{n,i} \leq M(t)e^{-ta}. \tag{2}$$

Therefore a good choice of t would be that value that minimizes the bound in (2)

- why is this?

Can minimize the bound by minimizing $\log(M(t)e^{-ta}) \; = \; \log(M(t)) - ta$.

Straightforward to check that minimizing value of t satisfies $\mu_t = a$ where $\mu_t := \mathsf{E}_t[S_n]$.

Applications From Insurance: Estimating Ruin Probabilities

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 $\mathsf{E}[X_1] > 0$ and the X_i's are IID with MGF, $M_X(t)$, then $P(\tau_a < \infty) = 1$.

The case of interest is then when $\mathsf{E}[X_1] \leq 0$. We obtain

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

where $\psi(t) := \log(M_X(t))$ is the cumulant generating function.

Estimating Ruin Probabilities

Note that if $\mathsf{E}_t[X_1] > 0$ then $\tau_a < \infty$ almost surely and so we obtain

$$\theta = \mathsf{E}_t\left[e^{-tS_{\tau_a} + \tau_a \psi(t)}\right].$$

In fact, importance sampling this way ensures the simulation stops almost surely!

Question: How can we use $\psi(\cdot)$ to choose a good value of t?

This problem has direct applications to the estimation of ruin probabilities in the context of insurance risk.

Estimating Ruin Probabilities

e.g. 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.

Estimating Conditional Expectations

Importance sampling also very useful for computing conditional expectations when the event being conditioned upon is a rare event.

e.g. Suppose we wish to estimate $\theta = \mathsf{E}[h(\mathbf{X})|\mathbf{X} \in A]$ 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

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

so

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

Since A is a rare event we would be better off using a sampling density, g, that makes A more likely to occur.

Then we would have

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

Estimating Conditional Expectations

To estimate θ using importance sampling, we generate $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with density g, and set

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

In contrast to our usual estimators, $\hat{\theta}_{n,i}$ is no longer an average of n IID random variables but instead, it is the **ratio** of two such averages

- has implications for computing approximate confidence intervals for θ
- in particular, confidence intervals should now be estimated using **bootstrap** techniques.

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

Bernoulli Mixture Models

Definition: Let $p < m$ and let $\mathbf{\Psi} = (\Psi_1, \dots, \Psi_p)^\top$ be a p-dimensional random vector.

Then we say the random vector $\mathbf{Y} = (Y_1, \dots, Y_m)^\top$ 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

$$P(Y_i = 1 \mid \mathbf{\Psi} = \psi) = p_i(\psi).$$

An Application to Portfolio Credit Risk

We consider a portfolio loss of the form $L = \sum_{i=1}^{m} e_i Y_i$

- e_i is the deterministic and positive exposure to the i^{th} credit
- Y_i is the default indicator with corresponding default probability, p_i.

Assume also that $\mathbf{Y}$ follows a Bernoulli mixture model.

Want to estimate $\theta := P(L \geq c)$ where $c >> \mathsf{E}[L]$.

Note that a good importance sampling distribution for θ should also work well for estimating 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 ...

Case 1: Independent Default Indicators

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

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

Let Q_t be the corresponding **tilted** probability measure so that

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

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.

Case 1: Independent Default Indicators

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 in the partial sum example suggests that we should take t equal to that value that solves

$$\mathsf{E}_t[L] = \sum_{i=1}^{m} q_{i,t} e_i = c.$$

This value can be found easily using numerical methods.

Case 2: Dependent Default Indicators

Suppose now that there is a p-dimensional factor vector, $\mathbf{\Psi}$.

We assume the default indicators are independent with default probabilities $p_i(\psi)$ **conditional** on $\mathbf{\Psi} = \psi$.

Suppose also that $\mathbf{\Psi} \sim \mathsf{MVN}_p(\mathbf{0}, \mathbf{\Sigma})$.

The Monte-Carlo scheme for estimating θ is to first simulate $\mathbf{\Psi}$ and to then simulate $\mathbf{Y}$ conditional on $\mathbf{\Psi}$.

Can apply importance sampling to the second step using our discussion of independent default indicators.

However, can also apply importance sampling to the first step, i.e. the simulation of $\mathbf{\Psi}$.

Case 2: Dependent Default Indicators

A natural way to do this is to simulate $\mathbf{\Psi}$ form the $\mathsf{MVN}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution for some $\boldsymbol{\mu} \in \mathbb{R}^p$.

Corresponding likelihood ratio, $r_{\boldsymbol{\mu}}(\mathbf{\Psi})$, is given by ratio of the two multivariate normal densities.

It satisfies

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

Case 2: How Do We Choose μ?

Recall the quantity of interest is $\theta := P(L \geq c) = \mathsf{E}[P(L \geq c \mid \mathbf{\Psi})]$.

Know from earlier discussion that we'd like to choose importance sampling density, $g^*(\mathbf{\Psi})$, so that

$$g^*(\mathbf{\Psi}) \propto \; P(L \geq c \mid \mathbf{\Psi}) \; \exp(-\frac{1}{2}\mathbf{\Psi}^\top \mathbf{\Sigma}^{-1}\mathbf{\Psi}). \tag{3}$$

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.

Maximum principle applied to the $\mathsf{MVN}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ distribution would then suggest taking $\boldsymbol{\mu}$ equal to the value of $\mathbf{\Psi}$ which maximizes the rhs of (3).

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

1. Generate $\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_n$ independently from the $\mathsf{MVN}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ 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 $\hat{\theta}_{n_1}^{IS}(\mathbf{\Psi}_i)$ be the corresponding estimator based on n_1 samples.

3. Full importance sampling estimator then given by

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

Stratified Sampling: A Motivating Example

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.

Stratified Sampling: A Motivating Example

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 $f_I(y)$.

Average payoff, $\theta := \mathsf{E}[Y]$, then estimated by

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

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 $\hat{\theta}_n$?

Question: What if f_g tended to produce very high payoffs and f_b tended to produce very low payoffs?

Question: Is there anything that we could have done to avoid this type of problem occurring?

Stratified Sampling: A Motivating Example

Know each ball color should be selected $1/4$ of the time so we could force this to hold by conducting four separate simulations, one each to estimate $\mathsf{E}[X|I = c]$ for $c = r, g, b, y$.

Note that

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

so an unbiased estimator of θ is obtained by setting

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

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

Question: How does Var $\left(\hat{\theta}_{st,n}\right)$ compare with Var $\left(\hat{\theta}_n\right)$?

To answer this we assume (for now) that $n_c = n/4$ for each c, and that Y_c is a sample from the density, f_c.

Stratified Sampling: A Motivating Example

Then a **fair** comparison of $\text{Var}(\hat{\theta}_n)$ with $\text{Var}(\hat{\theta}_{st,n})$ should compare

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

- $Y_1, \ Y_2, \ Y_3$ and Y_4 are IID samples from the original simulation algorithm
- Y_c's are independent with density $f_c(\cdot)$, for $c = r, g, b, y$.

Now recall the **conditional variance** formula which states

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

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

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

Stratified Sampling

This implies

$$\begin{aligned}\mathsf{Var}(Y_1 + Y_2 + Y_3 + Y_4) &= 4\,\mathsf{Var}(Y) \\ &\geq \mathsf{Var}(Y_r + Y_g + Y_b + Y_y).\end{aligned}$$

Can therefore conclude that using $\hat{\theta}_{st,n}$ leads to a variance reduction.

Variance reduction will be substantial if I accounts for a large fraction of the variance of Y.

Note also that computational requirements for computing $\hat{\theta}_{st,n}$ are similar to those required for computing $\hat{\theta}_n$.

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

The Stratified Sampling Algorithm

Want to estimate $\theta := \mathsf{E}[Y]$ where Y is a random variable.

Let W be another random variable that satisfies the following two conditions:

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

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

- note that Y and W should be **dependent** to achieve a variance reduction.

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

Notation

1. Let $\theta_j \ := \ \mathsf{E}[Y|W \in \Delta_j]$ and $\sigma_j^2 := \mathsf{Var}(Y|W \in \Delta_j)$.
2. 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 $(Y|W \in \Delta_j) \equiv (Y|I = j)$.

Therefore have

$$\theta_j \ = \ \mathsf{E}[Y|I = j] \ = \ \mathsf{E}[Y^{(j)}]$$

and

$$\sigma_j^2 \ = \ \mathsf{Var}(Y|I = j) \ = \ \mathsf{Var}(Y^{(j)}).$$

Stratified Sampling

In particular obtain

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

To estimate θ we only need to estimate the θ_i's since the p_i's are easily computed by condition 1.

And 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

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

Clear that $\hat{\theta}_{st,n}$ will be **unbiased** if each $\hat{\theta}_{i,n_i}$ is **unbiased.**

Obtaining a Variance Reduction

Would like to compare $\text{Var}(\hat{\theta}_n)$ with $\text{Var}(\hat{\theta}_{st,n})$.

First must choose $n_1, \ldots, n_m$ such that $n_1 + \ldots + n_m = n$.

Clearly, optimal to choose the n_i's so as to minimize $\text{Var}(\hat{\theta}_{st,n})$.

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

Then

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

Obtaining a Variance Reduction

But the usual simulation estimator has variance σ^2/n where $\sigma^2 := \mathsf{Var}(Y)$.

Therefore, need only show that $\sum_{j=1}^m p_j \sigma_j^2 < \sigma^2$ to prove the non-optimized stratification estimator has a lower variance than the usual raw estimator.

But the **conditional variance formula** implies

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

and the proof is complete!

Optimizing the Stratified Estimator

We know

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

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

This then implies

$$\text{Var}(\hat{\theta}_{st,n}) \quad = \quad p_1^2 \frac{\sigma_1^2}{n_1} + \ \ldots \ + p_m^2 \frac{\sigma_m^2}{n_m} \quad = \quad \sum_{j=1}^{m} \frac{p_j^2 \sigma_j^2}{n_j}. \tag{7}$$

To minimize $\text{Var}(\hat{\theta}_{st,n})$ must therefore solve the following **constrained optimization** problem:

$$\min_{n_j} \sum_{j=1}^{m} \frac{p_j^2 \sigma_j^2}{n_j} \quad \text{subject to} \quad n_1 + \ldots + n_m = n. \tag{8}$$

Optimizing the Stratified Estimator

Can easily solve (8) using a Lagrange multiplier to obtain

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

Minimized variance is given by

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

Note that the solution (9) makes intuitive sense:

- If p_j large then (other things being equal) makes sense to expend more effort simulating from stratum j.

- If σ_j^2 is large then (other things being equal) makes sense to simulate more often from stratum j so as to get a more accurate estimate of θ_j.

Stratification Simulation Algorithm for Estimating θ

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

Example: Pricing a European Call Option

Wish to price a European call option where we assume $S_t \sim GBM(r, \sigma^2)$.

Then

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

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's 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. We can easily computed $P(W \in \Delta)$ for $\Delta \subseteq \mathbb{R}$.
2. We can easily generate $(Y | W \in \Delta)$.

Therefore clear that we can estimate C_0 using X as a stratification variable.

Example: Pricing an Asian Call Option

The discounted payoff of an Asian call option is given by

$$Y := e^{-rT} \max\left(0, \frac{\sum_{i=1}^{m} S_{iT/m}}{m} - K\right) \tag{10}$$

– its price therefore given by $C_a = \mathsf{E}[\,Y\,]$.

Now each $S_{iT/m}$ may be expressed as

$$S_{iT/m} = S_0 \; \exp\left((r - \sigma^2/2)\frac{iT}{m} + \sigma\sqrt{\frac{T}{m}}(X_1 + \;\ldots\; + X_i)\right) \tag{11}$$

where the X_i's are IID $\mathsf{N}(0,1)$.

Can therefore write $C_a = \mathsf{E}\left[h(X_1, \ldots, X_m)\right]$ where $h(.)$ given implicitly by (10) and (11).

Example: Pricing an Asian Call Option

Can estimate C_a using stratified sampling but must first choose a stratification variable, W.

One possible choice would be to set $W = X_j$ for some j.

But 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, i.e. must show that

(1) $P(W \in \Delta)$ is easily computed

(2) $(Y \,|\, W \in \Delta)$ is easily generated.

Computing $P(W \in \Delta)$

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

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

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

And if $\Delta = (-\infty, a]$, then $P(W \in \Delta) = \Phi\left(\frac{a}{\sqrt{m}}\right)$.

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

Need two results from the theory of multivariate normal random variables:

Result 1:

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

$$\mathbf{X} = \mathbf{C}^T\mathbf{Z} \tag{12}$$

  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

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

Therefore, to generate $\{(Z_1, \ldots, Z_m) | \sum_{i=1}^{m} a_i Z_i = w\}$ just need to generate $\mathbf{V}$ 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 (12).

Back to Generating $(Y|W \in \Delta)$

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:

Back to Generating $(Y|W \in \Delta)$

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

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

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

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 Result 2 and the comments that follow.