Stochastic Simulation

Jan-Pieter Dorsman[1] & Michel Mandjes[1,2,3]

[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam

[2]CWI, Amsterdam

[3]Eurandom, Eindhoven

University of Amsterdam,
Fall, 2018

Chapter V
VARIANCE REDUCTION METHODS

The simulation paradigm so far

Recall the way we went about simulating a performance measure $z := \mathbb{E}[Z]$:

- Perform independent samples $Z_1, \ldots, Z_R$.
- Compute the *crude Monte Carlo estimator* (CMC):

$$\hat{z}_R := \frac{1}{R} \sum_{r=1}^{R} Z_r$$

  for large R.

The simulation paradigm so far

Recall the way we went about simulating a performance measure $z := \mathbb{E}[Z]$:

- Perform independent samples $Z_1, \ldots, Z_R$.
- Compute the *crude Monte Carlo estimator* (CMC):

$$\hat{z}_R := \frac{1}{R} \sum_{r=1}^{R} Z_r$$

  for large R.

We found that this estimator is unbiased (i.e. $\mathbb{E}[\hat{z}_R] = z$), which is a good thing.

The simulation paradigm so far

Recall the way we went about simulating a performance measure $z := \mathbb{E}[Z]$:

- Perform independent samples $Z_1, \ldots, Z_R$.
- Compute the *crude Monte Carlo estimator* (CMC):

$$\hat{z}_R := \frac{1}{R} \sum_{r=1}^{R} Z_r$$

  for large R.

We found that this estimator is unbiased (i.e. $\mathbb{E}[\hat{z}_R] = z$), which is a good thing.

Moreover, the use of confidence intervals gave us an idea of how reliable our estimate is. With probability α, the true value of z will be in the interval

$$\left(\hat{z}_R - q_\alpha \frac{\sigma}{\sqrt{R}}, \hat{z}_R + q_\alpha \frac{\sigma}{\sqrt{R}} \right)$$

Possible issues with Crude Monte Carlo sampling

- Q: Is the use of a CMC estimator always a guarantee for success?

Possible issues with Crude Monte Carlo sampling

- Q: Is the use of a CMC estimator always a guarantee for success?
- A: Well, let's think about it.

Possible issues with Crude Monte Carlo sampling

- Q: Is the use of a CMC estimator always a guarantee for success?
- A: Well, let's think about it.

There are some caveats.

Possible issues with Crude Monte Carlo sampling

- Q: Is the use of a CMC estimator always a guarantee for success?
- A: Well, let's think about it.

There are some caveats.

- In some applications it may take an enormous computation time, amount of memory or other resource to generate a single replication, say in the order of hours, let alone R of them.
- The value of z might be extremely small leading to long computation times or precision errors (e.g. the estimation of $\mathbb{P}(\mathcal{N}(0,1) > 1000)$...).

In these cases, the CMC estimator can simply require a too large R in order for the confidence interval to be of an acceptable width.

Possible issues with Crude Monte Carlo sampling

- Q: Is the use of a CMC estimator always a guarantee for success?
- A: Well, let's think about it.

There are some caveats.

- In some applications it may take an enormous computation time, amount of memory or other resource to generate a single replication, say in the order of hours, let alone R of them.
- The value of z might be extremely small leading to long computation times or precision errors (e.g. the estimation of $\mathbb{P}(\mathcal{N}(0,1) > 1000)$...).

In these cases, the CMC estimator can simply require a too large R in order for the confidence interval to be of an acceptable width.

We could therefore think of so-called *variance reduction methods* to decrease the variance of the CMC estimator.

Possible issues with VRMs

In the decision of whether to use VRMs, it is wise to take a few things in account:

- For VRMs to be worthwhile, the reduction of variance should be substantial. If the variance is only cut by 25%, the reduction of the width of the confidence interval is only $1 - \sqrt{0.75} = 13.4\%$ with the same number of replications. It may not be worth to go through the hassle.
- Also, computing alternative estimators may take more computation time, possibly actually negating the positive effects of variance reduction.
- And then we didn't think of implementation time, the implementation of alternative estimators usually takes more time as well.

Possible issues with VRMs

In the decision of whether to use VRMs, it is wise to take a few things in account:

- For VRMs to be worthwhile, the reduction of variance should be substantial. If the variance is only cut by 25%, the reduction of the width of the confidence interval is only $1 - \sqrt{0.75} = 13.4\%$ with the same number of replications. It may not be worth to go through the hassle.
- Also, computing alternative estimators may take more computation time, possibly actually negating the positive effects of variance reduction.
- And then we didn't think of implementation time, the implementation of alternative estimators usually takes more time as well.

However, we're still going to talk about variance reduction methods, because sometimes **they really are a life saver**, as we will see.

Chapter V.2
Control Variates

Control variates

Definition: W is called a control variate for Z if Z and W are strongly correlated (either positively or negatively), and $\mathbb{E}[W]$ is known.

If such a control variate exists, they can be used for variance reduction.

Control variates

Simulation algorithm:

1. Generate replications of (Z, W) and compute the empirical means $\hat{z}$ and $\hat{w}$.
2. Compute the control variate estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ for an appropriate value of α.

The control variate estimator is the alternative proposed estimator for $z = \mathbb{E}[Z]$. Is this is a good estimator?

Naive control variate algorithm

Q: Is the estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ unbiased?

Naive control variate algorithm

Q: Is the estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ unbiased?

A: $\mathbb{E}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = z$, so yes. In that sense, we're good.

Naive control variate algorithm

Q: Is the estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ unbiased?

A: $\mathbb{E}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = z$, so yes. In that sense, we're good.

Q: What is the variance of this estimator?

Naive control variate algorithm

Q: Is the estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ unbiased?

A: $\mathbb{E}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = z$, so yes. In that sense, we're good.

Q: What is the variance of this estimator?

A: $\mathrm{Var}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = \mathrm{Var}[\hat{z}] + \alpha^2 \mathrm{Var}[\hat{w}] + 2\alpha \mathrm{Cov}[\hat{z}, \hat{w}]$.

Naive control variate algorithm

Q: Is the estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ unbiased?

A: $\mathbb{E}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = z$, so yes. In that sense, we're good.

Q: What is the variance of this estimator?

A: $\mathrm{Var}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = \mathrm{Var}[\hat{z}] + \alpha^2 \mathrm{Var}[\hat{w}] + 2\alpha \mathrm{Cov}[\hat{z}, \hat{w}]$.

Q: So, is this variance better than the crude Monte Carlo estimator?

Naive control variate algorithm

Q: Is the estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ unbiased?

A: $\mathbb{E}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = z$, so yes. In that sense, we're good.

Q: What is the variance of this estimator?

A: $\mathrm{Var}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = \mathrm{Var}[\hat{z}] + \alpha^2 \mathrm{Var}[\hat{w}] + 2\alpha \mathrm{Cov}[\hat{z}, \hat{w}]$.

Q: So, is this variance better than the crude Monte Carlo estimator?

A: It all depends on α! No general answer possible. For $\alpha = 0$, there is no difference at all.

Naive control variate algorithm

Q: Is the estimator $\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])$ unbiased?

A: $\mathbb{E}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = z$, so yes. In that sense, we're good.

Q: What is the variance of this estimator?

A: $\mathrm{Var}[\hat{z} + \alpha(\hat{w} - \mathbb{E}[W])] = \mathrm{Var}[\hat{z}] + \alpha^2 \mathrm{Var}[\hat{w}] + 2\alpha \mathrm{Cov}[\hat{z}, \hat{w}]$.

Q: So, is this variance better than the crude Monte Carlo estimator?

A: It all depends on α! No general answer possible. For $\alpha = 0$, there is no difference at all.

The million dollar question now is: how should we choose α?

Control variates

We want to minimise the estimator variance

$$\mathrm{Var}\left[\hat{z}\right] + \alpha^2 \mathrm{Var}\left[\hat{w}\right] + 2\alpha \mathrm{Cov}\left[\hat{z}, \hat{w}\right].$$

This is a quadratic polynomial, hence optimised for

$$\alpha = -\frac{\mathrm{Cov}\left[\hat{z}, \hat{w}\right]}{\mathrm{Var}\left[\hat{w}\right]} = -\frac{\mathrm{Cov}\left[Z, W\right]}{\mathrm{Var}\left[W\right]}.$$

Check that the latter equation holds!

Control variates

We want to minimise the estimator variance

$$\mathrm{Var}\left[\hat{z}\right] + \alpha^2 \mathrm{Var}\left[\hat{w}\right] + 2\alpha \mathrm{Cov}\left[\hat{z}, \hat{w}\right].$$

This is a quadratic polynomial, hence optimised for

$$\alpha = -\frac{\mathrm{Cov}\left[\hat{z}, \hat{w}\right]}{\mathrm{Var}\left[\hat{w}\right]} = -\frac{\mathrm{Cov}\left[Z, W\right]}{\mathrm{Var}\left[W\right]}.$$

Check that the latter equation holds!

Using this α, the variance reduces to

$$\frac{\mathrm{Var}\left[Z\right]}{R} + \left(\frac{\mathrm{Cov}\left[Z, W\right]}{\mathrm{Var}\left[W\right]}\right)^2 \frac{\mathrm{Var}\left[W\right]}{R} - 2\left(\frac{\mathrm{Cov}\left[Z, W\right]}{\mathrm{Var}\left[W\right]}\right) \frac{\mathrm{Cov}\left[Z, W\right]}{R}$$

$$= \frac{\mathrm{Var}\left[Z\right]}{R}\left(1 - \frac{\mathrm{Cov}\left[Z, W\right]^2}{\mathrm{Var}\left[Z\right]\mathrm{Var}\left[W\right]}\right).$$

Control variates

If $\rho = \text{Corr}[Z, W] = \frac{\text{Cov}[Z,W]}{\sqrt{\text{Var}[Z]\text{Var}[W]}}$ is the Pearson correlation coefficient between Z and W, then this equals $\frac{\text{Var}[Z]}{R}(1 - \rho^2)$. This is always smaller than $\text{Var}[\hat{z}] = \frac{\text{Var}[Z]}{R}$!

Great! Are we done now?

Control variates

If $\rho = \text{Corr}\,[Z, W] = \frac{\text{Cov}[Z,W]}{\sqrt{\text{Var}[Z]\text{Var}[W]}}$ is the Pearson correlation coefficient between Z and W, then this equals $\frac{\text{Var}[Z]}{R}(1 - \rho^2)$. This is always smaller than $\text{Var}\,[\hat{z}] = \frac{\text{Var}[Z]}{R}$!

Great! Are we done now? No!

Control variates

If $\rho = \text{Corr}[Z, W] = \frac{\text{Cov}[Z,W]}{\sqrt{\text{Var}[Z]\text{Var}[W]}}$ is the Pearson correlation coefficient between Z and W, then this equals $\frac{\text{Var}[Z]}{R}(1-\rho^2)$. This is always smaller than $\text{Var}[\hat{z}] = \frac{\text{Var}[Z]}{R}$!

Great! Are we done now? No! We don't know $\text{Cov}[Z, W]$ and $\text{Var}[W]$, so we can't compute α. As before, use **unbiased** sample estimates for (co-)variances:

$$s_Z^2 = \frac{1}{R-1}\sum_{r=1}^{R}(Z_r - \hat{z})^2, \qquad s_W^2 = \frac{1}{R-1}\sum_{r=1}^{R}(W_r - \hat{w})^2,$$

$$s_{ZW}^2 = \frac{1}{R-1}\sum_{r=1}^{R}(Z_r - \hat{z})(W_r - \hat{w}).$$

Using $\alpha = -\frac{s_{ZW}^2}{s_W^2}$, the variance of the estimator will asymptotically behave as $\frac{\text{Var}[Z]}{R}(1-\rho^2)$.

Variance reduction by control variates

Some remarks:

- This method is only useful when we can find a control variate W, which has a substantial correlation with Z.

Variance reduction by control variates

Some remarks:

- This method is only useful when we can find a control variate W, which has a substantial correlation with Z.
- This method has similarities with classic linear regression. Suppose we have data $(Z_1, W_1), (Z_2, W_2), \ldots, (Z_R, W_R)$ and we assume that $Z = m + \beta W = m' + \beta(W - \mathbb{E}[W])$.

  We would then fit the parameters

$$m' = \hat{z}, \qquad \beta = \frac{\sum_{r=1}^{R}(Z_r - \hat{z})(W_r - \hat{w})}{\sum_{r=1}^{R}(W_r - \hat{w})^2} = -\alpha.$$

  ρ^2 would be the squared coefficient of determination.

Variance reduction by control variates

Some remarks:

- This method is only useful when we can find a control variate W, which has a substantial correlation with Z.
- This method has similarities with classic linear regression. Suppose we have data $(Z_1, W_1), (Z_2, W_2), \ldots, (Z_R, W_R)$ and we assume that $Z = m + \beta W = m' + \beta(W - \mathbb{E}[W])$.

  We would then fit the parameters

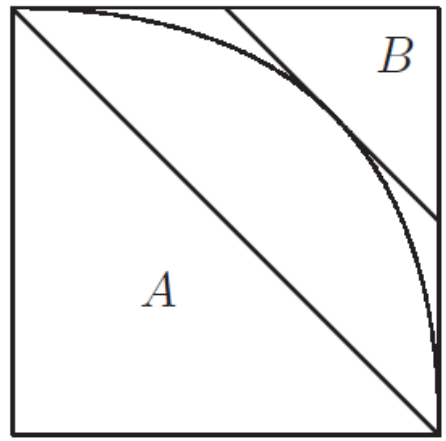
$$m' = \hat{z}, \qquad \beta = \frac{\sum_{r=1}^{R}(Z_r - \hat{z})(W_r - \hat{w})}{\sum_{r=1}^{R}(W_r - \hat{w})^2} = -\alpha.$$

  ρ^2 would be the squared coefficient of determination.
- This similarity is not entirely one-to-one: in linear regression, we require residuals to be normal and Z and W to be linearly correlated. We make no such assumptions here.

Control variates: Example of application

In the homework, you have found a way to simulate π. Suppose that $Z = 4\mathbb{1}_{\{U_1^2+U_2^2\leq 1\}}$. Then,

$$\mathbb{E}[Z] = 4\mathbb{P}\left(U_1^2 + U_2^2 \leq 1\right) = \pi.$$



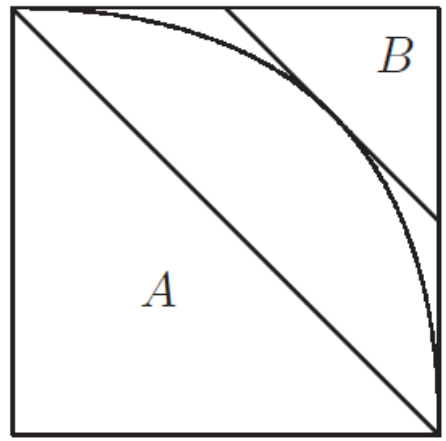
Suppose we wish to simulate π. Possible control variates:

- $W_1 = \mathbb{1}_{\{U_1+U_2\leq 1\}}$. Essential: $\mathbb{E}[W_1] = \frac{1}{2}$ known.
- $W_2 = \mathbb{1}_{\{U_1+U_2\geq \sqrt{2}\}}$. Essential: $\mathbb{E}[W_2] = (2-\sqrt{2})^2/2$ known.

Control variates: Example of application

In the homework, you have found a way to simulate π. Suppose that $Z = 4\mathbb{1}_{\{U_1^2+U_2^2\leq 1\}}$. Then,

$$\mathbb{E}[Z] = 4\mathbb{P}\left(U_1^2 + U_2^2 \leq 1\right) = \pi.$$



Suppose we wish to simulate π. Possible control variates:

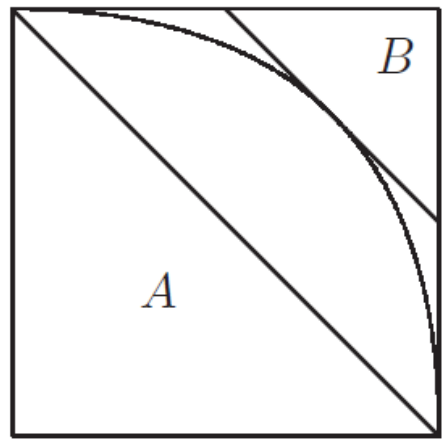
- $W_1 = \mathbb{1}_{\{U_1+U_2\leq 1\}}$. Essential: $\mathbb{E}[W_1] = \frac{1}{2}$ known.
- $W_2 = \mathbb{1}_{\{U_1+U_2\geq \sqrt{2}\}}$. Essential: $\mathbb{E}[W_2] = (2-\sqrt{2})^2/2$ known.

Which one to use?

Control variates: Example of application

In the homework, you have found a way to simulate π. Suppose that $Z = 4\mathbb{1}_{\{U_1^2+U_2^2\leq 1\}}$. Then,

$$\mathbb{E}[Z] = 4\mathbb{P}\left(U_1^2 + U_2^2 \leq 1\right) = \pi.$$



Suppose we wish to simulate π. Possible control variates:

- $W_1 = \mathbb{1}_{\{U_1+U_2\leq 1\}}$. Essential: $\mathbb{E}[W_1] = \frac{1}{2}$ known.
- $W_2 = \mathbb{1}_{\{U_1+U_2\geq \sqrt{2}\}}$. Essential: $\mathbb{E}[W_2] = (2-\sqrt{2})^2/2$ known.

Which one to use? Clearly W_2, since
$|\mathrm{Corr}[Z, W_2]| > |\mathrm{Corr}[Z, W_1]|$.

Multiple control variates

We could also use both of them! When using multiple controls, the single control variate W can be replaced by

$$\boldsymbol{W} = (W_1, \ldots, W_p).$$

The multiple-control estimator is then given by

$$\hat{z} + \sum_{i=1}^{p} \alpha_i (\hat{w}_i - \mathbb{E}[W_i]) = \hat{z} + \boldsymbol{\alpha}^T (\hat{\boldsymbol{w}} - \mathbb{E}[\boldsymbol{W}]).$$

Multiple control variates

Let the covariance matrix of $(Z, \boldsymbol{W})$ be given by

$$\begin{pmatrix} \sigma^2 & \boldsymbol{\Sigma}_{Z\boldsymbol{W}} \\ \boldsymbol{\Sigma}_{\boldsymbol{W}Z} & \boldsymbol{\Sigma}_{\boldsymbol{W}\boldsymbol{W}} \end{pmatrix}.$$

The variance of the estimator $\hat{z} + \boldsymbol{\alpha}^T(\hat{\boldsymbol{w}} - \mathbb{E}[\boldsymbol{W}])$ is given by

$$\frac{1}{R}\left(\mathrm{Var}[Z] + \boldsymbol{\alpha}^T\boldsymbol{\Sigma}_{\boldsymbol{W}\boldsymbol{W}}\boldsymbol{\alpha} + 2\boldsymbol{\Sigma}_{Z\boldsymbol{W}}\boldsymbol{\alpha}\right).$$

Multiple control variates

Let the covariance matrix of $(Z, \boldsymbol{W})$ be given by

$$\begin{pmatrix} \sigma^2 & \boldsymbol{\Sigma}_{Z\boldsymbol{W}} \\ \boldsymbol{\Sigma}_{\boldsymbol{W}Z} & \boldsymbol{\Sigma}_{\boldsymbol{W}\boldsymbol{W}} \end{pmatrix}.$$

The variance of the estimator $\hat{z} + \boldsymbol{\alpha}^T(\hat{\boldsymbol{w}} - \mathbb{E}[\boldsymbol{W}])$ is given by

$$\frac{1}{R}\left(\mathrm{Var}[Z] + \boldsymbol{\alpha}^T\boldsymbol{\Sigma}_{\boldsymbol{W}\boldsymbol{W}}\boldsymbol{\alpha} + 2\boldsymbol{\Sigma}_{Z\boldsymbol{W}}\boldsymbol{\alpha}\right).$$

One can show that this expression is minimised by choosing $\boldsymbol{\alpha} = -\boldsymbol{\Sigma}_{\boldsymbol{W}\boldsymbol{W}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{W}Z}$. Then, the variance becomes

$$\frac{1}{R}(\mathrm{Var}[Z] - \boldsymbol{\Sigma}_{Z\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{W}\boldsymbol{W}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{W}Z}) = \frac{\mathrm{Var}[Z]}{R}(1 - \rho^2_{Z\boldsymbol{W}}).$$

The coefficient $\rho^2_{Z\boldsymbol{W}} = \boldsymbol{\Sigma}_{Z\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{W}\boldsymbol{W}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{W}Z}/\sigma^2_Z$ is the *multiple squared correlation coefficient*. Think of it as the squared coefficient of determination in linear regression!

Multiple control variates

Again, we will have to use the sample estimates

$$\boldsymbol{S}_{Z\boldsymbol{W}} = \frac{1}{R-1}\sum_{r=1}^{R}(Z_r - \hat{z})(\boldsymbol{W}_r - \hat{\boldsymbol{W}}),$$

$$\boldsymbol{S}_{\boldsymbol{W}\boldsymbol{W}} = \frac{1}{R-1}\sum_{r=1}^{R}(\boldsymbol{W}_r - \hat{\boldsymbol{W}})^T(\boldsymbol{W}_r - \hat{\boldsymbol{W}}),$$

so that the estimate becomes

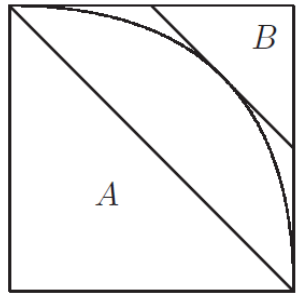
$$\hat{z} - (\boldsymbol{S}_{\boldsymbol{W}\boldsymbol{W}}^{-1}\boldsymbol{S}_{\boldsymbol{W}Z})^T(\hat{\boldsymbol{w}} - \mathbb{E}[\boldsymbol{W}])$$

with variance

$$\frac{1}{R}\left(s_Z^2 - \boldsymbol{S}_{Z\boldsymbol{W}}\boldsymbol{S}_{\boldsymbol{W}\boldsymbol{W}}^{-1}\boldsymbol{S}_{\boldsymbol{W}Z}\right).$$

Again, there exist many similarities with linear regression!

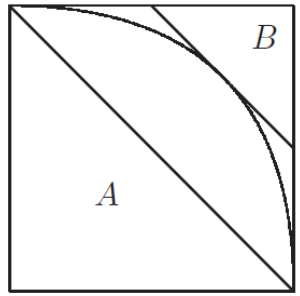
Control variates: Back to our example



Candidate control variates:

- $W_1 = \mathbb{1}_{\{U_1+U_2\leq 1\}}$. $\mathbb{E}[W_1] = \frac{1}{2}$ known.
- $W_2 = \mathbb{1}_{\{U_1+U_2\geq \sqrt{2}\}}$. $\mathbb{E}[W_2] = (2-\sqrt{2})^2/2$ known.
- $W_3 = (U_1+U_2-1)\mathbb{1}_{\{1<U_1+U_2<\sqrt{2}\}}$ with $\mathbb{E}[W_3] = (\sqrt{2}-1)^2/2 - (\sqrt{2}-1)^3/3$.

Control variates: Back to our example



Candidate control variates:

- $W_1 = \mathbb{1}_{\{U_1+U_2\leq 1\}}$. $\mathbb{E}[W_1] = \frac{1}{2}$ known.
- $W_2 = \mathbb{1}_{\{U_1+U_2\geq \sqrt{2}\}}$. $\mathbb{E}[W_2] = (2-\sqrt{2})^2/2$ known.
- $W_3 = (U_1+U_2-1)\mathbb{1}_{\{1<U_1+U_2<\sqrt{2}\}}$ with $\mathbb{E}[W_3] = (\sqrt{2}-1)^2/2 - (\sqrt{2}-1)^3/3$.

We get the following values of $1-\rho^2$ when using various subsets of the three control variates:

1	2	3	1,2	1,3	2,3	1,2,3
0.727	0.242	0.999	0.222	0.620	0.181	0.175

Note that using W_3 alone is useless, but e.g. using W_1 and W_3 reduces variance much more than just using W_1!

Chapter V.3
Antithetic Sampling

Antithetic sampling

The idea of the method is as follows. Using classic Crude Monte Carlo, we generate i.i.d. replicates

$$Z_1, Z_2, \ldots, Z_R$$

and estimate $\mathbb{E}[Z]$ by $\hat{z}_R = \frac{1}{R}\sum_{i=1}^{R} Z_i$. We have $\mathrm{Var}[\hat{z}_R] = \frac{\mathrm{Var}[Z]}{R}$.

Antithetic sampling

The idea of the method is as follows. Using classic Crude Monte Carlo, we generate i.i.d. replicates

$$Z_1, Z_2, \ldots, Z_R$$

and estimate $\mathbb{E}[Z]$ by $\hat{z}_R = \frac{1}{R}\sum_{i=1}^{R} Z_i$. We have $\mathrm{Var}[\hat{z}_R] = \frac{\mathrm{Var}[Z]}{R}$.

Under antithetic sampling, we generate i.i.d. pairs

$$(Z_1, Z_2), (Z_3, Z_4), \ldots, (Z_{R-1}, Z_R),$$

where

- the Z_i are all distributed like Z is.
- the pairs are mutually independent, but the Z_i within a pair are dependent.

We then still compute the unbiased estimator

$$\hat{z}_{ant} = \frac{1}{R}\sum_{i=1}^{R} Z_i.$$

Antithetic sampling

Let's check the variance of the estimator.

$$\begin{aligned}
\mathrm{Var}\left[\hat{z}_{ant}\right] &= \mathrm{Var}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] \\
&= \frac{1}{2R}\mathrm{Var}\left[Z_1 + Z_2\right] \\
&= \frac{1}{2R}(2\mathrm{Var}\left[Z\right] + 2\mathrm{Cov}\left[Z_1, Z_2\right]) \\
&= \frac{\mathrm{Var}\left[Z\right]}{R}(1 + \mathrm{Corr}\left[Z_1, Z_2\right]).
\end{aligned}$$

Antithetic sampling

Let's check the variance of the estimator.

$$\begin{aligned}
\mathrm{Var}\left[\hat{z}_{ant}\right] &= \mathrm{Var}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] \\
&= \frac{1}{2R}\mathrm{Var}\left[Z_1 + Z_2\right] \\
&= \frac{1}{2R}(2\mathrm{Var}\left[Z\right] + 2\mathrm{Cov}\left[Z_1, Z_2\right]) \\
&= \frac{\mathrm{Var}\left[Z\right]}{R}(1 + \mathrm{Corr}\left[Z_1, Z_2\right]).
\end{aligned}$$

So, beneficial if $\mathrm{Corr}\left[Z_1, Z_2\right]$ negative, and as small as possible!
But how do we achieve this?

Example: uniform antithetic sampling

Often, the quantity of interest has the form $Z = g(U)$ for some monotone function g, where U is a standard-uniform r.v.

U and $1 - U$ are identically distributed and perfectly negatively correlated:

$$\begin{aligned}\mathrm{Corr}\left[U, 1-U\right] &= \frac{\mathbb{E}\left[U(1-U)\right] - \mathbb{E}\left[U\right]\mathbb{E}\left[1-U\right]}{\mathrm{Var}\left[U\right]} \\ &= \frac{\left(\frac{1}{2} - \frac{1}{3}\right) - \frac{1}{2}\frac{1}{2}}{\frac{1}{12}} = -1.\end{aligned}$$

Example: uniform antithetic sampling

Theorem: $\mathrm{Corr}\left[g(U), g(1-U)\right]$ is negative.
Proof:

- Let U and U' be two standard-uniform copies, and let $C = \mathrm{Cov}\left[g(U) - g(U'), g(1-U) - g(1-U')\right]$.
- Note that $C = \mathbb{E}\left[(g(U) - g(U'))(g(1-U) - g(1-U'))\right]$, which due to the monotonicity of g must be negative.
- Note also that $C = \mathrm{Cov}\left[g(U), g(1-U)\right] + \mathrm{Cov}\left[g(U'), g(1-U')\right] = 2\mathrm{Cov}\left[g(U), g(1-U)\right]$.
- As $\mathrm{Cov}\left[g(U), g(1-U)\right]$ must now be negative, $\mathrm{Corr}\left[g(U), g(1-U)\right]$ is too.

Thus, to estimate $\mathbb{E}\left[Z\right]$, $\frac{2}{R}\sum_{r=1}^{R/2}(g(U_r) + g(1-U_r))$ has less variance than the crude Monte Carlo estimator $\frac{1}{R}\sum_{r=1}^{R} g(U_r)$.

Example: uniform antithetic sampling

If $g(x) = x^2$, we have

$$\mathrm{Corr}\left[g(U), g(1-U)\right] = \frac{\mathbb{E}\left[U^2(1-U)^2\right] - \mathbb{E}\left[U^2\right]\mathbb{E}\left[(1-U)^2\right]}{\mathrm{Var}\left[U^2\right]}$$
$$\frac{(\frac{1}{3} - \frac{1}{2} + \frac{1}{5}) - \frac{1}{9}}{\frac{1}{5} - \frac{1}{9}} = -\frac{7}{8}.$$

For $g(x) = x^n$, we have

$$\left|\mathrm{Corr}\left[g(U), g(1-U)\right]\right| \leq \frac{\mathbb{E}\left[U^n(1-U)^n\right] + \mathbb{E}\left[U^n\right]^2}{\mathrm{Var}\left[U^n\right]}$$
$$\leq \frac{\frac{1}{4^n} + 1/(n+1)^2}{1/(2n+1) - 1/(n+1)^2}.$$

This expression vanishes as n grows large. Antithetic sampling does not always make a big difference!

Other example: Gaussian antithetic sampling

Often, the quantity of interest is $Z = g(X)$ for some monotone function g, where X is a normal r.v. with mean μ and variance σ^2.

X and $2\mu - X$ are perfectly negatively correlated:

$$\begin{aligned}\mathrm{Corr}\left[X, 2\mu - X\right] &= \frac{\mathbb{E}\left[X(2\mu - X)\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[2\mu - X\right]}{\mathrm{Var}\left[X\right]} \\ &= \frac{(2\mu^2 - \sigma^2 - \mu^2) - \mu^2}{\sigma^2} = -1.\end{aligned}$$

Again, we can prove that $\mathrm{Corr}\left[g(X), g(2\mu - X)\right]$ is negative.

Gaussian antithetic sampling is widely applied in financial engineering for option pricing.

Some remarks

- When possible, antithetic sampling is not very hard to implement,
- but often, the obtained variance reduction is not dramatic.

Some remarks

- When possible, antithetic sampling is not very hard to implement,
- but often, the obtained variance reduction is not dramatic.

To illustrate the latter, consider the i.i.d. continuous random variables Z_1, Z_2 with marginal CDF $F(z)$ and joint CDF $F(z_1, z_2)$.

We know that

$$\mathrm{Corr}\left[Z_1, Z_2\right] = \frac{\mathbb{E}\left[Z_1 Z_2\right] - \left(\mathbb{E}\left[Z_1\right]\right)^2}{\mathrm{Var}\left[Z_1\right]},$$

where we can write

$$\mathbb{E}\left[Z_1 Z_2\right] = \int_0^\infty \int_0^\infty \left[1 - F(z_1) - F(z_2) + F(z_1, z_2)\right] dz_1 dz_2.$$

Some remarks

$$\mathbb{E}[Z_1 Z_2] = \int_0^\infty \int_0^\infty [1 - F(z_1) - F(z_2) + F(z_1, z_2)] dz_1 dz_2.$$

By standard theory on copula's, we know that

$$F(z_1, z_2) \geq (F(z_1) + F(z_2) - 1)^+$$

This is the *Fréchet-Hoeffding lower bound* and can be attained by choosing

$$Z_1 = F^{\leftarrow}(U), \qquad Z_2 = F^{\leftarrow}(1 - U).$$

Some remarks

Thus, for example, when Z_1 and Z_2 are exponentially (λ) distributed (recall: $F^{\leftarrow}(z) = -\log(1-z)/\lambda$), we have

$$\begin{aligned}\operatorname{Cov}[Z_1, Z_2] &= \mathbb{E}[Z_1 Z_2] - (\mathbb{E}[Z_1])^2 \\ &\geq \frac{\mathbb{E}[\log(1-U)\log U] - 1}{\lambda^2} \approx \frac{-0.645}{\lambda^2},\end{aligned}$$

so that $\operatorname{Corr}[Z_1, Z_2] = \operatorname{Corr}[F^{\leftarrow}(U), F^{\leftarrow}(1-U)] \approx -0.645$.

Some remarks

Thus, for example, when Z_1 and Z_2 are exponentially (λ) distributed (recall: $F^{\leftarrow}(z) = -\log(1-z)/\lambda$), we have

$$\begin{aligned}\operatorname{Cov}[Z_1, Z_2] &= \mathbb{E}[Z_1 Z_2] - (\mathbb{E}[Z_1])^2 \\ &\geq \frac{\mathbb{E}[\log(1-U)\log U] - 1}{\lambda^2} \approx \frac{-0.645}{\lambda^2},\end{aligned}$$

so that $\operatorname{Corr}[Z_1, Z_2] = \operatorname{Corr}[F^{\leftarrow}(U), F^{\leftarrow}(1-U)] \approx -0.645$.

Q: What does this mean?

Some remarks

Thus, for example, when Z_1 and Z_2 are exponentially (λ) distributed (recall: $F^{\leftarrow}(z) = -\log(1-z)/\lambda$), we have

$$\begin{aligned}\mathrm{Cov}\left[Z_1, Z_2\right] &= \mathbb{E}\left[Z_1 Z_2\right] - \left(\mathbb{E}\left[Z_1\right]\right)^2 \\ &\geq \frac{\mathbb{E}\left[\log(1-U)\log U\right] - 1}{\lambda^2} \approx \frac{-0.645}{\lambda^2},\end{aligned}$$

so that $\mathrm{Corr}\left[Z_1, Z_2\right] = \mathrm{Corr}\left[F^{\leftarrow}(U), F^{\leftarrow}(1-U)\right] \approx -0.645$.

Q: What does this mean?

A: It means that when one wants to simulate the expectation of an exponential random variable, the method of antithetic sampling only nets you a variance reduction of at most 64.5%.

Some remarks

Thus, for example, when Z_1 and Z_2 are exponentially (λ) distributed (recall: $F^{\leftarrow}(z) = -\log(1-z)/\lambda$), we have

$$\begin{aligned}\mathrm{Cov}\left[Z_1, Z_2\right] &= \mathbb{E}\left[Z_1 Z_2\right] - \left(\mathbb{E}\left[Z_1\right]\right)^2 \\ &\geq \frac{\mathbb{E}\left[\log(1-U)\log U\right] - 1}{\lambda^2} \approx \frac{-0.645}{\lambda^2},\end{aligned}$$

so that $\mathrm{Corr}\left[Z_1, Z_2\right] = \mathrm{Corr}\left[F^{\leftarrow}(U), F^{\leftarrow}(1-U)\right] \approx -0.645$.

Q: What does this mean?

A: It means that when one wants to simulate the expectation of an exponential random variable, the method of antithetic sampling only nets you a variance reduction of at most 64.5%.

A: Meaning that the width reduction of the confidence interval is about 40%($\sqrt{0.355} \approx 0.6$).

Some remarks

Thus, for example, when Z_1 and Z_2 are exponentially (λ) distributed (recall: $F^{\leftarrow}(z) = -\log(1-z)/\lambda$), we have

$$\begin{aligned}\mathrm{Cov}\left[Z_1, Z_2\right] &= \mathbb{E}\left[Z_1 Z_2\right] - \left(\mathbb{E}\left[Z_1\right]\right)^2 \\ &\geq \frac{\mathbb{E}\left[\log(1-U)\log U\right] - 1}{\lambda^2} \approx \frac{-0.645}{\lambda^2},\end{aligned}$$

so that $\mathrm{Corr}\left[Z_1, Z_2\right] = \mathrm{Corr}\left[F^{\leftarrow}(U), F^{\leftarrow}(1-U)\right] \approx -0.645$.

Q: What does this mean?

A: It means that when one wants to simulate the expectation of an exponential random variable, the method of antithetic sampling only nets you a variance reduction of at most 64.5%.

A: Meaning that the width reduction of the confidence interval is about 40%($\sqrt{0.355} \approx 0.6$).

A: This is not dramatic... but why not implement it anyway?!

Chapter V.4
Conditional Monte Carlo

Conditional Monte Carlo

The idea of the method is as follows. Using classic Crude Monte Carlo, we generate i.i.d. replicates

$$Z_1, Z_2, \ldots, Z_R$$

and estimate $\mathbb{E}[Z]$ by $\hat{z}_R = \frac{1}{R}\sum_{i=1}^{R} Z_i$. We have $\mathrm{Var}[\hat{z}_R] = \frac{\mathrm{Var}[Z]}{R}$.

Conditional Monte Carlo

The idea of the method is as follows. Using classic Crude Monte Carlo, we generate i.i.d. replicates

$$Z_1, Z_2, \ldots, Z_R$$

and estimate $\mathbb{E}[Z]$ by $\hat{z}_R = \frac{1}{R}\sum_{i=1}^{R} Z_i$. We have $\mathrm{Var}[\hat{z}_R] = \frac{\mathrm{Var}[Z]}{R}$.

Suppose now, however, that we would replace Z_i by $\mathbb{E}[Z_i \mid W_i]$. Thus, our estimator is $\frac{1}{R}\sum_{i=1}^{R} \mathbb{E}[Z_i \mid W_i] = \mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]$ for some random variable W as an estimator:

- $\mathbb{E}\left[\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]\right] = \mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] = \mathbb{E}[Z]$ by the law of total expectation.

Conditional Monte Carlo

The idea of the method is as follows. Using classic Crude Monte Carlo, we generate i.i.d. replicates

$$Z_1, Z_2, \ldots, Z_R$$

and estimate $\mathbb{E}[Z]$ by $\hat{z}_R = \frac{1}{R}\sum_{i=1}^{R} Z_i$. We have $\mathrm{Var}[\hat{z}_R] = \frac{\mathrm{Var}[Z]}{R}$.

Suppose now, however, that we would replace Z_i by $\mathbb{E}[Z_i \mid W_i]$. Thus, our estimator is $\frac{1}{R}\sum_{i=1}^{R} \mathbb{E}[Z_i \mid W_i] = \mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]$ for some random variable W as an estimator:

- $\mathbb{E}\left[\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]\right] = \mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] = \mathbb{E}[Z]$ by the law of total expectation. Unbiased estimator!

Conditional Monte Carlo

The idea of the method is as follows. Using classic Crude Monte Carlo, we generate i.i.d. replicates

$$Z_1, Z_2, \ldots, Z_R$$

and estimate $\mathbb{E}[Z]$ by $\hat{z}_R = \frac{1}{R}\sum_{i=1}^{R} Z_i$. We have $\mathrm{Var}[\hat{z}_R] = \frac{\mathrm{Var}[Z]}{R}$.

Suppose now, however, that we would replace Z_i by $\mathbb{E}[Z_i \mid W_i]$. Thus, our estimator is $\frac{1}{R}\sum_{i=1}^{R} \mathbb{E}[Z_i \mid W_i] = \mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]$ for some random variable W as an estimator:

- $\mathbb{E}\left[\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]\right] = \mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] = \mathbb{E}[Z]$ by the law of total expectation. Unbiased estimator!
- But what happens with $\mathrm{Var}\left[\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]\right]$?

Conditional Monte Carlo

Theorem: (Law of total variance) For any two random variables U, W, $\mathrm{Var}[U] = \mathbb{E}[\mathrm{Var}[U \mid W]] + \mathrm{Var}[\mathbb{E}[U \mid W]]$.

Proof

- $\mathrm{Var}[U \mid W] = \mathbb{E}\left[U^2 \mid W\right] - (\mathbb{E}[U \mid W])^2$. Taking expectations leads to

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[U \mid W]] &= \mathbb{E}\left[\mathbb{E}\left[U^2 \mid W\right]\right] - \mathbb{E}\left[(\mathbb{E}[U \mid W])^2\right] \\ &= \mathbb{E}\left[U^2\right] - \mathbb{E}\left[(\mathbb{E}[U \mid W])^2\right]\end{aligned}$$

- For the other term,

$$\begin{aligned}\mathrm{Var}[\mathbb{E}[U \mid W]] &= \mathbb{E}\left[(\mathbb{E}[U \mid W])^2\right] - (\mathbb{E}[\mathbb{E}[U \mid W]])^2 \\ &= \mathbb{E}\left[(\mathbb{E}[U \mid W])^2\right] - (\mathbb{E}[U])^2\end{aligned}$$

- Adding leads to $\mathbb{E}\left[U^2\right] - (\mathbb{E}[U])^2 = \mathrm{Var}[U]$.

Conditional Monte Carlo

Theorem: (Law of total variance) For any two random variables U, W, $\mathrm{Var}[U] = \mathbb{E}[\mathrm{Var}[U \mid W]] + \mathrm{Var}[\mathbb{E}[U \mid W]]$.

Corollary: Since variances (and expectations of them) are non-negative, we must have that

$$\mathrm{Var}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] \geq \mathrm{Var}\left[\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]\right].$$

We just found an unbiased estimator, $\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]$, which is guaranteed to be a variance reductor!

Conditional Monte Carlo

Theorem: (Law of total variance) For any two random variables U, W, $\mathrm{Var}[U] = \mathbb{E}[\mathrm{Var}[U \mid W]] + \mathrm{Var}[\mathbb{E}[U \mid W]]$.

Corollary: Since variances (and expectations of them) are non-negative, we must have that

$$\mathrm{Var}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] \geq \mathrm{Var}\left[\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]\right].$$

We just found an unbiased estimator, $\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]$, which is guaranteed to be a variance reductor!

Are we done now?

Conditional Monte Carlo

Theorem: (Law of total variance) For any two random variables U, W, $\text{Var}[U] = \mathbb{E}[\text{Var}[U \mid W]] + \text{Var}[\mathbb{E}[U \mid W]]$.

Corollary: Since variances (and expectations of them) are non-negative, we must have that

$$\text{Var}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i\right] \geq \text{Var}\left[\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]\right].$$

We just found an unbiased estimator, $\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid W\right]$, which is guaranteed to be a variance reductor!

Are we done now? NO! How to choose W?

Examples

Choosing W is very much situation-dependent. Consider the following example.

Recall that one can estimate π by computing $\frac{1}{R}\sum_{i=1}^{R} Z_i$, where $Z_i = 4\mathbb{1}_{\{(U_1^{(i)})^2+(U_2^{(i)})^2\leq 1\}}$ and $U_1^{(i)}, U_2^{(i)}$ are i.i.d. uniform on $[0,1]$.

Examples

Choosing W is very much situation-dependent. Consider the following example.

Recall that one can estimate π by computing $\frac{1}{R}\sum_{i=1}^{R} Z_i$, where $Z_i = 4\mathbb{1}_{\{(U_1^{(i)})^2+(U_2^{(i)})^2\leq 1\}}$ and $U_1^{(i)}, U_2^{(i)}$ are i.i.d. uniform on $[0,1]$.

We could, however, condition on (the samples of) U_1!

Examples

Choosing W is very much situation-dependent. Consider the following example.

Recall that one can estimate π by computing $\frac{1}{R}\sum_{i=1}^{R} Z_i$, where $Z_i = 4\mathbb{1}_{\{(U_1^{(i)})^2+(U_2^{(i)})^2\leq 1\}}$ and $U_1^{(i)}, U_2^{(i)}$ are i.i.d. uniform on $[0,1]$.

We could, however, condition on (the samples of) U_1!

$$\mathbb{E}\left[\frac{1}{R}\sum_{i=1}^{R} Z_i \mid U_1^{(1)}, U_1^{(2)}, \ldots, U_1^{(R)}\right] = \frac{1}{R}\sum_{i=1}^{R} \mathbb{E}\left[Z_i \mid U_1^{(i)}\right]$$

$$= \frac{4}{R}\sum_{i=1}^{R} \mathbb{P}\left((U_1^{(i)})^2 + (U_2^{(i)})^2 \leq 1 \mid U_1^{(i)}\right)$$

$$= \frac{4}{R}\sum_{i=1}^{R} \mathbb{P}\left(U_2^{(i)} \leq \sqrt{1-(U_1^{(i)})^2} \mid U_1^{(i)}\right) = \frac{4}{R}\sum_{i=1}^{R} \sqrt{1-(U_1^{(i)})^2}$$

Examples

So in this case, using conditional Monte Carlo, we would simply generate R uniform samples on $[0, 1]$ and compute

$$\frac{4}{R} \sum_{i=1}^{R} \sqrt{1 - (U_1^{(i)})^2}.$$

Examples

So in this case, using conditional Monte Carlo, we would simply generate R uniform samples on $[0, 1]$ and compute

$$\frac{4}{R} \sum_{i=1}^{R} \sqrt{1 - (U_1^{(i)})^2}.$$

Two advantages:

- We only need half the uniform samples.
- Less sampling in this case means less variance: more than three times as small!

Examples

So in this case, using conditional Monte Carlo, we would simply generate R uniform samples on $[0, 1]$ and compute

$$\frac{4}{R} \sum_{i=1}^{R} \sqrt{1 - (U_1^{(i)})^2}.$$

Two advantages:

- We only need half the uniform samples.
- Less sampling in this case means less variance: more than three times as small!

This is quite worth it!

Examples

Another example: suppose X_1 and X_2 are i.i.d. with known distribution F.

Examples

Another example: suppose X_1 and X_2 are i.i.d. with known distribution F.

Problem: Convolution of F is unknown, so we wish to simulate $\mathbb{P}(X_1 + X_2 \leq x)$.

Crude Monte Carlo:

Examples

Another example: suppose X_1 and X_2 are i.i.d. with known distribution F.

Problem: Convolution of F is unknown, so we wish to simulate $\mathbb{P}(X_1 + X_2 \leq x)$.

Crude Monte Carlo: Sample $\mathbb{1}_{\{X_1+X_2\leq x\}}$ and take averages.

Conditional Monte Carlo:

Examples

Another example: suppose X_1 and X_2 are i.i.d. with known distribution F.

Problem: Convolution of F is unknown, so we wish to simulate $\mathbb{P}(X_1 + X_2 \leq x)$.

Crude Monte Carlo: Sample $\mathbb{1}_{\{X_1+X_2\leq x\}}$ and take averages.

Conditional Monte Carlo: Sample $\mathbb{E}\left[\mathbb{1}_{\{X_1+X_2\leq x\}}|X_1\right] = F(x - X_1)$ and take averages.

Again, considerable variance reduction.

Examples

We can also extend this to higher-fold convolutions and densities.

Examples

We can also extend this to higher-fold convolutions and densities.

For example, the density $f^{*n}(x)$ of $S_n := \sum_{i=1}^{n} X_i$ can be (conditionally) estimated by

$$f(x - S_{n-1})$$

given S_{n-1}. So, for simulation of the n-fold convolution of X, which is say Pareto distributed:

$$f(x) = \frac{\alpha}{(1+x)^{\alpha+1}} \mathbb{1}_{\{x \geq 0\}}$$

with parameter $\alpha = \frac{3}{2}$, we

Examples

We can also extend this to higher-fold convolutions and densities.

For example, the density $f^{*n}(x)$ of $S_n := \sum_{i=1}^{n} X_i$ can be (conditionally) estimated by

$$f(x - S_{n-1})$$

given S_{n-1}. So, for simulation of the n-fold convolution of X, which is say Pareto distributed:

$$f(x) = \frac{\alpha}{(1+x)^{\alpha+1}} \mathbb{1}_{\{x \geq 0\}}$$

with parameter $\alpha = \frac{3}{2}$, we

- generate $R = 10.000$ replicates of S_{n-1}

Examples

We can also extend this to higher-fold convolutions and densities.

For example, the density $f^{*n}(x)$ of $S_n := \sum_{i=1}^{n} X_i$ can be (conditionally) estimated by

$$f(x - S_{n-1})$$

given S_{n-1}. So, for simulation of the n-fold convolution of X, which is say Pareto distributed:

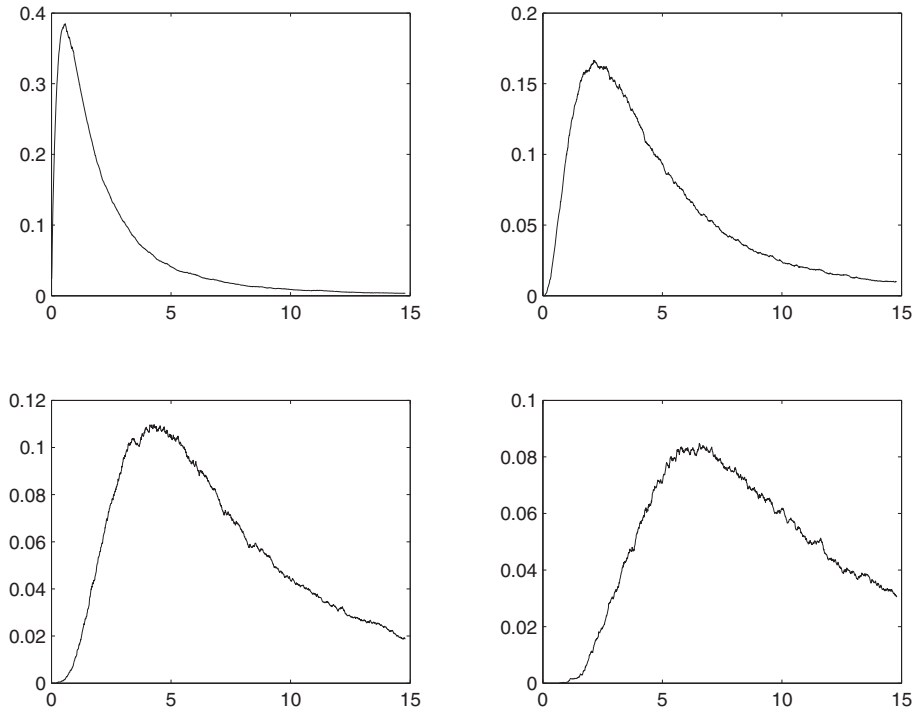
$$f(x) = \frac{\alpha}{(1+x)^{\alpha+1}} \mathbb{1}_{\{x \geq 0\}}$$

with parameter $\alpha = \frac{3}{2}$, we

- generate $R = 10.000$ replicates of S_{n-1}
- Calculate the corresponding values for $f(x - S_{n-1})$ for any desired x.

Yet another example

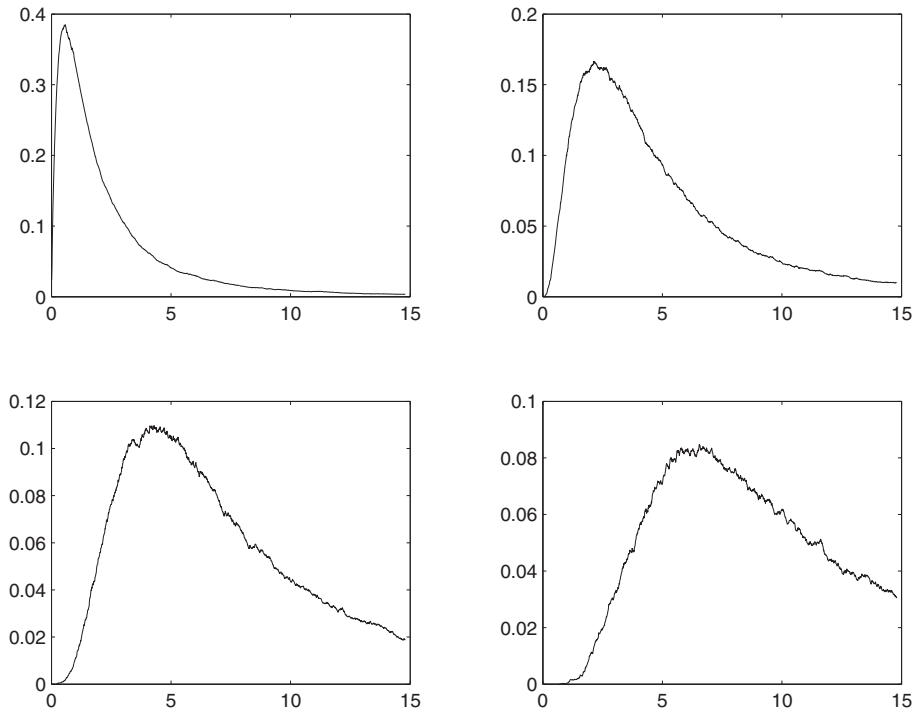
For $n = 2, 4, 6, 8$ and $0 < x < 15$, this leads to:



- In this particular case, conditional MC works better than empirical PDFs or even kernel estimations!

Yet another example

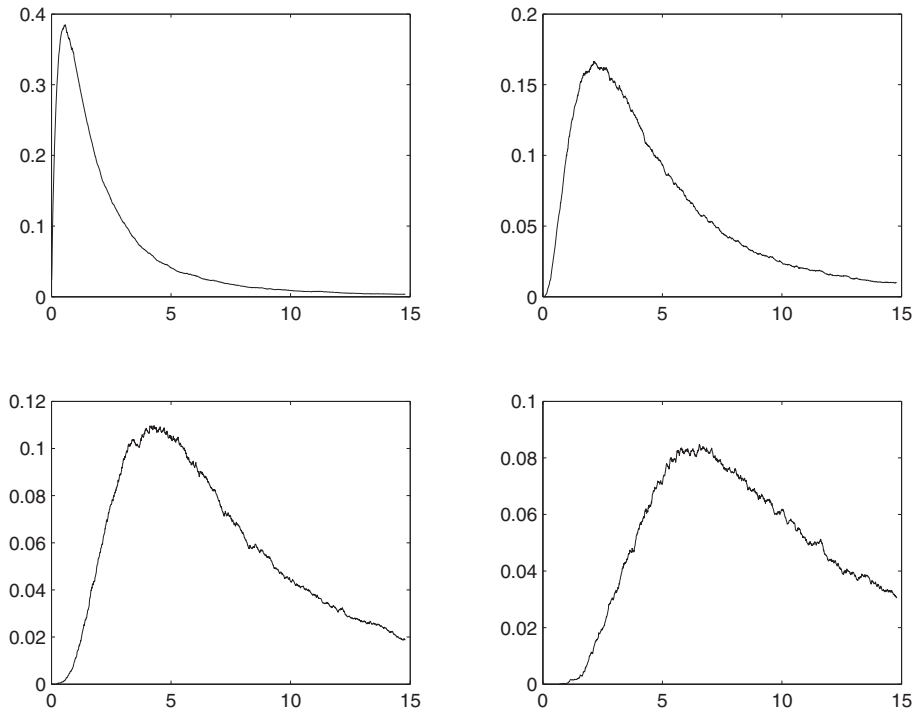
For $n = 2, 4, 6, 8$ and $0 < x < 15$, this leads to:



- In this particular case, conditional MC works better than empirical PDFs or even kernel estimations!
- Jaggedness because of discontinuity of $f(x)$ at $x = 0$.

Yet another example

For $n = 2, 4, 6, 8$ and $0 < x < 15$, this leads to:



- In this particular case, conditional MC works better than empirical PDFs or even kernel estimations!
- Jaggedness because of discontinuity of $f(x)$ at $x = 0$.
- Jaggedness appears to increase in n, but that's because for larger n, S_n has less probability mass below 15.

Chapter V.5
Splitting

Splitting

Let's say we want to estimate $\mathbb{E}[Z] = \phi(X, Y)$.

Splitting

Let's say we want to estimate $\mathbb{E}[Z] = \phi(X, Y)$.

Naively, we would want to generate samples of $Z_i = \phi(X_i, Y_i)$ and use the crude Monte Carlo estimator

$$\frac{1}{R} \sum_{i=1}^{R} \phi(X_i, Y_i).$$

Splitting

However, suppose now that

- samples from X are computationally much harder to generate than from Y, and
- the value of $\phi(X, Y)$ is much more influenced by Y than by X.

Splitting

However, suppose now that

- samples from X are computationally much harder to generate than from Y, and
- the value of $\phi(X, Y)$ is much more influenced by Y than by X.

Q: Can we do something smarter in this case?

Splitting

However, suppose now that

- samples from X are computationally much harder to generate than from Y, and
- the value of $\phi(X, Y)$ is much more influenced by Y than by X.

Q: Can we do something smarter in this case?

A: Yes, we'll borrow a variance reduction technique from physics. The idea is to simply reuse samples of X a number of S times. I.e., we use the estimator

$$\frac{1}{RS} \sum_{r=1}^{R} \sum_{s=1}^{S} \phi(X_r, Y_{rs}).$$

Splitting

However, suppose now that

- samples from X are computationally much harder to generate than from Y, and
- the value of $\phi(X, Y)$ is much more influenced by Y than by X.

Q: Can we do something smarter in this case?

A: Yes, we'll borrow a variance reduction technique from physics. The idea is to simply reuse samples of X a number of S times. I.e., we use the estimator

$$\frac{1}{RS} \sum_{r=1}^{R} \sum_{s=1}^{S} \phi(X_r, Y_{rs}).$$

Splitting

So, we regard the splitting estimator

$$\frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S}\phi(X_r, Y_{rs})$$

Splitting

So, we regard the splitting estimator

$$\frac{1}{RS}\sum_{r=1}^{R}\sum_{s=1}^{S}\phi(X_r, Y_{rs})$$

- Needs less replicates of X, so far less computation time.
- About the same variance if $\mathrm{Corr}\,[\phi(X_r, Y_{r1}), \phi(X_r, Y_{r2})]$ is close to zero.
- When $Z = \phi(X, Y)$ is more influenced by Y than by X, this correlation coefficient will indeed be small!

Splitting

Are we done now?

Splitting

Are we done now? NO! How to choose S?

Splitting

Are we done now? NO! How to choose S?

In deciding how to choose S, we cannot just compare variances anymore, since computation time now plays a role.

Let's say that

- a is the time to generate a sample of X, b a sample of Y, $a >> b$.
- $Z = \phi(X, Y)$ and $\tilde{Z} = \frac{1}{S} \sum_{s=1}^{S} \phi(X, Y_s)$.
- The time to get one estimate of $Z = a + b$. The time to get one estimate of $\tilde{Z} = a + Sb$.
- In a given time t, one gets $\frac{t}{a+b}$ estimates of Z, and $\frac{t}{a+Sb}$ estimates of $\tilde{Z}$.

Splitting

Recall that the width of a confidence interval of an estimate for z is generally given by:

$$\left(\hat{z} - q_\alpha \frac{\sigma}{\sqrt{R}}, \hat{z} + q_\alpha \frac{\sigma}{\sqrt{R}}\right),$$

where R is the number of estimates.

Splitting

Recall that the width of a confidence interval of an estimate for z is generally given by:

$$\left(\hat{z} - q_\alpha \frac{\sigma}{\sqrt{R}}, \hat{z} + q_\alpha \frac{\sigma}{\sqrt{R}}\right),$$

where R is the number of estimates.

Both Z and $\tilde{Z}$ are unbiased estimators, so we wish to minimise $\frac{\sigma}{\sqrt{R}}$.

Splitting

Recall that the width of a confidence interval of an estimate for z is generally given by:

$$\left(\hat{z} - q_\alpha \frac{\sigma}{\sqrt{R}}, \hat{z} + q_\alpha \frac{\sigma}{\sqrt{R}}\right),$$

where R is the number of estimates.

Both Z and $\tilde{Z}$ are unbiased estimators, so we wish to minimise $\frac{\sigma}{\sqrt{R}}$.

Concluding: we need to compare

$$e := (a + b)\text{Var}\left[Z\right]$$

(corresponding to one iterate of crude Monte Carlo) with

$$\tilde{e} := (a + Sb)\text{Var}\left[\tilde{Z}\right]$$

(one 'splitting' iterate).

Splitting

Comparing $e := (a+b)\mathrm{Var}[Z]$ with $\tilde{e} = (a+Sb)\mathrm{Var}\left[\tilde{Z}\right]$.

Note that [AG] has got errors in this analysis!
Suppose that $\mathrm{Var}[Z] = 1$ and $\rho = \mathrm{Corr}[\phi(X_r, Y_{rs_1}), \phi(X_r, Y_{rs_2})]$.

Then, since $\tilde{Z} = \frac{1}{S}\sum_{s=1}^{S}\phi(X, Y_s)$:

$$\begin{aligned}
\mathrm{Var}\left[\tilde{Z}\right] &= \frac{1}{S^2}\left(\mathrm{Cov}\left[\sum_{s=1}^{S}\phi(X, Y_s), \sum_{s=1}^{S}\phi(X, Y_s)\right]\right) \\
&= \frac{1}{S^2}\left(\sum_{r=1}^{S}\sum_{s=1}^{S}\mathrm{Cov}\left[\phi(X, Y_r), \phi(X, Y_s)\right]\right) \\
&= \frac{1}{S^2}\left(S + (S^2 - S)\rho\right).
\end{aligned}$$

Therefore,

$$e = a + b \text{ versus } \tilde{e} = (a+Sb)\left(\frac{1}{S} + (1-\frac{1}{S})\rho\right).$$

Splitting

To choose S, we compare

$$e = a + b \text{ versus } \tilde{e} = (a + Sb)\left(\frac{1}{S} + (1 - \frac{1}{S})\rho\right).$$

The $\tilde{e}$-approximation $(a + Sb)\left(\frac{1}{S} + \rho\right)$ (for large S) is minimised by

$$S = \sqrt{\frac{a}{\rho b}},$$

Splitting

To choose S, we compare

$$e = a + b \text{ versus } \tilde{e} = (a + Sb)\left(\frac{1}{S} + (1 - \frac{1}{S})\rho\right).$$

The $\tilde{e}$-approximation $(a + Sb)\left(\frac{1}{S} + \rho\right)$ (for large S) is minimised by

$$S = \sqrt{\frac{a}{\rho b}},$$

so that the optimal efficiency is approximated by

$$\left(a + \sqrt{\frac{ab}{\rho}}\right)\left(\sqrt{\frac{\rho b}{a}} + \rho\right) = a\left(\sqrt{\frac{b}{a}} + \sqrt{\rho}\right)^2.$$

Splitting

To choose S, we compare

$$e = a + b \text{ versus } \tilde{e} = (a + Sb)\left(\frac{1}{S} + (1 - \frac{1}{S})\rho\right).$$

The $\tilde{e}$-approximation $(a + Sb)\left(\frac{1}{S} + \rho\right)$ (for large S) is minimised by

$$S = \sqrt{\frac{a}{\rho b}},$$

so that the optimal efficiency is approximated by

$$\left(a + \sqrt{\frac{ab}{\rho}}\right)\left(\sqrt{\frac{\rho b}{a}} + \rho\right) = a\left(\sqrt{\frac{b}{a}} + \sqrt{\rho}\right)^2.$$

This leads to

$$\frac{\tilde{e}}{e} = \frac{a}{a + b}\left(\sqrt{\frac{b}{a}} + \sqrt{\rho}\right)^2.$$

Splitting

To choose S, we compare

$$e = a + b \text{ versus } \tilde{e} = (a + Sb)\left(\frac{1}{S} + (1 - \frac{1}{S})\rho\right).$$

The $\tilde{e}$-approximation $(a + Sb)\left(\frac{1}{S} + \rho\right)$ (for large S) is minimised by

$$S = \sqrt{\frac{a}{\rho b}},$$

so that the optimal efficiency is approximated by

$$\left(a + \sqrt{\frac{ab}{\rho}}\right)\left(\sqrt{\frac{\rho b}{a}} + \rho\right) = a\left(\sqrt{\frac{b}{a}} + \sqrt{\rho}\right)^2.$$

This leads to

$$\frac{\tilde{e}}{e} = \frac{a}{a + b}\left(\sqrt{\frac{b}{a}} + \sqrt{\rho}\right)^2.$$

Conclusion: with the right choice of S, and small values of $\frac{b}{a}$ and ρ, splitting nets you the desired precision faster!

Example

Splitting in the context of discrete event simulation.

- A restaurant opens at 10.00; customers spend an exponentially distributed amount of time dining.
- Not many arrivals until noon, but especially between 12.30 and 13.30 there are many, many arrivals.
- We want to estimate the expected number of lost customers between noon and 14.00.

Example

Splitting in the context of discrete event simulation.

- A restaurant opens at 10.00; customers spend an exponentially distributed amount of time dining.
- Not many arrivals until noon, but especially between 12.30 and 13.30 there are many, many arrivals.
- We want to estimate the expected number of lost customers between noon and 14.00.

How to take X and Y?

Example

Splitting in the context of discrete event simulation.

- A restaurant opens at 10.00; customers spend an exponentially distributed amount of time dining.
- Not many arrivals until noon, but especially between 12.30 and 13.30 there are many, many arrivals.
- We want to estimate the expected number of lost customers between noon and 14.00.

How to take X and Y?

- Number of lost customers more influenced by arrivals than by number of occupied tables at noon.
- Arrivals after noon not correlated with occupancy at noon.

Example

Splitting in the context of discrete event simulation.

- A restaurant opens at 10.00; customers spend an exponentially distributed amount of time dining.
- Not many arrivals until noon, but especially between 12.30 and 13.30 there are many, many arrivals.
- We want to estimate the expected number of lost customers between noon and 14.00.

How to take X and Y?

- Number of lost customers more influenced by arrivals than by number of occupied tables at noon.
- Arrivals after noon not correlated with occupancy at noon.

  - Choose X to be number of occupied tables at noon.
  - Choose Y to be the arrival times and service times of customers arriving between noon and 14:00.

Another example

Suppose you want to estimate the expectation of a functional $Z = Z(B)$, where B is a standard Brownian motion in [0,1].

The crude way of simulating B:

- Divide $[0, 1]$ up in 1024 grid regions.
- Generate $B(\frac{1}{1024}), B(\frac{2}{1024}), \ldots$ by summing 1024 i.i.d. normal samples.

Another example

Suppose you want to estimate the expectation of a functional $Z = Z(B)$, where B is a standard Brownian motion in [0,1].

The crude way of simulating B:

- Divide $[0, 1]$ up in 1024 grid regions.
- Generate $B(\frac{1}{1024}), B(\frac{2}{1024}), \ldots$ by summing 1024 i.i.d. normal samples.
- Interpolate linearly.

Another example

Splitting approach:

- Use e.g. 16 grid regions.
- Generate $B(\frac{1}{16}), B(\frac{2}{16}), \ldots, B(1)$ by summing 16 i.i.d. normal samples.
- Replace linear interpolations by Brownian bridges $W_{i,c}(t)$. These are Brownian motions constrained not only by $W(0) = 0$, but also by a value $W(\frac{1}{16}) = c$, where c is the difference between the sampled B-points:

$$W_i(t) = B_i(t) + 16t(c - B_i(\frac{1}{16})),$$

where the B_i are independent Brownian motions. These can be sampled, each using $\frac{1024}{16} = 64$ grid regions.

Another example

Splitting approach:

- Use e.g. 16 grid regions.
- Generate $B(\frac{1}{16}), B(\frac{2}{16}), \ldots, B(1)$ by summing 16 i.i.d. normal samples.
- Replace linear interpolations by Brownian bridges $W_{i,c}(t)$. These are Brownian motions constrained not only by $W(0) = 0$, but also by a value $W(\frac{1}{16}) = c$, where c is the difference between the sampled B-points:

$$W_i(t) = B_i(t) + 16t(c - B_i(\frac{1}{16})),$$

  where the B_i are independent Brownian motions. These can be sampled, each using $\frac{1024}{16} = 64$ grid regions.

  - Choose X to be 16 Brownian bridges. These do not have to be sampled often.
  - Choose Y to be the numbers $B(\frac{1}{16}), B(\frac{2}{16}), \ldots, B(1)$.

Chapter V.6
Common Random Numbers

Common Random Numbers

Let's say we want to compute the difference between the means z', z'' of two random variables Z' and Z'' with $\mathrm{Var}[Z'] = \mathrm{Var}[Z''] = \sigma^2$ that behave in some sense similar to the random input $U_1, \ldots, U_n$.

Common Random Numbers

Let's say we want to compute the difference between the means z', z'' of two random variables Z' and Z'' with $\mathrm{Var}[Z'] = \mathrm{Var}[Z''] = \sigma^2$ that behave in some sense similar to the random input $U_1, \ldots, U_n$.

Naively, we would just

- sample Z' using input $U_1, \ldots U_n$ (e.g. uniform samples used by the inversion method, discrete-event simulation, etc).

Common Random Numbers

Let's say we want to compute the difference between the means z', z'' of two random variables Z' and Z'' with $\mathrm{Var}[Z'] = \mathrm{Var}[Z''] = \sigma^2$ that behave in some sense similar to the random input $U_1, \ldots, U_n$.

Naively, we would just

- sample Z' using input $U_1, \ldots U_n$ (e.g. uniform samples used by the inversion method, discrete-event simulation, etc).
- likewise sample Z'' using input $U_{n+1}, \ldots, U_{2n}$,
- compute $Z' - Z''$.

say R times (every time using new U's), and then take sample means.

This estimator would yield a variance of $\frac{2\sigma^2}{R}$.

Can we do better?

Common Random Numbers

Let's say we want to compute the difference between the means z', z'' of two random variables Z' and Z'' with $\mathrm{Var}[Z'] = \mathrm{Var}[Z''] = \sigma^2$ that behave in some sense similar to the random input $U_1, \ldots, U_n$.

Naively, we would just

- sample Z' using input $U_1, \ldots U_n$ (e.g. uniform samples used by the inversion method, discrete-event simulation, etc).
- likewise sample Z'' using input $U_{n+1}, \ldots, U_{2n}$,
- compute $Z' - Z''$.

say R times (every time using new U's), and then take sample means.

This estimator would yield a variance of $\frac{2\sigma^2}{R}$.

Can we do better? Maybe! We could just use $U_1, \ldots U_n$ for the estimate of Z'' as well!

Common Random Numbers

Thus, common random numbers advocates replicating

$$Z'(U_1, \ldots, U_n) - Z''(U_1, \ldots, U_n)$$

R times and computing the sample mean.

Common Random Numbers

Thus, common random numbers advocates replicating

$$Z'(U_1, \ldots, U_n) - Z''(U_1, \ldots, U_n)$$

R times and computing the sample mean.

This estimator has variance

$$\frac{2\sigma^2 - 2\mathrm{Cov}\left[Z'(U_1, \ldots, U_n), Z''(U_1, \ldots, U_n)\right]}{R},$$

- Z' and Z'' behave similar to the random input, which yields positive, significant covariance: high variance reduction!
- This method needs less input samples (the U_i), improving computation time.

Common Random Numbers

Thus, common random numbers advocates replicating

$$Z'(U_1, \ldots, U_n) - Z''(U_1, \ldots, U_n)$$

R times and computing the sample mean.

This estimator has variance

$$\frac{2\sigma^2 - 2\mathrm{Cov}\left[Z'(U_1, \ldots, U_n), Z''(U_1, \ldots, U_n)\right]}{R},$$

- Z' and Z'' behave similar to the random input, which yields positive, significant covariance: high variance reduction!
- This method needs less input samples (the U_i), improving computation time.
- Method also works when $\mathrm{Var}\left[Z'\right] \neq \mathrm{Var}\left[Z''\right]$, but variance reduction may be less considerable.
- This is in some sense the opposite of antithetic sampling.

Common Random Numbers

Thus, common random numbers advocates replicating

$$Z'(U_1, \ldots, U_n) - Z''(U_1, \ldots, U_n)$$

R times and computing the sample mean.

This estimator has variance

$$\frac{2\sigma^2 - 2\mathrm{Cov}\left[Z'(U_1, \ldots, U_n), Z''(U_1, \ldots, U_n)\right]}{R},$$

- Z' and Z'' behave similar to the random input, which yields positive, significant covariance: high variance reduction!
- This method needs less input samples (the U_i), improving computation time.
- Method also works when $\mathrm{Var}\left[Z'\right] \neq \mathrm{Var}\left[Z''\right]$, but variance reduction may be less considerable.
- This is in some sense the opposite of antithetic sampling.

Caution: do NOT apply this technique when $Z'(U_1, \ldots, U_n)$ and $Z''(U_1, \ldots, U_n)$ might be negatively correlated!

Example

- A factory has two identical machines. At some point in time n jobs are ready to be processed. Their processing times are i.i.d. random variables $U_1, \ldots, U_n$.
- The process manager investigates two processing policies:
  - SJF: shortest job first (when a machine becomes free, it is allocated the job with the shortest processing time);
  - LJF: longest job first (as above for the longest processing time).

Example

- A factory has two identical machines. At some point in time n jobs are ready to be processed. Their processing times are i.i.d. random variables $U_1, \ldots, U_n$.
- The process manager investigates two processing policies:
  - SJF: shortest job first (when a machine becomes free, it is allocated the job with the shortest processing time);
  - LJF: longest job first (as above for the longest processing time).
- The makespan is the total processing time to complete all n jobs on the two machines.
- Question: what is the average difference in makespan for e.g. Weibull distributed processing times?

Example

It is intuitive that SJF and LJF are best tested using the same processing time samples. Think of unpaired versus paired testing in statistics.

Example

It is intuitive that SJF and LJF are best tested using the same processing time samples. Think of unpaired versus paired testing in statistics.

Therefore, we use common random numbers. Let $Z'(U_1, \ldots, U_n)$ be the makespan under SJF, and let $Z''(U_1, \ldots, U_n)$ be the makespan under LJF.

Then, for example:

- When $U_1 = 1, U_2 = 2, U_3 = 5$,
  $Z'(U_1, U_2, U_3) = 6$ and $Z''(U_1, U_2, U_3) = 5$
- When $U_1 = 2, U_2 = 9, U_3 = 12, U_4 = 14$,
  $Z'(U_1, U_2, U_3, U_4) = 23$ and $Z''(U_1, U_2, U_3, U_4) = 21$.

Example

Proposed strategy:

1. Sample R times n Weibull processing times.
2. Using these numbers, compute the R replicates of $Z'(U_1, \ldots, U_n) - Z''(U_1, \ldots, U_n)$.
3. Averaging the R replicates nets you an estimate of the average distance in makespan for n jobs.

Example

Proposed strategy:

1. Sample R times n Weibull processing times.
2. Using these numbers, compute the R replicates of $Z'(U_1, \ldots, U_n) - Z''(U_1, \ldots, U_n)$.
3. Averaging the R replicates nets you an estimate of the average distance in makespan for n jobs.

- Of course, $\text{Corr}\left[Z'(U_1, \ldots, U_n), Z''(U_1, \ldots, U_n)\right]$ is very large, thus so is $\text{Cov}\left[Z'(U_1, \ldots, U_n), Z''(U_1, \ldots, U_n)\right]$.
- In this case, variance reductions of about 99% can be achieved by using common random numbers, while saving the need to sample n more Weibull samples per replication.
- Needless to say, common random numbers is a must-use technique in this case!

Chapter V.7
Stratification

Stratification

Suppose we want to estimate $\mathbb{E}[Z]$, where $Z = g(X)$. We know by now that we can use the naive estimator

$$\frac{1}{R}\sum_{i=1}^{R} Z_i = \frac{1}{R}\sum_{i=1}^{R} g(X_i).$$

Variance of this estimator can be reduced by **stratification**.

The idea behind the method of stratification is to partition the sample space Ω into regions $\Omega_1, \ldots, \Omega_S$ called strata.

Stratification

Suppose we want to estimate $\mathbb{E}[Z]$, where $Z = g(X)$. We know by now that we can use the naive estimator

$$\frac{1}{R}\sum_{i=1}^{R} Z_i = \frac{1}{R}\sum_{i=1}^{R} g(X_i).$$

Variance of this estimator can be reduced by **stratification**.

The idea behind the method of stratification is to partition the sample space Ω into regions $\Omega_1, \ldots, \Omega_S$ called strata.

Aim: eliminate as much of the variation **between strata** as possible. (And not within strata as [AG] claims!).

The strata are often obtained by dividing the range of one or more important random variables driving the simulation (e.g. X).

Stratification

For example, when $Z = g(X)$, where X is standard uniformly distributed, we could divide the range $[0, 1]$ of X in the strata

$$\Omega_s = \{(s-1)/S \leq X < s/S\}.$$

Moreover, when Z_s is a random variable having the distribution of Z conditioned on Ω_s, we have

$$\mathbb{P}(Z_s \in A) = \mathbb{P}(Z \in A \mid \Omega_s) = \frac{\mathbb{P}(Z \in A, \Omega_s)}{\mathbb{P}(\Omega_s)}.$$

Stratification

This suggests the following estimation strategy:

1. Divide the total number of R replicates into $R_1, \ldots, R_S$.
2. For each $s \in \{1, \ldots, S\}$, simulate R_s replicates of Z_s.
3. Estimate z_s by the empirical average $\hat{z}_s$.
4. Compute the following estimate for $\mathbb{E}[Z]$:

$$\hat{z}_{str} = \sum_{s=1}^{S} p_s \hat{z}_s,$$

where $p_s = \mathbb{P}(\Omega_s)$.

Stratification

This suggests the following estimation strategy:

1. Divide the total number of R replicates into $R_1, \ldots, R_S$.
2. For each $s \in \{1, \ldots, S\}$, simulate R_s replicates of Z_s.
3. Estimate z_s by the empirical average $\hat{z}_s$.
4. Compute the following estimate for $\mathbb{E}[Z]$:

$$\hat{z}_{str} = \sum_{s=1}^{S} p_s \hat{z}_s,$$

where $p_s = \mathbb{P}(\Omega_s)$.

Q: Is this a better estimator for $\mathbb{E}[Z]$ than the crude Monte Carlo estimator?

Stratification

This suggests the following estimation strategy:

1. Divide the total number of R replicates into $R_1, \ldots, R_S$.
2. For each $s \in \{1, \ldots, S\}$, simulate R_s replicates of Z_s.
3. Estimate z_s by the empirical average $\hat{z}_s$.
4. Compute the following estimate for $\mathbb{E}[Z]$:

$$\hat{z}_{str} = \sum_{s=1}^{S} p_s \hat{z}_s,$$

where $p_s = \mathbb{P}(\Omega_s)$.

Q: Is this a better estimator for $\mathbb{E}[Z]$ than the crude Monte Carlo estimator?

A: As usual, we need to check biasedness and variance.

Stratification

Note that

$$\begin{aligned}\mathbb{E}\left[\hat{z}_{str}\right] &= \mathbb{E}\left[\sum_{s=1}^{S} p_s \hat{z}_s\right] \\ &= \sum_{s=1}^{S} p_s \frac{1}{R_s} \sum_{j=1}^{R_s} \mathbb{E}\left[Z \mid \Omega_s\right] \\ &= \sum_{s=1}^{S} p_s \mathbb{E}\left[Z \mid \Omega_s\right] = \mathbb{E}\left[Z\right].\end{aligned}$$

Hence, the estimator is unbiased, which is a good thing.

Stratification

Note that

$$\begin{aligned}\mathbb{E}\left[\hat{z}_{str}\right] &= \mathbb{E}\left[\sum_{s=1}^{S} p_s \hat{z}_s\right] \\ &= \sum_{s=1}^{S} p_s \frac{1}{R_s} \sum_{j=1}^{R_s} \mathbb{E}\left[Z \mid \Omega_s\right] \\ &= \sum_{s=1}^{S} p_s \mathbb{E}\left[Z \mid \Omega_s\right] = \mathbb{E}\left[Z\right].\end{aligned}$$

Hence, the estimator is unbiased, which is a good thing.

How about the variance?

Stratification

We know from Chapter 3 that

$$(\hat{z}_s - z_s) \to \mathcal{N}\left(0, \frac{\sigma_s^2}{R_s}\right)$$

as $R_s \to \infty$, where $\sigma_s^2 = \text{Var}[Z_s] = \frac{1}{R_s}\text{Var}[Z \mid \Omega_s]$.

Stratification

We know from Chapter 3 that

$$(\hat{z}_s - z_s) \to \mathcal{N}\left(0, \frac{\sigma_s^2}{R_s}\right)$$

as $R_s \to \infty$, where $\sigma_s^2 = \text{Var}[Z_s] = \frac{1}{R_s}\text{Var}[Z \mid \Omega_s]$.

Hence,

$$p_s(\hat{z}_s - z_s) \to \mathcal{N}\left(0, \frac{p_s^2\sigma_s^2}{R_s}\right)$$

and

$$(\hat{z}_{str} - z) \to \mathcal{N}\left(0, \sum_{s=1}^{S} \frac{p_s^2\sigma_s^2}{R_s}\right)$$

or

$$\sqrt{R}(\hat{z}_{str} - z) \to \mathcal{N}\left(0, \sum_{s=1}^{S} \frac{p_s^2\sigma_s^2}{R_s/R}\right)$$

as $R \to \infty$ in such a way that the R_s/R have non-zero limits.

Stratification

$$\sqrt{R}(\hat{z}_{str} - z) \to \mathcal{N}\left(0, \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s/R}\right)$$

As usual, replacing σ_s^2 by the sample variance gives us a way to evaluate the confidence interval.

Stratification

$$\sqrt{R}(\hat{z}_{str} - z) \to \mathcal{N}\left(0, \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s/R}\right)$$

As usual, replacing σ_s^2 by the sample variance gives us a way to evaluate the confidence interval.

The variance itself is given by

$$\text{Var}\left[\hat{z}_{str}\right] = \text{Var}\left[\sum_{s=1}^{S} p_s \hat{z}_s\right] = \sum_{s=1}^{S} p_s^2 \text{Var}\left[\hat{z}_s\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s}.$$

Stratification

$$\sqrt{R}(\hat{z}_{str} - z) \to \mathcal{N}\left(0, \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s/R}\right)$$

As usual, replacing σ_s^2 by the sample variance gives us a way to evaluate the confidence interval.

The variance itself is given by

$$\text{Var}\left[\hat{z}_{str}\right] = \text{Var}\left[\sum_{s=1}^{S} p_s \hat{z}_s\right] = \sum_{s=1}^{S} p_s^2 \text{Var}\left[\hat{z}_s\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s}.$$

Q: Is $\text{Var}\left[\hat{z}_{str}\right] < \text{Var}\left[\hat{z}\right]$?

Stratification

$$\sqrt{R}(\hat{z}_{str} - z) \to \mathcal{N}\left(0, \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s/R}\right)$$

As usual, replacing σ_s^2 by the sample variance gives us a way to evaluate the confidence interval.

The variance itself is given by

$$\text{Var}\left[\hat{z}_{str}\right] = \text{Var}\left[\sum_{s=1}^{S} p_s \hat{z}_s\right] = \sum_{s=1}^{S} p_s^2 \text{Var}\left[\hat{z}_s\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s}.$$

Q: Is $\text{Var}\left[\hat{z}_{str}\right] < \text{Var}\left[\hat{z}\right]$?

A: That depends on the actual choice of R_s.

Stratification

First thought: we could adopt *proportional allocation*, i.e. take the R_s such that $p_s = \frac{R_s}{R}$.

Then,

$$\text{Var}\left[\hat{\bar{z}}_{str}\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s} = \sum_{s=1}^{S} \frac{p_s \sigma_s^2}{R}.$$

Stratification

First thought: we could adopt *proportional allocation*, i.e. take the R_s such that $p_s = \frac{R_s}{R}$.

Then,

$$\mathrm{Var}\left[\hat{z}_{str}\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s} = \sum_{s=1}^{S} \frac{p_s \sigma_s^2}{R}.$$

Note that $\sum_{s=1}^{S} p_s \sigma_s^2$ can be interpreted as $\mathbb{E}\left[\mathrm{Var}\left[Z_J \mid J\right]\right]$, where J satisfies $\mathbb{P}\left(J = s\right) = p_s$ for $s = 1, \ldots, S$, so that

$$\begin{aligned}
\mathrm{Var}\left[\hat{z}_{str}\right] &= \frac{1}{R}\mathbb{E}\left[\mathrm{Var}\left[Z_J \mid J\right]\right] \leq \frac{1}{R}\left(\mathbb{E}\left[\mathrm{Var}\left[Z_J \mid J\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[Z_J \mid J\right]\right]\right) \\
&= \frac{1}{R}\mathrm{Var}\left[Z_J\right] = \mathrm{Var}\left[\hat{z}\right].
\end{aligned}$$

Stratification

First thought: we could adopt *proportional allocation*, i.e. take the R_s such that $p_s = \frac{R_s}{R}$.

Then,

$$\mathrm{Var}\left[\hat{z}_{str}\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s} = \sum_{s=1}^{S} \frac{p_s \sigma_s^2}{R}.$$

Note that $\sum_{s=1}^{S} p_s \sigma_s^2$ can be interpreted as $\mathbb{E}\left[\mathrm{Var}\left[Z_J \mid J\right]\right]$, where J satisfies $\mathbb{P}\left(J = s\right) = p_s$ for $s = 1, \ldots, S$, so that

$$\begin{aligned}\mathrm{Var}\left[\hat{z}_{str}\right] &= \frac{1}{R}\mathbb{E}\left[\mathrm{Var}\left[Z_J \mid J\right]\right] \leq \frac{1}{R}\left(\mathbb{E}\left[\mathrm{Var}\left[Z_J \mid J\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[Z_J \mid J\right]\right]\right) \\ &= \frac{1}{R}\mathrm{Var}\left[Z_J\right] = \mathrm{Var}\left[\hat{z}\right].\end{aligned}$$

Thus, when choosing proportional allocation, there is always variance reduction!

Example with proportional allocation

Suppose again that $Z = g(X)$, where X is a standard uniform random variable, and we choose the strata

$$\Omega_s = \{\frac{s-1}{S} \leq X < \frac{s}{S}\}.$$

Then,

$$\begin{aligned}
\text{Var}\left[Z_s\right] &= \text{Var}\left[g(X) \mid \frac{s-1}{S} \leq X < \frac{s}{S}\right] \\
&= \text{Var}\left[g\left(\frac{s-1}{S} + \frac{X}{S}\right)\right] \\
&:= \text{Var}\left[g\left(\frac{s-1}{S}\right) + \frac{h(s,S,X)}{S}\right] \\
&= \text{Var}\left[\frac{h(s,S,X)}{S}\right] \leq \frac{1}{S^2}\mathbb{E}\left[h(s,S,X)^2\right]
\end{aligned}$$

Example with proportional allocation

$$\mathrm{Var}\left[Z_s\right] \leq \frac{1}{S^2}\mathbb{E}\left[h(s,S,X)^2\right]$$

Example with proportional allocation

$$\mathrm{Var}\left[Z_s\right] \leq \frac{1}{S^2}\mathbb{E}\left[h(s,S,X)^2\right]$$

If g' is smooth, then we must have that
$|h(s,S,X)| \leq \max_u\{|g'(u)|\} =: c$.

Example with proportional allocation

$$\mathrm{Var}\left[Z_s\right] \leq \frac{1}{S^2}\mathbb{E}\left[h(s,S,X)^2\right]$$

If g' is smooth, then we must have that $|h(s,S,X)| \leq \max_u\{|g'(u)|\} =: c$. Thus,

$$\mathrm{Var}\left[Z_s\right] \leq \frac{c^2}{S^2},$$

so that

$$\mathrm{Var}\left[\hat{z}_{str}\right] = \sum_{s=1}^{S} \frac{p_s\sigma_s^2}{R} \leq \sum_{s=1}^{S} \frac{c^2}{RS^3} = \frac{c^2}{RS^2}.$$

We can obtain variance reduction at rate S^{-2}!

Example with proportional allocation

$$\mathrm{Var}\left[Z_s\right] \leq \frac{1}{S^2}\mathbb{E}\left[h(s,S,X)^2\right]$$

If g' is smooth, then we must have that $|h(s,S,X)| \leq \max_u\{|g'(u)|\} =: c$. Thus,

$$\mathrm{Var}\left[Z_s\right] \leq \frac{c^2}{S^2},$$

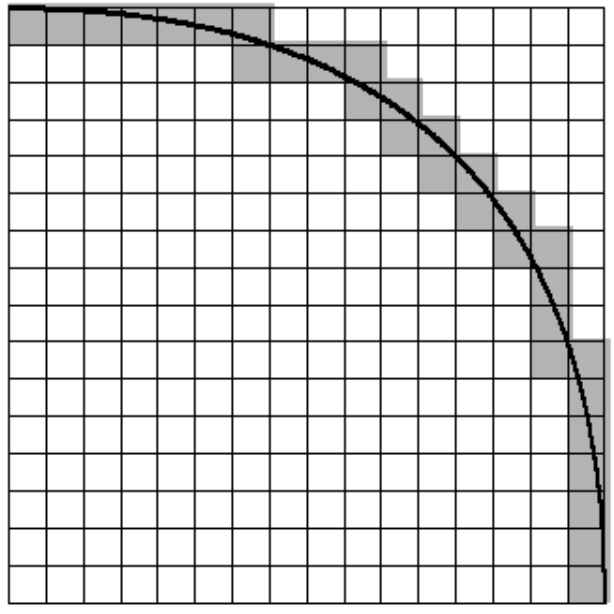
so that

$$\mathrm{Var}\left[\hat{z}_{str}\right] = \sum_{s=1}^{S} \frac{p_s\sigma_s^2}{R} \leq \sum_{s=1}^{S} \frac{c^2}{RS^3} = \frac{c^2}{RS^2}.$$

We can obtain variance reduction at rate S^{-2}! Is this always the case?

Another example with proportional allocation

Smoothness turns out to be essential. Consider $Z = \mathbb{1}_{\{U_1^2+U_2^2\leq 1\}}$ to estimate $\frac{\pi}{4}$.

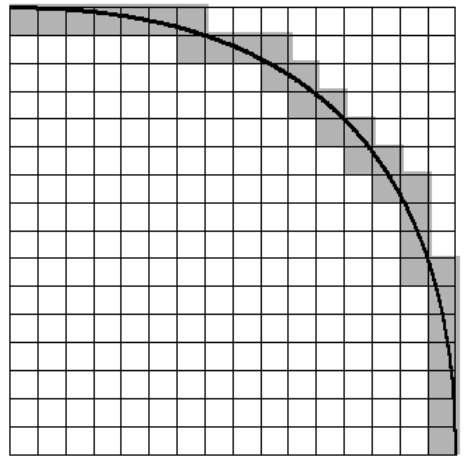


Consider the strata

$$\Omega_{ij} = \left\{ \frac{i-1}{16} < U_1 < \frac{i}{16}, \frac{j-1}{16} < U_2 < \frac{j}{16} \right\},$$

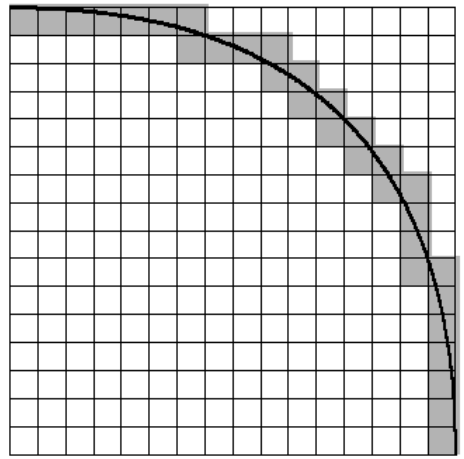
i.e. $S = 256$.

Another example with proportional allocation



In this figure, $\mathrm{Var}[Z_s]$ is zero in many cases because either $\mathbb{P}(Z_s = 0) = 1$ or $\mathbb{P}(Z_s = 1) = 1$.

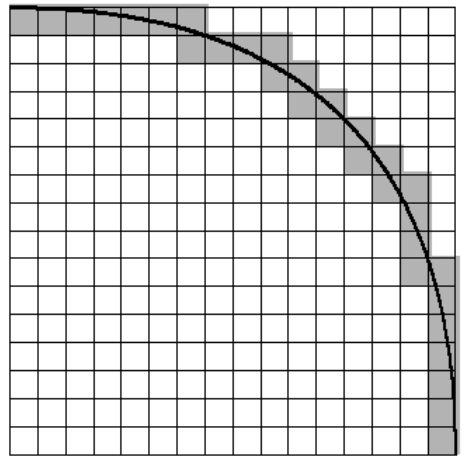
Another example with proportional allocation



In this figure, $\mathrm{Var}[Z_s]$ is zero in many cases because either $\mathbb{P}(Z_s = 0) = 1$ or $\mathbb{P}(Z_s = 1) = 1$.

For the grey cases (number in the order of $\sqrt{S}$), we have $p_s = \frac{1}{S}$.
Why?

Another example with proportional allocation

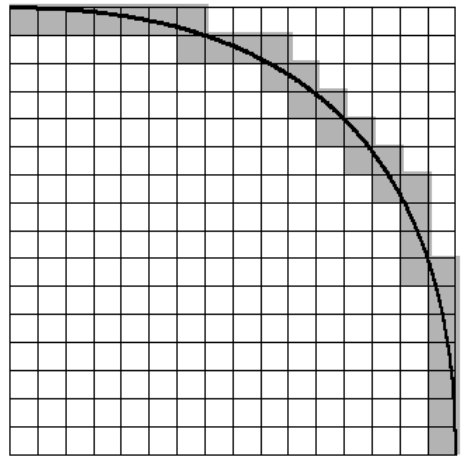


In this figure, $\mathrm{Var}[Z_s]$ is zero in many cases because either $\mathbb{P}(Z_s = 0) = 1$ or $\mathbb{P}(Z_s = 1) = 1$.

For the grey cases (number in the order of $\sqrt{S}$), we have $p_s = \frac{1}{S}$. Why?

Moreover, in these cases, $0 < \mathrm{Var}[Z_s] \leq 1$. Why?

Another example with proportional allocation



In this figure, $\mathrm{Var}[Z_s]$ is zero in many cases because either $\mathbb{P}(Z_s = 0) = 1$ or $\mathbb{P}(Z_s = 1) = 1$.

For the grey cases (number in the order of $\sqrt{S}$), we have $p_s = \frac{1}{S}$. Why?

Moreover, in these cases, $0 < \mathrm{Var}[Z_s] \leq 1$. Why?

Hence, $\mathrm{Var}[\hat{z}_{str}] = \sum_{s=1}^{S} \frac{p_s \sigma_s^2}{R} = O(\frac{\sqrt{S}}{S})$. Thus, we only have variance reduction at rate $S^{-\frac{1}{2}}$.

Stratification

Q: Proportional allocation brings variance reduction, but is it the best variance reductor?

A: Let's find out!

Problem:

$$\begin{aligned} \text{Min} \quad & \text{Var}\left[\hat{z}_{str}\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s} \\ \text{s.t.} \quad & R_1 + \ldots + R_S = R. \end{aligned}$$

Stratification

Q: Proportional allocation brings variance reduction, but is it the best variance reductor?

A: Let's find out!

Problem:

$$\begin{aligned} \text{Min} \quad & \text{Var}\left[\hat{z}_{str}\right] = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s} \\ \text{s.t.} \quad & R_1 + \ldots + R_S = R. \end{aligned}$$

Using Lagrange multipliers

$$f(R_1, \ldots, R_S; \lambda) = \sum_{s=1}^{S} \frac{p_s^2 \sigma_s^2}{R_s} + \lambda \left(\sum_{s=1}^{S} R_s - R \right)$$

leads for $s = 1, \ldots, S$ to

$$\frac{-p_s^2 \sigma_s^2}{(R_s^{opt})^2} = \lambda \text{ and } \sum_{s=1}^{S} R_s^{opt} = R.$$

Stratification

Solution:

$$R_s^{opt} = R \frac{p_s \sigma_s}{\sum_{t=1}^{S} p_t \sigma_t}.$$

Q: Nice, so should we always implement this, or is there a catch?

Stratification

Solution:

$$R_s^{opt} = R \frac{p_s \sigma_s}{\sum_{t=1}^{S} p_t \sigma_t}.$$

Q: Nice, so should we always implement this, or is there a catch?

A: We don't know the σ_s!

Stratification

Solution:

$$R_s^{opt} = R \frac{p_s \sigma_s}{\sum_{t=1}^{S} p_t \sigma_t}.$$

Q: Nice, so should we always implement this, or is there a catch?

A: We don't know the σ_s! One way to go about this is to estimate them using a pilot run or to use an adaptive scheme.

Stratification

A variant of proportional allocation R_s: *poststratification*.

Main idea: allocate the R_s on-the-fly during the simulation, so that $R_s \approx Rp_s$.

Stratification

A variant of proportional allocation R_s: *poststratification*.

Main idea: allocate the R_s on-the-fly during the simulation, so that $R_s \approx R p_s$.

Let the simulation generate i.i.d. samples (Σ_r, Z_r), $r = 1, \ldots, R$, from a multi-variate r.v. (Σ, Z) such that

- $\mathbb{P}(\Sigma = s) = p_s$ (and not, as [AG] claims, p_r!).
- Z conditioned on $\Sigma = s$ has the same distribution as Z_s.

Then, we estimate $\mathbb{E}[Z_s]$ by the empirical average

$$\hat{z}_s = \frac{\sum_{r:\Sigma_r = s} Z_r}{R_s},$$

Stratification

Then, we can use

$$\hat{z}_{poststr} = \sum_{s=1}^{S} p_s \hat{z}_s.$$

Pro's:

Stratification

Then, we can use

$$\hat{z}_{poststr} = \sum_{s=1}^{S} p_s \hat{z}_s.$$

Pro's:

- In the limit, the same CLT as for proportional allocation holds!
- Usable when hard to sample directly from Z_s, but knowledge on the p_s can be used.

Con's:

Stratification

Then, we can use

$$\hat{z}_{poststr} = \sum_{s=1}^{S} p_s \hat{z}_s.$$

Pro's:

- In the limit, the same CLT as for proportional allocation holds!
- Usable when hard to sample directly from Z_s, but knowledge on the p_s can be used.

Con's:

- Still quite larger variance for small R, as the R_s are random variables themselves.

Stratification

The idea of stratification is omnipresent. Think of Latin hypercube sampling, which may be done using stratification.

Goal: sample R d-dimensional random variables $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_R \in [0,1]^d$.

$$(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_R) = \begin{bmatrix} V_{11} & \cdots & V_{1R} \\ \vdots & & \vdots \\ V_{d1} & \cdots & V_{dR} \end{bmatrix}$$

Stratification

The idea of stratification is omnipresent. Think of Latin hypercube sampling, which may be done using stratification.

Goal: sample R d-dimensional random variables $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_R \in [0,1]^d$.

$$(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_R) = \begin{bmatrix} V_{11} & \cdots & V_{1R} \\ \vdots & & \vdots \\ V_{d1} & \cdots & V_{dR} \end{bmatrix}$$

Stratification is usual, to make sure small numbers of sample cover most regions of $[0,1]^d$.

Stratification

$$\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_R) = \begin{bmatrix} V_{11} & \cdots & V_{1R} \\ \vdots & & \vdots \\ V_{d1} & \cdots & V_{dR} \end{bmatrix}$$

Stratification

$$\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_R) = \begin{bmatrix} V_{11} & \cdots & V_{1R} \\ \vdots & & \vdots \\ V_{d1} & \cdots & V_{dR} \end{bmatrix}$$

- Each row is stratified according to the strata

$$[0, \frac{1}{R}), [\frac{1}{R}, \frac{2}{R}), \ldots, [(R-1)/R, 1],$$

  so that each stratum is represented exactly once in each row.
- The representations in different rows are random according to a random permutation: in row j use permutation $\pi_j = (\pi_j(1), \ldots, \pi_j(R))$ of $(1, \ldots, R)$.

Stratification

$$\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_R) = \begin{bmatrix} V_{11} & \cdots & V_{1R} \\ \vdots & & \vdots \\ V_{d1} & \cdots & V_{dR} \end{bmatrix}$$

- Each row is stratified according to the strata
$$[0, \frac{1}{R}), [\frac{1}{R}, \frac{2}{R}), \ldots, [(R-1)/R, 1],$$
so that each stratum is represented exactly once in each row.
- The representations in different rows are random according to a random permutation: in row j use permutation $\pi_j = (\pi_j(1), \ldots, \pi_j(R))$ of $(1, \ldots, R)$.

Then,

$$V_{jr} = \frac{1}{R}(\pi_j(r) - 1 + U_{jr}),$$

where the U_{jr} are standard-uniformly distributed. There now is dependence between the $\boldsymbol{V}_r$! Often used when estimating $f(\boldsymbol{V})$, as it yields variance reduction.

Chapter V.8
Indirect Sampling

Indirect sampling

In some cases, some parts of the expectation $\mathbb{E}[Z]$ can be evaluated analytically.

Sometimes, $\mathbb{E}[Z]$ may be related to the expectations of some other random variables.

When such phenomena may be exploited for variance reduction, this is called 'indirect estimation'. Let us see one example.

Indirect sampling

Let $T_1, T_2, \ldots$ be non-negative i.i.d. random variables, and let

$$Z := \sup\{n : S_n \leq t\}$$

be the number of renewals up to time t, where
$S_n := T_1 + \ldots + T_n$.

Goal: to simulate $\mathbb{E}[Z]$.

Indirect sampling

Note that

$$\tau := Z + 1 = \inf\{n : S_n > t\},$$

is a stopping time for $T_1, T_2, \ldots$, so that we have by Wald's identity:

$$\mathbb{E}[S_\tau] = \mathbb{E}[T_1]\,\mathbb{E}[\tau] = \mathbb{E}[T_1]\,(\mathbb{E}[Z] + 1).$$

Thus, we can estimate $\mathbb{E}[Z]$ by drawing replicates of S_τ, created by generating sample paths of the renewal process.

Q: Can we do better?

Indirect sampling

Note that

$$\tau := Z + 1 = \inf\{n : S_n > t\},$$

is a stopping time for $T_1, T_2, \ldots$, so that we have by Wald's identity:

$$\mathbb{E}[S_\tau] = \mathbb{E}[T_1]\,\mathbb{E}[\tau] = \mathbb{E}[T_1](\mathbb{E}[Z] + 1).$$

Thus, we can estimate $\mathbb{E}[Z]$ by drawing replicates of S_τ, created by generating sample paths of the renewal process.

Q: Can we do better?

A: Oh yes..., definitely!

Indirect sampling

We have

$$\mathbb{E}[S_\tau] = \mathbb{E}[T_1]\,\mathbb{E}[\tau] = \mathbb{E}[T_1](\mathbb{E}[Z] + 1).$$

Note that $\xi := S_\tau - t$ is the overshoot of an inter-event time after t.

Thus, we can use

$$\mathbb{E}[Z] = \frac{t + \mathbb{E}[\xi]}{\mathbb{E}[T_1]} - 1.$$

- Better to estimate the RHS, as estimating $\mathbb{E}[\xi]$ may computationally be easier than creating replicates of S_τ.
- Furthermore, great variance reduction. Suppose the $T_i \sim Exp(1)$ and $t = 50$. Then,

$$\mathrm{Var}[Z] = 50, \text{ but } \mathrm{Var}\left[\frac{t + \xi}{\mathbb{E}[T_1]} - 1\right] = 1.$$

In other words, indirect estimation leads to variance reduction by using our brains!

Chapter V.1
Importance Sampling

Importance sampling

- Suppose we want to estimate some performance measure $z = \mathbb{E}[Z] = \mathbb{E}[h(\boldsymbol{X})]$.
- We allow the output variable $h(\boldsymbol{X})$ to be a function of a random input vector $\boldsymbol{X} = (X_1, \ldots, X_n)$.
- For ease of discussion, let's say that X has pdf $f(\boldsymbol{x})$ and/or cdf $F(\boldsymbol{x})$.
- Thus, $z = \mathbb{E}[h(\boldsymbol{X})] = \int h(\boldsymbol{x}) f(\boldsymbol{x}) dx$.

Importance sampling

- Suppose we want to estimate some performance measure $z = \mathbb{E}[Z] = \mathbb{E}[h(\boldsymbol{X})]$.
- We allow the output variable $h(\boldsymbol{X})$ to be a function of a random input vector $\boldsymbol{X} = (X_1, \ldots, X_n)$.
- For ease of discussion, let's say that X has pdf $f(\boldsymbol{x})$ and/or cdf $F(\boldsymbol{x})$.
- Thus, $z = \mathbb{E}[h(\boldsymbol{X})] = \int h(\boldsymbol{x}) f(\boldsymbol{x}) dx$.

Suppose now that a crude Monte Carlo simulation of z is inefficient, because

a) it is difficult so simulate a random vector having density function $f(\boldsymbol{x})$, and/or

b) the variance of $h(\boldsymbol{X})$ is just very large, requiring too many replicates.

In either of these cases, importance sampling might be a lifesaver.

Importance sampling

What the book says:

Importance sampling obtains its variance reduction by modifying the sampling distribution $\mathbb{P}$ so that most of the sampling is done in that part of the state space that contributes the most to $z = \mathbb{E}[Z]$.

Importance sampling

What the book says:

Importance sampling obtains its variance reduction by modifying the sampling distribution $\mathbb{P}$ so that most of the sampling is done in that part of the state space that contributes the most to $z = \mathbb{E}[Z]$.

Specifically, if we choose a sampling (or importance) distribution $\tilde{\mathbb{P}}$ for which there exists a density (or Radon-Nikodym derivative) L such that

$$\mathbb{1}_{\{Z(\omega)\neq 0\}}\mathbb{P}(d\omega) = \mathbb{1}_{\{Z(\omega)\neq 0\}}L(\omega)\tilde{\mathbb{P}}(d\omega),$$

in the sense of equality of measures, then

$$z = \mathbb{E}[Z] = \tilde{\mathbb{E}}[ZL],$$

where $\tilde{\mathbb{E}}$ is the expectation associated with $\tilde{\mathbb{E}}$.

Importance sampling

What the book says:

Importance sampling obtains its variance reduction by modifying the sampling distribution $\mathbb{P}$ so that most of the sampling is done in that part of the state space that contributes the most to $z = \mathbb{E}[Z]$.

Specifically, if we choose a sampling (or importance) distribution $\tilde{\mathbb{P}}$ for which there exists a density (or Radon-Nikodym derivative) L such that

$$\mathbb{1}_{\{Z(\omega)\neq 0\}}\mathbb{P}(d\omega) = \mathbb{1}_{\{Z(\omega)\neq 0\}}L(\omega)\tilde{\mathbb{P}}(d\omega),$$

in the sense of equality of measures, then

$$z = \mathbb{E}[Z] = \tilde{\mathbb{E}}[ZL],$$

where $\tilde{\mathbb{E}}$ is the expectation associated with $\tilde{\mathbb{E}}$.

Output analysis is then performed precisely as for the crude MC method by generating R i.i.d. replicates of $Z_1L_1, \ldots, Z_RL_R$ from $\tilde{\mathbb{P}}$.

Importance sampling

The book is quite right, but I'll be explaining things a little differently, under the assumption that $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a joint density (or mass) function $f(\boldsymbol{x}) = f(x_1, \ldots, x_n)$.

Importance sampling

The book is quite right, but I'll be explaining things a little differently, under the assumption that $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a joint density (or mass) function $f(\boldsymbol{x}) = f(x_1, \ldots, x_n)$.

Supose that we want to estimate

$$z = \mathbb{E}\left[h(\boldsymbol{X})\right] = \int h(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}.$$

Instead of generating replicates of $h(\boldsymbol{X})$ and averaging, we note that

$$z = \int h(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_g \left[h(\boldsymbol{X}) \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})} \right] = \mathbb{E}_g \left[h(\boldsymbol{X}) L(\boldsymbol{X}) \right],$$

where $L(\boldsymbol{x}) = f(\boldsymbol{x})/g(\boldsymbol{x})$ is called is a likelihood ratio function.

Importance sampling

$$z = \int h(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_g \left[h(\boldsymbol{X}) \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})} \right] = \mathbb{E}_g \left[h(\boldsymbol{X}) L(\boldsymbol{X}) \right],$$

where $L(\boldsymbol{x}) = f(\boldsymbol{x})/g(\boldsymbol{x})$ is called a likelihood ratio function.

Importance sampling

$$z = \int h(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_g \left[h(\boldsymbol{X}) \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})} \right] = \mathbb{E}_g \left[h(\boldsymbol{X}) L(\boldsymbol{X}) \right],$$

where $L(\boldsymbol{x}) = f(\boldsymbol{x})/g(\boldsymbol{x})$ is called a likelihood ratio function.
This suggests the following estimation scheme:

1. Create R replicates of $\boldsymbol{X}$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_R$, as if they have density $g(\boldsymbol{x})$.
2. Compute the values $h(\boldsymbol{X}_i)L(\boldsymbol{X}_i)$, $i = 1, \ldots, R$.
3. Average these values to get an importance sampling estimate based on importance sampling density g.

Importance sampling

$$z = \int h(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_g \left[h(\boldsymbol{X}) \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})} \right] = \mathbb{E}_g \left[h(\boldsymbol{X}) L(\boldsymbol{X}) \right],$$

where $L(\boldsymbol{x}) = f(\boldsymbol{x})/g(\boldsymbol{x})$ is called a likelihood ratio function.
This suggests the following estimation scheme:

1. Create R replicates of $\boldsymbol{X}$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_R$, as if they have density $g(\boldsymbol{x})$.
2. Compute the values $h(\boldsymbol{X}_i)L(\boldsymbol{X}_i)$, $i = 1, \ldots, R$.
3. Average these values to get an importance sampling estimate based on importance sampling density g.

Note: works only if $g(\boldsymbol{x}) = 0$ implies $h(\boldsymbol{x})f(\boldsymbol{x}) = 0$.

Q: How to choose g?

Importance sampling

$$z = \int h(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_g \left[h(\boldsymbol{X}) \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})} \right] = \mathbb{E}_g \left[h(\boldsymbol{X}) L(\boldsymbol{X}) \right],$$

where $L(\boldsymbol{x}) = f(\boldsymbol{x})/g(\boldsymbol{x})$ is called a likelihood ratio function.
This suggests the following estimation scheme:

1. Create R replicates of $\boldsymbol{X}$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_R$, as if they have density $g(\boldsymbol{x})$.
2. Compute the values $h(\boldsymbol{X}_i)L(\boldsymbol{X}_i)$, $i = 1, \ldots, R$.
3. Average these values to get an importance sampling estimate based on importance sampling density g.

Note: works only if $g(\boldsymbol{x}) = 0$ implies $h(\boldsymbol{x})f(\boldsymbol{x}) = 0$.

Q: How to choose g?

A: Good question. The choice of g is critical in order to get substantial variance reduction. Taking the wrong g may lead to substantial variance increases.

Importance sampling

A simple example of importance sampling:

If X is Bin(n, p) distributed,

$$\begin{aligned}
\mathbb{E}\left[h(X)\right] &= \sum_{x=0}^{n} h(x) \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} h(x) \frac{p^x (1-p)^{n-x}}{q^x (1-q)^{n-x}} \binom{n}{x} q^x (1-q)^{n-x} \\
&= \mathbb{E}_g \left[h(X) \frac{p^X (1-p)^{n-X}}{q^X (1-q)^{n-X}} \right],
\end{aligned}$$

where X is Bin(n, q) distributed under density g.

Importance sampling

A simple example of importance sampling:

If X is Bin(n, p) distributed,

$$\begin{aligned}\mathbb{E}\left[h(X)\right] &= \sum_{x=0}^{n} h(x)\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=0}^{n} h(x)\frac{p^x(1-p)^{n-x}}{q^x(1-q)^{n-x}}\binom{n}{x}q^x(1-q)^{n-x} \\ &= \mathbb{E}_g\left[h(X)\frac{p^X(1-p)^{n-X}}{q^X(1-q)^{n-X}}\right],\end{aligned}$$

where X is Bin(n, q) distributed under density g.

Suppose we want to know $\mathbb{P}(X \leq 1)$ where X is Bin$(100, 1 - 10^{-8})$ distributed.

Importance sampling

A simple example of importance sampling:

If X is $\text{Bin}(n, p)$ distributed,

$$\begin{aligned}\mathbb{E}\left[h(X)\right] &= \sum_{x=0}^{n} h(x)\binom{n}{x}p^x(1-p)^{n-x} \\ &= \sum_{x=0}^{n} h(x)\frac{p^x(1-p)^{n-x}}{q^x(1-q)^{n-x}}\binom{n}{x}q^x(1-q)^{n-x} \\ &= \mathbb{E}_g\left[h(X)\frac{p^X(1-p)^{n-X}}{q^X(1-q)^{n-X}}\right],\end{aligned}$$

where X is $\text{Bin}(n, q)$ distributed under density g.

Suppose we want to know $\mathbb{P}(X \leq 1)$ where X is $\text{Bin}(100, 1 - 10^{-8})$ distributed.

Use this with e.g. $q = \frac{2}{100}$, since then we've made the region $\{X \leq 1\}$ 'more important' (more probability mass, but caution: not too much!).

Importance sampling

Another example.

If X is $\exp(\lambda)$ distributed,

$$\begin{aligned}\mathbb{E}\left[h(X)\right] &= \int_{x=0}^{\infty} h(x)\lambda e^{-\lambda x} dx \\ &= \int_{x=0}^{\infty} h(x)\frac{\lambda e^{-\lambda x}}{\mu e^{-\mu x}}\mu e^{-\mu x} dx \\ &= \mathbb{E}_g\left[h(X)\frac{\lambda e^{-\lambda X}}{\mu e^{-\mu X}}\right],\end{aligned}$$

where X is $\exp(\mu)$ distributed under density g.

Importance sampling

Our importance sampling estimator now is

$$\hat{z}_{IS} = \frac{1}{R} \sum_{i=1}^{R} h(\boldsymbol{X}_i) L(\boldsymbol{X}_i).$$

with $\boldsymbol{X}_i$ sampled from density g. How to choose g?

Importance sampling

Our importance sampling estimator now is

$$\hat{z}_{IS} = \frac{1}{R} \sum_{i=1}^{R} h(\boldsymbol{X}_i) L(\boldsymbol{X}_i).$$

with $\boldsymbol{X}_i$ sampled from density g. How to choose g?
We don't have to worry about biasedness. We already saw that

$$\mathbb{E}\left[\hat{z}_{IS}\right] = \mathbb{E}_g\left[h(\boldsymbol{X}) L(\boldsymbol{X})\right] = \mathbb{E}\left[h(\boldsymbol{X})\right] = z.$$

Unbiased estimator, regardless of the choice of g.

Importance sampling

Let's check the variance.

$$\sigma^2_{IS} = \mathrm{Var}_g\left[\frac{1}{R}\sum_{i=1}^{R} h(\boldsymbol{X}_i)L(\boldsymbol{X}_i)\right] = \frac{1}{R}\mathrm{Var}_g\left[h(\boldsymbol{X})L(\boldsymbol{X})\right]$$
$$= \frac{1}{R}\left(\mathbb{E}_g\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] - \mathbb{E}\left[h(\boldsymbol{X})\right]^2\right)$$
$$= \frac{1}{R}\left(\mathbb{E}\left[h^2(\boldsymbol{X})L(\boldsymbol{X})\right] - \mathbb{E}\left[h(\boldsymbol{X})\right]^2\right)$$

This number can be estimated by the sample variance as usual, and confidence intervals can be built.

Problem:

$$\min_g \left\{\mathbb{E}_g\left[H^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] : L(\boldsymbol{X}) = \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right\}.$$

Importance sampling

Theorem: Let g^* be given by

$$g^*(\boldsymbol{x}) = \frac{|h(\boldsymbol{x})|}{\mathbb{E}\left[|h(\boldsymbol{X})|\right]} f(\boldsymbol{x}).$$

Then, for any importance sampling density g:

$$\mathbb{E}_{g^*}\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] \leq \mathbb{E}_{g}\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right].$$

Importance sampling

Theorem: Let g^* be given by

$$g^*(\boldsymbol{x}) = \frac{|h(\boldsymbol{x})|}{\mathbb{E}\left[|h(\boldsymbol{X})|\right]} f(\boldsymbol{x}).$$

Then, for any importance sampling density g:

$$\mathbb{E}_{g^*}\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] \leq \mathbb{E}_g\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right].$$

Proof:

$$\begin{aligned}
\mathbb{E}_{g^*}\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] &= \mathbb{E}_{g^*}\left[h^2(\boldsymbol{X})\frac{f^2(X)\left(\mathbb{E}\left[|h(\boldsymbol{X})|\right]\right)^2}{|h(\boldsymbol{X})|^2 f^2(X)}\right] \\
&= \mathbb{E}_{g^*}\left[\left(\mathbb{E}\left[|h(\boldsymbol{X})|\right]\right)^2\right] = \left(\mathbb{E}\left[|h(\boldsymbol{X})|\right]\right)^2 \\
&= \left(\mathbb{E}_g\left[L(\boldsymbol{X})|h(\boldsymbol{X})|\right]\right)^2 \leq \mathbb{E}_g\left[\left(L(\boldsymbol{X})|h(\boldsymbol{X})|\right)^2\right].
\end{aligned}$$

Last line follows from Jensen's equality.

Importance sampling

Something peculiar is going on.

$$g^*(\boldsymbol{x}) = \frac{|h(\boldsymbol{x})|}{\mathbb{E}\left[|h(\boldsymbol{X})|\right]} f(\boldsymbol{x}).$$

Suppose that $h(\boldsymbol{x}) \geq 0$ for any x.

Then,

$$\begin{aligned}
\mathbb{E}_{g^*}\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] &= \mathbb{E}\left[h^2(\boldsymbol{X})L(\boldsymbol{X})\right] = \mathbb{E}\left[h^2(\boldsymbol{X})\frac{f(\boldsymbol{X})}{g^*(\boldsymbol{X})}\right] \\
&= \mathbb{E}\left[h(\boldsymbol{X})\mathbb{E}\left[h(\boldsymbol{X})\right]\right] = \mathbb{E}\left[h(\boldsymbol{X})\right]^2 \\
&= \mathbb{E}_{g^*}\left[h(\boldsymbol{X})L(\boldsymbol{X})\right]^2
\end{aligned}$$

This means that sampling from density g^* yields a zero-variance estimator! Awesome!

Importance sampling

Something peculiar is going on.

$$g^*(\boldsymbol{x}) = \frac{|h(\boldsymbol{x})|}{\mathbb{E}\left[|h(\boldsymbol{X})|\right]} f(\boldsymbol{x}).$$

Suppose that $h(\boldsymbol{x}) \geq 0$ for any x.

Then,

$$\begin{aligned}
\mathbb{E}_{g^*}\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] &= \mathbb{E}\left[h^2(\boldsymbol{X})L(\boldsymbol{X})\right] = \mathbb{E}\left[h^2(\boldsymbol{X})\frac{f(\boldsymbol{X})}{g^*(\boldsymbol{X})}\right] \\
&= \mathbb{E}\left[h(\boldsymbol{X})\mathbb{E}\left[h(\boldsymbol{X})\right]\right] = \mathbb{E}\left[h(\boldsymbol{X})\right]^2 \\
&= \mathbb{E}_{g^*}\left[h(\boldsymbol{X})L(\boldsymbol{X})\right]^2
\end{aligned}$$

This means that sampling from density g^* yields a zero-variance estimator! Awesome!

- Q: Where is the catch?

Importance sampling

Something peculiar is going on.

$$g^*(\boldsymbol{x}) = \frac{|h(\boldsymbol{x})|}{\mathbb{E}\left[|h(\boldsymbol{X})|\right]} f(\boldsymbol{x}).$$

Suppose that $h(\boldsymbol{x}) \geq 0$ for any x.

Then,

$$\begin{aligned}
\mathbb{E}_{g^*}\left[h^2(\boldsymbol{X})L^2(\boldsymbol{X})\right] &= \mathbb{E}\left[h^2(\boldsymbol{X})L(\boldsymbol{X})\right] = \mathbb{E}\left[h^2(\boldsymbol{X})\frac{f(\boldsymbol{X})}{g^*(\boldsymbol{X})}\right] \\
&= \mathbb{E}\left[h(\boldsymbol{X})\mathbb{E}\left[h(\boldsymbol{X})\right]\right] = \mathbb{E}\left[h(\boldsymbol{X})\right]^2 \\
&= \mathbb{E}_{g^*}\left[h(\boldsymbol{X})L(\boldsymbol{X})\right]^2
\end{aligned}$$

This means that sampling from density g^* yields a zero-variance estimator! Awesome!

- Q: Where is the catch?
- A: We don't know $\mathbb{E}\left[|h(\boldsymbol{X})|\right]$ or $\mathbb{E}\left[h(\boldsymbol{X})\right]$, that's actually the thing we want to estimate. Bummer...

Importance sampling

Fret not. An often used successful 'change of measure' is given by the technique of exponential tilting, where we choose

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}) = e^{\boldsymbol{\theta}^T \boldsymbol{x} - \kappa(\boldsymbol{\theta})} f(\boldsymbol{x}),$$

where $\kappa(\boldsymbol{\theta}) = \log \mathbb{E}\left[e^{\boldsymbol{\theta}^T \boldsymbol{X}}\right]$.

Importance sampling

Fret not. An often used successful 'change of measure' is given by the technique of exponential tilting, where we choose

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}) = e^{\boldsymbol{\theta}^T \boldsymbol{x} - \kappa(\boldsymbol{\theta})} f(\boldsymbol{x}),$$

where $\kappa(\boldsymbol{\theta}) = \log \mathbb{E}\left[e^{\boldsymbol{\theta}^T \boldsymbol{X}}\right]$.

For one-dimensional densities, this implies

$$g_\theta(x) = \frac{e^{\theta x}}{\mathbb{E}\left[e^{\theta X}\right]} f(x).$$

Importance sampling

For one-dimensional densities,

$$g_\theta(x) = \frac{e^{\theta x}}{\mathbb{E}\left[e^{\theta X}\right]} f(\boldsymbol{x}).$$

Examples:

- When X has a Gamma density with parameters α and λ, i.e. $f(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x}/\Gamma(\alpha)$, then

$$\mathbb{E}\left[e^{\theta X}\right] = \int_0^\infty e^{\theta x} \lambda^\alpha x^{\alpha-1} \frac{e^{-\lambda x}}{\Gamma(\alpha)} dx = \left(\frac{\lambda}{\lambda-\theta}\right)^\alpha,$$

  and $g_\theta(x) = (\lambda-\theta)^\alpha x^{\alpha-1} e^{-(\lambda-\theta)x}/\Gamma(\alpha)$: Gamma distribution with shifted scale parameter $\lambda - \theta$.

Importance sampling

For one-dimensional densities,

$$g_\theta(x) = \frac{e^{\theta x}}{\mathbb{E}\left[e^{\theta X}\right]} f(\boldsymbol{x}).$$

Examples:

- The $\mathrm{N}(\mu, \sigma^2)$ distribution gets tilted into the $\mathrm{N}(\mu + \theta\sigma^2, \sigma^2)$ distribution.
- The $\mathrm{bin}(n, \alpha)$ distribution gets tilted into the $\mathrm{bin}(n, \frac{\alpha e^\theta}{1-\alpha+\alpha e^\theta})$ distribution.
- The $\mathrm{Poisson}(\mu)$ distribution gets tilted into the $\mathrm{Poisson}(\mu e^\theta)$ distribution.
- Multidimensional: the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{C})$ distribution gets tilted into the $\mathrm{N}(\boldsymbol{\mu}+\boldsymbol{C\theta}, \boldsymbol{C})$ distribution.

Importance sampling

Why exponential tilting? Let's see an example.

Suppose $X_1, \ldots, X_n$ are i.i.d with common density $f(x)$ and suppose that the importance distribution preserves the i.i.d. property but changes $f(x)$ to $g_\theta(x) = \frac{e^{\theta x}}{\mathbb{E}\left[e^{\theta X}\right]} f(x)$. Then, note that

$$L(\boldsymbol{X}) = \frac{f^n(\boldsymbol{X})}{g_\theta^n(\boldsymbol{X})} = \prod_{i=1}^{n} \frac{f(X_i)}{g_\theta(X_i)} = e^{-\theta S_n} (\mathbb{E}\left[e^{\theta X_1}\right])^n,$$

where $S_n = \sum_{i=1}^n X_i$. This is a very simple form!

Importance sampling

Why exponential tilting? Let's see an example.

Suppose $X_1, \ldots, X_n$ are i.i.d with common density $f(x)$ and suppose that the importance distribution preserves the i.i.d. property but changes $f(x)$ to $g_\theta(x) = \frac{e^{\theta x}}{\mathbb{E}\left[e^{\theta X}\right]} f(x)$. Then, note that

$$L(\boldsymbol{X}) = \frac{f^n(\boldsymbol{X})}{g_\theta^n(\boldsymbol{X})} = \prod_{i=1}^{n} \frac{f(X_i)}{g_\theta(X_i)} = e^{-\theta S_n} (\mathbb{E}\left[e^{\theta X_1}\right])^n,$$

where $S_n = \sum_{i=1}^n X_i$. This is a very simple form!

- Exponential tilting is often used in problems involving light-tailed r.v.'s.
- In this case, for instance, it can be used to make large values of $S_n = X_1 + \ldots + X_n$ more likely.
- If we aim for values of t or larger, a common choice is to choose θ such that $\mathbb{E}_{g_\theta}[S_n] = t$. Why?

Importance sampling: stopped processes

- Suppose $X_1, X_2, \ldots$ are i.i.d. with common density $f(x)$.
- Assume $Z = h(X_1, \ldots, X_\tau)\mathbb{1}_{\{\tau<\infty\}}$.
- τ is a stopping time adapted to the X_i, e.g. $\tau = \inf\{n : \sum_{i=1}^n |X_i| > t\}$.
- Let g be an importance sampling density and $L_n(\boldsymbol{X}) := \prod_{i=1}^n f(X_i)/g(X_i)$.

Then

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{n=1}^{\infty} \mathbb{E}\left[Z\mathbb{1}_{\{\tau=n\}}\right] = \sum_{n=1}^{\infty} \mathbb{E}_g\left[ZL_n\mathbb{1}_{\{\tau=n\}}\right] \\ &= \mathbb{E}_g\left[ZL_\tau\right].\end{aligned}$$

Importance sampling: stopped processes

- Suppose $X_1, X_2, \ldots$ are i.i.d. with common density $f(x)$.
- Assume $Z = h(X_1, \ldots, X_\tau)\mathbb{1}_{\{\tau<\infty\}}$.
- τ is a stopping time adapted to the X_i, e.g. $\tau = \inf\{n : \sum_{i=1}^n |X_i| > t\}$.
- Let g be an importance sampling density and $L_n(\boldsymbol{X}) := \prod_{i=1}^n f(X_i)/g(X_i)$.

Then

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{n=1}^{\infty} \mathbb{E}\left[Z\mathbb{1}_{\{\tau=n\}}\right] = \sum_{n=1}^{\infty} \mathbb{E}_g\left[ZL_n\mathbb{1}_{\{\tau=n\}}\right] \\ &= \mathbb{E}_g\left[ZL_\tau\right].\end{aligned}$$

Thus, IS-estimate of $\mathbb{E}[Z]$ can be obtained as follows:

- Generate $X_1, X_2, \ldots$ under density g until τ.
- Compute $h(X_1, \ldots, X_\tau)L_\tau(\boldsymbol{X})$.
- Repeat these steps R times and average.

Importance sampling: discrete-time Markov Chains

- Let $\{X_s : s \in \mathbb{N}\}$ be a time-inhomogeneous DTMC on a countable state space.

Importance sampling: discrete-time Markov Chains

- Let $\{X_s : s \in \mathbb{N}\}$ be a time-inhomogeneous DTMC on a countable state space.
- Define

$$p_0(x) := \mathbb{P}(X_0 = x) \text{ and } p_n(x, y) = \mathbb{P}(X_n = y \mid X_{n-1} = x).$$

- One can use importance sampling, so that $\{X_s : s \in \mathbb{N}\}$ remains a Markov chain after the 'change of measure'.

Importance sampling: discrete-time Markov Chains

- Let $\{X_s : s \in \mathbb{N}\}$ be a time-inhomogeneous DTMC on a countable state space.
- Define

$$p_0(x) := \mathbb{P}(X_0 = x) \text{ and } p_n(x, y) = \mathbb{P}(X_n = y \mid X_{n-1} = x).$$

- One can use importance sampling, so that $\{X_s : s \in \mathbb{N}\}$ remains a Markov chain after the 'change of measure'. E.g.,

$$q_0(x) := \mathbb{P}_g(X_0 = x) \text{ and } q_n(x, y) = \mathbb{P}_g(X_n = y \mid X_{n-1} = x).$$

- When τ is a stopping time, $X_1, \ldots, X_\tau$ has associated likelihood ratio

$$L(X_1, \ldots, X_\tau) = \frac{p_0(X_0)}{q_0(X_0)} \prod_{n=1}^{\tau} \frac{p_n(X_{n-1}, X_n)}{q_n(X_{n-1}, X_n)}.$$

- Importance sampling now possible for replication of $h(X_1, \ldots, X_\tau)$.

Importance sampling: continuous-time Markov Chains

- Let $\{X(t), t \geq 0\}$ be a time-homogeneous CTMC on a countable state space.
- Define

$$p_0(x) := \mathbb{P}(X(0) = x), \quad a(x, y)dt := \mathbb{P}(X(t + dt) = y | X(t) = x).$$

Importance sampling: continuous-time Markov Chains

- Let $\{X(t), t \geq 0\}$ be a time-homogeneous CTMC on a countable state space.
- Define

$$p_0(x) := \mathbb{P}(X(0) = x), \quad a(x, y)dt := \mathbb{P}(X(t + dt) = y | X(t) = x).$$

- Let $J(t)$ be the number of state transitions in $[0, t]$, and let $T_1, T_2, \ldots$ be the transition epochs.
- Note that $T_{j+1} - T_j$ is exponentially distributed with rate $a(x) = \sum_y a(x, y)$.
- One can use importance sampling, so that $\{X_s : s \in \mathbb{N}\}$ remains a Markov chain after the 'change of measure'.

Importance sampling: continuous-time Markov Chains

- Let $\{X(t), t \geq 0\}$ be a time-homogeneous CTMC on a countable state space.
- Define

$$p_0(x) := \mathbb{P}(X(0) = x), \quad a(x,y)dt := \mathbb{P}(X(t+dt) = y | X(t) = x).$$

- Let $J(t)$ be the number of state transitions in $[0, t]$, and let $T_1, T_2, \ldots$ be the transition epochs.
- Note that $T_{j+1} - T_j$ is exponentially distributed with rate $a(x) = \sum_y a(x,y)$.
- One can use importance sampling, so that $\{X_s : s \in \mathbb{N}\}$ remains a Markov chain after the 'change of measure'. E.g.,

$$q_0(x) := \mathbb{P}_g(X(0) = x), \quad b(x,y)dt = \mathbb{P}_g(X(t+dt) = y | X(t) = x)$$

Importance sampling: continuous-time Markov Chains

- When τ is a stopping time, $X(0), T_1, X(T_1), \ldots, T_{J(\tau)}, X(T_{J(\tau)})$ has associated likelihood ratio

$$L = \frac{p_0(X(0))}{q_0(X(0))} \prod_{j=1}^{J(\tau)} \frac{a\left(X(T_{j-1}), X(T_j)\right) b\left(X(T_{j-1})\right)}{b\left(X(T_{j-1}), X(T_j)\right)) a\left(X(T_{j-1})\right)}$$

$$\times \prod_{j=1}^{J(\tau)} \frac{a\left(X(T_{j-1})\right)) e^{-a\left(X(T_{j-1})\right)(T_j - T_{j-1})}}{b\left(X(T_{j-1})\right)) e^{-b\left(X(T_{j-1})\right)(T_j - T_{j-1})}}$$

or, in short,

$$L = \frac{p_0(X(0))}{q_0(X(0))} \prod_{j=1}^{J(\tau)} \frac{a\left(X(T_{j-1}), X(T_j)\right)) e^{-a\left(X(T_{j-1})\right)(T_j - T_{j-1})}}{b\left(X(T_{j-1}), X(T_j)\right)) e^{-b\left(X(T_{j-1})\right)(T_j - T_{j-1})}}.$$

- Importance sampling now possible for replication of $h(X(0), T_1, X(T_1), \ldots, T_{J(\tau)}, X(T_{J(\tau)}))$.

Importance sampling: compound Poisson processes

- Let $X(t) = \sum_{i=1}^{N(t)} Y_i$, where the Y_i are i.i.d. with density f, and $\{N(t), t \geq 0\}$ is a Poisson(λ) process.
- Let T_i be the time between the $i-1$-st and the i-th jump of the Poisson process.
- Let $R(t) = t - \sum_{i=1}^{N(t)} T_i$ be the time between t and the last event prior to that.
- Using importance sampling, one can make $X(t)$ a compound Poisson process with parameters $\tilde{f}$ and $\tilde{\lambda}$.

Importance sampling: compound Poisson processes

- Let $X(t) = \sum_{i=1}^{N(t)} Y_i$, where the Y_i are i.i.d. with density f, and $\{N(t), t \geq 0\}$ is a Poisson(λ) process.
- Let T_i be the time between the $i-1$-st and the i-th jump of the Poisson process.
- Let $R(t) = t - \sum_{i=1}^{N(t)} T_i$ be the time between t and the last event prior to that.
- Using importance sampling, one can make $X(t)$ a compound Poisson process with parameters $\tilde{f}$ and $\tilde{\lambda}$.
- The likelihood ratio for $Y_1, T_1, \ldots, Y_{N(t)}, T_{N(t)}, R(t)$ is

$$L = \frac{e^{-\lambda R(t)}}{e^{-\tilde{\lambda} R(t)}} \prod_{i=1}^{N(t)} \frac{f(Y_i)}{\tilde{f}(Y_i)} \frac{\lambda e^{-\lambda T_i}}{\tilde{\lambda} e^{-\tilde{\lambda} T_i}} = \left(\frac{\lambda}{\tilde{\lambda}}\right)^{N(t)} e^{-(\lambda - \tilde{\lambda})t} \prod_{i=1}^{N(t)} \frac{f(Y_i)}{\tilde{f}(Y_i)}.$$

- And again, importance sampling now possible for replication of $h(Y_1, T_1, \ldots, Y_{N(t)}, T_{N(t)}, R(t))$!