Chapter 5

Variance Reduction Techniques

Main Topics

- What is variance reduction?
- Why should we reduce the variance?
- How do we reduce the variance?

0. Derivative Pricing and High-Dimensional Integrals

In mathematical finance and financial engineering, the prices of financial derivatives and Greeks can often be expressed as mathematical expectations under the risk-neutral measure.

- The value of a European financial derivative at time zero is

$$\mathrm{Price} = \mathbb{E}\,[g(\cdot)],$$

  $g(\cdot)$ — discounted payoff. For European arithmetic Asian option:

$$\mathrm{Price} = \mathbb{E}\,[e^{-rT}\max(S_A - K, 0)],$$

  S_A — the arithmetic average of the underlying asset prices.

- The pathwise estimate for the *delta* of arithmetic Asian option:

$$\mathrm{Delta} = \mathbb{E}\left[e^{-rT}\mathbf{I}_{\{S_A > K\}}(\mathbf{S})\,\frac{S_A}{S_0}\right],$$

  $\mathbf{I}_{\{\cdot\}}(\mathbf{S})$ — an indicator function (Note: $\Delta = \frac{\partial \mathrm{Price}}{\partial S_0}$).

- The purpose of many stochastic simulations is to estimate the mathematical expectations of some cost functions.

0. Derivative Pricing and High-Dimensional Integrals

After suitable transformations, the mathematical expectations can be transformed into a high-dimensional integration over $[0,1]^d$:

$$I_d(f) = \int_{[0,1]^d} f(\mathbf{x})d\mathbf{x}.$$

Challenges:

- The dimension can be huge (hundreds or thousands)!
  Source of dimensionality:
  - Number of time steps in discretization;
  - Number of state variables (risk factors).
- The function can be discontinuous.

Only in rare cases do explicit solutions exist (Black-Scholes formula). In most cases we have to use numerical methods (PDE, Simulation).

0. Derivative Pricing and High-Dimensional Integrals

- Discounted Payoff: $g(S_{t_1}, \ldots, S_{t_d})$, where $S_{t_1}, \cdots, S_{t_d}$ are the prices of the asset at $t_j = j\Delta t, j = 1, \ldots, d, \Delta t = T/d$, T is expiration date.
- Black-Scholes model:

$$dS_t = rS_t dt + \sigma S_t dB_t.$$

- Based on risk-neutral valuation, the value of derivative at time 0 is

$$\mathbf{E}[g(S_{t_1}, \ldots, S_{t_d})].$$

- The analytical solution to the SDE:

$$S_t = S_0 \exp\left((r - \frac{\sigma^2}{2})t + \sigma B_t\right).$$

  Simulating stock prices reduces to simulating Brownian motion (BM).
- MC estimate $= \frac{1}{n}\sum_{k=1}^{n} g^{(k)}$, where $g^{(k)}$ is the discounted payoff of the k-th path.

0. Derivative Pricing and High-Dimensional Integrals

Let $(B_{t_1}, \ldots, B_{t_d})^T =: \mathbf{x}$. Then $\mathbf{x} \sim N(\mathbf{0}, \mathbf{C})$ with $C_{i,j} = \min(t_i, t_j) = \Delta t \min(i, j)$. Let $g(S_{t_1}, \ldots, S_{t_d}) =: G(\mathbf{x}), \mathbf{x} \sim N(\mathbf{0}, \mathbf{C})$. The derivative price can be written as a Gaussian integral:

$$V(G) = \mathbf{E}(G(\mathbf{x})) = \int_{\mathbb{R}^d} G(\mathbf{x})\, P(\mathbf{x}; N(\mathbf{0}, \mathbf{C})\, d\mathbf{x}. \tag{1}$$

By setting $\mathbf{x} = A\mathbf{z}$ with $AA^T = \mathbf{C}$, the Gaussian integral is transformed to

$$V(G) = \int_{\mathbb{R}^d} G(A\mathbf{z})\, P(\mathbf{z}; N(\mathbf{0}, \mathbf{I})\, d\mathbf{z}.$$

The change of variables $\mathbf{x} = A\mathbf{z}$ with $AA^T = \mathbf{C}$ is equivalent to a PGM of BM

$$(B_1, \ldots, B_d)^T = A\,(z_1, \ldots, z_d)^T, \quad (z_1, \ldots, z_d)^T \sim N(\mathbf{0}, I). \tag{2}$$

A key insight is that the matrix A can be arbitrary as long as $AA^T = \mathbf{C}$.

Construction of BM $\Longleftrightarrow$ Change of variables $\Longleftrightarrow$ Decomposition of matrix $\mathbf{C} = AA^T$.

0. Derivative Pricing and High-Dimensional Integrals

$$V(G) = \int_{R^d} G(AZ)\, P(Z; N(0, I))\, dZ$$

$$= \int_{R^d} G(AZ) \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2} Z^T Z} dZ$$

$$= \int_{(0,1)^d} G(A\Phi^{-1}(U))\, dU$$

where $\Phi^{-1}(U) = (\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_d))$, **and**

Φ **is the cdf of standard normal distribution.**

1. Expectation and Integration

- **One dimension**

  If X is a RV uniformly distributed on [0,1], then the expectation of a function f(X) is equal to its integral:

$$E[f(X)] = \int_0^1 f(x)dx =: I[f].$$

- **Generalization to high dimension**

  If X is a R.V. uniformly distributed on d-dimensional unit cube $[0,1]^d$, then

$$E[f(X)] = \int_{[0,1]^d} f(x)dx =: I[f].$$

- Thus, calculating expectations (e.g. option prices) is connected to numerical integration, often in very high dimensions.

2. Deterministic Quadrature

- For one-dimensional integration I[f], we may use quadrature:

$$I[f] = \int_0^1 f(x)dx \approx \sum_{j=0}^{m} w_j f(x_j) =: Q[f].$$

where w_j are the weights and x_j are the nodes.

- The commonly used ones are
  - Trapeziodal rule,
  - Simpson's rule,
  - Gaussian quadrature.

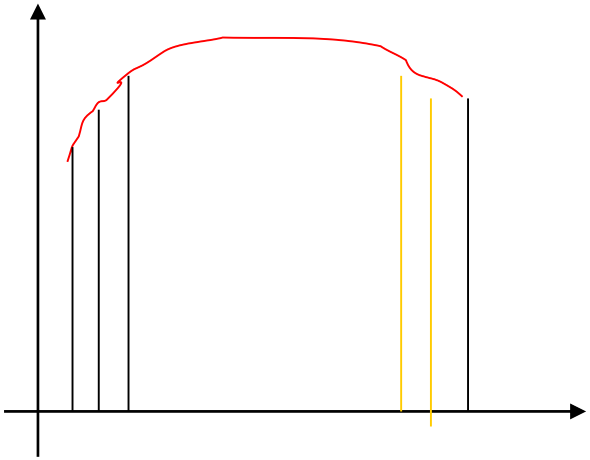
2. Deterministic Quadrature

- **Trapeziodal rule:**

$$\int_0^1 f(x)dx \approx \sum_{j=0}^{m} w_j f(\frac{j}{m}),$$

$$w_0 = w_m = 1/(2m), w_j = 1/m \ \ \textbf{for} \ \ j = 1,\ldots,m\text{-}1.$$

$$\textbf{Error}: O(m^{-2}), \ \ \textbf{if} \ \ f \in \mathbf{C}^2[0,1].$$



2. Deterministic Quadrature

For multi-dim integration, direct generalization of one-dim quadrature leads to **product rule**

Using one-dim rule

$$I[f] = \int_{[0,1]^d} f(x)dx \approx \sum_{i_1=1}^{m_1} w_{i_1}^1 \underline{\int_{[0,1]^{d-1}} f(x_{i_1}^1, x_2, ..., x_d)dx_2 \cdots dx_d}$$

$$\approx \sum_{i_1=1}^{m_1} \cdots \sum_{i_d=1}^{m_d} w_{i_1}^1 \dots w_{i_d}^d f(x_{i_1}^1, ..., x_{i_d}^d) =: Q[f].$$

- **Caution:**

  A product rule builds nodes taking the Cartesian product of node sets along each dimension. The regular grid is going to be impractical for large dimension (say, for d > 10).

2. Deterministic Quadrature

- Doubling the number of points in the one-dim integration rule multiplies the number of function evaluations in the product rule by 2^d.
- Calculation cost increases exponentially in dimension for required accuracy.
- This is known as curse of dimensionality.

Ex: Product Trapeziodal Rule

$$\int_{[0,1]^d} f(x)dx \approx \sum_{i_1=0}^{m} \cdots \sum_{i_d=0}^{m} w_{i_1} \ldots w_{i_d} f(\frac{i_1}{m},\ldots,\frac{i_d}{m}).$$

Total number of points : $N = (m+1)^d$.

Error : $O(m^{-2}) \propto O(N^{-2/d})$.

Let $N^{-2/d} = \varepsilon$ **(given accuracy), then** $\mathbf{N} = \varepsilon^{-d/2}$.

The number of points required to achieve a given precision grows exponentially with dimension d

--- the curse of dimensionality.

- This motivates the Monte Carlo Integration.

3. Monte Carlo Integration

MC approximates the expectation/integral E[f(X)] by

$$Q_N[f] = \frac{1}{N}\sum_{i=1}^{N} f(x_i),\ x_i \sim U(0,1)^d .$$

Key features:

MC points always have clusters and gaps.
What is random is where they appear.

- Unbiasedness

$$E[Q_N(f)] = I(f).$$

- Convergence (by the strong law of large numbers):

$$\lim_{N\to\infty} Q_N[f] = I[f]\ \ (\textbf{with probabilit y}\ 1).$$

3. Monte Carlo Integration

- **Error:** By Central Limit Theorem

$$\frac{I[f]-Q_N[f]}{\sigma/\sqrt{N}} \sim_d N(0,1),$$

$$\lim_{N\to\infty} \mathbf{Pr}\left(\frac{I[f]-Q_N[f]}{\sigma/\sqrt{N}} \le z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

$$\Rightarrow \mathbf{Pr}\left(\left|\frac{I[f]-Q_N[f]}{\sigma/\sqrt{N}}\right| \le 1.96\right) \approx 0.95$$

$\boxed{2.58 \qquad 0.99}$

where $\sigma^2 = \mathrm{Var}(f) = I(f^2) - [I(f)]^2$ is the variance.

Additional feature (easy to estimate error)

- The variance σ^2 can be estimated:

$$\overline{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(f(x_i) - Q_N[f]\right)^2.$$

Beware: in 5% of cases the estimate is outside the interval

- 95% (approximate) confidence interval:

$$\left(Q_N[f] - 1.96\frac{\overline{\sigma}}{\sqrt{N}}, \quad Q_N[f] + 1.96\frac{\overline{\sigma}}{\sqrt{N}} \right)$$

i.e.,

$$|\,\mathbf{Error}\,| < 1.96 N^{-1/2}\sigma \ \mathbf{with\ \ probability}\ 0.95$$

$$\mathbf{Put}\ \ 1.96 N^{-1/2}\overline{\sigma} = \varepsilon \Rightarrow N = \left(\frac{1.96\,\overline{\sigma}}{\varepsilon}\right)^2.$$

3. Monte Carlo Integration

- **RMSE** (Root-mean-square-error) $Q_N[f] = \frac{1}{N}\sum_{i=1}^{N} f(x_i),\ x_i \sim U(0,1)^d.$

$$\mathrm{RMSE} = \sqrt{\mathrm{E}\left[\mathrm{I(f)} - \mathrm{Q_N[f]}\right]^2}$$

$$= \sqrt{\mathrm{Var}(\mathrm{Q_N[f]})} = \sqrt{\sigma^2/\mathrm{N}}$$

--- for any dimension d.

MC error is $O(N^{-1/2})$, independent of dimension

--- breaks the curse of dimensionality.

However, MC does not deliver extreme accuracy.

Comparison with Product Rules

If the integrand if sufficiently smooth, then a product trapezoidal rule with

$$m_1 = m_2 = \ldots = m_d = m + 1 = N^{1/d}$$

points in each direction (with total number of nodes $N = (m+1)^d$) has

$$\textbf{Error} \propto \boldsymbol{m}^{-2} \propto \boldsymbol{N}^{-2/\boldsymbol{d}} \quad \textbf{(compare with O(N}^{\textbf{-1/2}}\textbf{) of MC)}$$

This scales better than MC for d<4, but worse for d>4.

MC is better at handling high dimensional problems (with low smoothness).

4. Variance Reduction Techniques

- Monte Carlo starts as a very simple method.
- Much of the complexity in MC practice comes from trying
  - to reduce the variance,
  - to reduce the number of samples that have to be simulated to achieve a given accuracy.

4. Variance Reduction Techniques

- Consider $$I[f] = \int_{[0,1]^d} f(x)dx.$$

$$\mathbf{RMSE} = \sqrt{E\big[I(f) - Q_N[f]\big]^2} = \sqrt{\frac{\sigma^2(f)}{N}}.$$

To reduce RMSE or reduce size of confidence interval

- **Increasing N; or**
- **Decreasing the variance.**

However, to cut RMSE by a factor of 10, we must raise N by a factor of 100.

Or twice as much accuracy requires 4 times as many samples.

4. Variance Reduction Techniques

- Original (Crude) MC:

$$E_p(f(x)) = \int_{(0,1)^d} f(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i) = Q_N[f],\ x_i \sim p = U(0,1)^d.$$

- Basic idea of VRT:

- Replacing the integrand f(·) by another one g(·);
- Changing the uniform distribution p by another probability distribution q

$$\text{keeping } E_q(g(x)) = E_p(f(x)),$$

$$\text{but with } Var_q(g) < Var_p(f)$$

(with the same expectation but with smaller variance)

Using MC to function g will give smaller RMSE.

4. Variance Reduction Techniques

RMSE of MC:

$$\mathbf{RMSE} = \sqrt{\boldsymbol{E}\left[\boldsymbol{I}(\boldsymbol{f}) - \boldsymbol{Q}_N[\boldsymbol{f}]\right]^2} = \sqrt{\sigma^2(\boldsymbol{f})/\boldsymbol{N}}.$$

If you can reduce variance by a factor of K, you may get the same accuracy by using only 1/K samples.

- Sometimes VRTs many reduces the variance by many thousand fold;
- Sometimes the reduction may be modest;
- Sometimes the variance may be even raised.

Measuring Efficiency of MC

Let the cost of each evaluation of f(·) is c.

The cost to get RMSE $\sqrt{\sigma^2/N} \le \varepsilon$ is: $cN = c\,\sigma^2/\varepsilon^2$.

The relative efficiency of new method to the old one:

$$\mathrm{EFF} := \frac{c_{old}\,\sigma^2_{old}}{c_{new}\,\sigma^2_{new}} = \frac{\sigma^2_{old}}{\sigma^2_{new}} \times \frac{c_{old}}{c_{new}}.$$

At any fixed level of accuracy, the old method takes EFF times as much as the new one.

We win by lowering the variance σ^2_{new}

(unless cost c_{new} goes up more).

We can even gain by raising the variance σ^2_{new}

if c_{new} goes down more.

How to judge one MC scheme is better than another?

The relative efficiency of new method to old one :

$$\mathbf{EFF} = \frac{c_{old}\,\sigma^2_{old}}{c_{new}\,\sigma^2_{new}} = \frac{\sigma^2_{old}}{\sigma^2_{new}} \times \frac{c_{old}}{c_{new}}$$

If two estimates require approximately the same computing time, then the smaller the variance, the larger the efficiency and the better the estimate.

Note:

To find more efficient estimates than crude MC, we need to find ways of getting q-fold reduction in variance, while restricting the increase in computing time to a factor no larger than q. There are a lot of tricks.

4. Variance Reduction Techniques

- Commonly used variance reduction techniques

① Antithetic variables
② Control variates
③ Stratified sampling
④ Conditional MC
⑤ Importance sampling
⑥ Latin hypercube sampling
⑦ ……

- We will focus on **How and Why** each technique reduces the variance.
- In MC practice, much of the complexity comes from trying to reduce the variance, to reduce the number of samples to achieve a given accuracy.

5. Antithetic Variables

If the vector $X=(X_{1,}...,X_d)$ is uniformly distributed on $[0,1]^d$, so is the vector $(1-X_{1,}...,1-X_d)$. Let

$$g(X_1,...,X_d)=\frac{1}{2}\left[f(X_1,...,X_d)+f(1-X_1,...,1-X_d)\right].$$

Then

$$\int_{[0,1]^d} f(x)dx=\int_{[0,1]^d} g(x)dx.$$

$$\textbf{Crude MC}:\quad Q_N[f]=\frac{1}{N}\sum_{i=1}^{N} f(x_i),\quad x_i\sim U(0,1)^d;$$

$$\textbf{AV Estimate}:\ Q_N^{AV}[f]=\frac{1}{N}\sum_{i=1}^{N}\frac{f(x_i)+f(1-x_i)}{2},\quad x_i\sim U(0,1)^d.$$

g(Xi)

5. Antithetic Variables

- The variance of g is given by

$$\operatorname{var}(\boldsymbol{g}) = \frac{1}{4}\left[\operatorname{var}(\boldsymbol{f}(\boldsymbol{x})) + 2\operatorname{cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(1-\boldsymbol{x})) + \operatorname{var}(\boldsymbol{f}(1-\boldsymbol{x}))\right]$$

$$= \frac{1}{2}\left[\operatorname{var}(\boldsymbol{f}(\boldsymbol{x})) + \underline{\operatorname{cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(1-\boldsymbol{x}))}\right] \le \operatorname{var}(\boldsymbol{f}).$$

--- The variance is always reduced!

- **However,** the cost of each function evaluation of g is doubled, thus we have net benefit only if the variance is reduced at least by a factor of 2, or

  cov (f(x), f(1-x)) < 0.

This depends on f

In this case: $\operatorname{var}(g) \le \frac{1}{2}\operatorname{var}(f).$

Two Extreme Cases

- If $f(x)$ is linear w.r.t. its variables, then the function $g(x)$ is a constant (i.e., the variance is reduced to zero), and is integrated exactly by MC. This is the optimal case. AV cancels some linear structure.
- If $f(x)$ is symmetric w.r.t. ½ (for each variable), then $f(x) = f(1-x)$. Then $var(g) = var(f)$.
  This is the worse case (the computing time is doubled, but no variance reduction is achieved).
- **General assessment**

AV is usually not very helpful (especially for symmetric functions), but can be good in some particular cases where the functions are nearly linear.

Theorem:

If f(x_1,…,x_d) is a monotone function of each of its arguments, then

$$\mathrm{cov}[f(X_1,\ldots,X_d), f(1-X_1,\ldots,1-X_d)] \le 0.$$

In this case, the variance of AV estimate is reduced by at least a factor of 2.

We win if $\mathrm{cov}[f(X_1,\ldots,X_d), f(1-X_1,\ldots,1-X_d)] \le 0.$

Note: X, $1-X$ **have the same distribution,**

but are perfectively negatively correlated.

When f is monotone, this negative correlation is preserved.

Lemma:

If $X_1, \ldots, X_d$ are independent, then for any increasing functions f_1 and f_2 of d variables,

$$E[f_1(X) f_2(X)] \geq E[f_1(X)]\, E[f_2(X)],$$

$$\textbf{or}\ \operatorname{cov}\big(f_1(X), f_2(X)\big) \geq 0, \quad X = (X_1, \ldots, X_d)$$

Proof. We prove the case for d=1. Since f_1 and f_2 are both increasing, we have

$$\big[f_1(x) - f_1(y)\big]\big[f_2(x) - f_2(y)\big] \geq 0, \ \forall x, y.$$

This implies that for any RV X, Y,

$$E\Big(\big[f_1(X) - f_1(Y)\big]\big[f_2(X) - f_2(Y)\big]\Big) \geq 0.$$

Proof (cont.)

or equivalently,

$$E[f_1(X)f_2(X)] + E[f_1(Y)f_2(Y)] \geq E[f_1(X)f_2(Y)] + E[f_1(Y)f_2(X)]$$

Now suppose X and Y are i.i.d., then

$$E[f_1(X)f_2(X)] = E[f_1(Y)f_2(Y)]$$

$$E[f_1(X)f_2(Y)] = E[f_1(Y)f_2(X)] = E[f_1(X)]E[f_2(X)].$$

We obtain the result for d=1:

$$E[f_1(X)f_2(X)] \geq E[f_1(X)]\,E[f_2(X)].$$

The proof is by induction on d for d>1. (please try)

Proof of the Theorem:

Assume (without loss of generality) that f is increasing in its first r variables and decreasing in its final d-r. Let

$$f_1(x_1,\ldots,x_d) = f(x_1,\ldots,x_r,1-x_{r+1},\ldots,1-x_d);$$

$$f_2(x_1,\ldots,x_d) = -f(1-x_1,\ldots,1-x_r,x_{r+1},\ldots,x_d).$$

Then f_1 and f_2 are both increasing. By the Lemma

$$\operatorname{cov}\big(f_1(X_1,\ldots,X_d), f_2(X_1,\ldots,X_d)\big) \geq 0.$$

or, equivalently,

$$\operatorname{cov}\big(f(X_1,\ldots,X_r,1-X_{r+1},\ldots,1-X_d), f(1-X_1,\ldots,1-X_r,X_{r+1},\ldots,X_d)\big) \leq 0.$$

implying that

$$\operatorname{cov}[f(X_1,\ldots,X_d), f(1-X_1,\ldots,1-X_d)] \leq 0.$$

(since the concerned random vectors have the same joint distributions)

Remark:

AV Estimate : $Q_N^{AV}[f] = \frac{1}{N}\sum_{i=1}^{N}\frac{f(x_i)+f(1-x_i)}{2},\ x_i \sim U(0,1)^d.$

Break f into "even" and "odd" parts (w.r.t. 1/2):

$$f(x) = \frac{f(x_i)+f(1-x_i)}{2} + \frac{f(x_i)-f(1-x_i)}{2} =: f_{\mathbf{even}}(x) + f_{\mathbf{odd}}(x)$$

Since $\int f_{\mathbf{even}}(x)\cdot f_{\mathbf{odd}}(x)dx = 0,$ **thus**

$$\mathrm{var}(f) = \mathrm{var}(f_{\mathbf{even}}) + \mathrm{var}(f_{\mathbf{odd}}).$$

So AV eliminates the variance contributi on of $f_{\mathbf{odd}}$,

and AV is extremely benefical for functions that are

primarily odd functions, having $\mathrm{var}(f_{\mathbf{odd}}) >> \mathrm{var}(f_{\mathbf{even}})$.

Example:

- Estimate $I[f]=E[f]=\int_0^1 e^x dx,\ f(x)=e^x.$

  Of course, we know that I[f] = e-1.
- **What is the improvement by using AV?**

  Var(f) = E[f^2] - (E[f])^2 = 0.2420.

  Let g = 1/2 [f(x) + f(1-x)].

  Var(g) =1/2 var(f) + 1/2 cov(f(x), f(1-x))=0.0039.
- **A large variance reduction is achieved!**

  (the variance reduction factor is about 62, and efficiency is increased by about a factor of 31).

Note: The integrand f(x) is increasing.

Applications of AV in Finance

In finance, to use AV, whenever you use a random vector $u \in [0,1]^d$ to simulate a stock path, use the vector $1-u \in [0,1]^d$ to simulate an antithetic path.

Multiple assets at multiple dates:

$$(U_1,\ldots,U_{nd}) \Rightarrow (Z_1,\ldots,Z_{nd}) \Rightarrow \begin{pmatrix} B_{t1}^1,\ldots,B_{tn}^1 \\ \ldots \\ B_{t1}^d,\ldots,B_{tn}^d \end{pmatrix} \Rightarrow (S_{t1}^1,\ldots,S_{t1}^d,\ldots S_{tn}^1,\ldots,S_{tn}^d)$$

Normal

Antithetic Path

$$(1-U_1,\ldots,1-U_{nd}) \Rightarrow (-Z_1,\ldots,-Z_{nd}) \Rightarrow \begin{pmatrix} \overline{B}_{t1}^1,\ldots,\overline{B}_{tn}^1 \\ \ldots \\ \overline{B}_{t1}^d,\ldots,\overline{B}_{tn}^d \end{pmatrix} \Rightarrow (\overline{S}_{t1}^1,\ldots,\overline{S}_{t1}^d,\ldots \overline{S}_{tn}^1,\ldots,\overline{S}_{tn}^d)$$

Example: European call

$$\mathbf{For\ i = 1\ to\ N}$$

$$\mathbf{generate\ a\ sample\ Z_i} \sim N(0,1);$$

$$\mathbf{set\ S_i^+} = S_0 \exp((r - \sigma^2/2)T + \sigma\sqrt{T}\,\mathbf{Z_i});$$

$$\mathbf{S_i^-} = S_0 \exp((r - \sigma^2/2)T - \sigma\sqrt{T}\,\mathbf{Z_i});$$

$$\mathbf{set\ V_i^+} = e^{-rT} \max(\mathbf{S_i^+} - K,\ 0);$$

$$\mathbf{V_i^-} = e^{-rT} \max(\mathbf{S_i^-} - K,\ 0);$$

$$\mathbf{set\ V_i} = (\mathbf{V_i^+} + \mathbf{V_i^-})/2;$$

$$\mathbf{end}$$

$$\mathbf{Set\ Price} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{i=1}^{N} \mathbf{V_i}.$$

Antithetic Path

5. Antithetic Variables

- **General assessment about AV:**
  1. Very simple to use.
  2. Usually not very useful.
  3. Can be good in particular case when the function is nearly linear.
  4. Can be used in QMC as well (latter).

- The best way to see if it helps is to do it.
- Partial antithetics, flipping just some components of x also works.

6. Control Variates

We want $\int f(x)\,dx$

We know $\int g(x)\,dx$

Some connection, e.g., $f(x) \approx g(x)$.

How to use the knowledge of $g(x)$?

6. Control Variates

If there is another integrand g which is similar with f, and for which we know the exact value of I[g], we may write

$$\int f(x)dx = \int f(x)\,dx - \left(\int g(x)dx - I[g]\right)$$

$$= \int \big[f(x) - g(x)\big]dx + I[g].$$

We use MC for the first integral and obtain an estimator

$$Q_{\mathbf{diff}}[f] = \frac{1}{N}\sum_{i=1}^{N}\big[f(x_i) - g(x_i)\big] + I[g],\; x_i \sim U(0,1)^d.$$

6. Control Variates

More generally, if there is another integrand g which is similar with f, and for which we know the exact value of I[g], we may write

$$\int f(x)dx = \int f(x)\,dx - b\left(\int g(x)dx - I[g]\right)$$

$$= \int \left[f(x) - bg(x)\right]dx + b\,I[g], \qquad \textbf{(*)}$$

where b is a parameter (to be chosen).

We use MC for the first integral and obtain CV estimate:

$$Q_{\mathbf{CV}}[f] = \frac{1}{N}\sum_{i=1}^{N}\left[f(x_i) - bg(x_i)\right] + b\,I[g],\ x_i \sim U(0,1)^d.$$

$$\int f(x)dx = \int f(x)\,dx - b\left(\int g(x)dx - I[g]\right)$$
$$= \int \left[f(x) - bg(x)\right]dx + b\,I[g], \qquad (*)$$

6. Control Variates

- Clearly, CV estimate is unbiased (based on (*)). Its RMSE is determined by the variance of (f-bg)

$$\mathbf{Var}(f - bg) = \mathbf{Var}(f) - 2b\mathbf{Cov}(f, g) + b^2\mathbf{Var}(g).$$

- To minimize the variance, the optimal value for b is obtained by letting

$$\frac{d}{db}\mathbf{Var(f - bg)} = 2b\mathbf{Var(g)} - \mathbf{2Cov(f, g)} = \mathbf{0}$$

$$\Rightarrow b^* = \mathbf{Cov}(f, g) / \mathbf{Var(g)}.$$

- The resulting variance is

$$\mathbf{Var}(f - b^* g) = \mathbf{Var}(f)\,(1 - \rho^2).$$

(ρ is the correlation between f and g)

6. Control Variates

$$\frac{d}{db}\mathbf{Var(f - bg)} = 2b\mathbf{Var(g) - 2Cov(f, g) = 0}$$

$$\Rightarrow b^* = \mathbf{Cov}(f, g) / \mathbf{Var(g)}.$$

The optimal parameter b* can be estimated from data

$$b^* \approx \frac{\sum_{i=1}^{N}(f(x_i) - \overline{f})(g(x_i) - \overline{g})}{\sum_{i=1}^{N}(g(x_i) - \overline{g})^2},$$

$\overline{f}, \overline{g}$: **sample mean.**

$$\mathbf{Var}(f - b^* g) = \mathbf{Var}(f)(1-\rho^2).$$

6. Control Variates

The challenge

is to choose a good control variable g, such that

(1) The function g is well correlated with f , i.e. $|\rho| \approx 1$.
If $\rho = \pm 1$, then CV estimate has zero variance.

The stronger the correlation, the larger the improvement.

(2) The integral of g is known.

Remark:

The relative efficiency of CV to crude MC:

$$\mathrm{EFF} = \frac{\sigma_{old}^2}{\sigma_{cv}^2} \times \frac{c_{old}}{c_{cv}} = \frac{\mathrm{var}(f)}{\mathrm{var}(f)(1-\rho^2)} \times \frac{c_f}{(c_f + c_g)}$$

$$= \frac{1}{1-\rho^2} \times \frac{c_f}{(c_f + c_g)}.$$

If $(1-\rho^2)(c_f + c_g) < c_f$, then CV improves the efficiency.

When $c_f = c_g$, then we need

$$|\rho| > \sqrt{1/2} \approx 0.71$$

in order to benefit from CV.

6. Control Variates: Multiple Controls

- We want to estimate E[f(X)]=:E[Y].
- m controls: $Z_i = g_i(X)$ with known mean.
  Let $G = (Z_1, \ldots, Z_m)^T$, μ_G is its mean vector.
- CV estimate:

$$Q_{CV} = Y - b^T(G - \mu_G) = (Y - b^T G) - b^T \mu_G$$

It is unbiased and its variance is

$$\mathrm{var}(Q_{cv}) = \sigma^2 + b^T \Sigma_{GG} b - 2b^T \Sigma_{GY}$$

This is minimized when $b = b^* = \Sigma_{GG}^{-1} \Sigma_{GY}$.

Leading to a variance $\mathrm{var}(Q_{cv}^*) = \sigma^2 - \Sigma_{GY}^T \Sigma_{GG}^{-1} \Sigma_{GY}$

Example:

Consider the pricing of arithmetic Asian option under the Black-Scholes model. The payoff is

$$f_A = \max\left(0, \sum_{i=1}^{d} w_i\, S(t_i) - K\right), \sum_{i=1}^{d} w_i = 1.$$

There is no analytic expression for price of the option.

We may use geometric Asian option as control variable:

$$f_G = \max\left(0, \Pi_{i=1}^{d} S(t_i)^{w_i} - K\right).$$

Geometrical average has a lognormal distribution

The pricing of geometric Asian option is analytically tractable (try to get a formula using the BS model).

Example:

It is also possible to combine CV with AV

For $k = 1$ to N

Generate a sample path : $(S_0^{(k)}, S_{t_1}^{(k)}, \ldots, S_{t_d}^{(k)})$

Calculate $A_k = e^{-rT}\left(\sum_{i=1}^{d} w_i S_{t_i}^{(k)} - K\right)^+,$

$$G_k = e^{-rT}\left(\prod_{i=1}^{d}\left(S_{t_i}^{(k)}\right)^{w_i} - K\right)^+.$$

Set $\quad Y_k = A_k - b^*(G_k - P_G)$

end

$$\text{Price} = \frac{1}{N}\sum_{k=1}^{N} Y_k.$$

P_G is the price of the geometric Asian option, b* can be calculated as indicated above.

6. Control Variates

General assessment

- easy to implement
- can be very effective, depending on the applications
- but requires careful choice of control variate in each case

7. Stratified Sampling

Basic idea:

- To achieve a more regular sampling in the most important dimension.

- Consider

$$I[f] = \int_0^1 f(x)dx.$$

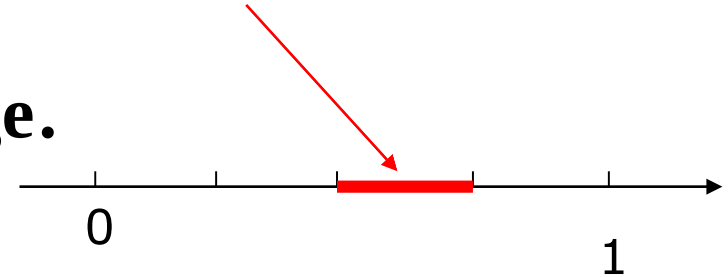
- Instead of taking N samples, drawn from uniform distribution on [0, 1], break the interval [0,1] into M strata of equal width and take L samples from each.

7. Stratified Sampling

- Break the interval [0,1] into M strata of equal width and take L samples from each.
- Let X_{ij} to be the ith sample from strata j, define

$$\overline{Y_j} = \frac{1}{L}\sum_{i=1}^{L} f(X_{ij}) = \textbf{average from strata } j,$$

$$\overline{Y}_{SS} = \frac{1}{M}\sum_{j=1}^{M} \overline{Y_j} = \textbf{overall average.}$$



The estimate is unbiased

Denote Y = f(X) and

$$\mu = E[f(X)], \quad \sigma^2 = \mathrm{var}[f(X)].$$

$$\mu_j = E[f(X) \mid X \in \textbf{strata } j],$$

$$\textbf{--- conditional expectation}$$

$$\sigma_j^2 = \mathrm{var}[f(X) \mid X \in \textbf{strata } j],$$

$$\textbf{--- conditional variance.}$$

Under the stratified sampling, the estimate $\overline{Y}_{SS}$ **is unbiased**, since

The estimate is unbiased

Under the stratified sampling, the estimate $\overline{Y}_{SS}$ **is unbiased**, since

$$E[\overline{Y}_{SS}] = \frac{1}{M}\sum_{j=1}^{M} E[\overline{Y}_j]$$

$$= \frac{1}{M}\sum_{j=1}^{M}\frac{1}{L}\sum_{i=1}^{L} E[f(X_{ij})] = \frac{1}{M}\sum_{j=1}^{M}\mu_j$$

$$= \sum_{j=1}^{M} E(f(X) \mid X \in \textbf{strata } j)\cdot P(X \in \textbf{strata } j) = \mu.$$

The expectation is the same for each i

全期望公式: Law of total probability

$\overline{Y}_j = \frac{1}{L}\sum_{i=1}^{L} f(X_{ij}) = \textbf{average from strata } j,$

The variance is reduced

$\overline{Y}_{SS} = \frac{1}{M}\sum_{j=1}^{M}\overline{Y}_j = \textbf{overall average.}$

The variance of $\overline{Y}_{SS}$ is (N=ML is total number of samples)

$$\mathrm{var}(\overline{Y}_{SS}) = \frac{1}{M^2}\sum_{j=1}^{M}\mathrm{var}(\overline{Y}_j) = \frac{1}{M^2 L}\sum_{j=1}^{M}\sigma_j^2 = \frac{1}{MN}\sum_{j=1}^{M}\sigma_j^2.$$

Without stratified sampling, $\mathrm{var}(\overline{Y}_{MC}) = \boldsymbol{\sigma}^2 / N$ with

$$\sigma^2 = E[f^2(X)] - \mu^2 = \sum_j E[f^2(X) \mid X \in \textbf{strata } j] \cdot \underbrace{P(X \in \textbf{strata } j)}_{1/M} - \mu^2$$

$$= \frac{1}{M}\sum_j\left(\mu_j^2 + \sigma_j^2\right) - \mu^2 = \frac{1}{M}\sum_j\left((\mu_j - \mu)^2 + \sigma_j^2\right) \geq \frac{1}{M}\sum_j \sigma_j^2.$$

$$\Rightarrow \frac{\sigma^2}{N} \geq \frac{1}{NM}\sum_j \sigma_j^2.$$

Thus stratified sampling **reduces the variance**.

More General Case

- Breaking the interval [0,1] into M strata, each strata has width p_j and take n_j samples from each strata

$$\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} f(X_{ij}), \quad \overline{Y}_{SS} = \sum_{j=1}^{M} \overline{Y}_j \, p_j$$

- Again, the estimate **is unbiased**, since

$$E[\overline{Y}_{SS}] = \sum_{j=1}^{M} p_j E[\overline{Y}_j] = \sum_{j=1}^{M} p_j \frac{1}{n_j}\sum_{i=1}^{n_j} E[f(X_{ij})] = \sum_{j=1}^{M} p_j \mu_j$$

$$= \sum_{j=1}^{M} E(f(X) \mid X \in \textbf{strata } j) \cdot \underline{P(X \in \textbf{strata } j)} = \mu.$$

$$\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} f(X_{ij}), \quad \overline{Y}_{SS} = \sum_{j=1}^{M} \overline{Y}_j \, p_j$$

Proportional Allocation

- The variance is given by

$$\operatorname{var}(\overline{Y}_{SS}) = \sum_{j=1}^{M} p_j^2 \operatorname{var}(\overline{Y}_j) = \sum_{j=1}^{M} \frac{p_j^2}{n_j}\sigma_j^2 = \frac{1}{N}\sum_{j=1}^{M} \frac{p_j^2}{q_j}\sigma_j^2, \quad q_j = \frac{n_j}{N}.$$

- **Proportional allocation** of samples:

$$\operatorname{var}(\overline{Y}_{PS}) = \frac{1}{N}\sum_{j=1}^{M} p_j \sigma_j^2, \quad (\textbf{put } n_j = p_j N, \ \ \textbf{i.e.,} q_j = p_j).$$

- By the **conditional variance formula** (Y = f(X))

$$\sum_j p_j \sigma_j^2 = E[\operatorname{var}(Y \mid X)] \le \operatorname{var}[Y] = \sigma^2. \quad (*)$$

Thus stratified sampling with proportional allocation reduces variance. (Devide both sides of (*) by N).

Optimal Allocation

- Consider $\min \ \mathrm{var}(\overline{Y}_{SS}) = \frac{1}{N}\sum_{j=1}^{M}\frac{p_j^2}{q_j}\sigma_j^2$

$$\textbf{s.t.} \sum_{j=1}^{M} q_j = 1, \ \ q_j \geq 0.$$

- This yields optimal allocation:

$$q_j^* = p_j\sigma_j \Big/ \sum_{i=1}^{M} p_i\sigma_i, \ \ j = 1,\ldots,M.$$

We should allocate more samples where the variance is larger (with high variability).

The resulting minimum variance:

$$\mathrm{var}(\overline{Y}_{\mathbf{OPT}}) = \frac{1}{N}\left(\sum_{j=1}^{M} p_j\sigma_j\right)^2 =: \frac{1}{N}\overline{\sigma}^2.$$

Variance Decomposition

$$\operatorname{var}(\overline{Y}_{MC}) = \frac{\sigma^2}{N} = \frac{1}{N}\left(\sum_j p_j\left(\mu_j^2 + \sigma_j^2\right) - \mu^2\right) = \frac{1}{N}\sum_j p_j\left((\mu_j - \mu)^2 + \underline{\sigma_j^2}\right)$$

$$\operatorname{var}(\overline{Y}_{PS}) = \frac{1}{N}\sum_j p_j \sigma_j^2 = \frac{1}{N}\left(\sum_j p_j\left(\sigma_j - \overline{\sigma}\right)^2 + \overline{\sigma}^2\right)$$

$$\text{since}: \sum_j p_j\left(\sigma_j - \overline{\sigma}\right)^2 = \sum_j p_j \sigma_j^2 - \overline{\sigma}^2.$$

$$\Rightarrow \operatorname{var}(\overline{Y}_{MC}) = \frac{1}{N}\left[\underline{\sum_j p_j(\mu_j - \mu)^2} + \underline{\sum_j p_j\left(\sigma_j - \overline{\sigma}\right)^2} + \overline{\sigma}^2\right].$$

- Proportional allocation removes the first term, the optimal allocation further removes the second term.

Summary of Stratified Sampling

- **Idea:**

- Breaking the range of integration into several pieces.
- Applying crude MC sampling to each piece separately.

- **Effect:**

If the stratification is well carried out (proportional allocation or optimal allocation), the variance will be smaller than the crude MC.

Ex. Stratified sampling for European call

$\mathbf{For\ j = 1,\dots, N}$

$\mathbf{Generate\ U_j \sim U(0,1)\ \ and\ compute}$

$$\overline{\mathbf{Z}}_j = \Phi^{-1}\left(\frac{j-1}{N} + \frac{\mathbf{U_j}}{N}\right)$$

Divide the unit interval into N little intervals, each of length 1/N, and generate randomly one point in each interval

$$\mathbf{S_T = S_0 \exp((r - \sigma^2/2)T + \sigma\sqrt{T}\,\overline{Z}_j},$$

$$\mathbf{f_j = e^{-rT}\max(S_T - K, 0)}$$

end

$$\mathbf{Price} = \frac{\mathbf{1}}{\mathbf{N}}\sum_{j=1}^{N}\mathbf{f_j}$$

Ex. Stratified sampling for European call

Test case: European call

$r = 0.05,\ \sigma = 0.5,\ T = 1,\ S_0 = 110,\ K = 100,\ N = 10^4$ samples

M	L	MC error bound
1	10000	1.39
10	1000	0.55
100	100	0.21
1000	10	0.07

Crude MC

Break the interval [0,1] into M strata of equal width and take L samples from each.

7. Stratified Sampling

Stratified Sampling in high dimension

- For a d-dim application, split each dimension of the [0, 1]d hypercube into M strata producing M^d sub-cubes.
- One generalization of stratified sampling is to generate L points in each of these hypercubes.
- However, the total number of points is LM^d which for large d would force M to be very small in practice.
- Instead, one may use a method called Latin Hypercube Sampling (latter).

8. Variance Reduction by Conditioning

- **We want to estimate** $\theta = E[X]$.

  Suppose there is another RV Y, such that

  $$E[X|Y] =: g(Y)$$

  is known (analytically tractable). We have

  $$E\,[E[X|Y]] = E[X],$$

  thus g(Y) is also an unbiased estimate of θ. Since

  $$Var(X) = E\,[Var\,(X|Y)] + Var\,(E\,[X|Y]),$$

  thus we have

  next slide

  $$Var\,(X) >= Var\,(g(Y)),$$

So using g(Y) **reduces the variance.**

Conditional Variance Formula

(*) $\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}(E[X|Y])$

Proof. $\mathrm{Var}(X|Y) = E[X^2|Y] - (E[X|Y])^2$.

Taking expectations of both sides:

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

Also,

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

Adding these two equations, we obtain (*).

Note: We have used $E[E[X^2|Y]] = E[X^2]$;
$E[E[X|Y]] = E[X]$.

8. Variance Reduction by Conditioning

- The final estimate for E[x] is

$$\frac{1}{N}\sum_{i=1}^{N} E[X \mid Y_i] = \frac{1}{N}\sum_{i=1}^{N} g(Y_i).$$

We actually sample Y, not X

Remark

To be able to reduce the variance via conditional MC, we must have another random variable Y that satisfying the following:

(1) Y can be easily simulated;

(2) $g(Y) := E[X|Y]$ can be

--- computed exactly (analytically) or

--- can be approximately with high accuracy.

Example:

Let $S_N = X_1 + \cdots + X_N$, $X_1, \ldots, X_N \sim N(0,1)$.

Calculate $E[I_{\{S_N > K\}}]$.

$$\begin{aligned}
E[I_{\{S_N > K\}}] &= E[I_{\{X_1 + \cdots + X_N > K\}}] \\
&= E\{E[I_{\{X_1 + \cdots + X_N > K\}} \mid X_1, \ldots, X_{N-1}]\} \\
&= E\{E[I_{\{X_N > K - X_1 - \cdots - X_{N-1}\}} \mid X_1, \ldots, X_{N-1}]\} \\
&= E\{P[X_N > K - X_1 - \cdots - X_{N-1} \mid X_1, \ldots, X_{N-1}]\} \\
&= E\{1 - \Phi(K - X_1 - \cdots - X_{N-1})\}.
\end{aligned}$$

This smooths the integrand and saves one dimension, both aspects are valuable and crucial for QMC.

Remarks

The benefits of conditional MC

- Conditional reduces the variance;
- Saving one dimension;
- Smoothing the integrand
  (the resulting integrand is smoother).

The third one is crucial for QMC methods.

Ex. Forward Start Option

Forward start options are options that will start at some time in the future. Often structured so that strike price equals asset price at time T_1

$$0 \longrightarrow T1 \longrightarrow T2$$

Consider a forward start at the money European call option that starts at T_1 and matures at T_2. To value this option, consider an regular European call option that starts at 0 and matures at $T_2 - T_1$.

From B-S formula : $c = S_0 N(d_1) - S_0 e^{-r(T_2 - T_1)} N(d_2) \quad (S_0 = K)$

$\Rightarrow$ the price c is proportional to the strike price S_0.

d1 and d2 do not depend on S0

The value of the forward start option at T_1 is cS_1/S_0.

Using risk-neutral valuation, its value at T_0 is

$$e^{-rT_1} E\left(c\, S_1/S_0\right) = e^{-rT_1} E\left(c\, S_1/S_0\right) = c \qquad (\text{Note}: E(S_1) = S_0 e^{-rT_1})$$

Ex. Forward Start Option

Crude MC: For $j = 1,\ldots, N$

Generate $Z_1 \sim N(0,1)$ and compute

$$S_{T_1} = S_0 \exp((r - \sigma^2/2)T_1 + \sigma\sqrt{T_1}\, Z_1,$$

$$\text{Set } K = \lambda S_{T_1} + 1$$

The strike is set as a factor of the level of stock at that time plus 1.

Generate $Z_2 \sim N(0,1)$ and compute

$$S_T = S_{T_1} \exp((r - \sigma^2/2)(T - T_1) + \sigma\sqrt{T - T_1}\, Z_2,$$

discount payoff $f_j = e^{-rT} \max(S_T - K, 0)$,

end

$$\text{Price} = \frac{1}{N}\sum_{j=1}^{N} f_j$$

Ex. Forward Start Option

Conditional MC:

For $j = 1, \ldots, N$

Generate $Z_1 \sim N(0,1)$ and compute

$$S_{T_1} = S_0 \exp((r - \sigma^2/2)T_1 + \sigma\sqrt{T_1}\, Z_1,$$

$$\text{Set } K = \lambda\, S_{T_1} + 1$$

$$f_j = e^{-rT_1} * \underline{\text{BSFormula}(\, S_{T_1}, K)}$$

end

$$\text{Price} = \frac{1}{N}\sum_{j=1}^{N} f_j$$

Conditional on the level of stock at T1, we can price the option from T1 to T using BS formula (analytically tractable).

9. Importance Sampling (IS)

- Importance sampling is more complicated than other variance reduction methods.
- Done well, it can turn a problem from intractable to easy.
- It can also give infinite variance.

9. Importance Sampling (IS)

Consider $\mu = E_p[f(X)] = \int_A f(x)p(x)dx$,

where $f(x) = 0$ outside of the region A and $\Pr(x \in A)$ is tiny.

Examples :

rare event, small probability, deep out-of-money option, probability of network failture, ...

Probability that an insurance company fails ...

The idea :

Arrange for $x \in A$ to happen more often, then adjust for bias.

9. Importance Sampling

Consider the problem of estimating

$$\mu = E_p[f(X)] = \int f(x)p(x)dx.$$

A crude MC estimate is

$$Q_N = \frac{1}{N}\sum_{i=1}^{N} f(x_i), \quad x_i \sim p(x).$$

We may write

$$\mu = E_p[f(X)] = \int f(x)p(x)dx = \int \frac{f(x)p(x)}{q(x)} q(x)dx = E_q\left[\frac{f(x)p(x)}{q(x)}\right],$$

where q(x) is another density, such that

$$p(x) > 0 \Rightarrow q(x) > 0.$$

9. Importance Sampling

Thus the expectation $\mu = E_p[f(X)]$ can also be estimated by

$$Q_N^{IS} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)\frac{p(x_i)}{q(x_i)}, \; x_i \sim q(x).$$

The weight p(x)/q(x) is the **likelihood ratio or Randon-Nikodym derivative**.

Clearly, this estimate is **unbiased.**

Is the variance reduced?

9. Importance Sampling

The variances (without or with IS) are, respectively,

$$\mathrm{var}_{\boldsymbol{p}}(\boldsymbol{f}(\boldsymbol{x})) = \int \boldsymbol{f}^2(\boldsymbol{x})\boldsymbol{p}(\boldsymbol{x})\boldsymbol{dx} - \mu^2.$$

$$\mathrm{var}_{\boldsymbol{q}}\left(\frac{\boldsymbol{f}(\boldsymbol{x})\boldsymbol{p}(\boldsymbol{x})}{\boldsymbol{q}(\boldsymbol{x})}\right) = \int \left(\frac{\boldsymbol{f}(\boldsymbol{x})\boldsymbol{p}(\boldsymbol{x})}{\boldsymbol{q}(\boldsymbol{x})}\right)^2 \boldsymbol{q}(\boldsymbol{x})\boldsymbol{dx} - \mu^2.$$

- The second one can be **smaller or larger** than the first one, depending on the choice of q(x).
- We should determine an IS distribution that minimizes the IS estimate.
- Successful IS lies in the art of selecting a good IS density q(x).

An Ideal Case

- Suppose f(x) is **nonnegative**, we choose q(x) as

$$q(x) = \frac{1}{\mu} f(x) p(x).$$

Indeed, q(x) is a density. Moreover, with this choice,

$$\operatorname{var}_q\left(\frac{f(x)p(x)}{q(x)}\right) = \int \left(\frac{f(x)p(x)}{q(x)}\right)^2 q(x)dx - \mu^2 = 0.$$

That is, we obtain **zero-variance** estimate:

$$Q_N^{IS} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)\frac{p(x_i)}{q(x_i)} = \mu,\ x_i \sim q(x).$$

9. Importance Sampling

- However, this optimal choice of density

$$q(x) = \frac{1}{\mu} f(x) p(x)$$

requires the knowledge of μ .
Thus it is useless in practice (since μ is unknown).

- Nevertheless, this optimal choice does provide **useful guidance:**

Try to find a density q(x), which mimics the behavior of the product of f(x) and p(x).

9. Importance Sampling

The integrand f(x) is not necessarily nonnegative.

Theorem:

$$\min_{\boldsymbol{q}}\left\{\operatorname{var}_{\boldsymbol{q}}\left(\frac{\boldsymbol{f}(\boldsymbol{x})\boldsymbol{p}(\boldsymbol{x})}{\boldsymbol{q}(\boldsymbol{x})}\right)\right\}=\left(\int|\boldsymbol{f}(\boldsymbol{x})|\boldsymbol{p}(\boldsymbol{x})\boldsymbol{dx}\right)^2-\mu^2.$$

which occurs when

$$q(x)=\frac{|f(x)|\,p(x)}{\int|f(x)|\,p(x)dx}. \qquad (*)$$

Proof. $\operatorname{var}_q\left(\frac{f(x)p(x)}{q(x)}\right) = E_q\left(\frac{f(x)p(x)}{q(x)}\right)^2 - \mu^2$

$$\geq \left(E_q\left[\frac{|f(x)|p(x)}{q(x)}\right]\right)^2 - \mu^2 \text{ (Jensen's inequality)}$$

$$= \left(\int |f(x)|p(x)dx\right)^2 - \mu^2.$$

When q(x) is chosen as in (*),

$$\boxed{q(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx}. \quad (*)}$$

$$E_q\left(\frac{f(x)p(x)}{q(x)}\right)^2 = \int \frac{f^2(x)p^2(x)}{q(x)}dx$$

$$= \int |f(x)|p(x)dx \cdot \int \frac{f^2(x)p^2(x)}{|f(x)|p(x)}dx = \left(\int |f(x)|p(x)dx\right)^2.$$

Remarks

(1) Finding a good importance sampler is an art and a science.

(2) Success is not assured.

(3) Small q (·) is problematic.

For safety, take q (·) heavier tailed than p (·).

$$Q_N^{IS} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)\frac{p(x_i)}{q(x_i)},\; x_i \sim q(x).$$

Rare Event and IS

Suppose we want to estimate

$$\theta = P(X \in A) = E_p[f(X)] = \int f(x)p(x)dx,$$

$$f(x) = 1_{\{X \in A\}}, \quad \{X \in A\} \textbf{ is a rare event.}$$

The integrand is zero almost everywhere.

Crude MC estimate is

$$Q_N = \frac{1}{N}\sum_{i=1}^{N} 1_{\{x_i \in A\}}, \; x_i \sim p(x).$$

Many samples will be **wasted**,
as the event {X in A} rarely occurs.

Rare Event and IS

- Use IS

$$Q_N^{IS} = \frac{1}{N}\sum_{i=1}^{N} 1_{\{x_i \in A\}} \frac{p(x_i)}{q(x_i)},\ x_i \sim q(x).$$

- The optimal IS density is

$$q(x) = \frac{1}{\theta} p(x), x \in A, \ \textbf{or} \ \ q(x) = \frac{1}{\theta} I_{\{x \in A\}} p(x).$$

- We should look for an IS density that approximates the optimal IS density, i.e., choosing q(x) which makes the event {X in A} more likely to happen (see example on deep-out-of money option).

Another Use of IS

Crude MC estimate is

$$\mu = E_p[f(X)] = \int f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i),\ x_i \sim p(x).$$

The sampling from p(x) can be quite difficult. Find a density q, from which **it is easy** to sample and use

$$\mu = E_p[f(X)] = E_q\left[\frac{f(x)p(x)}{q(x)}\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)\underbrace{\frac{p(x_i)}{q(x_i)}}_{\text{校正因子}},\quad x_i \sim q(x).$$

We sample from a wrong distribution q(·), and **correct** them by using the likelihood ratio.

Remarks

- The same samples from q can be used for different p and f.
- The variance of the new estimate **can be larger** than the original one.
- Here IS method is not used as a variance reduction device, but as a sampling technique.

9. Importance Sampling

- **Importance-Sampling**-based estimate

$$Q_N^{IS} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)w(x_i), \quad w(x_i) = \frac{p(x_i)}{q(x_i)}, \; x_i \sim q(x).$$

- A modified estimate (**weighted estimate**):

$$Q_N^{W} = \frac{\cancel{1/N}\sum_{i=1}^{N} f(x_i)w(x_i)}{\cancel{1/N}\sum_{i=1}^{N} w(x_i)}, \; x_i \sim q(x).$$

9. Importance Sampling

- When the integration domain is [0,1]^d, we may choose q(x) to be the uniform distribution. This leads to **Weighted Uniform Sampling**:

$$Q_N^{WUS} = \frac{\sum_{i=1}^{N} f(x_i) p(x_i)}{\sum_{i=1}^{N} p(x_i)}, \quad x_i \sim U[0,1]^d.$$

Remarks

- The weighted estimate and weighted uniform sampling estimate are slightly biased, but we find them often have smaller mean squared error.
- In weighted estimate, we only need to know the ratio $w(x) = p(x)/q(x)$ up to a multiplicative constant (this is important in Bayesian statistics, where the posterior distribution is known in shape, not the normalizing constant).

- In summary, there are two reasons to use IS:
  - Variance reduction;
  - Direct sampling from the target density is difficult.

Ex. Compute Probability

$$E\{I_{\{Z>C\}}\} \quad (\text{given } Z \sim N(0,1),\ c \text{ is a constant, say } c = 8)$$

$$= \int_{-\infty}^{+\infty} I_{\{z>C\}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Choose another normal as IS density

$$= \int_{-\infty}^{+\infty} I_{\{z>C\}} \frac{\frac{1}{\sqrt{2\pi}} e^{-z^2/2}}{\frac{1}{\sqrt{2\pi}} e^{-(z-\mu)^2/2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-(z-\mu)^2/2} dz$$

Likelihhod Ratio

$$= \int_{-\infty}^{+\infty} I_{\{z>C\}} \frac{e^{-z^2/2}}{e^{-(z-\mu)^2/2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-(z-\mu)^2/2} dz$$

$$= E_\mu \left\{ I_{\{Y>C\}} e^{\mu^2/2-\mu Y} \right\}, \quad \text{where} \quad Y \sim N(\mu,1).$$

Question:

How to choose the drift to minimize the variance of the new estimate?

The idea : The optimal density

$$q(x) = \frac{1}{\theta} p(x), x \in A, \text{ or } q(x) = \frac{1}{\theta} I_{\{x \in A\}} p(x).$$

Choose Normal as IS density, with mean

$$\mu = \arg\max_x I_{\{x \in A\}} p(x)$$

$$= \arg\max_x I_{\{x>8\}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$= \arg\max_{x>8} e^{-x^2/2} \quad = 8.$$

Mode matching:

The mode of q(·) coincides with that of $I_{\{x>8\}}$ p(x)

Ex. Deep out-of-money option

Consider a **deep out-of-money** call option ($S_0 << K$). Under the risk-neutral measure,

$$S_T = S_0 e^Z, Z \sim N\left(\left(r - \frac{\sigma^2}{2}\right)T, \sigma\sqrt{T}\right) =: N(\alpha, b^2).$$

The option price is approximated by **crude MC:**

$$\textbf{Price} = e^{-rT} E\left(\max(0, S_T - K)\right)$$

$$= e^{-rT} \int \max(0, S_0 e^z - K) p(z) dz$$

$$\approx e^{-rT} \frac{1}{N} \sum_{i=1}^{N} \max(0, S_0 e^{Z_i} - K), Z_i \sim N(\boldsymbol{\alpha}, b^2).$$

IS for deep-out-of-money option

- In crude MC, most function values are zeros for deep out-of-money options ($S_0 << K$).
- Under the risk-neutral measure, the expected stock value at T is $S_0 e^{rT}$ (usually much less than K).
- To use IS, we choose a drift in order to increase the probability that the payoff is positive. One possible is such that expected value of S_T to be K:

$$S_0 e^{\mu T} = K \Rightarrow \mu = \frac{1}{T}\log\left(\frac{K}{S_0}\right).$$

- Let $Y \sim N\left(\left(\mu - \frac{\sigma^2}{2}\right)T, \sigma\sqrt{T}\right) =: N(\beta, b^2).$

IS for deep-out-of-money option

Let p(z) be the density of Z, i.e., $N(\alpha, b^2)$

Let q(z) be the density of Y, i.e., $N(\beta, b^2)$

$$\textbf{Price} = e^{-rT} \int \max(0, S_0 e^z - K) p(z) dz$$

$$= e^{-rT} \int \max(0, S_0 e^z - K) \frac{p(z)}{q(z)} q(z) dz$$

$$\approx e^{-rT} \frac{1}{N} \sum_{i=1}^{N} \max(0, S_0 e^{Y_i} - K) \frac{p(Y_i)}{q(Y_i)},$$

$$Y_i \sim N(\boldsymbol{\beta}, b^2).$$

Example (Glasserman et al)

- In Black-Scholes model, the price of a derivative can often be written as a Gaussian integral:

$$\text{Price} = \int_{R^d} G(z)\, p(z)\, dz,$$

$$\text{where } p(z) = (2\pi)^{-d/2} \exp(-\frac{1}{2} z^T z),$$

and G(·) is related to the discounted payoff.

- Crude MC estimate for the price is

$$Q_N = \frac{1}{N} \sum_{i=1}^{N} G(z_i), \quad z_i \sim p(z).$$

Example (Glasserman et al)

Since $$\text{Price} = \int_{R^d} G(z)\, p(z)\, dz = \int_{R^d} G(z) \frac{p(z)}{q(z)} q(z)\, dz,$$

thus we have an **IS-based estimate:**

$$Q_N^{IS} = \frac{1}{N} \sum_{i=1}^{N} G(z_i) \frac{p(z_i)}{q(z_i)},\ z_i \sim q(z).$$

The optimal (zero-variance) density

$$q(z) = \frac{1}{\text{price}} G(z)\, p(z).$$

The price is unknown.

Thus the optimal density is impractical.

A Solution

Restricting IS density in class of multivariate normal

$$N(\boldsymbol{\alpha}, I_d) \textbf{ for some } \alpha \in R^d.$$

And choosing the shift α to be the solution (denoted by z*) of the optimization problem

$$\max_{z \in R^d} \{G(z) p(z)\}, \text{ or equivalently, } (*)$$

Mode matching

$$\max_{z \in R^d} \{F(z) - \frac{1}{2} z^T z\}, F(z) := \log G(z).$$

Log payoff

Interpretation

- We approximate the optimal (zero-variance) density

$$q(z) = \frac{1}{\text{price}} G(z)\, p(z)$$

by a normal density $N(\alpha, I_d)$ whose **mode coincides** with that of the optimal density, which occurs at the solution to

$$\max_{z \in R^d} \{G(z)\, p(z)\}, \text{ or equivalently}: \qquad (*)$$

$$\max_{z \in R^d} \{F(z) - \frac{1}{2} z^T z\}, F(z) := \log G(z).$$

Mode matching

Alternative Interpretation

Consider

integrand

$$G(z)\frac{p(z)}{q(z)} = G(z)\exp(-\alpha^T z + \alpha^T \alpha / 2),\ z \sim N(\alpha, I_d).$$

or $\underline{\exp(F(\alpha + z) - \alpha^T z - \alpha^T \alpha / 2)},\ z \sim N(0, I_d).$

Note that the α found by solving (*) satisfies the first-order condition (under some condition)

$$\nabla F(\alpha) = \alpha^T,$$

with ∇F **denoting the gradient of F.**

Alternative Interpretation

Using $F(\alpha + z) \approx F(\alpha) + \nabla F(\alpha) z$, **we have**

$$\exp(F(\alpha + z) - \alpha^T z - \alpha^T \alpha / 2)$$

$$\approx \exp(F(\alpha) + \underline{\nabla F(\alpha) z - \alpha^T z} - \alpha^T \alpha / 2)$$

$$= \exp(F(\alpha) - \alpha^T \alpha / 2),$$

which is constant and has zero variance.

Thus if the log payoff F is close to linear, the choice of density can be expected to eliminate much of the variance!

Questions

- Can we choose the importance density q(z) in the class of more general multivariate normal densities $N(\alpha, \Sigma)$?

- If yes, how to choose good parameters α, Σ ?
  If no, why?

Ex: Binary Option

Under the BS model, consider the price of a binary call option with payoff

$$\text{Payoff} = I\{S_T \geq K\}.$$

$$\text{Note that} \quad S_T = S_0 \exp((r - \frac{1}{2}\sigma^2)T + \sigma B_T).$$

The plain MC scheme is straightforward.

Problem:

When K is large, the probability that $S_T < K$ can be very low, only a few of the samples will reach the strike price K (this is a rare event), which leads to a poor estimate.

Ex: Binary Option

The price can be written as

$$\text{Price} = E[e^{-rT} I\{S_T \geq K\}].$$

Since

$$S_T = S_0 \exp((r - \frac{1}{2}\sigma^2)T + \sigma\sqrt{T}\,Z),\ Z \sim N(0,1),$$

thus

$$\text{Price} = E[e^{-rT} \underline{I\{Z \geq b\}}] =: E[G(Z)], \quad \text{where}$$

$$b = \frac{\log(K/S_0) - (r - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} \quad \underline{(\text{If } S_0 << K, \text{then } b >> 0)}.$$

Ex: Binary Option

Let $p(z)$ be the density of $N(0,1)$, then

$$G(z)p(z) = \begin{cases} 0, & \text{if } z < b, \\ e^{-rT}p(z), & \text{if } z \geq b. \end{cases}$$

The mode of $G(z)p(z)$ is

$$z^* = \max(b,0).$$

The method of mode matching suggests

$N(z^*, 1)$ as the IS density, $q(z)$.

Ex: Binary Option

IS-based estimate

$$\text{Price} = E[e^{-rT} I\{Z \geq b\}] = E[G(Z)]$$

$$= \int_{-\infty}^{+\infty} e^{-rT} I\{z \geq b\} p(z) dz$$

$$= \int_{-\infty}^{+\infty} e^{-rT} I\{z \geq b\} \frac{p(z)}{q(z)} q(z) dz$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} e^{-rT} I\{y_i \geq b\} \frac{p(y_i)}{q(y_i)}, \quad y_i \sim q(y)$$

i.e., $N(z^*, 1)$

Remarks

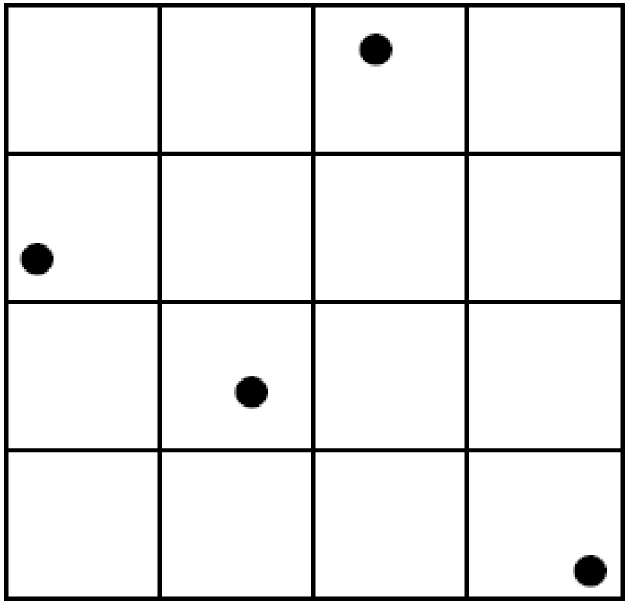
- The IS method tries to reduce the variance of a MC estimate by changing the probability measure from which paths are generated.

- IS method has the highest potential of variance reduction, sometimes even by orders of magnitude.

- However, if IS distribution is not chosen carefully, it can also increase the variance and can even produce infinite variance.

10. Latin Hypercube Sampling

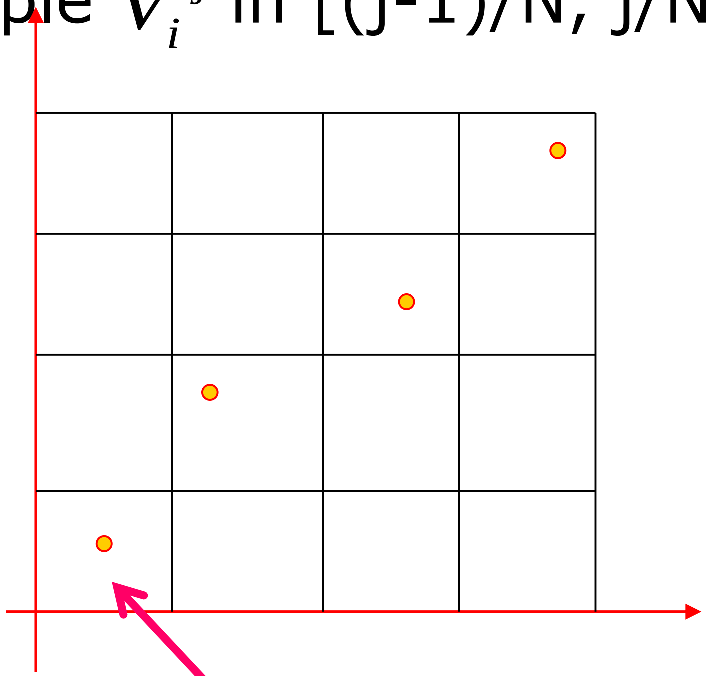
- Generate N points, dimension-by-dimension, using 1D stratified sampling with 1 value per stratum, assigning them randomly to the N points to give precisely one point in each stratum.



10. Latin Hypercube Sampling

- For each dimension i=1,…,d, break the interval [0,1] into N strata: [(j-1)/N, j/N), j=1, …, N.
- Independently generate a sample V_i^j in [(j-1)/N, j/N) We obtain

$$
\begin{array}{cccc}
V_1^1 & V_2^1 & \dots & V_d^1 \\
V_1^2 & V_2^2 & \dots & V_d^2 \\
\dots & \dots & \dots & \dots \\
V_1^N & V_2^N & \dots & V_d^N
\end{array}
$$



- Each row gives the coordinates of a points in [0,1)^d. For ex., the first row identifies a point in [0,1/N)^d

Permutation

- For **each column**, random permute the entries:

$$
\begin{matrix}
\boxed{\begin{matrix} V_1^1 \\ V_1^2 \\ \cdots \\ V_1^N \end{matrix}} & \begin{matrix} V_2^1 \\ V_2^2 \\ \cdots \\ V_2^N \end{matrix} & \begin{matrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{matrix} & \begin{matrix} V_d^1 \\ V_d^2 \\ \cdots \\ V_d^N \end{matrix}
\end{matrix}
\longrightarrow
\begin{matrix}
\boxed{\begin{matrix} V_1^{\pi_1(1)} \\ V_1^{\pi_1(2)} \\ \cdots \\ V_1^{\pi_1(N)} \end{matrix}} & \begin{matrix} V_2^{\pi_2(1)} \\ V_2^{\pi_2(2)} \\ \cdots \\ V_2^{\pi_2(N)} \end{matrix} & \begin{matrix} \cdots \\ \cdots \\ \cdots \\ \cdots \end{matrix} & \begin{matrix} V_d^{\pi_d(1)} \\ V_d^{\pi_d(2)} \\ \cdots \\ V_d^{\pi_d(N)} \end{matrix}
\end{matrix}
$$

where $\pi_1, \ldots, \pi_d$ are independent permutations of {1, …, N}. $\pi(i)$ denotes the value to which *i* is mapped by the permutation π.

- Each row gives the coordinates of a points in [0,1)^d.

Important property: The projections of these points on each dimension are the same as before.
(the numbers in each column remain the same)

Definition of LHS

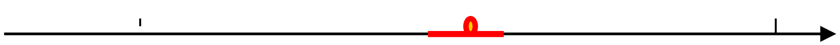
- Mathematically, we define

$$V_i^j = \frac{\pi_i(j) - 1 + U_i^j}{N}, i = 1, \ldots, d;\ j = 1, \ldots, N,$$

where $\pi_1, \ldots, \pi_d$ are independent permutations of 1, …, N; U_i^j are i.i.d. U[0,1) . Then the point set

$$\left\{ \left(V_1^j, \ldots, V_d^j \right),\ j = 1, \ldots, N \right\}$$

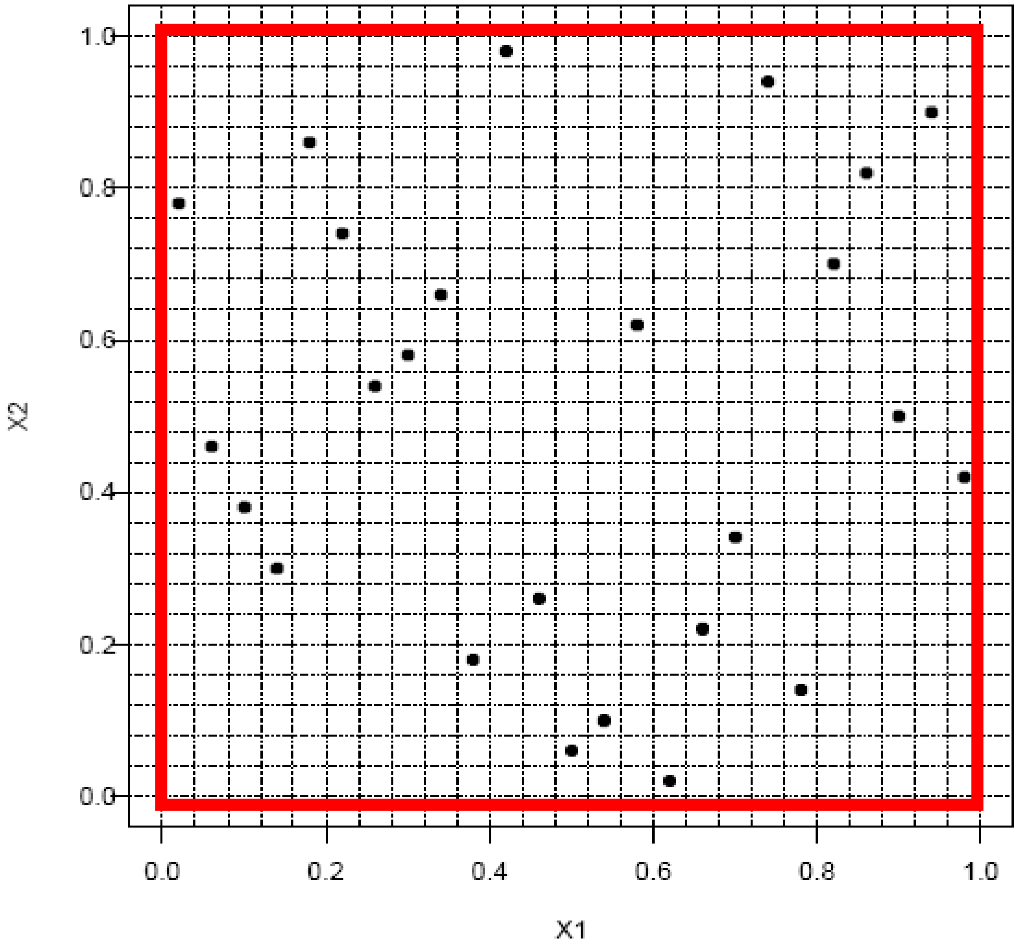
is a **LHS** set.



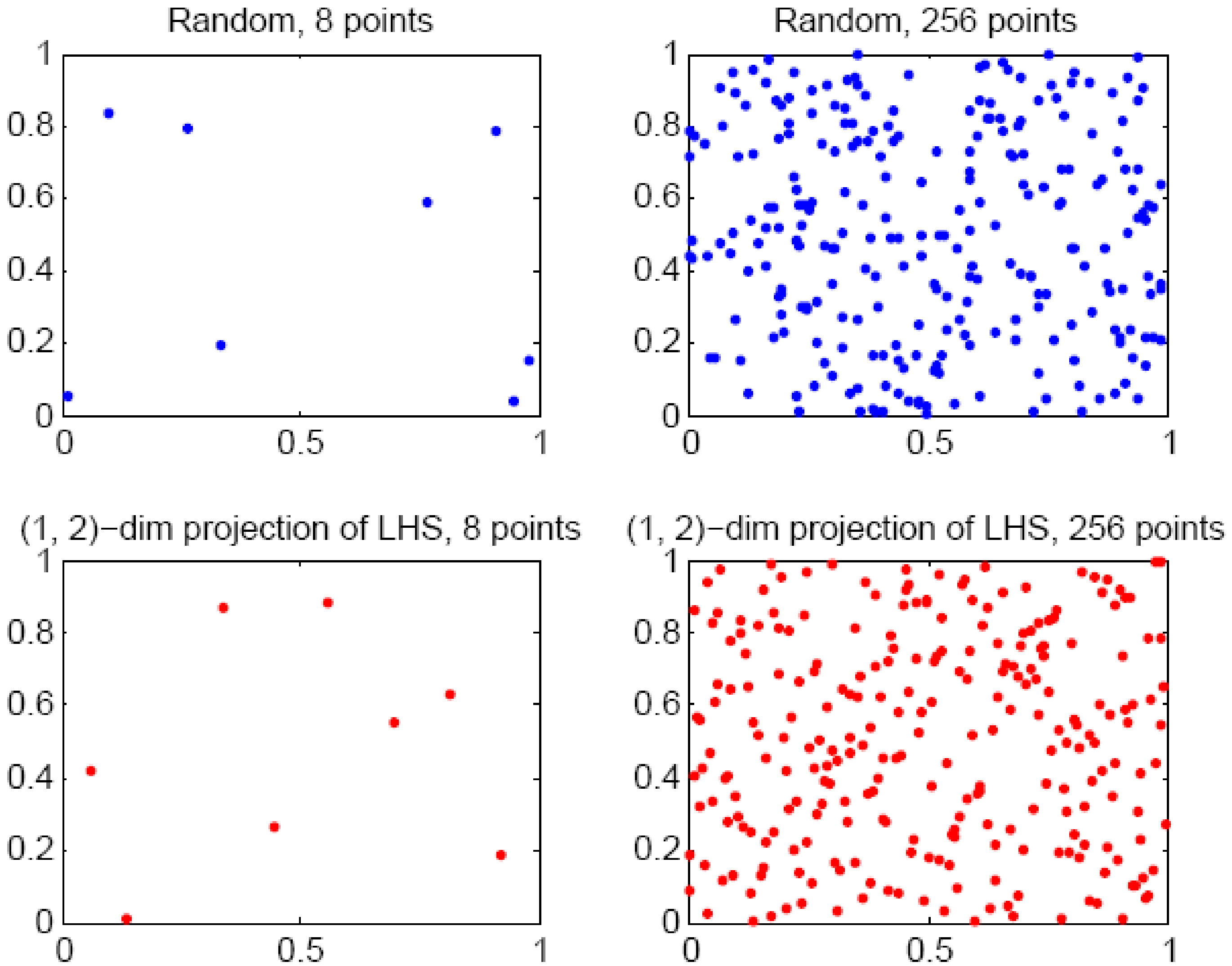
- In each dimension *i*,
  - **randomly choose a strata,**
  - **then randomly choose a point in the strata.**

10. Latin Hypercube Sampling

One point per row, one per column



Random points vs LHS points



10. Latin Hypercube Sampling

- If we use the LHS set to estimate an integral:

$$\int_{[0,1)^d} f(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(V^j), V^j = (V_1^j,\ldots,V_d^j),$$

Since each of the points is uniformly distributed over the hypercube,

$$E\left[\frac{1}{N}\sum_{i=1}^{N} f(V^j)\right] = \int_{[0,1)^d} f(x)dx.$$

- The fact that the points are not independently generated does not affect the expectation, only the variance (Reduced?)

10. Latin Hypercube Sampling

$$\int_{[0,1)^d} f(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(V^j), V^j = (V_1^j,\ldots,V_d^j),$$

The variance of the LHS can be much smaller than that of crude MC.

We can prove that LHS eliminates the variance due to the **additive part** of function f(x) in **ANOVA decomposition** (see next chapter).

Thus LHS is most effective for integrands that nearly separate into a sum of univariate functions.

10. Latin Hypercube Sampling

LHS results

1st Never much worse than Monte Carlo (Owen)

$$V_{LHS}(\hat{I}) \leq \frac{n}{n-1} V_{MC}(\hat{I})$$

2nd Additive part of f removed from error (Stein)

$$\begin{aligned} V_{LHS}(\hat{I}) &\doteq \frac{1}{n}\sigma^2(f - f_{\text{Add}}) \\ &= \frac{1}{n}\left(\sigma^2(f) - \sigma^2(f_{\text{Add}})\right) \end{aligned}$$

10. Latin Hypercube Sampling

- Latin Hypercube is very effective when function can be decomposed into a sum of 1D functions.
- Hard to predict which variance reduction approach will be most effective.

- **Advice:**

When facing a new class of applications, try each one, and don't forget we can sometimes combine different techniques (e.g. stratified sampling with antithetic variables, or Latin Hypercube with importance sampling).

11. Common Random Numbers

Suppose that X and Y are two RVs. Our goal is to compare the means E[f(X)] and E[g(Y)], where f and g are two functions. We may estimate the difference

$$E\,[f(X)] - E[g(Y)].$$

Crude MC:

$$Q_{CMC} = \frac{1}{N}\sum_{i=1}^{N} f(x_i) - \frac{1}{N}\sum_{i=1}^{N} g(y_i);$$

where x_i and y_i are simulated independently.

$$\text{Variance}: \ Var(Q_{CMC}) = \frac{1}{N}Var(f) + \frac{1}{N}Var(g).$$

Idea of CRN

$$\text{Note}: \underline{E[f-g]} \approx \frac{1}{N}\sum_{i=1}^{N}[f(x_i) - g(y_i)] =: Q_2;$$

where x_i and y_i are simulated suitably.

$$\text{Variance}: \mathrm{Var}(Q_2) = \frac{1}{N}\mathrm{Var}(f-g), \text{ where}$$

$$\mathrm{Var}(f-g) = \mathrm{Var}(f(X)) + \mathrm{Var}(g(Y)) - 2\mathrm{Cov}(f(X), g(Y)).$$

▷ Simulating X and Y independently makes covariance zero.

▷ CRN attemps to reduce the variance by introducing positive dependence between f(X) and g(Y).

Idea of CRN

If $\mathrm{Cov}(f(X), g(Y)) > 0$, then

$$\begin{aligned} \mathrm{Var}(f - g) &= \mathrm{Var}(f(X)) + \mathrm{Var}(g(Y)) - 2\mathrm{Cov}(f(X), g(Y)) \\ &< \mathrm{Var}(f(X)) + \mathrm{Var}(g(Y)). \end{aligned}$$

Thus the variance is indeed reduced.

Idea of CRN

- The CRN technique is a popular and useful VRT which applies when we are comparing two or more alternative configurations (of a system) instead of investigating a single configuration.
- CRN requires synchronization of the random number streams, which ensures that in addition to using the same random numbers to simulate all configurations, a specific random number used for a specific purpose in one configuration is used for exactly the same purpose in all other configurations.

Ex：Calculating Delta of an Option

$$\Delta = \frac{\partial C}{\partial S} \approx \frac{C(S+\Delta S) - C(S-\Delta S)}{2\Delta S}.$$

中心差分

The option prices $C(S+\Delta S)$ and $C(S-\Delta S)$ should not be estimated independently.

Instead, in calculating $C(S+\Delta S)$ and $C(S-\Delta S)$ via MC by CRN, we use the same RV at each time step and for each path starting from $S_0+\Delta S$ and $S_0-\Delta S$.

Two questions:

- When does CRN method work?
- When is CRN optimal? (in the sense that no other mechanism introduces greater positive dependence)

For the simulator:

- If the same random numbers are used, will the variance be reduce?
- Is this the best one can do?

Further reading

P. Glasserman and D. Yao. Some guidelines and guarantees for common random numbers. Management Science (1992).

More Methods

- Moment matching
- Weighted Monte Carlo
- Combined variance reduction techniques

- We may combine different variance reduction techniques, but care must be taken when we are doing so.
- We may also use variance reduction techniques in quasi-Monte Carlo (next chapter).

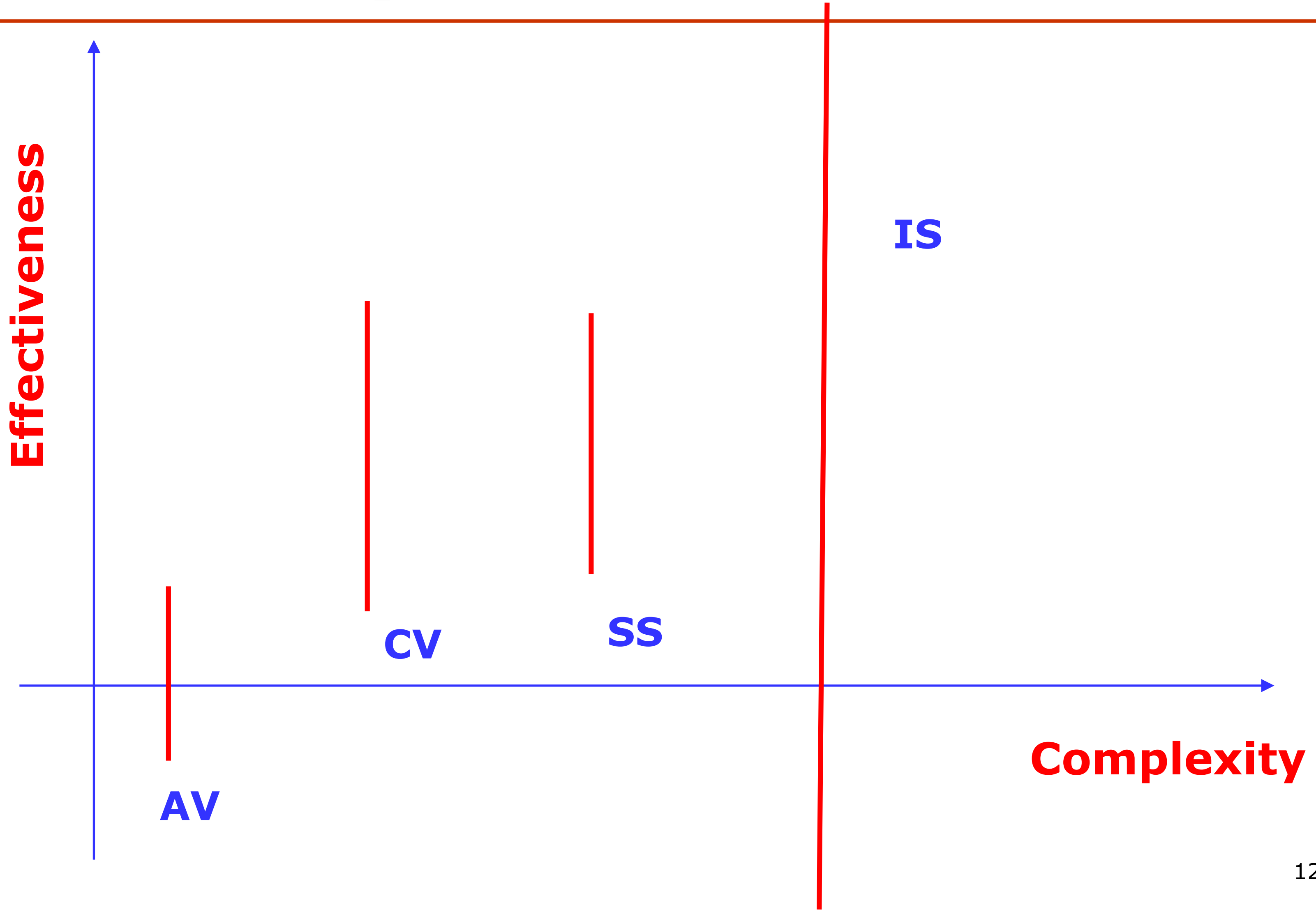
Summary

1-4: Replace the integrand by another one with smaller variance

1. Antithetic variables --- generic and easy to implement but limited effectiveness.
2. Control variates --- easy to implement and can be very effective but require careful choice of CV.
3. Conditional MC --- not easy to implement.
4. Importance sampling --- very useful but again needs to be fine-tuned for each applications.
5. Stratified sampling --- fairly generic but a bit more complex to implement and needs good knowledge of "important dimension".
6. Latin hypercube sampling --- easy to implement and very useful for nearly additive functions.

5 and 6: More regular sampling

Summary



Summary

Which method to use, and which method is the best?

Some simple advice:

- If there is an obvious good control variate at hand, then use it. Often, AV is beneficial before CV is applied.

- For computing expectation where rare event plays an important role, importance sampling is usually the preferable method, sometimes the only one that works.

Further Reading

- Boyle, P., M. Broadie, P. Glasserman. 1997. Monte Carlo methods for security pricing. J. Econom. Dynam. Control, 21, 1267-1321.

  (About all variance reduction techniques)

- Glasserman, P., P. Heidelberger, P. Shahabuddin. 1999. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. Math. Finance, 9, 117-152.

  (About importance sampling and stratification)

Assignments 3

1. Prove the next **Lemma** for d>1:

   If $X_1, \ldots, X_d$ are independent, then for any increasing functions f_1 and f_2 of d variables,

$$\mathbf{E[f_1(X)f_2(X)] \geq E[f_1(X)]\, E[f_2(X)],}$$

$$\mathbf{or\ \ cov\left(f_1(X), f_2(X)\right) \geq 0,}$$

$$\mathbf{where\ \ X = (X_1, \ldots, X_d).}$$

Assignments 3

2. Show that in the CV method, var(f-bg) < var(f), if and only if b lies between 0 and b*, where b* is the optimal parameter which minimizes the variance of the CV estimate (note that b* may be negative).

3. (1) Show that for the function

$$f(x) = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d + b$$

the AV estimate has zero variance.

(2) What is the effect of using AV to function

$$f(x) = \sum_{j=1}^{d} \left(1 - 2x_j\right)^2 \text{ ?}$$

Assignments 3

4. Suppose one wishes to use MC simulation to estimate the value of

$$E[f(X)],$$

where $f(X)$ is a twice-differentiable function (with a continuous second derivative) and X is Normally distributed with zero mean and variance $c << 1$.

Show that standard MC estimator with N samples has variance which is $O(c^2/N)$, whereas the use of antithetic variables reduces the variance to $O(c^4/N)$.

Hint：Use Taylor expansion and Let $X=cy$, where $y \sim N(0,1)$.

Assignments 3

5. Suppose there are two control variates g1 and g2 with known expectations, and they are to be used by computing the average of

$$f - b_1 (g_1 - E[g_1]) - b_2 (g_2 - E[g_2])$$

for N independent samples to get an estimate for E[f].

How would you choose the values for b_1 and b_2 to minimize the variance of this estimator?

Assignments 3:

6. (1) Under the framework of Black-Scholes model, derive an analytical formula for the price of geometric Asian option with a payoff

$$f_G = \max\left(0,\ \Pi_{i=1}^{n} S(t_i)^{w_i} - K\right),$$

where $w_i = 1/n.$

6. (2) Use geometric Asian option as an control variable (CV), write program to price arithmetic Asian option (and compute the variance reduction factor comparing with crude MC).

Assignments 3:

7. Show that stratified estimator with proportional allocation has a variance no larger than that of crude MC estimate. Show that optimal allocation gives smaller variance than proportional allocation.

8. Is the LHS estimate unbiased? Why?
   Is the WUS estimate unbiased? Why?
   (LHS --- Latin hypercube sampling;
   WUS --- Weighted Uniform Sampling)

Assignments 3:

9. Consider estimating $\theta = \int_0^1 4x^3 dx.$

(1) Using standard simulation method to estimate θ.

(2) Using antithetic variable technique to estimate θ.

(3) Construct a control variable estimate of θ.

(4) Using stratification, construct another estimate of θ.

(5) Can you combine the above methods to improve the results?

Assignments 3:

10 **Consider the problem of estimating**

$$\theta = P(Z > b),$$

where $Z \sim N(0,1)$ **and b is a positive constant.**

(1) Estimate θ **via simulation without doing IS.**

(2) Estimate θ **by doing IS with a new random variable** $Y \sim N(\mu, 1)$ **with some appropriat e chice for** μ **(how to choose** μ**?)**

Projects

1. Try to use **each VRT** for option pricing (say, Asian options), and compare their efficiency with crude MC (compute their variance reduction factors).

(Note: You may combine Quasi-Monte Carlo methods with dimension reduction techniques - next chapter)

2. Try to use LHS combining with Brownian bridge or PCA for pricing options (say, Asian options).

Projects

3. (1) Under the multi-dimensional framework of Black-Scholes model, derive an analytical formula for the price of geometric basket option with payoff

$$f_G = \max\left(0,\ \Pi_{i=1}^{n} S_i(T)^{w_i} - K\right),$$

$$\text{where } w_i = 1/n.$$

where $S_i(T)$ is the price of the i-th stock at time T.

(2) Use geometric basket option as a CV, write program to price arithmetic basket option with payoff

$$f_A = \max\left(0,\ \sum_{i=1}^{n} w_i S_i(T) - K\right),$$

$$\text{where } w_i = 1/n.$$

Projects

4. Try to use importance sampling method to price deep out-of-money options or for interest rate derivatives.

References:

Glasserman, P., P. Heidelberger, P. Shahabuddin. 1999. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. Math. Finance, 9, 117-152.

(Note: You may combine Importance Sampling with Quasi-Monte Carlo methods)

Projects

5. Try to use some other variance reduction methods listed below and study their applications in finance:

- Moment matching
- Weighted Monte Carlo
- Combined variance reduction techniques

6. Try to use some variance reduction techniques beyond Black-Scholes model (say, stochastic volatility model or jump diffusion model).

7. Try to use some variance reduction techniques for American options.

End of Chapter 5