S6880 #13
Variance Reduction Methods
Outline

1. Variance Reduction Methods
   - Variance Reduction Methods

2. Importance Sampling
   - Importance Sampling

3. Control Variates
   - Control Variates
   - Cauchy Example Revisited

4. Antithetic Variates
   - Antithetic Variates

5. Conditioning
   - Conditioning
## General Techniques

Reduction of simulation cost $\propto$ variance reduction.

1. Importance Sampling
2. Control and antithetic variates
3. Conditioning
4. Use of special features of a problem
A note about hit-or-miss Monte Carlo integration

⇒ less efficient than original Monte Carlo

\[ \theta = \int_a^b \phi(x) dx \quad \phi(x) \leq c \text{ over } [a, b] \]

draw uniform sample from \( U([a, b] \times [0, c]) \)

\[ \hat{\theta} = c(b - a) \times \{ \text{proportion of sample items falling in shaded area} \} \]

\( Var(\text{hit-or-miss Monte-Carlo}) \geq Var(\text{original Monte-Carlo}) \)
Outline

1. Variance Reduction Methods
   - Variance Reduction Methods

2. Importance Sampling
   - Importance Sampling

3. Control Variates
   - Control Variates
   - Cauchy Example Revisited

4. Antithetic Variates
   - Antithetic Variates

5. Conditioning
   - Conditioning
What is Importance Sampling

Want: Estimate $\theta = \int \phi(x)f(x)dx = E_f \phi(X)$, $X \sim$ p.d.f. $f$
estimated by $\hat{\theta}_f = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$.

Method: Instead of sampling from $f$, we sample from another p.d.f. $g$. i.e., $X_1, \ldots, X_n \overset{i.i.d.}{\sim} g$. $\hat{\theta}_g = \frac{1}{n} \sum_{i=1}^{n} \psi(X_i)$, where $\psi = \frac{f}{g}$. Note that

$$\theta = \int \phi(x)f(x)dx = \int \left[ \phi(x) \frac{f(x)}{g(x)} \right] g(x)dx$$

$$= \int \psi(x)g(x)dx$$

$$= E \psi(X), \quad X \sim g.$$
Unbiasedness of Importance Sampling Estimator

Note that

1. \( \hat{\theta}_g \) is unbiased (since \( E\hat{\theta}_g = E\psi(X) = \theta \)).

2. 

\[
\text{Var}(\hat{\theta}_g) = \frac{\text{Var}_g(\psi(X))}{n} = \frac{1}{n} \int (\psi(x) - \theta)^2 g(x) dx
\]

\[
= \frac{1}{n} \int \left( \phi(x) \frac{f(x)}{g(x)} - \theta \right)^2 g(x) dx.
\]

which can be made small if we choose \( g \) such that \( \psi = \phi \frac{f}{g} \) is closed to a constant.

\[\Rightarrow \text{Choose } g \text{ such that } \]

\[
\begin{cases}
g \approx c \cdot \phi f \\
g \text{ is easier to sample than } f
\end{cases}
\]
Cauchy Example, continued

\[ \theta = P(\text{Cauchy} > 2), \quad f \sim \text{Cauchy}, \quad \phi(x) = l(x > 2). \]

So, want \( g \propto \frac{1}{\pi(1+x^2)} l(x > 2). \)

Try \( g(x) \equiv \frac{2}{x^2} l(x > 2) \equiv \text{p.d.f. of } \frac{2}{U(0,1)} \equiv \text{p.d.f. of } \frac{1}{U(0,1/2)}. \)

Then \( \psi(x) = \phi(x) \frac{f(x)}{g(x)} = \frac{1}{2\pi} \frac{x^2}{1+x^2} l(x > 2) \) which is a (monotone) increasing function over \((2, \infty)\) with values bounded by \( \frac{4}{5} \) and 1.

\[ \implies \approx \text{constant.} \]

Can show \( \text{Var}_g(\psi(X)) \approx 9.3 \times 10^{-5} \) and hence

\[ \text{Var}(\hat{\theta}_g) \approx \frac{9.3 \times 10^{-5}}{n}. \]
Note about Cauchy Example

\[ \theta = \int_2^\infty \psi(x)g(x)dx = \int_2^\infty \frac{1}{2\pi} \frac{x^2}{1 + x^2} g(x)dx \]

\[ = \int_2^\infty \frac{1}{2\pi} \frac{1}{1 + x^{-2}} g(x)dx \]

\[ = \int_{1/2}^{0} \frac{1}{2\pi} \frac{1}{1 + y^2} g(y^{-1})(-y^{-2})dy \quad \text{(let } y = x^{-1}) \]

\[ = \int_0^{1/2} \frac{1}{\pi(1 + y^2)} \left[ \frac{1}{2} g(y^{-1})y^{-2} \right] dy \]

\[ = \int_0^{1/2} \frac{1}{\pi(1 + y^2)}dy \quad \text{(Note: } g(y^{-1}) = 2y^2) \]

\[ = \text{Monte Carlo integration } \int_0^{1/2} f(y)dy \]

see class notes 12.
Outline

1. Variance Reduction Methods
   - Variance Reduction Methods

2. Importance Sampling
   - Importance Sampling

3. Control Variates
   - Control Variates
   - Cauchy Example Revisited

4. Antithetic Variates
   - Antithetic Variates

5. Conditioning
   - Conditioning
What are Control Variates

Goal: Estimate $\theta = E\phi(X)$.

Method: Find $\psi(\cdot)$ such that $\mu_\psi = E\psi(X) = \int_{\mathbb{R}} \psi(x)f(x)dx$ is known and easy-to-calculate. Idea is to choose $\psi(\cdot)$ so that $\psi \approx \phi$, $\psi(X)$ is a control variate.

\[
\theta = E\phi(X) = E[\phi(X) - \psi(X)] + E\psi(X) = E[\phi(X) - \psi(X)] + \mu_\psi.
\]

Then $\hat{\theta} = \mu_\psi + \frac{1}{n} \sum_{i=1}^{n} [\phi(X_i) - \psi(X_i)]$ where $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f$.

\[
\text{Var}(\hat{\theta}) = \frac{1}{n} \text{Var}[\phi(X) - \psi(X)] \ll \frac{1}{n} \text{Var}\phi(X) \text{ if } \psi \approx \phi.
\]
How to Find Control Variates

Example: \( X_1, \ldots, X_n \) \( \text{i.i.d.} \sim F \) with mean \( \mu \) (known).

Want to estimate

\[
\theta = E\phi(X) = E\phi(X_1, \ldots, X_n) \\
= E(\text{median}) = EX_{(m+1)} \text{ if } n = 2m + 1.
\]

Choose \( \psi(X) = \bar{X} \) = the sample mean, then \( E\psi(X) = \mu (= \mu_\psi) \).

Then estimate \( \theta \) by

\[
\hat{\theta} = \mu + \frac{1}{N} \sum_{i=1}^{N} [\phi(X^{(i)}) - \psi(X^{(i)})], \text{ where}
\]

\[
X^{(i)} = i^{th} \text{ sample of size } n = 2m + 1
\]

\[
= (X_1^{(i)}, \ldots, X_n^{(i)}) \text{ i.i.d.} \sim F \text{ for } i = 1, \ldots, N.
\]

\( \hat{\theta} \) often has smaller variance than \( \frac{1}{N} \sum_{i=1}^{N} \phi(X^{(i)}) \).
General Result

In general, $\psi(x)$ can be taken as an approximation of $\phi(x)$ based on some basis $\{b_k(x)\}$.

That is, $\psi(x) = \sum_{k=1}^{K} a_k b_k(x)$, where $Eb_k(X) = \mu_k$ and is known.

If $a_k$’s are unknown, then estimate $a_k$’s by least square estimates based on initial sample and obtain $\hat{a}_k$’s.

Use $\hat{\psi}(x) = \sum_{k=1}^{K} \hat{a}_k b_k(x)$ and estimate $\theta$ by

$$\hat{\theta} = \sum_{k=1}^{K} \hat{a}_k \mu_k + \frac{1}{n} \sum_{i=1}^{n} [\phi(X^{(i)}) - \hat{\psi}(X^{(i)})].$$
Cauchy Example, revisited 1

The first few terms of the Taylor expansion of \( \frac{1}{1+x^2} \) suggest that we can consider \( x^2 \) and \( x^4 \).

\[
\theta = P(X > 2) = \frac{1}{2} - P(0 \leq X \leq 2) = \frac{1}{2} - \int_0^2 \frac{1}{\pi} \frac{1}{1 + x^2} \, dx
\]

\[
= \frac{1}{2} - \int_0^2 2f(x)g(x) \, dx, \quad f(x) = \frac{1}{\pi} \frac{1}{1 + x^2} = \text{p.d.f. of Cauchy}
\]

\[
= \frac{1}{2} - \frac{2}{\pi} E\phi(X) \quad g(x) = \frac{1}{2} I_{[0,2]}(x) = \text{p.d.f. of U}(0,2)
\]

where \( \phi(x) = \frac{1}{1+x^2} \) with \( X \sim U(0,2) \). If we approximate \( \phi(x) \) by

\[
\psi(x) = a_0 + a_2 x^2 + a_4 x^4:
\]

sample \( x_1, \cdots, x_5 \sim U(0,2) \) (initial sample) yielded:
Cauchy Example, revisited 1, continued

<table>
<thead>
<tr>
<th>obs</th>
<th>$x$</th>
<th>$\phi(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.261</td>
<td>.386</td>
</tr>
<tr>
<td>2</td>
<td>0.548</td>
<td>.769</td>
</tr>
<tr>
<td>3</td>
<td>0.207</td>
<td>.959</td>
</tr>
<tr>
<td>4</td>
<td>0.850</td>
<td>.581</td>
</tr>
<tr>
<td>5</td>
<td>1.842</td>
<td>.228</td>
</tr>
</tbody>
</table>

regress $y = \phi(x)$ on $x^2$, $x^4$ \[ \hat{a}_0 = .943, \hat{a}_2 = -.513, \hat{a}_4 = .090. \]

That is, $\hat{\psi}(x) = .943 - .513x^2 + .090x^4$. Then

\[
\hat{\theta} = \frac{1}{2} - \frac{2}{\pi}[(.943 - .513x^2 + .090x^4) + \frac{1}{n}\sum_{i=1}^{n}[\phi(x_i) - \hat{\psi}(x_i)]]
\]

\[
= \frac{1}{2} - \frac{2}{\pi}[.943 - .513 \frac{4}{3} + .090 \frac{16}{5} + \frac{1}{n}\sum_{i=1}^{n}[\phi(x_i) - \hat{\psi}(x_i)]]
\]

here $EX^k = \int_0^2 x^k \frac{1}{2} dx = \frac{2^k}{k+1}$ so $EX^2 = \frac{4}{3}$, $EX^4 = \frac{16}{5}$. 

(WMU) S6880 #13
Cauchy Example, revisited 1, continued

\[ \text{Var}(\hat{\theta}) = \frac{4}{\pi^2 n} \text{Var}(\phi(X) - \hat{\psi}(X)) = \frac{4}{\pi^2 n} (.001774410) \approx 7.2 \times 10^{-4}/n \]

Plotting \( \phi(x) \) and \( \hat{\psi}(x) \) suggests that we would do better by including terms in \( x \) and \( x^3 \). Regress \( y \) on \( x, x^2, x^3, x^4 \) \( \Rightarrow \hat{a}_0 = 1.012, \hat{a}_1 = -0.069, \hat{a}_2 = -1.076, \hat{a}_3 = .813, \hat{a}_4 = -1.181 \) with variance \( 1.11 \times 10^{-5}/n \).
Cauchy Example, revisited 1, continued

Note: can increase the precision by increasing the initial sample size. For instance, an initial sample of size 10 yielded 0.036, 1.642, 1.829, 1.420, 0.532, 1.974, 1.245, 1.615, 0.303, 0.630. Fit $\phi(x)$ on polynomial of degree 4 on $x$ yielded

$$\hat{\psi}(x) = 1.005 - 0.081x - 0.962x^2 + 0.681x^3 - 0.141x^4$$

with

$$\text{Var}(\hat{\theta}) \approx 6.57 \times 10^{-6}/n.$$
Cauchy Example, revisited 2

\[ \theta = P(X > 2) = \frac{1}{2} Ef(Y), \quad Y \sim U(0, \frac{1}{2}) \]

\[ = \frac{1}{2\pi} E\phi(Y) \quad \text{where} \quad \phi(y) = \frac{1}{1 + y^2}, \quad Y \sim U(0, \frac{1}{2}) \]

Approximate \( \phi(y) \) by \( \psi(y) = a_0 + a_1 y + a_2 y^2 + a_3 y^3 + a_4 y^4: \)

<table>
<thead>
<tr>
<th>obs.</th>
<th>( y )</th>
<th>( z = \phi(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.214</td>
<td>.956</td>
</tr>
<tr>
<td>2</td>
<td>.075</td>
<td>.994</td>
</tr>
<tr>
<td>3</td>
<td>.325</td>
<td>.904</td>
</tr>
<tr>
<td>4</td>
<td>.445</td>
<td>.835</td>
</tr>
<tr>
<td>5</td>
<td>.225</td>
<td>.952</td>
</tr>
<tr>
<td>6</td>
<td>.310</td>
<td>.912</td>
</tr>
<tr>
<td>7</td>
<td>.232</td>
<td>.949</td>
</tr>
<tr>
<td>8</td>
<td>.485</td>
<td>.810</td>
</tr>
<tr>
<td>9</td>
<td>.175</td>
<td>.970</td>
</tr>
<tr>
<td>10</td>
<td>.499</td>
<td>.801</td>
</tr>
</tbody>
</table>

where \( y \)'s were sampled from \( U(0, \frac{1}{2}) \).

Regress \( z \) on \( y, y^2, y^3, y^4: \)

\( \hat{a}_0 = .997, \hat{a}_1 = .056, \hat{a}_2 = -1.376, \)

\( \hat{a}_3 = 1.078, \hat{a}_4 = - .246. \) That is,

\( \hat{\psi}(y) = .997 + .056y - 1.376y^2 + 1.078y^3 - .246y^4. \)
Cauchy Example revisited 2, continued

\[ EY^k = \int_0^{1/2} y^k \cdot 2 dy = \frac{2^{-k}}{k+1}, \text{ so} \]

\[ \hat{\theta} = \frac{1}{2\pi} \left[ .997 + .056 \frac{2^{-1}}{1+1} - 1.376 \frac{2^{-2}}{2+1} + 1.078 \frac{2^{-3}}{3+1} - .246 \frac{2^{-4}}{4+1} + \frac{1}{n} \sum_{i=1}^{n} (\phi(y) - \hat{\psi}(y)) \right]. \]

\[ \text{Var}(\hat{\theta}) = \frac{1}{4\pi^2 n} \text{Var}(\phi(Y) - \hat{\psi}(Y)), \text{ where } Y \sim U(0, \frac{1}{2}) \]

\[ \approx 8.64 \times 10^{-9} / n \]

\[ \Rightarrow \approx \text{constant}. \text{ Taking } n = 1 \text{ (sample of size 1!)} \text{ gives an accurate enough approximation to } \theta. \]
Variance Reduction Methods

1. Importance Sampling
2. Control Variates
3. Antithetic Variates
4. Conditioning
Antithetic Variates Method

- **Goal:** Estimate $\theta = E\phi(X)$.
- **Method:** Find $\psi$ such that $\psi(X)$ has the same distributions as $\phi(X)$ such that $\text{cor}(\phi, \psi) < 0$. Then $\theta = \frac{1}{2} E[\phi(X) + \psi(X)]$ and estimate $\theta$ by $\hat{\theta} = \frac{1}{2n} \sum_{i=1}^{n} [\phi(X_i) + \psi(X_i)]$.

$$
\text{Var}(\hat{\theta}) = \frac{1}{4n} \text{Var}[\phi(X) + \psi(X)]
= \frac{1}{4n} \left[ \text{Var}\phi(X) + \text{Var}\psi(X) + 2\text{cov}(\phi(X), \psi(X)) \right]
= \frac{1}{2n} \left[ \text{Var}\phi(X) + \text{cov}(\phi(X), \psi(X)) \right]
= \frac{\text{Var}\phi(X)}{2n} [1 + \text{cor}(\phi(X), \psi(X))].
$$

Consequently, there are gains (in reduction of variance) since $\text{cor}(\phi(X), \psi(X)) < 0$. 
How to Choose Antithetic Variates

Theorem

(Ripley) If \( U \sim U(0, 1) \) then

\[
\begin{align*}
E\phi(U) &= E\phi(1 - U) \\
Var\phi(U) &= Var\phi(1 - U)
\end{align*}
\]

If, in addition, \( \phi \) is monotone, then

\[
\text{cor}(\phi(U), \phi(1 - U)) < 0.
\]
Remarks on the Theorem

1. Reasons for monotone: $\theta = E\phi(X) = EZ$, $Z = \phi(X) \sim F_Z$ (since $F_Z$ is monotone, so is $F_Z^{-1}$). $\theta = EZ = EF_Z^{-1}(U)$. So we can choose $\psi(X) = F_Z^{-1}(1 - U) = F_Z^{-1}(1 - F_Z(Z))$.

2. Theorem (Ripley, about generalization): If $X$ is symmetric about $c$, then can choose $\psi(X) = \phi(2c - X)$.

3. Use other method if $F_Z^{-1}$ is hard to calculate or $\phi$ is not monotone.
Example

φ is symmetric about $\frac{1}{2}$, $\text{cor}(\phi(U), \phi(1 - U)) \approx 1$. Better use

$$g(x) = \begin{cases} 
\frac{1}{2} + x, & x < \frac{1}{2} \\
\frac{3}{2} - x, & x > \frac{1}{2} 
\end{cases}$$

ψ(X) = φ(g(X)). Then $E\psi(X) = E\phi(X)$, for $X \sim U(0, 1)$. 

![Graph of φ and g(x)](image-url)
Cauchy Example, revisited 2

\[ \theta = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1 + x^2)} \, dx = \frac{1}{2} - Eh(X), \quad h(x) = \frac{2}{\pi(1 + x^2)}, \quad X \sim U(0, 2) \]

\[ = \frac{1}{2} - Eh(2U) = \frac{1}{2} - E\phi(U) \]

by letting \( X = 2U, \ U \sim U(0, 1), \ \phi(U) = h(2U) = \frac{2}{\pi(1+4U^2)}. \)

So, we can estimate \( \theta \) by

\[ \hat{\theta} = \frac{1}{2} - \frac{1}{2} \frac{1}{n} \sum_{i=1}^{n} [\phi(U_i) + \phi(1 - U_i)] \]

since \( \phi \) is monotone.

\[ \text{Var}(\hat{\theta}) = \frac{1}{4n} \text{Var}[\phi(U) + \phi(1 - U)] \approx (5.9 \times 10^{-4})/n. \]
Outline

1. Variance Reduction Methods
   - Variance Reduction Methods

2. Importance Sampling
   - Importance Sampling

3. Control Variates
   - Control Variates
   - Cauchy Example Revisited

4. Antithetic Variates
   - Antithetic Variates

5. Conditioning
   - Conditioning
About Conditioning

Recall:
1. \( E[X|Y = y] = h(y) \), a function of \( y \).
2. \( EX = E[E[X|Y]] \)
3. \( \text{Var}(X) = \text{Var}(E[X|Y]) + E[\text{Var}(X|Y)] \) so that \( \text{Var}(E[X|Y]) \leq \text{Var}(X) \) with the equality iff \( X = g(Y) \) (in which case \( \text{Var}(X|Y = y) = \text{Var}(g(y)|y) = 0! \))

Goal: Estimate \( \theta = E\phi(X) \).

Method: If there exists \( Y \) such that \( \psi(y) = E[\phi(X)|Y = y] \) can be easily calculated, then estimate \( \theta \) by \( \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i) \) with variance

\[
\text{Var}(\hat{\theta}) = \frac{1}{n} \text{Var}(\psi(Y)) = \frac{1}{n} \text{Var}(E[\phi(X)|Y]) \\
\leq \frac{1}{n} \text{Var}(\phi(X)).
\]
Example

Suppose $X = \frac{Z}{W}$, $Z \sim N(0, 1)$, $W > 0$, and $W$ and $Z$ are independent (recall “swindle”). The goal is to estimate $\theta = P(\bar{X}_n > c)$, where $\bar{X}_n$ is the sample mean of the sample $X_1, \ldots, X_n$.

No conditioning (naive!):

$$\hat{\theta} = \frac{1}{N} \# \{ \bar{X}_n^{(j)} > c, 1 \leq j \leq N \},$$

where $\bar{X}_n^{(j)}$ is the sample mean of the $j^{th}$ sample $\{ X_1^{(j)}, \ldots, X_n^{(j)} \}$, $1 \leq j \leq N$. 
Example, continued

Conditioning:

Denote $\mathbf{W} = (W_1, \ldots, W_n)^t$, $\mathbf{w} = (w_1, \ldots, w_n)$.

Given $\mathbf{W} = \mathbf{w}$,

$$X_n = \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i}{W_i} \sim \mathcal{N}\left(0, \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{w_i^2}\right) \equiv \mathcal{N}(0, \sigma^2(\mathbf{w})),$$

where $\sigma^2(\mathbf{w}) = \sum_{i=1}^{n} \frac{1}{w_i^2}$

So, $P(X_n > c | \mathbf{W} = \mathbf{w}) = 1 - \Phi\left(\frac{c}{\sigma(\mathbf{w})}\right) \equiv \psi(\mathbf{w})$, where $\Phi(\cdot)$ is the standard normal c.d.f.

Therefore, a better estimate is $\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \psi(W^{(j)})$, where $W^{(j)}$ denotes the $j^{th}$ sample $\{W_1^{(1)}, \ldots, W_n^{(j)}\}$ from $\mathbf{W}$, $1 \leq j \leq N$. 