Importance Sampling

STAT 525

9/11/18

Importance Sampling: A Motivating Example

- Suppose $X \sim$ Bernoulli($\frac{1}{6}$), i.e., $\pi(1) = \frac{1}{6}$ and $\pi(0) = \frac{5}{6}$.

  Want to estimate $E_\pi(I_{\{X=1\}}) = P(X = 1)$.

- Naive Monte Carlo:
  - Draw i.i.d. samples $x^{(1)}, \cdots, x^{(m)}$ from $\pi(x)$.
  - Estimate μ by

$$\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} I_{\{x_i=1\}}.$$

$$Var(\hat{\mu}_m) = \frac{1}{m} Var(I_{\{X_i=1\}}) = \frac{1}{m} \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36m}.$$

- Another method: Rewrite

$$E_\pi(I_{\{X=1\}}) = \int I_{\{x=1\}}\pi(x)dx \quad = \quad \int I_{\{x=1\}}\frac{\pi(x)}{g(x)}g(x)dx$$
$$= \quad E_g\left(I_{\{X=1\}}\frac{\pi(X)}{g(X)}\right).$$

Choose $g(x)$ to be Bernoulli$(\frac{1}{2})$.

  - Draw i.i.d. samples $x^{(1)}, \cdots, x^{(m)}$ from $g(x)$.
  - Estimate μ by

$$\hat{\mu}_m^* = \frac{1}{m}\left(\sum_{i=1}^m I_{\{x_i=1\}}\frac{\pi(x_i)}{g(x_i)}\right) = \frac{1}{m}\left(\sum_{i=1}^m I_{\{x_i=1\}}\frac{1}{3}\right).$$

$$Var(\hat{\mu}_m^*) = \frac{1}{m}Var_g\left(I_{\{X_i=1\}}\frac{1}{3}\right) = \frac{1}{m}\cdot\frac{1}{9}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{36m}.$$

- Note that

$$Var(\hat{\mu}_m^*) = \frac{1}{36m} < \frac{5}{36m} = Var(\hat{\mu}_m).$$

- The 2nd method used the idea of drawing samples from another distribution $g(x)$ which concentrates on “region of importance”, so as to reduce the variance of Monte Carlo estimates.

Importance Sampling (IS)

Procedure: Write $\mu = \int h(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \int \left[h(\mathbf{x})\frac{\pi(\mathbf{x})}{g(\mathbf{x})}\right] g(\mathbf{x})d\mathbf{x}$.

- Draw $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}$ i.i.d. from a *proposal distribution* $g(\mathbf{x})$;
- Calculate the *importance weights*

$$w^{(i)} = \frac{\pi(\mathbf{x}^{(i)})}{g(\mathbf{x}^{(i)})}, \text{ for } i = 1, \cdots, m.$$

- Estimate μ by

$$\hat{\mu} = \frac{w^{(1)}h(\mathbf{x}^{(1)}) + \cdots + w^{(m)}h(\mathbf{x}^{(m)})}{m}$$

Optimal $g(\mathbf{x})$

- Theorem: The choice of $g(\mathbf{x})$ that minimizes the variance of the estimate $\hat{\mu}$ is
$$g^*(\mathbf{x}) = \frac{|h(\mathbf{x})|\pi(\mathbf{x})}{\int |h(\mathbf{z})|\pi(\mathbf{z})d\mathbf{z}}.$$

- But in practice, it's often hard to sample from $g^*(\mathbf{x})$ directly. We hope to choose $g(\mathbf{x})$ as "close" in shape to $|h(\mathbf{x})|\pi(\mathbf{x})$ as possible.

More Comments on Choosing $g(\mathbf{x})$

- $g(\mathbf{x})$ should be easy to sample from.
- The support of $g(\mathbf{x})$ should include the support of $h(\mathbf{x})\pi(\mathbf{x})$. Usually just make sure it includes the support of $\pi(\mathbf{x})$.
- Usually choose $g(\mathbf{x})$ with heavier tail than $\pi(\mathbf{x})$ to keep the variance of the estimate small.

When the Normalizing Constant of π Is Unknown

- Normalizing constant of $\pi(\mathbf{x})$ doesn't have to be known. If $\pi(\mathbf{x}) \propto l(\mathbf{x})$, just replace $\pi(\mathbf{x})$ by $l(\mathbf{x})$ in the algorithm, and estimate μ by

$$\tilde{\mu} = \frac{w^{(1)} h(\mathbf{x}^{(1)}) + \cdots + w^{(m)} h(\mathbf{x}^{(m)})}{w^{(1)} + \cdots + w^{(m)}}.$$

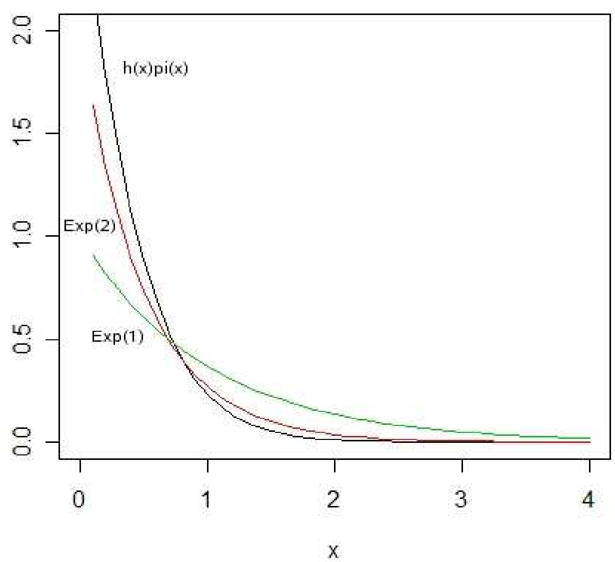
  where $w^{(i)} = l(\mathbf{x}^{(i)})/g(\mathbf{x}^{(i)})$.

- This estimator is biased, but still converges to μ.

- The variance of $\tilde{\mu}$ can be approximated by

$$\frac{Var_g\left[\frac{h(\mathbf{X})l(\mathbf{X})}{g(\mathbf{X})}\right]+\mu^2 Var_g\left[\frac{l(\mathbf{X})}{g(\mathbf{X})}\right]-2\mu Cov_g\left(\frac{h(\mathbf{X})l(\mathbf{X})}{g(\mathbf{X})},\frac{l(\mathbf{X})}{g(\mathbf{X})}\right)}{mE_g^2\left[\frac{l(\mathbf{X})}{g(\mathbf{X})}\right]}.$$

- Example 1. Suppose X $\sim$Exp(1). Want to estimate $E\left(e^{-X+\cos X}\right)$.



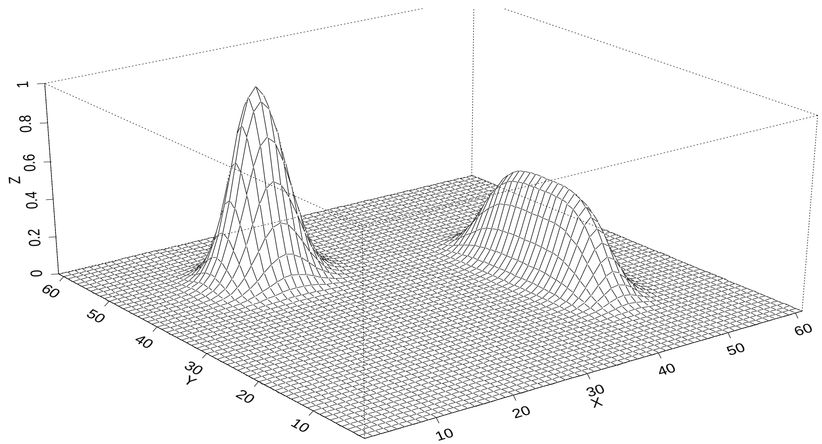
- Comparison: Based on 100 samples:

  Naive Monte Carlo: $\hat{\mu} = 1.21 \pm 0.09$.

  IS with Exp(2) as proposal: $\hat{\mu} = 1.37 \pm 0.03$.

- Example 2. On $[-1,1] \times [-1,1]$,

$$h(x,y) = 0.5e^{-90(x-0.5)^2-45(y+0.1)^2} + e^{-45(x+0.4)^2-60(y-0.5)^2}.$$



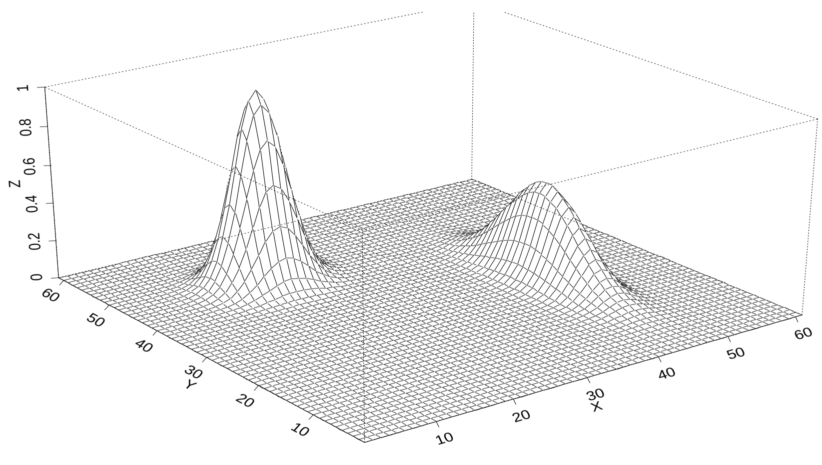
- Want to compute
$\mu = \int_{-1}^{1}\int_{-1}^{1} h(x,y)dxdy = E_\pi[4h(X,Y)]$, where
$(X,Y) \sim$ Unif$[-1,1] \times [-1,1]$.

- Importance sampling: Use proposal distribution $g(x, y)$

$$g(x,y) \propto 0.5e^{-90(x-.5)^2-10(y+.1)^2} + e^{-45(x+.4)^2-60(y-.5)^2},$$

with $(x, y) \in [-1, 1] \times [-1, 1]$. This is a truncated mixture of normal distributions:

$$.464N\left[\begin{pmatrix} .5 \\ -.1 \end{pmatrix}, \begin{pmatrix} \frac{1}{180} & 0 \\ 0 & \frac{1}{20} \end{pmatrix}\right] + .536N\left[\begin{pmatrix} -.4 \\ .5 \end{pmatrix}, \begin{pmatrix} \frac{1}{90} & 0 \\ 0 & \frac{1}{120} \end{pmatrix}\right]$$



- Comparison: Based on 2500 samples:

  Naive Monte Carlo: $\hat{\mu} = 0.1307 \pm 0.009$.

  Importance sampling: $\hat{\mu} = 0.1259 \pm 0.0005$.

- Importance sampling is $\left(\frac{0.009}{0.0005}\right)^2 = 324$ times more efficient than naive Monte Carlo in this example.

Another Scenario for Using Importance Sampling

- If $\pi(\mathbf{x})$ is too complicated to sample from directly, importance sampling procedure may be used.

- In this case, we often want the proposal $g(\mathbf{x})$ to be close to $\pi(\mathbf{x})$. A "rule of thumb" for evaluating efficiency is *effective sample size*:

$$ESS = \frac{m}{1 + cv^2},$$

where the *coefficient of variation (cv)* is

$$cv^2 = \frac{Var_g[w(\mathbf{x})]}{E_g^2[w(\mathbf{x})]}.$$

- Want cv^2 to be small.

Comparison of Rejection Sampling with Importance Sampling

	Rejection Sampling	Importance Sampling
Generate iid samples from $\pi(\mathbf{x})$?	Yes	No
Can be used to estimate $E_\pi\{h(X)\}$?	Yes	Yes
Requires $\pi(\mathbf{x}) \leq cg(\mathbf{x})$?	Yes	No

- The instrumental density $g(\mathbf{x})$ for rejection sampling can also be used as the proposal distribution for importance sampling.

References

- Section 2.5 of Jun Liu's *Monte Carlo Strategies in Scientific Computing*.