Variance Reduction

Apurva Nakade

2025-04-01

Table of contents

Confidence Intervals 1
Relative Confidence Intervals 2
Markov Chain Monte Carlo (MCMC) 2
Estimating Integrals 4
Importance Sampling 5
Remarks 7
Antithetic and Control Random Variates 8
Antithetic Variates 8
Control Variates 9

In this notebook we will discuss two methods of variance reduction: **importance sampling** and **antithetic variates**. These methods are used to reduce the variance of an estimator, which can lead to more accurate and efficient estimates. We'll start by reviewing the basic concepts of estimation theory.

Confidence Intervals

We have seen two sampling methods that generate independent samples:

1. Inverse transform sampling
2. Acceptance-rejection sampling

When using these methods, we get a good control over the uncertainty of the simulation results.

An estimator is a function of the data that is used to estimate an unknown parameter. The common situation is when ℓ is the expectation of a function of a random variable X, i.e.,

$\ell = E[f(X)]$. A simple estimator of ℓ is the sample mean, which is given by

$$\widehat{\ell} = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \tag{1}$$

where $X_1, X_2, \ldots, X_N$ are independent and identically distributed (i.i.d.) samples from the distribution of X. The sample mean $\widehat{\ell}$ is an unbiased estimator of ℓ, meaning that $E[\widehat{\ell}] = \ell$. By the central limit theorem, as N increases, the distribution of $\widehat{\ell}$ approaches a normal distribution with mean ℓ and variance $\sigma^2(\widehat{\ell}) = \sigma^2(X)/N$, where $\sigma^2(X)$ is the variance of $f(X)$ and N is the number of samples. The confidence interval (CI) for ℓ is given by

$$\left[\widehat{\ell} - z_{1-\alpha/2}\frac{S}{\sqrt{N}}, \widehat{\ell} + z_{1-\alpha/2}\frac{S}{\sqrt{N}}\right],$$

where $z_{1-\alpha/2}$ is the critical value from the standard normal distribution corresponding to a significance level of α, and S is the sample standard deviation of $f(X)$. The width of the confidence interval is given by

$$\text{Width of CI} = 2z_{1-\alpha/2}\frac{S}{\sqrt{N}}.$$

Relative Confidence Intervals

It is common practice in simulation to use and report the relative widths of the confidence interval, defined as

$$\text{Relative Width} = \frac{\text{Width of CI}}{\widehat{\ell}} = 2z_{1-\alpha/2}\frac{S}{\widehat{\ell}\sqrt{N}}.$$

The relative width of the confidence interval is a measure of the precision of the estimate.

Markov Chain Monte Carlo (MCMC)

When the random variables X_i are not independent, the variance of the sum is given by

$$\text{Var}(\widehat{\ell}) = \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}(f(X_i)) + \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}\text{Cov}(f(X_i), f(X_j)).$$

There are n^2 terms in the covariance sum, and in general, we cannot guarantee that the covariance is small. So the above confidence interval is not valid. In the case of ergodic Markov chains, the central limit theorem for Markov chains allows us to estimate the CI as

$$\left[\hat{\ell} - z_{1-\alpha/2}\frac{S_{eff}}{\sqrt{N_{eff}}}, \hat{\ell} + z_{1-\alpha/2}\frac{S_{eff}}{\sqrt{N_{eff}}}\right],$$

where N_{eff} is the effective sample size. We will not discuss the details of effective sample size here.

Example: Estimating π. In the first example we saw for estimating π, we used the Monte Carlo method to estimate π by simulating random points in a square and counting the number of points that fall within a circle. The estimator we used was

$$\hat{\ell} = \frac{4}{N}\sum_{i=1}^{N} I(X_i),$$

where $I(X_i)$ is an indicator function that is 1 if the point X_i is inside the circle and 0 otherwise. The variance of this estimator is given by

$$\operatorname{Var}(\hat{\ell}) = \frac{16}{N}\left(\frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)\right).$$

Note: Using the variable M inside the estimator was a mistake in the worksheet.

Example: Estimation of Rare-Event Probabilities. Consider estimation of the tail probability $\ell = P(X > \gamma)$ of some random variable X for a large number γ. We can use the following estimator:

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} I_{>\gamma}(X_i),$$

where $I_{>\gamma}(X_i)$ is an indicator function that is 1 if $X_i > \gamma$ and 0 otherwise. The variance of this estimator is given by

$$\operatorname{Var}(\hat{\ell}) = \frac{1}{N}\left(\ell(1 - \ell)\right).$$

The relative width of the confidence interval is given by

$$\text{Relative Width} = \frac{2z_{1-\alpha/2}\sqrt{\ell(1-\ell)}}{\hat{\ell}\sqrt{N}} \approx \frac{2z_{1-\alpha/2}}{\sqrt{N}}\sqrt{\frac{1-\ell}{\ell}} \approx \frac{2z_{1-\alpha/2}}{\sqrt{N\ell}}$$

When ℓ is small, the relative width of the confidence interval is large. This means that we need a large number of samples to get a good estimate of ℓ.

Example: Magnetization in a 2D Ising Model. In a previous worksheet, we used Gibbs sampling and Metropolis–Hastings sampling to estimate the magnetization of a 2D Ising model. We used the simple estimator

$$\hat{M} = \frac{1}{N} \sum_{i=1}^{N} M(\sigma_i),$$

where $M(\sigma_i)$ is the magnetization of the i-th sample.

The samples σ_i are not independent, and so the variance of the estimator will be

$$\mathrm{Var}(\hat{M}) = \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}(M(\sigma_i)) + \frac{2}{N^2} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathrm{Cov}(M(\sigma_i), M(\sigma_j)),$$

which leads to much larger confidence intervals than we would expect from independent samples.

Estimating Integrals

A very common application of Monte Carlo methods is to estimate integrals. The integral

$$\ell = \int_a^b f(x) dx$$

of a function $f(x)$ over the interval $[a, b]$ can be estimated using the estimator:

$$\hat{\ell} = \frac{b-a}{N} \sum_{i=1}^{N} f(X_i),$$

where X_i are i.i.d. samples from the uniform distribution on $[a, b]$. One can check that this is an unbiased estimator,

$$\begin{aligned} E[\hat{\ell}] &= \frac{b-a}{N} \sum_{i=1}^{N} E[f(X_i)] \\ &= \frac{b-a}{N} \sum_{i=1}^{N} \int_a^b f(x) \frac{1}{b-a} dx \\ &= \frac{1}{N} \sum_{i=1}^{N} \ell \\ &= \ell. \end{aligned}$$

The variance of this estimator is given by

$$\operatorname{Var}(\hat{\ell}) = \frac{(b-a)^2}{N^2} \sum_{i=1}^{N} \operatorname{Var}(f(X_i)) = \frac{(b-a)^2}{N} \cdot \sigma^2(f(X)),$$

where $\sigma^2(f(X))$ is the variance of $f(X)$, and N is the number of samples.

We will return to this example later when we discuss the method of antithetic and control random variables.

Importance Sampling

Importance sampling is a method to reduce the variance of an estimator by changing the distribution from which we sample. The idea is to reduce the number of low probability events that contribute to the variance of the estimator.

For example, when estimating the tail probability $\ell = P(X > \gamma)$, if ℓ is small, then most of the samples will be in the region $X \leq \gamma$, which contributes little to the estimate. However, we cannot just sample from the tail of the distribution as this would provide us no information about the rest of the distribution. Importance sampling allows us to sample from tail but then "fix" the estimate by weighting the samples appropriately.

Definition: Importance Sampling. Let X be a random variable with probability density function (pdf) $p(x)$, and let $q(x)$ be a proposal pdf such that $q(x) > 0$ for all x in the support of $p(x)$. The importance sampling estimator of $\ell = E[f(X)]$ is given by

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^{N} f(X_i) \frac{p(X_i)}{q(X_i)},$$

where $X_1, X_2, \ldots, X_N$ are i.i.d. samples from the distribution with pdf $q(x)$.

For clarity, we'll denote the estimator in Equation 1 as $\hat{\ell}_{crude}$ and the importance sampling estimator as $\hat{\ell}_{IS}$.

Suppose $N = 1$ so that the estimator is given by

$$\hat{\ell}_{IS} = f(X) \frac{p(X)}{q(X)}.$$

Note that here $X \sim q(x)$ and NOT $p(x)$. We can check that this estimator is unbiased:

$$\begin{aligned} E_q[\hat{\ell}_{IS}] &= E_q\left[f(X)\frac{p(X)}{q(X)}\right] \\ &= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\ &= \int f(x)p(x)dx \\ &= E[f(X)] \\ &= \ell. \end{aligned}$$

However,

$$\mathrm{Var}(\hat{\ell}_{IS}) \neq \mathrm{Var}(\hat{\ell}_{crude}).$$

This allows us to reduce the variance of the estimator by choosing $q(x)$ appropriately.

Suppose $f(X)$ is a non-negative function. Then if we choose

$$q(x) \propto p(x)f(x),$$

then the importance sampling estimator for $N = 1$

$$\hat{\ell}_{IS} = f(X)\frac{p(X)}{q(X)}$$

is a constant and has zero variance! When H is not non-negative, we can show that

$$q(x) \propto p(x)|f(x)|$$

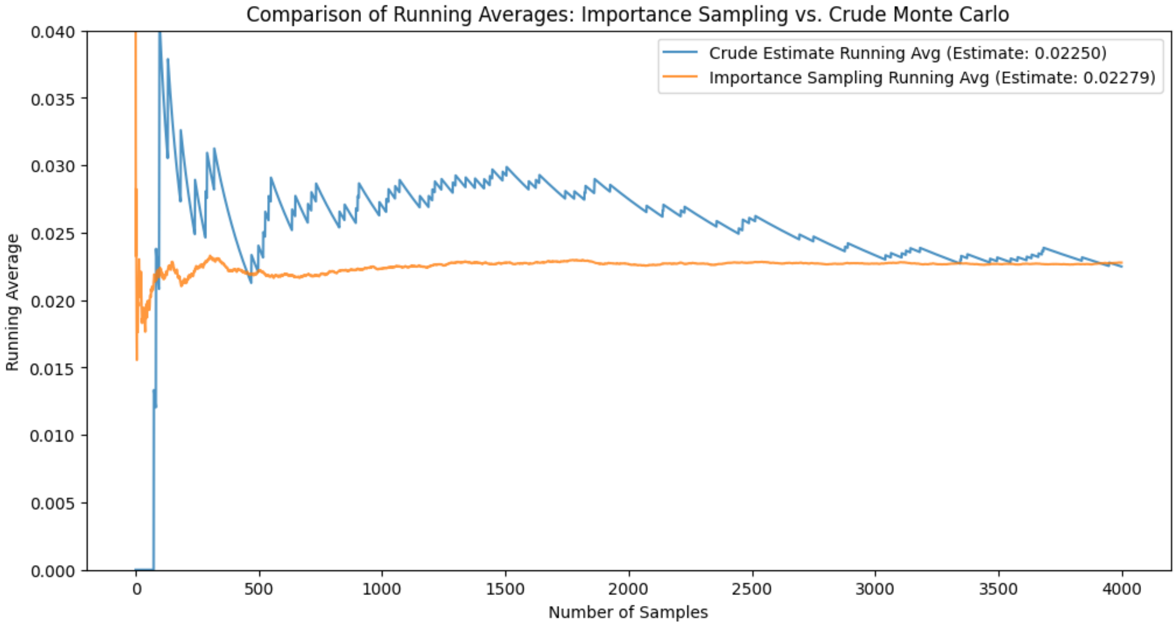
minimizes the variance of the estimator $\hat{\ell}_{IS}$.

However, note that our goal is to estimate $\ell = E[f(X)]$, which means that we do not know $f(x)$ in advance. So we cannot choose this $q(x)$ in advance. Even if we could, we might not be able to sample from $q(x)$ easily. In practice, we choose $q(x)$ to be a distribution that is easy to sample from and that is "close" to $p(x)$ in some sense.

Example: Importance Sampling for Rare Events. Consider the problem of estimating the tail probability $\ell = P(X > \gamma)$ for a random variable X with standard normal distribution. We can use importance sampling to estimate this probability by choosing a proposal distribution $q(x)$ that is concentrated in the tail region. One such distribution is the exponential distribution with parameter λ, which has pdf

$$q(x) = \lambda e^{-\lambda(x-2)}, \quad x \geq 2.$$

The plots below show the running averages of the crude and importance sampling estimators for $N = 2000$ samples. The importance sampling estimator is much more stable and converges to the true value of ℓ much faster than the crude estimator.



```
Variance of Importance Sampling Estimate: 0.00000
Relative Error of Importance Sampling Estimate: 0.01228
Variance of Crude Monte Carlo Estimate: 0.00001
Relative Error of Crude Monte Carlo Estimate: 0.10422
```

Remarks

1. The optimal choice of the proposal distribution $q(x)$ is not always easy to find. Even if we can find it, we may not be able to sample from it easily. For importance sampling algorithm, we need to be able to sample from $q(x)$. Often, we use a distribution that is easy to sample from and that is "close" to $|f(x)|p(x)$ in some sense.
2. Unlike rejection sampling and MCMC methods, for importance sampling we need to know the normalizing constant of the proposal distribution $q(x)$ in order to compute the weights. This means that we are fairly limited in the choice of $q(x)$. Some common choices are the exponential distribution, the normal distribution, and the uniform distribution, and a mixture of these distributions.
3. In order for the estimator to be well-defined, we need to ensure that $q(x) > 0$ for all x in the support of $f(x)p(x)$. As in the case of rare-event estimation, the support of $f(x)p(x)$ may be very small compared to the support of $p(x)$. We only need $q(x)$ to be positive in this smaller region.

Antithetic and Control Random Variates

In this section, we will discuss two methods of variance reduction that are based on the idea of using correlated random variables: **antithetic variables** and **control variates**. Recall that the variance of a sum of two random variables is given by

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y).$$

Oftentimes, having correlated random variables in undesirable as it reduces to an increase in variance and a decrease in the effective sample size. However, in some cases, we can use this correlation to our advantage.

Antithetic Variates

Antithetic variates are pairs of random variables that are negatively correlated. If we can find an estimator that uses sums to two antithetic random variables, we can reduce its variance.

Consider the example of estimating the integral from Section . Suppose $f(x)$ is a monotonic function over $[a,b]$. Then you'll show on the homework that if $X \sim U(a,b)$ then

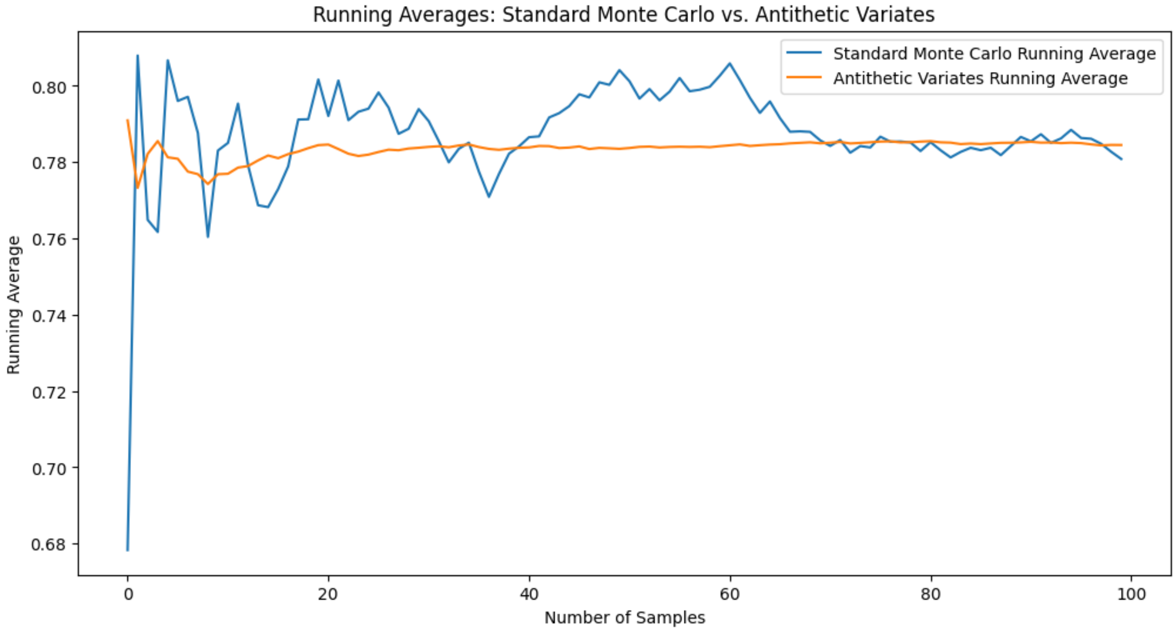
$$\mathrm{cov}(f(X), f(a+b-X)) \leq 0.$$

We can see intuitively why this is happening - if $f(x)$ is increasing the $f(b-x)$ is decreasing and vice versa, and hence the two are negatively correlated. In this case, we can reduce the variance of the crude estimator by instead using

$$\hat{\ell}_{anti} = \frac{(b-a)}{N} \sum_{i=1}^{2N} (f(X) + f(a+b-X))$$

Note that if $X \sim U(a,b)$ then so is $b-X$ and so $\hat{\ell}_{anti}$ is an unbiased estimator.

Example: Antithetic Variates. Consider the problem of estimating the integral $\ell = \int_0^1 (1+x^2)^{-1} dx$ using antithetic variates. The plots below show the running averages of the crude and antithetic variate estimators for $N = 100$ samples. The antithetic variate estimator achieves a $50x$ reduction in variance compared to the crude estimator.



```
Standard Monte Carlo Estimate: 0.779911, Variance: 1.308160e-04
Antithetic Variates Estimate: 0.784493, Variance: 2.260517e-06
Variance Reduction Factor: 57.87x
```

Control Variates

We'll talk about control variates in the next notebook.