Simulation - Lectures 5 - Normalized importance sampling

Lecture version: Tuesday 18$^{\text{th}}$ February, 2020, 09:44

Robert Davies

Part A Simulation and Statistical Programming

Hilary Term 2020

Recap from previous lecture

- Importance sampling is an approach for Monte Carlo with a target $p(x)$ and a proposal distribution $q(x)$
- We calculate the importance weight $w(x) = p(x)/q(x)$, and calculate the average of $\phi(x)w(x)$
- Importance sampling requires $q(x)$ covers $p(x)\phi(x)$, and with lower variance estimators being more desirable, and achievable when the proposal is concentrated towards $|\phi(x)|p(x)$
- Today we focus on two useful cases of importance sampling: **rare event estimation** and **normalized importance sampling**

Outline

Rare event estimation using exponential tilting

Importance sampling in high dimension

Normalised Importance Sampling

Normal Monte Carlo for rare events is impractical

- One important class of applications of IS is for problems in which we estimate the probability for a rare event. In such scenarios, we may be able to sample from p directly and use Monte Carlo, but it is inefficient.
- Consider for example $X \sim p$ with $\phi(X) = 1$ if $X > x_0$, *i.e.* $\mathbb{P}(X > x_0) = \mathbb{E}_p(\mathbb{I}[X > x_0]) = \theta$
- If $\theta \ll 1$, we may not get any samples $X_i > x_0$ even for moderately large n, and our estimate $\hat{\theta}_n = \sum_i \mathbb{I}(X_i > x_0)/n$ is simply zero.
- Though are estimator is still unbiased, it is impractical, with a variance that is too large
- By using IS, we can actually reduce the variance of our estimator.

We can get a proposal by exponentially tilting a normal target

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a scalar normal random variable and we want to estimate $\theta = \mathbb{P}(X > x_0)$ for some $x_0 \gg \mu + 3\sigma$.
- If p is the pdf of X then

$$q(x) = \frac{p(x)e^{tx}}{M_p(t)}$$

  is called an exponentially tilted version of p where $M_p(t) = \mathbb{E}_p(e^{tX})$ is the moment generating function of X.
- For many standard pdfs, the exponentially tilted pdf is in the same family as p, with different parameters
- For p the pdf of a Gaussian variable with mean μ and variance σ^2,

$$q(x) \propto e^{-(x-\mu)^2/2\sigma^2} e^{tx} = e^{-(x-\mu-t\sigma^2)^2/2\sigma^2} e^{\mu t + t^2\sigma^2/2}$$

  so we have

$$q(x) = \mathcal{N}(x; \mu + t\sigma^2, \sigma^2), \quad M_p(t) = e^{\mu t + t^2\sigma^2/2}.$$

Constructing our specific proposal

- The IS weight function is $p(x)/q(x) = e^{-tx} M_p(t)$ so

$$w(x) = e^{-t(x-\mu-t\sigma^2/2)}.$$

- We take samples $Y_i \sim \mathcal{N}(\mu + t\sigma^2, \sigma^2)$, and form our IS estimator for $\theta = \mathbb{P}(X > x_0)$

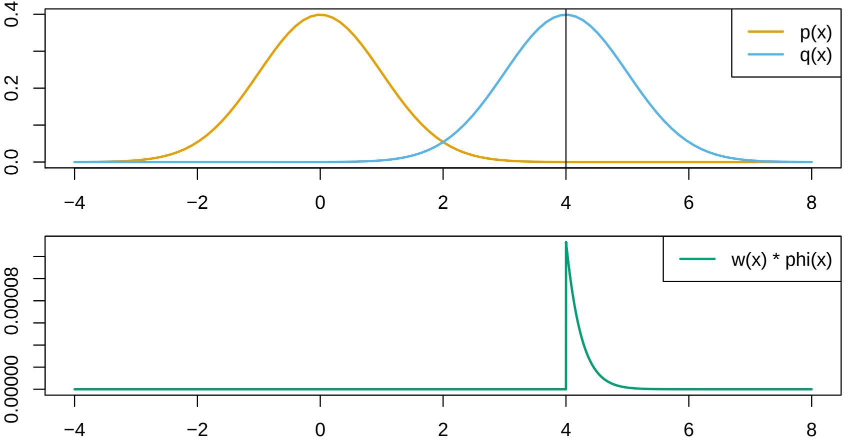
$$\hat{\theta}_n^{\mathsf{IS}} = \frac{1}{n} \sum_{i=1}^{n} w(Y_i) \mathbb{I}(Y_i > x_0)$$

  since $\phi(Y_i) = \mathbb{I}(Y_i > x_0)$.

- We have not said how to choose t. The point here is that we want samples in the region of interest. We choose the mean of the tilted distribution so that it equals x_0, this ensure we have samples in the region of interest; that is $\mu + t\sigma^2 = x_0$, or $t = (x_0 - \mu)/\sigma^2$.

Original and exponentially tilted densities

- $p(x) = N(x; 0, 1)$ and $q(x) = N(x; t, 1)$, $x_0 = t = 4$



Optimal tilting

- We selected t such that $\mu + t\sigma^2 = x_0$ somewhat heuristically.
- In practice, we might be interested in selecting the t value which minimizes the variance of $\hat{\theta}_n^{\mathsf{IS}}$ where

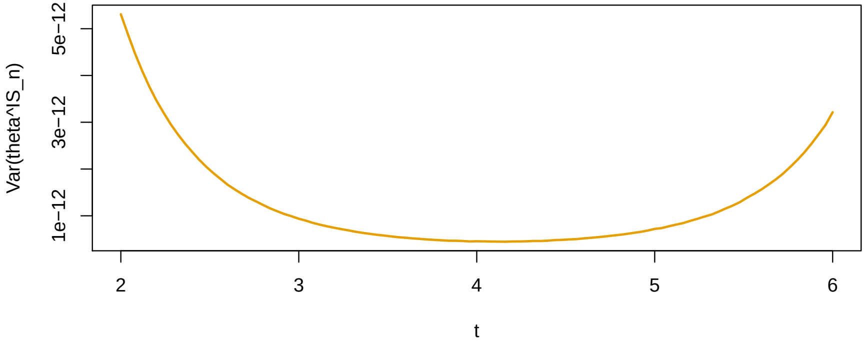
$$\begin{aligned}\mathbb{V}(\hat{\theta}_n^{\mathsf{IS}}) &= \frac{1}{n}\left(\mathbb{E}_p\left(w(X)\mathbb{I}(X > x_0)\right) - \mathbb{E}_p\left(\mathbb{I}(X > x_0)\right)^2\right) \\ &= \frac{1}{n}\left(\mathbb{E}_p\left(w(X)\mathbb{I}(X > x_0)\right) - \theta^2\right).\end{aligned}$$

- Hence we need to minimize $\mathbb{E}_p\left(w(X)\mathbb{I}(X > x_0)\right)$ w.r.t t where

$$\begin{aligned}\mathbb{E}_p\left(w(X)\mathbb{I}(X > x_0)\right) &= \int_{x_0}^{\infty} p(x) e^{-t(x-\mu-t\sigma^2/2)} dx \\ &= M_p(t) \int_{x_0}^{\infty} p(x) e^{-tx} dx\end{aligned}$$

Optimal Tilted Densities

- Here we see the variance $\mathbb{V}(\hat{\theta}_n^{\mathsf{IS}})$ for different values of t for $n = 10,000$



Estimate t using importance sampling

Calculate $M_p(t) \int_{x_0}^{\infty} p(x)e^{-tx}dx$ using importance sampling

```
calc_int <- function(t) {
    y <- rnorm(1000000, mean = 4, sd = 1)
    p <- dnorm(y, mean = 0, sd = 1)
    q <- dnorm(y, mean = 4, sd = 1)
    w <- p / q
    phi <- as.integer(y > 4) * exp(-t * y)
    is <- mean(w * phi)
    mu <- 0
    sigma <- 1
    mgf <- exp(mu * t + sigma **2 * t ** 2 /2)
    return(mgf * is)
}
```

Outline

Rare event estimation using exponential tilting

Importance sampling in high dimension

Normalised Importance Sampling

Importance sampling in high dimension

- Purely for illustration, consider that we want to estimate

$$\theta = \mathbb{E}_p(1) = 1$$

where the target pdf is a d-dimensional Gaussian

$$p(x_1, ..., x_d) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\sum_{k=1}^{d} x_k^2\right).$$

- Consider the proposal density

$$q(x_1, ..., x_d) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{d} x_k^2\right).$$

- We have

$$w(x) = \frac{p(x_1, ..., x_d)}{q(x_1, ..., x_d)} = \sigma^d \exp\left(-\frac{1}{2}(1-\sigma^{-2})\sum_{k=1}^{d} x_k^2\right).$$

Importance Sampling in High Dimension

- For $Y_i \sim q$, $\hat{\theta}_n^{\mathsf{IS}} = \frac{1}{n}\sum_{i=1}^{n} w(Y_i)$ is a consistent estimate of $\theta = 1$.
- The estimator has finite variance for $\sigma^2 > \frac{1}{2}$, with

$$\mathbb{V}\left(\hat{\theta}_n^{\mathsf{IS}}\right) = \frac{\mathbb{V}_q\left(w(Y_1)\right)}{n} = \frac{1}{n}\left(\left(\frac{\sigma^4}{2\sigma^2 - 1}\right)^{d/2} - 1\right)$$

  with $\frac{\sigma^4}{2\sigma^2-1} > 1$ for $\sigma^2 > \frac{1}{2}$, $\sigma^2 \neq 1$.
- Variance of the IS estimator grows exponentially with the dimension d.

Outline

Rare event estimation using exponential tilting

Importance sampling in high dimension

Normalised Importance Sampling

Normalised Importance Sampling

- In most practical scenarios,

$$p(x) = \tilde{p}(x)/Z_p \text{ and } q(x) = \tilde{q}(x)/Z_q$$

  where $\tilde{p}(x), \tilde{q}(x)$ are known but $Z_p = \int_\Omega \tilde{p}(x)dx$, $Z_q = \int_\Omega \tilde{q}(x)dx$ are unknown or difficult to compute.
- The previous IS estimator is not applicable as it requires evaluating $w(x) = p(x)/q(x)$.
- An alternative IS estimator can be proposed based on the following alternative IS identity.
- **Proposition**. Let $Y \sim q$ and $X \sim p$ be continuous or discrete rv on Ω. Assume $p(x) > 0 \Rightarrow q(x) > 0$, then for any function $\phi : \Omega \to \mathbb{R}$ we have

$$\mathbb{E}_p(\phi(X)) = \frac{\mathbb{E}_q(\phi(Y)\tilde{w}(Y))}{\mathbb{E}_q(\tilde{w}(Y))}$$

  where $\tilde{w} : \Omega \to \mathbb{R}^+$ is the importance weight function

$$\tilde{w}(x) = \tilde{p}(x)/\tilde{q}(x).$$

Normalised Importance Sampling

- Proof: Observe that

$$\begin{aligned}\mathbb{E}_q(\tilde{w}(Y)) &= \int \frac{\tilde{p}(x)}{\tilde{q}(x)} q(x)dx \\ &= \int \frac{p(x)}{q(x)} \frac{Z_q}{Z_p} q(x)dx \\ &= \frac{Z_q}{Z_p}\end{aligned}$$

and noting that $\tilde{w} = w\frac{Z_q}{Z_p}$ we have that

$$\frac{\mathbb{E}_q(\phi(Y)\tilde{w}(Y))}{\mathbb{E}_q(\tilde{w}(Y))} = \mathbb{E}_q(\phi(Y)w(Y))$$

- Remark: Even if we are interested in a simple function ϕ, we do need $p(x) > 0 \Rightarrow q(x) > 0$ to hold instead of $p(x)\phi(x) \neq 0 \Rightarrow q(x) > 0$ for the previous IS identity.

Normalised Importance Sampling

An alternate version of the proof

- Proof: We have

$$
\begin{aligned}
\mathbb{E}_p(\phi(X)) &= \int_\Omega \phi(x)p(x)dx \\
&= \frac{\int_\Omega \phi(x)\frac{p(x)}{q(x)}q(x)dx}{\int_\Omega \frac{p(x)}{q(x)}q(x)dx} \\
&= \frac{\int_\Omega \phi(x)\tilde{w}(x)q(x)dx}{\int_\Omega \tilde{w}(x)q(x)dx} \\
&= \frac{\mathbb{E}_q(\phi(Y)\tilde{w}(Y))}{\mathbb{E}_q(\tilde{w}(Y))}.
\end{aligned}
$$

Normalised Importance Sampling Pseudocode

1. **Inputs:**
   - Function to draw samples from q
   - Function $\tilde{w}(x) = \tilde{p}(x)/\tilde{q}(x)$
   - Function ϕ
   - Number of samples n
2. **For** $i = 1, \ldots, n$:
   - 2.1 Draw $y_i \sim q$.
   - 2.2 Compute $\tilde{w}_i = \tilde{w}(y_i)$.
3. **Return**

$$\frac{\sum_{i=1}^{n} \tilde{w}_i \phi(y_i)}{\sum_{i=1}^{n} \tilde{w}_i}.$$

Normalised Importance Sampling Estimator

Proposition

Let q and p be pdf or pmf on Ω, with $q(x) \propto \widetilde{q}(x)$ and $p(x) \propto \widetilde{p}(x)$. Assume $p(x) > 0 \Rightarrow q(x) > 0$. Let $X \sim p$, and $\phi : \Omega \to \mathbb{R}$ such that $\theta = \mathbb{E}_p(\phi(X))$ exists. Let $Y_1, ..., Y_n$ be a sample of independent random variables distributed according to q then the normalized importance sampling estimator, defined by

$$\hat{\theta}_n^{\text{NIS}} = \frac{\frac{1}{n}\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)}{\frac{1}{n}\sum_{i=1}^n \tilde{w}(Y_i)} = \frac{\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)}{\sum_{i=1}^n \tilde{w}(Y_i)},$$

with $\tilde{w}(x) = \frac{\widetilde{p}(x)}{\widetilde{q}(x)}$.

- This estimator is consistent.

- Remark: It is easy to show that $\hat{A}_n = \frac{1}{n}\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)$ (resp. $\hat{B}_n = \frac{1}{n}\sum_{i=1}^n \tilde{w}(Y_i)$) is an unbiased and consistent estimator of $A = \mathbb{E}_q\left(\phi(Y)\tilde{w}(Y)\right)$ (resp. $B = \mathbb{E}_q\left(\tilde{w}(Y)\right)$). However $\hat{\theta}_n^{\text{NIS}}$, which is a ratio of estimates, is biased for finite n.

Normalised Importance Sampling Estimator

- Proof strong consistency (not examinable). The strong law of large numbers yields

$$\mathbb{P}\left(\lim_{n\to\infty}\hat{A}_n \to A\right) = \mathbb{P}\left(\lim_{n\to\infty}\hat{B}_n \to B\right) = 1$$

This implies

$$\mathbb{P}\left(\lim_{n\to\infty}\hat{A}_n \to A, \lim_{n\to\infty}\hat{B}_n \to B\right) = 1$$

and

$$\mathbb{P}\left(\lim_{n\to\infty}\frac{\hat{A}_n}{\hat{B}_n} \to \frac{A}{B}\right) = 1.$$

Example Revisited: Gamma Distribution

- We are interested in estimating $\mathbb{E}_p(\phi(X))$ where $X \sim$Gamma(α, β) using samples from a Gamma(a, b) distribution; i.e.

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad q(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

- Suppose we do not remember the expression of the normalising constant for the Gamma, so that we use

$$\tilde{p}(x) = x^{\alpha-1} e^{-\beta x}, \; \tilde{q}(x) = x^{a-1} e^{-bx}$$
$$\Rightarrow \tilde{w}(x) = x^{\alpha-a} e^{-(\beta-b)x}$$

- Practically, we simulate $Y_i \sim$Gamma(a, b), for $i = 1, 2, ..., n$ then compute

$$\begin{aligned} \tilde{w}(Y_i) &= Y_i^{\alpha-a} e^{-(\beta-b)Y_i}, \\ \hat{\theta}_n^{\mathsf{NIS}} &= \frac{\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)}{\sum_{i=1}^n \tilde{w}(Y_i)}. \end{aligned}$$

Recap

- Importance sampling is particularly useful for rare events
- It can also be used for unnormalized proposals and targets, in which case, one additionally calculates a denominator as the average of the normalized importance weights