COMP 480/580 — Probabilistic Algorithms and Data Structure Oct 9, 2025

Lecture 14

Lecturer: Anshumali Shrivastava

Scribe By: Shizuka Dara (sd205), Shichen Tang (st167), Anika Patel (ap105)

1 Importance Sampling

$X \sim \pi$, π a known distribution, ϕ any function, we want to know

$$E(\phi(X)).$$

Naive approach: sample $x_1, \ldots, x_n$ from π,

$$\hat{\theta} = \sum_{i=1}^{n} \phi(x_i).$$

Problem 1: Sometimes we don't know how to sample from π.

But there are other problems as well: For example: $x \sim N(0,1)$, want to estimate

$$\begin{aligned} \theta &= P(x > 5) \\ &= E(I_{[x \geq 5]}). \quad = \frac{1}{n} \sum_{i=1}^{n} I_{[x \geq 5]}. \end{aligned}$$

However, this number is extremely small ($\sim 2.8 \times 10^{-7}$, so the normal way of sampling gives an estimate with huge error.

Solution: Importance sampling. Use a proposal distribution p, $x_1, \ldots, x_n \sim p$, and let

$$w_i = \frac{\pi(x_i)}{p(x_i)}, i = 1, 2, \ldots, n,$$

then the output is

$$\hat{\theta}^{IS} = \frac{1}{n} \sum_{i=1}^{n} w_i(x_i)\phi(x_i).$$

And we will show

$$E(\hat{\theta}^{IS}) = E(\theta).$$

Proof.

$$\begin{aligned} E(\hat{\theta}^{IS}) &= \frac{1}{n} \sum_{i=1}^{n} E\left(\frac{\pi(x_i)}{p(x_i)}\phi(x_i)\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\int_t \frac{\pi(t)}{p(t)}\phi(t)p(t)\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\int_t \pi(t)\phi(t)\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} E[\pi(x_i)p(x_i)] \\ &= E(\phi(x)) = \theta. \end{aligned}$$

□

We want to learn the variance of this new estimator. Recall that

$$Var(X) = E[X^2] - (E[X])^2.$$

So, we can calculate

$$\begin{aligned} Var(\hat{\theta}^{IS}) &= E[w_i^2(x)\phi^2(x)] - \theta^2 \\ &= E\left[\frac{\pi^2(x)}{p^2(x)}\phi^2(x)\right] - \theta^2 \end{aligned}$$

If $p(x)$ is small, this variance explodes. Therefore, to control the variance of $\hat{\theta}^{IS}$, we want to choose the proposal density p carefully. Another real world problem: In Bayesian analysis, we may not know the normalization constants, for example, if we know

$$\pi(x) \propto e^{-x^4/2} + 0.15e^{-(x-6)^2/0.12^2},$$

It could be hard to find the constant such that $\int \pi(x) = 1$. Then, you cannot directly apply importance sampling because the ratio $\pi(x)/p(x)$ cannot be calculated. So we introduce self normalized importance sampling estimator.

2 Normalized (Self-Normalized) Importance Sampling

Sometimes, $\pi(x)$ is only known up to a constant, or it's numerically more stable to use normalized weights:

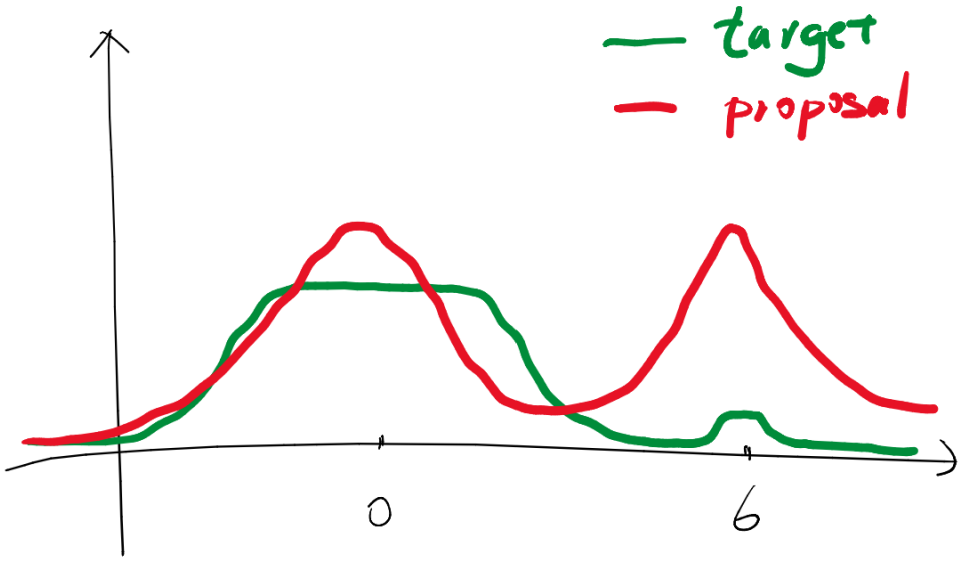
$$\hat{\theta}^{SNIS} = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}\phi(x_i),$$

where

$$\tilde{w}_i = \frac{w_i}{\sum_{i=1}^{n} w_i}.$$

This estimator is **biased but consistent**, meaning:

$$\mathbb{E}[\widehat{\theta}_{SNIS}] \neq \theta, \quad \text{but } \widehat{\theta}_{SNIS} \to \theta \text{ as } n \to \infty.$$



For example:

To sample from the target distribution described above, we can use

$$p = \alpha N(0,1) + (1-\alpha)N(6,1),$$

where $0 < \alpha < 1$. This ensures that all possible values of the target distribution are properly sampled.

$$ESS = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}.$$

Interpretation:

- $ESS \approx n$ (If the proposal distribution is a perfect match for target, all the importance weights will be roughly equal. In this case, the ESS will be close to actual sample size, n)
- $ESS \approx 1$ (If your proposal distribution is a poor match, you might get a situation where one sample has a massive weight and the other 999 have tiny weights. Your final estimate will be dominated by that single sample. Here, the ESS will be close to 1)

In practice, $ESS \approx 50\% \times n$ already means that p is pretty good.

Now let's try to estimate

$$\theta = P(x > 5),$$

where $x \sim N(0,1)$. We know that $\theta \approx 2.8 \times 10^{-7}$, so we need to be careful choosing p. Note that we are interested in the region $[5, \infty)$, we want to find a distribution whose samples more likely falls in that region. We can take a shift: $p \sim N(3,1)$, then

$$w(x) = e^{(3x-2)}.$$

Using importance sampling with such p is 60% to 70% more effective than the naive approach.

Another sampling scheme: Dependent sampling: given a stream of iid samples, create $x_1, x_2, \ldots$ such that x_{i+1} depends on x_i, but distribution of x_i becomes stationary when $i \to \infty$. This will be the main idea of MCMC which will be discussed in the next few lectures.

3 Choosing a Good Proposal

The proposal $p(x)$ should place enough probability mass in regions where $\phi(x)\pi(x)$ is large. For example, in the tail probability case $\theta = P(X > 5)$, we could use:

$$p(x) = N(3,1),$$

since this distribution samples more often in the region $x > 5$.

For this case,

$$w(x) = \frac{\pi(x)}{p(x)} = e^{(3x-2)}.$$

This shifts more samples into the region of interest, giving a more accurate estimate. The proposal distribution $p(x)$ should overlap well with the important region of $\pi(x)$ contributing most to θ. If $p(x)$ covers those regions poorly, variance increases drastically.