1 Importance Sampling

In many applications we want to compute $\mu = E(f(\boldsymbol{X}))$ where $f(\boldsymbol{x})$ is **nearly zero outside a region** A for which $P(\boldsymbol{X} \in A)$ is **small**. The set A may have **small volume**, or it may be in the tail of the distribution of $\boldsymbol{X}$. A **plain** Monte Carlo sample from the distribution of $\boldsymbol{X}$ could **fail to have even one point** inside the region A. It is clear intuitively that we must get **some samples** from the **interesting or important** region. We do this by sampling from a distribution that **over-weights the important region**, hence the name importance sampling. Having oversampled the important region, we **have to adjust our estimate** somehow to account for having sampled from this other distribution.

1.1 Basic Importance Sampling

Suppose that our problem is to find $\mu = E(f(\boldsymbol{X})) = \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ where p is a probability density function on $\mathcal{D} \subset \mathbb{R}^d$. We take $p(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \notin \mathcal{D}$. If q is a positive probability density function on $\mathbb{R}^d$, then

$$\mu = \int_{\mathcal{D}} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathcal{D}} \frac{f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})} q(\boldsymbol{x})d\boldsymbol{x} = E_q\left(\frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}\right),$$

where $E_q(\cdot)$ denotes **expectation** for $\boldsymbol{X} \sim q$. We also write $E_q(\cdot)$ and $Var_q(\cdot)$ for expectation and variance, respectively, when $\boldsymbol{X} \sim q$. Our **original goal** then is to find $E_p(f(\boldsymbol{X}))$. By making a **multiplicative adjustment** to f we **compensate** for sampling from q **instead** of p. The **adjustment factor** $p(\boldsymbol{x})/q(\boldsymbol{x})$ is called the **likelihood ratio**. The distribution q and p are called the **importance distribution** and the **nominal distribution**, respectively. The importance sampling estimate of $\mu = E_p(f(\boldsymbol{X}))$ is

$$\widehat{\mu}_{\text{imp}} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)} = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{X}_i),$$

where $h(\boldsymbol{x}) = \dfrac{f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})}$ and $\boldsymbol{X}_i \sim q$.

It is easy to see that $\widehat{\mu}_{\text{imp}}$ is **unbiased** for μ, as

$$E\left(\widehat{\mu}_{\text{imp}}\right) = E_q\left(h(\boldsymbol{X})\right) = \mu.$$

The **variance** of $\widehat{\mu}_{\text{imp}}$ can be expressed as σ_q^2/n, where

$$\sigma_q^2 = Var\left(h\left(\boldsymbol{X}\right)\right) = \int_{\mathcal{D}} \frac{f^2(\boldsymbol{x})p^2(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x} - \mu^2 = \int_{\mathcal{D}} \frac{(f(\boldsymbol{x})p(\boldsymbol{x}) - \mu q(\boldsymbol{x}))^2}{q(\boldsymbol{x})}d\boldsymbol{x}.$$

To construct a **confidence interval** for μ, we need to estimate σ_q^2. The natural variance estimator is

$$\widehat{\sigma}_q^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f(\boldsymbol{X}_i)p(\boldsymbol{X}_i)}{q(\boldsymbol{X}_i)} - \widehat{\mu}_{\text{imp}}\right)^2.$$

Therefore, an asymptotic 99% confidence interval for μ is $\widehat{\mu}_{\text{imp}} \mp 2.58\widehat{\sigma}_q^{\cancel{2}}/\sqrt{n}$.

Remark 1. The importance distribution q **does not have to be positive** everywhere. It is **enough** to have $q(\boldsymbol{x}) > 0$ **whenever** $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. †

Remark 2. The expression for the variance of $\widehat{\mu}_{\text{imp}}$ guides us in selecting a good importance sampling rule. The first expression of σ_q^2 suggests that a **better** q is **one that gives a smaller value** of $\int_{\mathcal{D}}(fp)^2/q d\boldsymbol{x}$.

The second integral expression of σ_q^2 illustrates **how importance sampling can succeed or fail**. The numerator in the **integrand is small** when $f(\boldsymbol{x})p(\boldsymbol{x}) - \mu q(\boldsymbol{x})$ is **close to zero**, that is, when $q(\boldsymbol{x})$ is **nearly proportional** to $f(\boldsymbol{x})p(\boldsymbol{x})$. From the denominator, we see that regions with **small values** of $q(\boldsymbol{x})$ **greatly magnify** whatever **lack of proportionality** appears in the numerator. †

Example 1. (Gaussian p and q: A word of caution) The effect of **light-tailed** q can be illustrated by this example. Suppose that $f(x) = x$, and $p(x) = \exp(-x^2/2)/\sqrt{2\pi}$. If $q(x) = \exp(-x^2/(2\sigma^2))/(\sigma\sqrt{2\pi})$ with $\sigma > 0$ then

$$\begin{aligned}
\sigma_q^2 &= \int_{-\infty}^{\infty} x^2 \frac{\left(\exp(-x^2/2)/\sqrt{2\pi}\right)^2}{\exp(-x^2/(2\sigma^2))/(\sigma\sqrt{2\pi})}dx \\
&= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 \exp(-x^2(2-\sigma^{-2})/2)dx \\
&= \begin{cases} \frac{\sigma}{(2-\sigma^{-2})^{3/2}} & \text{if } \sigma^2 > \frac{1}{2} \\ \infty & \text{otherwise.} \end{cases}
\end{aligned}$$

||

1.2 Self-normalized Importance Sampling

Sometimes we can **only compute an unnormalized version** of p, $p_u(\boldsymbol{x}) = cp(\boldsymbol{x})$, where $c > 0$ is **unknown**. Also suppose that we can compute $q_u(\boldsymbol{x}) = bq(\boldsymbol{x})$, where $b > 0$ might be **unknown**. If we are fortunate or clever enough to have $b = c$, then $p(\boldsymbol{x})/q(\boldsymbol{x}) = p_u(\boldsymbol{x})/q_u(\boldsymbol{x})$ and we can still use $\widehat{\mu}_{\text{imp}}$. Otherwise we may compute the **ratio** $w_u(\boldsymbol{x}) = p_u(\boldsymbol{x})/q_u(\boldsymbol{x}) = (c/b)p(\boldsymbol{x})/q(\boldsymbol{x})$ and consider the **self-normalized importance sampling estimator**

$$\tilde{\mu}_{\text{imp}} = \frac{\sum_{i=1}^{n} f(\boldsymbol{X}_i)w_u(\boldsymbol{X}_i)}{\sum_{i=1}^{n} w_u(\boldsymbol{X}_i)} = \frac{\sum_{i=1}^{n} f(\boldsymbol{X}_i)w(\boldsymbol{X}_i)}{\sum_{i=1}^{n} w(\boldsymbol{X}_i)}.$$

In general $\tilde{\mu}_{\text{imp}}$ is a **biased** estimator of μ.

Theorem 1. *Let p be a probability density function on $\mathbb{R}^d$ and let $f(\boldsymbol{x})$ be a function such that $\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ exists. Suppose that $q(\boldsymbol{x})$ is a probability density function on $\mathbb{R}^d$ with $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$. Let $X_1, \ldots, X_n \sim q$ be independent and let $\tilde{\mu}_{imp}$ be the self-normalized importance sampling estimator. Then*

$$P\left(\lim_{n\to\infty} \tilde{\mu}_{imp} = \mu\right) = 1.$$

Proof. The proof is simple using strong law of large numbers.

Remark 3. The **self-normalized** importance sampler $\tilde{\mu}_{\text{imp}}$ **requires a stronger condition** on q than the unbiased importance sampler $\widehat{\mu}_{\text{imp}}$ does. We now need $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$ even if $f(\boldsymbol{x})$ is zero. †

1.3 Importance Sampling Diagnostic

Importance sampling uses **unequally weighted** observations. The **weights** are $w_i = p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i) \geq 0$ for $i = 1, \ldots, n$. In extreme settings, one of the w_i may be **vastly larger** than all the others and then we have **effectively only got one observation**. We would like to have a **diagnostic** to tell when the weights are problematic. It is even possible that $w_1 = w_2 = \ldots = w_n = 0$. In that case, importance sampling has clearly failed and we do not need a diagnostic to tell us so. Hence, we may **assume** that $\sum_{i=1}^n w_i > 0$.

Consider a hypothetical linear combination

$$S_w = \frac{\sum_{i=1}^n w_i Z_i}{\sum_{i=1}^n w_i},$$

where Z_i are independent random variables with common mean and common variance $\sigma^2 > 0$ and $w_i > 0$ are weights. The **unweighted average** of n_e independent random variables Z_i has variance σ^2/n_e. Setting $Var(S_w) = \sigma^2/n_e$ and solving for n_e yields the **effective sample size**

$$n_e = \frac{\left(\sum_{i=1}^n w_i\right)^2}{\sum_{i=1}^n w_i^2} = \frac{n\bar{w}^2}{\overline{w^2}}.$$

If the **weights are too imbalanced** then the result is similar to averaging only $n_e \ll n$ observations and might therefore be **unreliable**. The point at which n_e becomes alarmingly small is hard to specify, because it is application specific.