

Self-normalized Importance Sampling

Estimating expectations with unknown normalizing constants

Junqi Tang

Computational Statistics
University of Birmingham

Review of variance minimization

Choose a proposal $q(x; \eta)$ from a parameterized family.

$$\hat{\mu}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^n f(X_i) \frac{p(X_i)}{q(X_i; \eta)}, \quad X_i \stackrel{\text{iid}}{\sim} q(\cdot; \eta).$$

- Select η to reduce the estimator's variance.
- The objective usually involves an expectation.
- A sample average approximation gives a finite-sum objective.

We must also be able to sample from the chosen proposal.

Review of exponential tilting

For a natural exponential family,

$$p(x; \vartheta) = h(x) \exp\{\vartheta^\top x - A(\vartheta)\}.$$

An admissible linear exponential tilt gives

$$q_\theta(x) \propto e^{\theta^\top x} p(x; \vartheta_0) \implies q_\theta(x) = p(x; \vartheta_0 + \theta).$$

- The proposal stays in the same family.
- A sampler for that family can often be reused.

For $p = N(\mu, \Sigma)$, the tilted law is $N(\mu + \Sigma\theta, \Sigma)$.

A density known up to a constant

Suppose we can evaluate $p_0(x)$, but only know

$$\boxed{p(x) = c_p p_0(x)} \quad \text{or equivalently} \quad p(x) \propto p_0(x).$$

The density must integrate to one:

$$1 = c_p \int p_0(x) dx, \quad c_p = \left(\int p_0(x) dx \right)^{-1}.$$

- p_0 is known and nonnegative.
- The integral is positive and finite, but may be hard to compute.

The shape is available even when its normalizing constant is not.

A simple illustration with a normal density

Suppose we are given only

$$p_0(x) = e^{-x^2/2}, \quad x \in \mathbb{R}.$$

Here we can evaluate the integral explicitly:

$$\int_{-\infty}^{\infty} p_0(x) dx = \sqrt{2\pi}.$$

Therefore

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

For a complicated high-dimensional p_0 , this integral can be impractical.

Bayesian posteriors often have this form

With observed data y and unknown parameter θ ,

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}, \quad p(y) = \int p(y \mid \theta)p(\theta) d\theta.$$

- We may be able to evaluate the likelihood and prior.
- The integral $p(y)$ can be difficult to evaluate.

$$p(\theta \mid y) \propto \underbrace{p(y \mid \theta)p(\theta)}_{\text{computable unnormalized posterior}}.$$

The target and the proposal

We want

$$\mu = \mathbb{E}_p[f(X)], \quad p = c_p p_0.$$

Choose a proposal

$$q = c_q q_0.$$

- We can evaluate p_0 and q_0 , and can sample from q .
- If q is known exactly, set $q_0 = q$ and $c_q = 1$.
- Require $q > 0$ wherever $p > 0$, up to null sets.
- Assume $\mathbb{E}_p[|f(X)|] < \infty$.

The expectation as a ratio

Since p is a probability density,

$$\int p(x) dx = 1.$$

Therefore

$$\mu = \int f(x)p(x) dx = \frac{\int f(x)p(x) dx}{\int p(x) dx}.$$

We will estimate both integrals using the proposal q .

At this stage, the ratio is an exact identity.

A change of measure in both integrals

Multiply and divide each integrand by $q(x)$:

$$\mu = \frac{\int f(x) \frac{p(x)}{q(x)} q(x) dx}{\int \frac{p(x)}{q(x)} q(x) dx}.$$

Define $w(x) = p(x)/q(x)$. Then

$$\mu = \frac{\mathbb{E}_q[f(X)w(X)]}{\mathbb{E}_q[w(X)]}.$$

Although $\mathbb{E}_q[w(X)] = 1$, retaining the denominator will be useful.

Computable unnormalized weights

The ordinary importance weight is

$$w(x) = \frac{p(x)}{q(x)} = \frac{c_p}{c_q} \frac{p_0(x)}{q_0(x)}.$$

Define

$$w_0(x) = \frac{p_0(x)}{q_0(x)}.$$

- $w_0(x)$ is computable from the known functions.
- $w(x) = (c_p/c_q)w_0(x)$ contains an unknown common factor.

The unknown constants cancel

Substitute $w = (c_p/c_q)w_0$ into the exact ratio:

$$\begin{aligned}\mu &= \frac{\mathbb{E}_q[f(X)(c_p/c_q)w_0(X)]}{\mathbb{E}_q[(c_p/c_q)w_0(X)]} \\ &= \frac{(c_p/c_q)\mathbb{E}_q[f(X)w_0(X)]}{(c_p/c_q)\mathbb{E}_q[w_0(X)]} \\ &= \boxed{\frac{\mathbb{E}_q[f(X)w_0(X)]}{\mathbb{E}_q[w_0(X)]}}.\end{aligned}$$

Neither normalizing constant appears in the final expression.

The self-normalized importance sampling estimator

Generate $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} q$.

Replace each expectation by a sample average:

$$\hat{\mu}_{\text{SNIS}} = \frac{\frac{1}{n} \sum_{i=1}^n f(X_i) w_0(X_i)}{\frac{1}{n} \sum_{i=1}^n w_0(X_i)}.$$

The factors $1/n$ cancel:

$$\hat{\mu}_{\text{SNIS}} = \frac{\sum_{i=1}^n f(X_i) w_0(X_i)}{\sum_{i=1}^n w_0(X_i)}.$$

Normalized weights sum to one

Define sample-normalized weights

$$\bar{w}_i = \frac{w_0(X_i)}{\sum_{j=1}^n w_0(X_j)}.$$

Then

$$\bar{w}_i \geq 0, \quad \sum_{i=1}^n \bar{w}_i = 1, \quad \boxed{\hat{\mu}_{\text{SNIS}} = \sum_{i=1}^n \bar{w}_i f(X_i)}.$$

SNIS is a weighted average of the sampled function values.

Any common positive multiplier of the weights cancels.

The computational procedure

1. Generate independent samples $X_i \sim q$, for $i = 1, \dots, n$.
2. Compute $w_i = p_0(X_i)/q_0(X_i)$.
3. Check that $S = \sum_{i=1}^n w_i > 0$ and set $\bar{w}_i = w_i/S$.
4. Return $\hat{\mu}_{\text{SNIS}} = \sum_{i=1}^n \bar{w}_i f(X_i)$.

If all sampled weights are zero, the displayed ratio is undefined.

Sampling from the proposal remains necessary

Cancellation of c_q allows us to evaluate the estimator.

We still need a method that generates samples from q .

- A familiar normalized proposal often has an available sampler.
- An arbitrary unnormalized q_0 does not automatically provide one.
- Markov chain methods can work with unnormalized densities, but produce dependent samples.

The arguments here use independent samples from q .

Ordinary importance sampling is unbiased

When density ratios $w = p/q$ are available,

$$\hat{\mu}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^n f(X_i) w(X_i).$$

Under the support and integrability conditions,

$$\mathbb{E}[\hat{\mu}_{\text{IS}}] = \mathbb{E}_q[f(X)w(X)] = \mathbb{E}_p[f(X)] = \mu.$$

For SNIS, both the numerator and denominator are random.

$$\mathbb{E}\left[\frac{A_n}{B_n}\right] \text{ need not equal } \frac{\mathbb{E}[A_n]}{\mathbb{E}[B_n]}.$$

The one-sample argument

Take $n = 1$ and suppose $w_0(X_1) > 0$ almost surely.

$$\hat{\mu}_{\text{SNIS}} = \frac{f(X_1)w_0(X_1)}{w_0(X_1)} = f(X_1).$$

Since $X_1 \sim q$,

$$\mathbb{E}[\hat{\mu}_{\text{SNIS}}] = \mathbb{E}_q[f(X)],$$

provided this expectation exists. Our target is

$$\mu = \mathbb{E}_p[f(X)].$$

Whenever these expectations differ, the one-sample estimator is biased.

SNIS is generally biased at finite sample sizes

The one-sample argument disproves universal unbiasedness.

It does not establish bias for every possible p , q , f and n .

- If $q = p$, all weights are equal and SNIS becomes ordinary Monte Carlo.
- If $f(x) = a$ is constant, SNIS returns a whenever defined.
- Different distributions can give the same expectation of a particular f .

Consistency follows from the law of large numbers

Write

$$A_n = \frac{1}{n} \sum_{i=1}^n f(X_i) w_0(X_i), \quad B_n = \frac{1}{n} \sum_{i=1}^n w_0(X_i).$$

Under the stated assumptions, almost surely,

$$A_n \longrightarrow \mathbb{E}_q[f w_0], \quad B_n \longrightarrow \mathbb{E}_q[w_0] > 0.$$

Therefore

$$\boxed{\hat{\mu}_{\text{SNIS}} = \frac{A_n}{B_n} \longrightarrow \frac{\mathbb{E}_q[f w_0]}{\mathbb{E}_q[w_0]} = \mu.}$$

The estimator is strongly consistent.

Asymptotic unbiasedness needs another condition

Asymptotic unbiasedness means

$$\text{Bias}(\hat{\mu}_{\text{SNIS}}) = \mathbb{E}[\hat{\mu}_{\text{SNIS}}] - \mu \longrightarrow 0.$$

Almost-sure convergence alone does not justify taking expectations of limits.

A simple sufficient setting is:

- $|f(x)| \leq M$ for some finite constant M .
- The total weight is positive almost surely for every n .

Then $|\hat{\mu}_{\text{SNIS}}| \leq M$, so dominated convergence gives

$$\mathbb{E}[\hat{\mu}_{\text{SNIS}}] \longrightarrow \mu.$$

Accuracy depends on both bias and variance

Whenever the second moment is finite,

$$\text{MSE}(\hat{\mu}) = \text{Var}(\hat{\mu}) + \text{Bias}(\hat{\mu})^2.$$

- For an unbiased estimator, only the variance contributes.
- For SNIS, both finite-sample bias and variance can contribute.
- With bounded f and positive total weights, both vanish as $n \rightarrow \infty$.

This does not imply monotone improvement at every sample size.

What self-normalization makes possible

For $p = c_p p_0$ and $q = c_q q_0$, we can estimate

$$\mu = \mathbb{E}_p[f(X)]$$

using

$$\hat{\mu}_{\text{SNIS}} = \frac{\sum_{i=1}^n f(X_i) p_0(X_i) / q_0(X_i)}{\sum_{i=1}^n p_0(X_i) / q_0(X_i)}, \quad X_i \stackrel{\text{iid}}{\sim} q.$$

- Unknown normalizing constants cancel.
- Finite-sample bias is generally present.
- Consistency and asymptotic unbiasedness have distinct assumptions.