Statistics 700 Homework 3

Importance Sampling

Due date: 6:00 pm (EST) Oct. 17, 2017

ESS for importance sampling. In class, we talked about the effective sample (ESS) size of an importance sampling with m samples defined as

$$ESS(m) = \frac{m}{1 + var_g[w(x)]}.$$

This problem will walk you through the mathematical derivations and examples to help you build intuitions towards the definition of the effective sample size.

Suppose we are interested in evaluating

$$\mu = E_\pi[h(x)] = \int h(x)\pi(x)dx.$$

Ideally, we can obtain i.i.d. samples from $\pi(\cdot)$, denoted by $\{y^{(j)}\}_{1\leq j\leq m}$. Thus an estimate of μ is

$$\hat{\mu}^\star = \frac{\sum_{j=1}^m h(y^{(j)})}{m}.$$

However, in most real applications, we can not sample directly from $\pi(\cdot)$ thus importance sampling with a trial function $g(\cdot)$ is adopted: let $\{x^{(j)}\}_{1\leq j\leq m}$ be i.i.d. samples from $g(\cdot)$, then we estimate μ by

$$\hat{\mu} = \frac{\sum_{j=1}^m h(x^{(j)})w(x^{(j)})}{\sum_{j=1}^m w(x^{(j)})},$$

where $w(x) = \pi(x)/g(x)$ for all x. Define $H_w = h(x)w(x)$, $H = h(x)$, $W = w(x)$.

1. Show that $E_g(w(x)) = 1$ and $E_g(h(x)w(x)) = \mu$.

2. For m large, apply the delta method and show that

$$E_g(\hat{\mu}) \approx \mu - \frac{Cov_g(W, H_w)}{m} + \frac{\mu Var_g(W)}{m}; \tag{1}$$

$$Var_g(\hat{\mu}) \approx m^{-1}\left[\mu^2 Var_g(W) + Var_g(H_w) - 2\mu Cov_g(W, H_w)\right]. \tag{2}$$

When $E_\pi\left[(H-\mu)^2(W - E_\pi(W))\right]$ is small as compared to $Var_\pi(H)\left[Var_g(W)+1\right]$,

$$\frac{Var(\hat{\mu}^\star)}{Var(\hat{\mu})} \approx \frac{1}{1+Var_g(W)}.$$

3. Let $\overline{H} = m^{-1}\sum_{j=1}^m h(x^{(j)})w(x^{(j)})$, $\overline{W} = m^{-1}\sum_{j=1}^m w(x^{(j)})$, show that
   (a) $E(\overline{H}) = \mu$, $E(\overline{W}) = 1$, thus $\overline{H}$ is also an unbiased estimate of μ;
   (b) using the approximations given in Equations (1) and (2), compare the mean squared error of estimating μ using $\hat{\mu}$ and $\overline{H}$, $MSE := E_g(\star - \mu)^2$; under what condition is $\hat{\mu}$ better than $\overline{H}$ in terms of mean squared error?

4. Examples.
   (a) Suppose $\pi(x) \propto$ standard Gaussian density, $g(x) \propto$ student t distribution with 2 degrees of freedom, $h(x) = x$. Implement an importance sampler and calculate $ESS(m)$ for $m = 50, 100, 200, 500, 1000$.
   (b) Suppose $g(x) \propto$ standard Gaussian density, $\pi(x) \propto$ student t distribution with 2 degrees of freedom, $h(x) = x$. Implement an importance sampler and calculate $ESS(m)$ for $m = 50, 100, 200, 500, 1000$.
   (c) What do you find by comparing the results above?
   (d) Suppose the target density is

$$\pi(\mu, \sigma^2) \propto \sigma^{-5} \exp\left[-\frac{(\mu-1)^2+4}{2\sigma^2}\right],$$

   where $(\mu, \sigma^2) \in [-3, 5] \times [0.01, 50]$.
      i. Make a contour plot of the target density in the specified range.
      ii. Given the target density and the contour plot, how can you choose a good importance function? Design an importance sampling procedure and estimate the ESS for several different sample sizes. Compare your samples with that obtained from grid sampling.

iii. If μ is the only quantity of interest, does marginalization help with obtaining better importance samples? If so, can you verify it?

5. Optional (bonus points $\leq$5). Can you verify the validity of the definition of ESS via repeated simulations? You can use any density of your choice. Further, do you see examples when the definition of ESS does not work as desired? If so, can you come up with an intuitive / mathematical argument to justify?

Rejection control algorithm. In classed, we talked about the rejection control algorithm which has the following two steps:

1. For $j = 1, \ldots, m$, accept $x^{(j)}$ with probability

$$r^{(j)} = \min\left\{1, \frac{w^{(j)}}{c}\right\},$$

where $w^{(j)} = \pi(x^{(j)})/g(x^{(j)})$.

2. If the jth sample $x^{(j)}$ is accepted, its weight is updated to $w^{(*j)} = q_c w^{(j)}/r^{(j)}$, where

$$q_c = \int \min\left\{1, \frac{w(x)}{c}\right\} g(x)dx,$$

where $w(x) = \pi(x)/g(x)$.

This problem walks you through some properties of this algorithm.

1. Show that the modified trial distribution in the rejection control is

$$g^*(x) = q_c^{-1} \min\{g(x), \pi(x)/c\}.$$

2. Prove that this method produces random variables that follow the target distribution $\pi(\cdot)$.

3. Prove the following theorem by following the steps given below.

Theorem RC. The rejection control method indeed reduces the χ^2 distance between the target distribution and the modified trial distribution, i.e.

$$Var_{g^*}[\pi(x)/g^*(x)] \leq Var_g[\pi(x)/g(x)].$$

(a) Prove that

$$1 + Var_{g^*}[\pi(x)/g^*(x)] = q_c E_g \left[\max\left\{w(x), c\right\} w(x)\right].$$

(b) Prove that the two random variables $\min\left\{w(x), c\right\}$ and $\max\left\{w(x), c\right\} w(x)$ are positively correlated by showing that for any $w_1 > 0, w_2 > 0$,

$$h(w_1, w_2) = [\min\left\{w_1, c\right\} - \min\left\{w_2, c\right\}][w_1 \max\left\{w_1, c\right\} - w_2 \max\left\{w_2, c\right\}] \geq 0.$$

(c) Prove that

$$q_c E_g \left[\max\left\{w(x), c\right\} w(x)\right] \leq 1 + Var_g[\pi(x)/g(x)].$$

Hint: notice that $\max\left\{w(x), c\right\} \min\left\{w(x), c\right\} = cw(x)$.

4. Example. Suppose $\pi(x) \propto$ Gaussian density with mean 0 and standard deviation 0.3, $g(x) \propto$ student t distribution with 2 degrees of freedom.

    (a) Implement the importance sampling with and without rejection control.

    (b) Do you see improvement by using rejection control with different c?

    (c) What value of c do you choose to use finally and why?

    (d) Repeat the above procedures when $\pi(x) \propto$ Gaussian density with mean 0 and standard deviation 3.

    (e) Explain intuitively how to choose a c such that the rejection control importance sampling gives satisfactory results.

Optional Reading. Read one of the following papers and post your summary and thoughts on Canvas. Bonus points up to 5 will be rewarded.

1. Adaptive importance sampling in Monte Carlo integration, Oh, M.-S. and Berger, J. O. (1992), Journal of Statistical Computation and Simulation 41: 450-456.

2. Sequential imputation and Bayesian missing data problems, Kong, A., Liu, J. S., and Wong, W. H. (1994), Journal of the American Statistical Association 89, 278-288.

3. Truncated importance sampling, Ionides, E. L. (2008), Journal of Computational and Graphical Statistics 17, 295-311.