

Bayesian Theory and Computation

Lecture 8: Importance Sampling



Cheng Zhang

School of Mathematical Sciences, Peking University

March 25, 2022

- ▶ While Monte Carlo estimation is attractive for high dimension integration, it may suffer from lots of problems, such as rare events, and irregular integrands, etc.
- ▶ In this lecture, we will discuss various methods to improve Monte Carlo approaches, with an emphasis on variance reduction techniques

- ▶ The simple Monte Carlo estimator of $\int_a^b h(x)f(x)dx$ is

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^n h(x^{(i)})$$

where $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ are randomly sampled from f

- ▶ A potential problem is the **mismatch** of the concentration of $h(x)f(x)$ and $f(x)$. More specifically, if there is a region A of relatively small probability under $f(x)$ that dominates the integral, we would not get enough data from the **important** region A by sampling from $f(x)$
- ▶ Main idea: Get more data from A , and then correct the bias

- ▶ **Importance sampling** (IS) uses importance distribution $q(x)$ to adapt to the true integrands $h(x)f(x)$, rather than the target distribution $f(x)$
- ▶ By correcting for this bias, importance sampling can greatly reduce the variance in Monte Carlo estimation
- ▶ Unlike the rejection sampling, we do not need the envelop property
- ▶ The only requirement is that $q(x) > 0$ whenever

$$h(x)f(x) \neq 0$$

- ▶ IS also applies when $f(x)$ is not a probability density function

- Now we can rewrite $I = \mathbb{E}_f(h(x)) = \int_{\mathcal{X}} h(x)f(x) dx$ as

$$\begin{aligned} I = \mathbb{E}_f(h(x)) &= \int_{\mathcal{X}} h(x)f(x) dx \\ &= \int_{\mathcal{X}} h(x)\frac{f(x)}{q(x)}q(x)dx \\ &= \int_{\mathcal{X}} (h(x)w(x))q(x) \\ &= \mathbb{E}_q(h(x)w(x)) \end{aligned}$$

where $w(x) = \frac{f(x)}{q(x)}$ is the **importance weight** function

We can then approximate the original expectation as follows

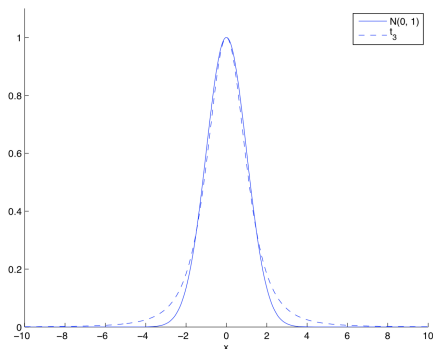
- ▶ Draw samples $x^{(1)}, \dots, x^{(n)}$ from $q(x)$
- ▶ Monte Carlo estimate

$$I_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^n h(x^{(i)}) w(x^{(i)})$$

where $w(x^{(i)}) = \frac{f(x^{(i)})}{q(x^{(i)})}$ are called importance ratios.

- ▶ Note that, now we only require sampling from q and do not require sampling from f

- ▶ We want to approximate a $\mathcal{N}(0, 1)$ distribution with $t(3)$ distribution



- ▶ We generate 500 samples and estimated $I = \mathbb{E}(x^2)$ as 0.97, which is close to the true value 1.

- Let $t(x) = h(x)w(x)$. Then $\mathbb{E}_q(t(X)) = I, X \sim q$

$$\mathbb{E}(I_n^{\text{IS}}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(t(x^{(i)})) = I$$

- Similarly, the variance is

$$\begin{aligned} \text{Var}_q(I_n^{\text{IS}}) &= \frac{1}{n} \text{Var}_q(t(X)) \\ &= \frac{1}{n} \int_{\mathcal{X}} \frac{(h(x)f(x))^2}{q(x)} dx - I^2 \end{aligned} \quad (1)$$

$$= \frac{1}{n} \int_{\mathcal{X}} \frac{(h(x)f(x) - Iq(x))^2}{q(x)} dx \quad (2)$$



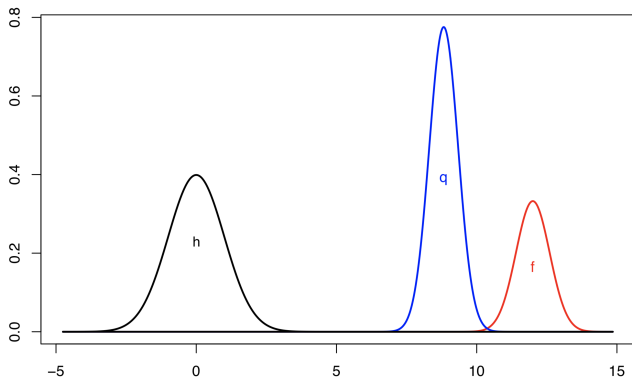
- ▶ Recall the convergence rate for Monte Carlo is

$$p\left(|\hat{I}_n - I| \leq \frac{\sigma}{\sqrt{n\delta}}\right) \geq 1 - \delta, \quad \forall \delta$$

For IS, $\sigma = \sqrt{\text{Var}_q(t(X))}$. A good importance distribution $q(x)$ would make $\text{Var}_q(t(X))$ small.

- ▶ What can we learn from equations (1) and (2)?
 - ▶ **Optimal choice:** $q(x) \propto h(x)f(x)$
 - ▶ $q(x)$ near 0 can be **dangerous**
 - ▶ **Bounding** $\frac{(h(x)f(x))^2}{q(x)}$ is useful theoretically





$$\text{Var}_q(t(X)) = 0$$

Gaussian h and $f \Rightarrow$ Gaussian optimal q lies between.



- ▶ When f or/and q are unnormalized, we can estimate the expectation as follows

$$I = \frac{\int_{\mathcal{X}} h(x) f(x) dx}{\int_{\mathcal{X}} f(x) dx} = \frac{\int_{\mathcal{X}} h(x) \frac{f(x)}{q(x)} q^*(x) dx}{\int_{\mathcal{X}} \frac{f(x)}{q(x)} q^*(x) dx}$$

where $q^*(x) = q(x)/c_q$

- ▶ Monte Carlo estimate

$$I_n^{\text{SNIS}} = \frac{\sum_{i=1}^n h(x^{(i)}) w(x^{(i)})}{\sum_{i=1}^n w(x^{(i)})}, \quad x^{(i)} \sim q(x)$$

- ▶ Requires a stronger condition: $q(x) > 0$ whenever $f(x) > 0$

- Unfortunately, I_n^{SNIS} is biased. However, the bias is asymptotically negligible.

$$\begin{aligned} I_n^{\text{SNIS}} &= \frac{1}{n} \sum_{i=1}^n h(x^{(i)}) f(x^{(i)}) / q(x^{(i)}) \bigg/ \frac{1}{n} \sum_{i=1}^n f(x^{(i)}) / q(x^{(i)}) \\ &\xrightarrow{p} \int_{\mathcal{X}} h(x) f(x) / q(x) q^*(x) dx \bigg/ \int_{\mathcal{X}} f(x) / q(x) q^*(x) dx \\ &= \int_{\mathcal{X}} h(x) f(x) dx \bigg/ \int_{\mathcal{X}} f(x) dx \\ &= I \end{aligned}$$



- ▶ We use delta method for the variance of SNIS, which is a ratio estimate

$$\text{Var}(I_n^{\text{SNIS}}) \approx \frac{\sigma_{q,\text{sn}}^2}{n} = \frac{\mathbb{E}_q(w(x)^2(h(x) - I)^2)}{n}$$

- ▶ We can rewrite the variance $\sigma_{q,\text{sn}}^2$ as

$$\begin{aligned}\sigma_{q,\text{sn}}^2 &= \int_{\mathcal{X}} \frac{f(x)^2}{q(x)} (h(x) - I)^2 dx \\ &= \int_{\mathcal{X}} \frac{(h(x)f(x) - If(x))^2}{q(x)} dx\end{aligned}$$

- ▶ For comparison, $\sigma_{q,\text{is}}^2 = \text{Var}_q(t(X)) = \int_{\mathcal{X}} \frac{(h(x)f(x) - If(x))^2}{q(x)} dx$
- ▶ No q can make $\sigma_{q,\text{sn}}^2 = 0$ (unless h is constant)

- ▶ The optimal density for self-normalized importance sampling has the form (Hesterberg, 1988)

$$q(x) \propto |h(x) - I|f(x)$$

- ▶ Using this formula we find that

$$\sigma_{q,\text{sn}}^2 \geq (\mathbb{E}_f(|h(x) - I|))^2$$

which is zero only for constant $h(x)$

- ▶ Note that the simple Monte Carlo has variance $\sigma^2 = \mathbb{E}_f((h(x) - I)^2)$, this means SNIS can not reduce the variance by

$$\frac{\sigma^2}{\sigma_{q,\text{sn}}^2} \leq \frac{\mathbb{E}_f((h(x) - I)^2)}{(\mathbb{E}_f(|h(x) - I|))^2}$$



- ▶ The importance weights in IS may be problematic, we would like to have a diagnostic to tell us when it happens.
- ▶ Unequal weighting raises variance (Kong, 1992). For IID Y_i with variance σ^2 and fixed weight $w_i \geq 0$

$$\text{Var} \left(\frac{\sum_i w_i Y_i}{\sum_i w_i} \right) = \frac{\sum_i w_i^2 \sigma^2}{(\sum_i w_i)^2}$$

- ▶ Write this as

$$\frac{\sigma^2}{n_e} \text{ where } n_e = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$$

- ▶ n_e is the **effective sample size** and $n_e \ll n$ if the weights are too imbalanced.



- ▶ Rejection Sampling requires bounded $w(x) = f(x)/q(x)$
- ▶ We also have to know a bound for the envelop distribution
- ▶ Therefore, importance sampling is generally easier to implement
- ▶ IS and SNIS require us to keep track of weights
- ▶ Plain IS requires normalized p/q
- ▶ Rejection sampling could be sample inefficient (due to rejections)



- ▶ Consider that $f(x) = p(x; \theta_0)$ is from a family of distributions $p_\theta(x)$, $\theta \in \Theta$
- ▶ A simple importance sampling distribution would be $q(x) = p(x; \theta)$ for some $\theta \in \Theta$.
- ▶ Suppose $f(x)$ belongs to an exponential family

$$f(x) = g(x) \exp(\eta(\theta_0)^T T(x) - A(\theta_0))$$

- ▶ Use $q(x) = g(x) \exp(\eta(\theta)^T T(x) - A(\theta))$, the IS estimate is

$$I_n^{\text{IS}} = \exp(A(\theta) - A(\theta_0)) \cdot \frac{1}{n} \sum_{i=1}^n h(x^{(i)}) \exp((\eta(\theta_0) - \eta(\theta))^T T(x^{(i)}))$$



- ▶ Suppose that we find the mode x^* of $k(x) = h(x)f(x)$
- ▶ We can use Taylor approximation

$$\log(k(x)) \approx \log(k(x^*)) - \frac{1}{2}(x - x^*)^T H^*(x - x^*)$$
$$k(x) \approx k(x^*) \exp\left(-\frac{1}{2}(x - x^*)^T H^*(x - x^*)\right)$$

which suggests $q(x) = \mathcal{N}(x^*, (H^*)^{-1})$

- ▶ This requires positive definite H^*
- ▶ Can be viewed as an IS version of the **Laplace approximation**



- ▶ Suppose we have K importance distributions q_1, \dots, q_K , we can combine them into a mixture of distributions with probability $\alpha_1, \dots, \alpha_K$, $\sum_i \alpha_i = 1$

$$q(x) = \sum_{i=1}^K \alpha_i q_i(x)$$

- ▶ IS estimate $I_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^n h(x^{(i)}) \frac{f(x^{(i)})}{\sum_{j=1}^K \alpha_j q_j(x^{(i)})}$
- ▶ An alternative. Suppose $x^{(i)}$ came from component $j(i)$, we could use

$$\frac{1}{n} \sum_{i=1}^n h(x^{(i)}) \frac{f(x^{(i)})}{q_{j(i)}(x^{(i)})}$$

Remark: This alternative is **faster** to compute, but has **higher** variance



- ▶ Designing importance distribution directly would be challenging. A better way would be to adapt some candidate distribution to our task through a learning process
- ▶ To do that, we first need to pick a family \mathcal{Q} of proposal distributions
- ▶ We have to choose a termination criterion, e.g., maximum steps, total number of observations, etc.
- ▶ Most importantly, we need a way to choose $q_{k+1} \in \mathcal{Q}$ based on the observed information

- ▶ Suppose now we have a family of distributions (e.g., exponential family) $q_\theta(x) = q(x; \theta)$, $\theta \in \Theta$
- ▶ Recall that the variance of IS estimate is

$$\frac{1}{n} \int_{\mathcal{X}} \frac{(h(x)f(x))^2}{q(x)} dx - I^2, \quad \text{therefore, we would like}$$

$$\theta = \arg \min_{\theta \in \Theta} \int_{\mathcal{X}} \frac{(h(x)f(x))^2}{q_\theta(x)} dx$$

- ▶ Variance based update

$$\theta^{(k+1)} = \arg \min_{\theta \in \Theta} \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{(h(x^{(i)})f(x^{(i)}))^2}{q_\theta(x^{(i)})^2}, \quad x^{(i)} \sim q_{\theta^{(k)}}$$

However, the optimization may be hard.



- ▶ Consider an exponential family

$$q_{\theta}(x) = g(x) \exp(\theta^T x - A(\theta))$$

- ▶ Now, replace variance by KL divergence

$$D_{KL}(k_* \| q_{\theta}) = \mathbb{E}_{k_*} \log \left(\frac{k_*(x)}{q_{\theta}(x)} \right)$$

- ▶ We seek θ to minimize

$$D_{KL}(k_* \| q_{\theta}) = \mathbb{E}_{k_*} (\log(k_*(x)) - \log(q(x; \theta)))$$

i.e., maximize

$$\mathbb{E}_{k_*} (\log(q(x; \theta)))$$



- Rewrite the negative cross entropy as

$$\begin{aligned}\mathbb{E}_{k_*}(\log(q(x; \theta))) &= \mathbb{E}_q \left(\frac{\log(q(x; \theta))k_*(x)}{q(x)} \right) \\ &= \frac{1}{I} \cdot \mathbb{E}_q \left(\frac{\log(q(x; \theta))h(x)f(x)}{q(x)} \right)\end{aligned}$$

- Update θ to maximize the above

$$\begin{aligned}\theta^{(k+1)} &= \arg \max_{\theta} \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{h(x^{(i)})f(x^{(i)})}{q(x^{(i)}; \theta^{(k)})} \log(q(x^{(i)}; \theta)) \\ &= \arg \max_{\theta} \frac{1}{n_k} \sum_{i=1}^k H_i \log(q(x^{(i)}; \theta)) \\ &= \arg \max_{\theta} \frac{1}{n_k} \sum_{i=1}^k H_i (\theta^T x^{(i)} - A(\theta))\end{aligned}$$



- ▶ The update often takes a simple moment matching form

$$\frac{\partial}{\partial \theta} A(\theta^{(k+1)}) = \frac{\sum_i H_i(x^{(i)})^T}{\sum_i H_i}$$

- ▶ Examples:

- ▶ $q_\theta = \mathcal{N}(\theta, I)$

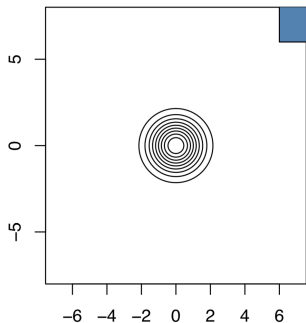
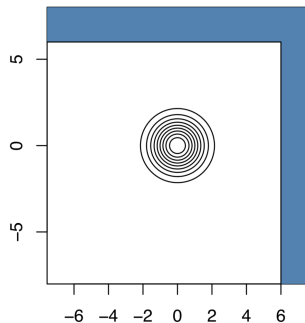
$$\theta^{(k+1)} = \frac{\sum_i H_i x^{(i)}}{\sum_i H_i}$$

- ▶ $q_\theta = \mathcal{N}(\theta, \Sigma)$

$$\theta^{(k+1)} = \Sigma^{-1} \frac{\sum_i H_i x^{(i)}}{\sum_i H_i}$$

- ▶ Other exponential family updates are typically closed form functions of sample moments

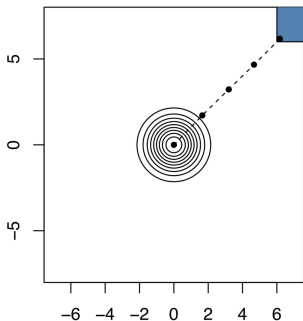
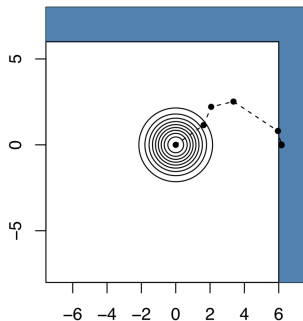


Gaussian, $\Pr(\min(x)>6)$ Gaussian, $\Pr(\max(x)>6)$ 

$$\theta_1 = (0, 0)^T$$

Take $K = 10$ steps with $n = 1000$ each



Gaussian, $\Pr(\min(x)>6)$ Gaussian, $\Pr(\max(x)>6)$ 

For $\min(x)$, $\theta^{(k)}$ heads Northeast, which is OK.

For $\max(x)$, $\theta^{(k)}$ heads North or East, and miss the other part completely, leading to underestimates of I by about 1/2



- ▶ The control variate strategy improves estimation of an unknown integral by relating the estimate to some correlated estimator with known integral
- ▶ A general class of unbiased estimators

$$I_{CV} = I_{MC} - \lambda(J_{MC} - J)$$

where $\mathbb{E}(J_{MC}) = J$. It is easy to show I_{CV} is unbiased, $\forall \lambda$

- ▶ We can choose λ to minimize the variance of I_{CV}

$$\hat{\lambda} = \frac{\text{Cov}(I_{MC}, J_{MC})}{\text{Var}(J_{MC})}$$

where the related moments can be estimated using samples from corresponding distributions

- ▶ Recall that IS estimator is

$$I_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^n h(x^{(i)})w(x^{(i)})$$

- ▶ Note that $h(x)w(x)$ and $w(x)$ are correlated and $\mathbb{E}w(x) = 1$, we can use the control variate

$$\bar{w} = \frac{1}{n} \sum_{i=1}^n w(x^{(i)})$$

and the importance sampling control variate estimator is

$$I_n^{\text{ISCV}} = I_n^{\text{IS}} - \lambda(\bar{w} - 1)$$

λ can be estimated from a regression of $h(x)w(x)$ on $w(x)$ as described before

- ▶ Consider estimation of $I = \mathbb{E}(h(X, Y))$ using a random sample $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$ drawn from f
- ▶ Suppose the conditional expectation $\mathbb{E}(h(X, Y)|Y)$ can be computed. Using $\mathbb{E}(h(X, Y)) = \mathbb{E}(\mathbb{E}(h(X, Y)|Y))$, the *Rao-Blackwellized estimator* can be defined as

$$I_n^{\text{RB}} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(h(x^{(i)}, y^{(i)})|y^{(i)})$$

- ▶ Rao-Blackwellized estimator gives smaller variance than the ordinary Monte Carlo estimator

$$\begin{aligned} \text{Var}(I_n^{\text{MC}}) &= \frac{1}{n} \text{Var}(\mathbb{E}(h(X, Y)|Y)) + \frac{1}{n} \mathbb{E}(\text{Var}(h(X, Y)|Y)) \\ &\geq \text{Var}(I_n^{\text{RB}}) \end{aligned}$$

follows from the conditional variance formula



- ▶ Suppose rejection sampling stops at a random time M with acceptance of the n th draw, yielding $x^{(1)}, \dots, x^{(n)}$ from all M proposals $y^{(1)}, \dots, y^{(M)}$
- ▶ The ordinary Monte Carlo estimator can be expressed as

$$I_n^{\text{MC}} = \frac{1}{n} \sum_{i=1}^M h(y^{(i)}) 1_{U_i \leq w(y^{(i)})}$$

- ▶ Rao-Blackwellization estimator

$$I_n^{\text{RB}} = \frac{1}{n} \sum_{i=1}^M h(y^{(i)}) t_i(Y)$$

where

$$t_i(Y) = \mathbb{E}(1_{U_i \leq w(y^{(i)})} | M, y^{(1)}, \dots, y^{(M)})$$



- ▶ Hesterberg, T. C. (1988). Advances in importance sampling. PhD thesis, Stanford University.
- ▶ Kong, A. (1992). A note on importance sampling using standardized weights. Technical Report 348, University of Chicago.