Stat 535 C - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca

1.1– Outline

- Importance Sampling.
- Normalized Importance Sampling.
- Importance Sampling versus Rejection Sampling.

2.1– Summary of Last Lecture

- Let $\pi(x)$ be a probability density on $\mathcal{X}$.

- Monte Carlo approximation is given by

$$\widehat{\pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(x) \text{ where } X^{(i)} \overset{\text{i.i.d.}}{\sim} \pi.$$

- For any $\varphi : \mathcal{X} \to \mathbb{R}$

$$E_{\widehat{\pi}_N}(\varphi(X)) = \frac{1}{N}\sum_{i=1}^{N} \varphi\left(X^{(i)}\right) \simeq E_{\pi}(\varphi(X))$$

and more precisely

$$E_X\left[E_{\widehat{\pi}_N}(\varphi(X))\right] = E_{\pi}(\varphi(X)) \text{ and } var_X\left(E_{\widehat{\pi}_N}(\varphi(X))\right) = \frac{var_{\pi}(\varphi(X))}{N}.$$

2.1– Summary of Last Lecture

- Direct methods feasible for standard distributions: inverse method, composition, etc.

- In case where $\pi \propto \pi^*$ does not admit any standard form, we can use a *proposal* distribution q on X where $q \propto q^*$.

- We need q to 'dominate' π; i.e.

$$C = \sup_{x \in \mathsf{X}} \frac{\pi^*(x)}{q^*(x)} < +\infty.$$

Consider $C' \geq C$. Then the accept/reject procedure proceeds as follows:

Accept/Reject procedure

1. Sample $Y \sim q$ and $U \sim \mathcal{U}(0,1)$.

2. If $U < \frac{\pi^*(Y)}{C' q^*(Y)}$ then return Y; otherwise return to step 1.

2.2– Accept Reject - Illustration

- This is a simple generic algorithm but it requires coming up with a bound C.

- Its performance typically degrade exponentially fast with the dimension of X.

- It seems you are wasting some information by rejecting samples.

- You need to wait a random time to obtain some samples from π.

- Is it possible to “recycle” these samples?

3.1– Importance Sampling

- Consider again the target distribution π and the proposal distribution q. We only require

$$\pi(x) > 0 \Rightarrow q(x) > 0.$$

- In this case, the Importance Sampling (IS) identity is

$$E_\pi(\varphi(X)) = \int_{\mathsf{X}} \varphi(x)\pi(x)dx = \int_{\mathsf{X}} \varphi(x)\frac{\pi(x)}{q(x)}q(x)dx = E_q(w(X)\varphi(X))$$

where the so-called Importance Weight is given by

$$w(x) = \frac{\pi(x)}{q(x)}$$

- This is a simple yet very flexible identity.

3.1– Importance Sampling

- Monte Carlo approximation of q is

$$\widehat{q}_N\left(x\right) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X^{(i)}}\left(x\right) \text{ where } X^{(i)} \overset{\text{i.i.d.}}{\sim} q.$$

- It follows that an estimate of $E_\pi(\varphi(X)) = E_q(w(X)\varphi(X))$ is

$$E_{\widehat{q}_N}(w(X)\varphi(X)) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)})\varphi(X^{(i)})$$

- It corresponds to the following approximation

$$\widehat{\pi}_N\left(x\right) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)})\delta_{X^{(i)}}\left(x\right)$$

- We have

$$E_X\left[E_{\widehat{q}_N}\left(w(X)\varphi\left(X\right)\right)\right]=E_\pi\left(\varphi\left(X\right)\right)$$

and

$$var_X\left(E_{\widehat{q}_N}\left(\varphi\left(X\right)\right)\right)=\frac{var_q\left(w(X)\varphi\left(X\right)\right)}{N}=\frac{E_\pi\left(w(X)\varphi^2\left(X\right)\right)-E_\pi^2\left(\varphi\left(X\right)\right)}{N}$$

- In practice, it is recommended to ensure

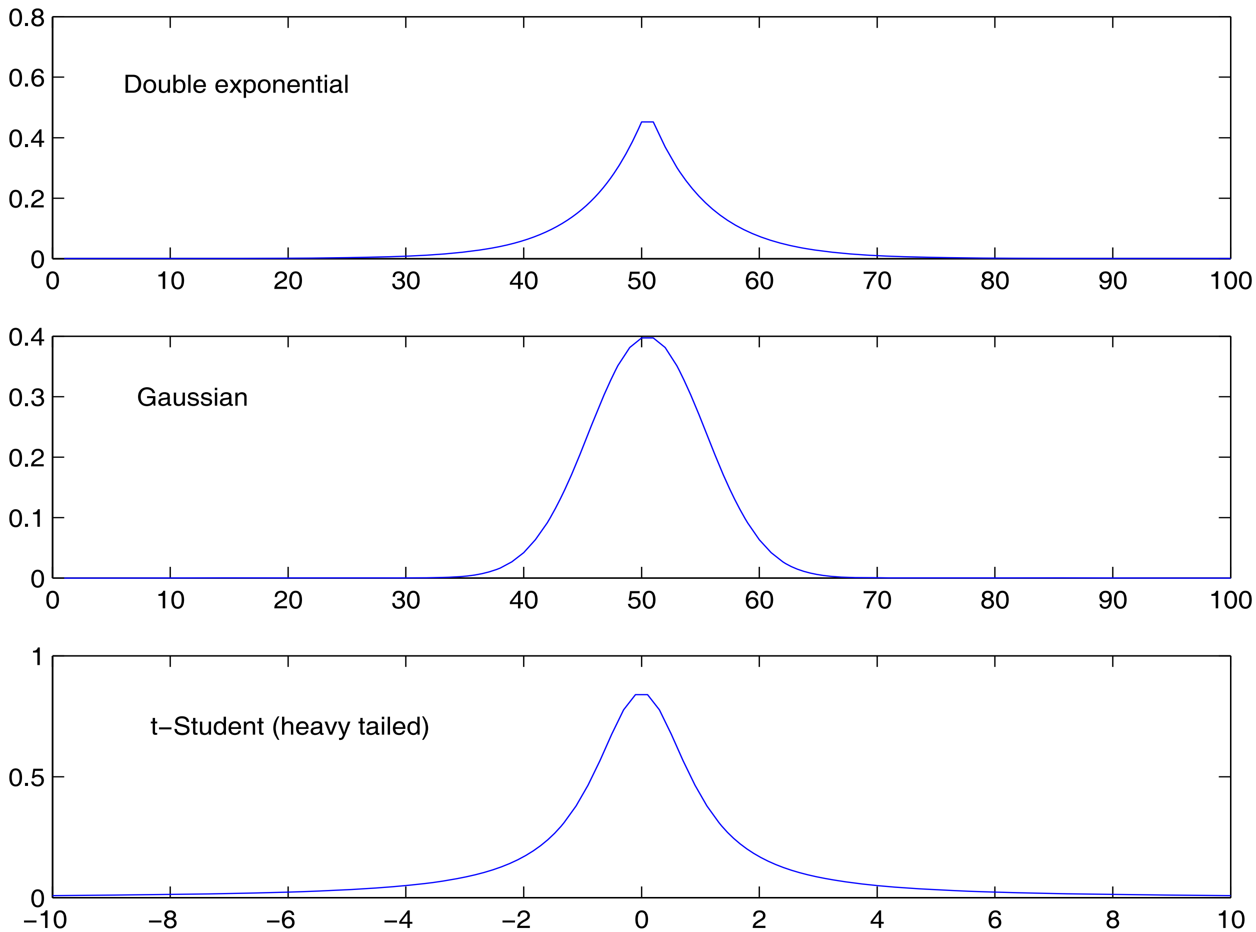
$$E_\pi\left(w(X)\right)=\int\frac{\pi^2\left(x\right)}{q\left(x\right)}dx<\infty.$$

- Even if it is not necessary, it is actually even better to ensure that

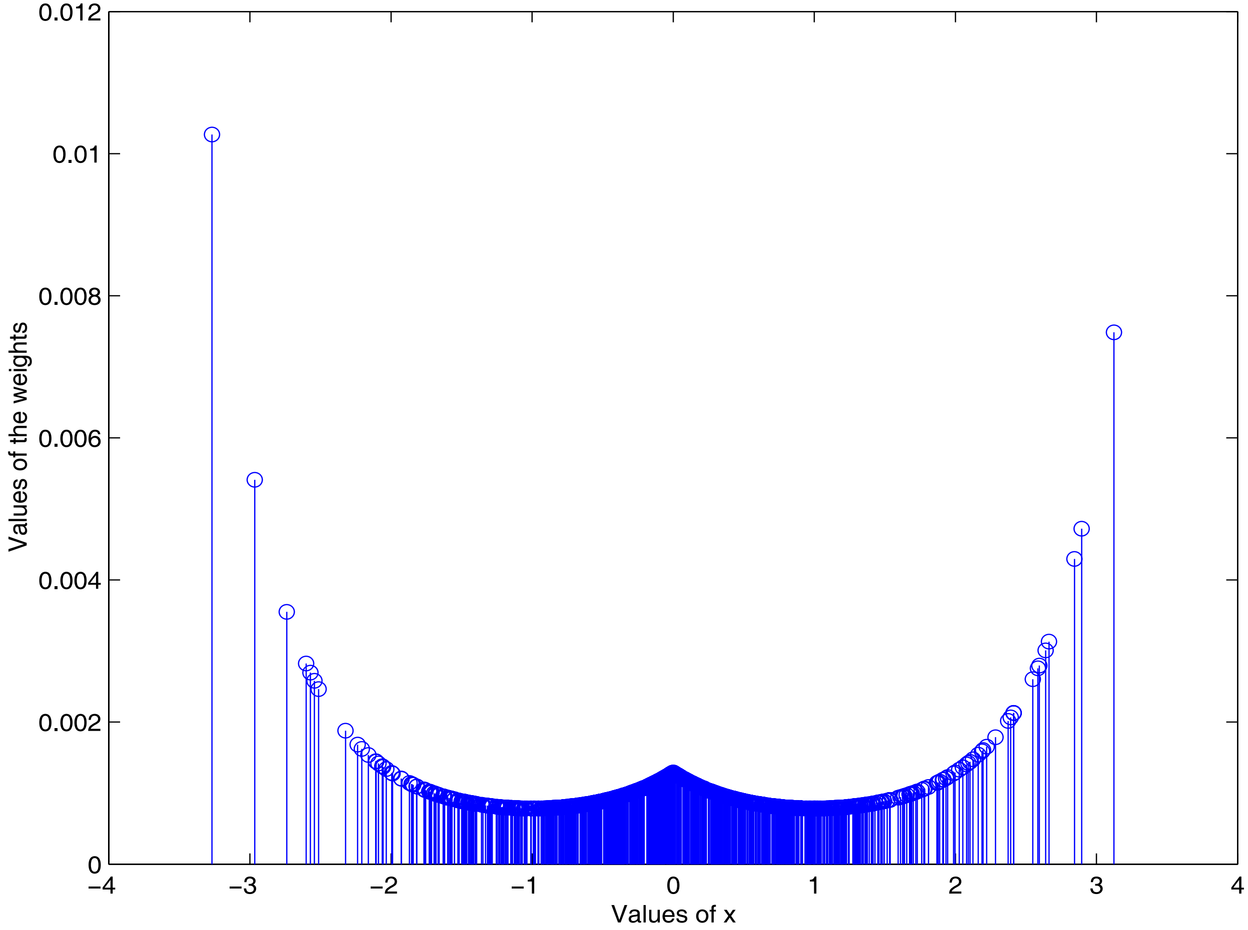
$$\sup_{x\in\mathcal{X}}w\left(x\right)<\infty.$$

3.2– Example

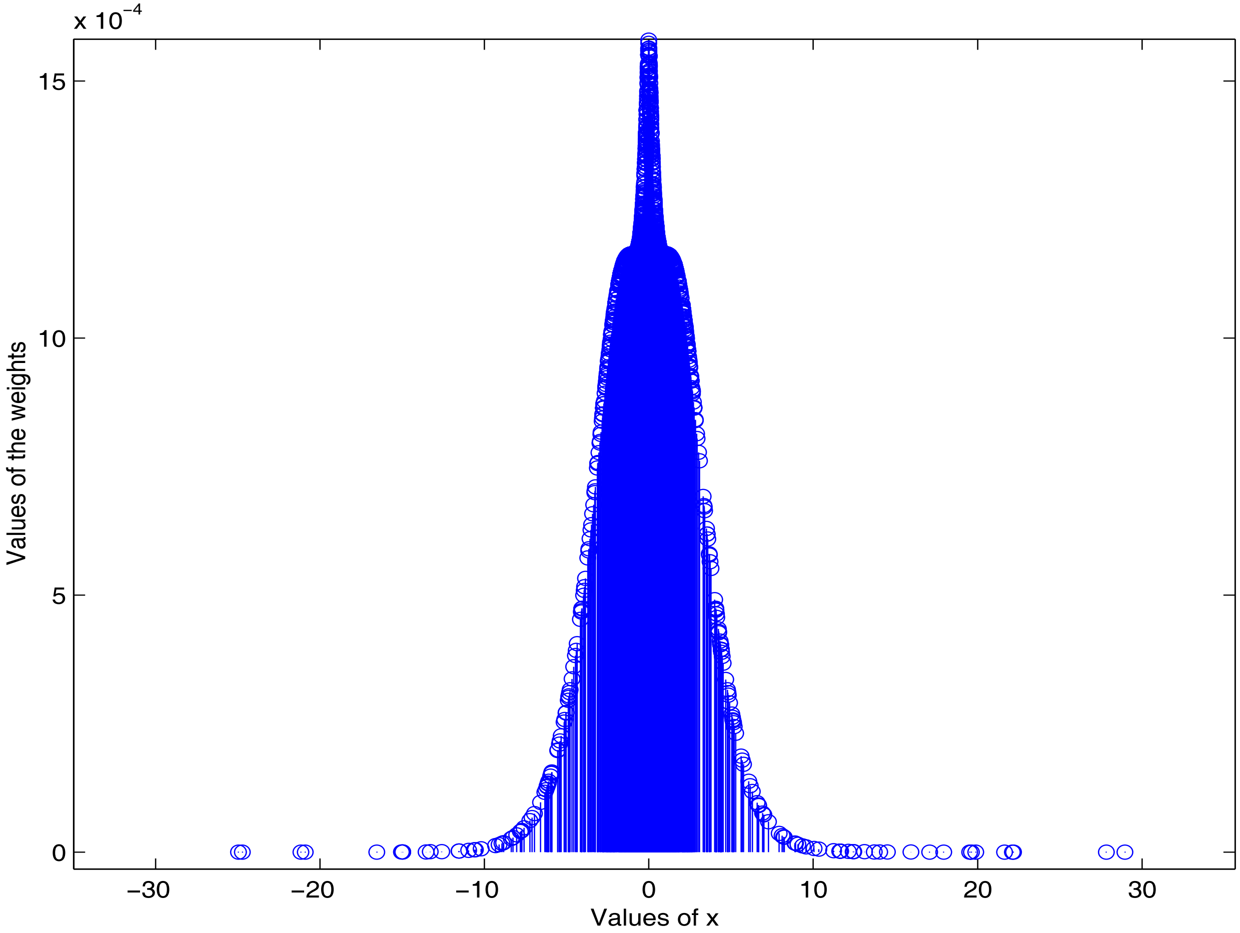
Target double exponential distributions and two IS distributions



IS approximation obtained using a Gaussian IS distribution



IS approximation obtained using a Student-t IS distribution



- For a given test function, one can minimize the IS variance using

$$q^{\mathrm{opt}}(x) = \frac{|\varphi(x)|\,\pi(x)}{\int_{\mathcal{X}} |\varphi(x)|\,\pi(x)\,dx}$$

Proof:

$$var_q(w(X)\varphi(X)) = \int q(x)\,\frac{\pi^2(x)}{q^2(x)}\varphi^2(x)\,dx - \left(\int \pi(x)\,\varphi(x)\,dx\right)^2$$

and

$$\int q(x)\,\frac{\pi^2(x)}{q^2(x)}\varphi^2(x)\,dx \geq \left(\int q(x)\,\frac{\pi(x)\,|\varphi(x)|}{q(x)}dx\right)^2 = \left(\int \pi(x)\,|\varphi(x)|\,dx\right)^2.$$

This lower bound is attained for $q^{\mathrm{opt}}(x)$.

3.4– Normalized Importance Sampling

• In most if not all applications we are interested in, standard IS cannot be used as the importance weights $w(x) = \pi(x)/q(x)$ cannot be evaluated in closed-form. In practice, we typically only know $\pi(x) \propto \pi^*(x)$ and $q(x) \propto q^*(x)$.

• Normalized IS identity is based on

$$\pi(x) = \frac{\pi^*(x)}{\int \pi^*(x)\, dx} = \frac{w^*(x)\, q^*(x)}{\int w^*(x)\, q^*(x)\, dx} = \frac{w^*(x)\, q(x)}{\int w^*(x)\, q(x)\, dx}$$

where

$$w^*(x) = \frac{\pi^*(x)}{q^*(x)}.$$

3.4– Normalized Importance Sampling

- For any test function φ, we can also write

$$E_{\pi}\left(\varphi\left(X\right)\right)=\frac{E_{q}\left(w^{*}\left(X\right)\varphi\left(X\right)\right)}{E_{q}\left(w^{*}\left(X\right)\right)}=\frac{E_{q}\left(w\left(X\right)\varphi\left(X\right)\right)}{E_{q}\left(w\left(X\right)\right)}.$$

- Given a Monte Carlo approximation of q; $\widehat{q}_{N}\left(x\right)=\frac{1}{N}\sum_{i=1}^{N}\delta_{X^{(i)}}\left(x\right)$ then

$$\widehat{\pi}_{N}\left(x\right)=\sum_{i=1}^{N}W^{(i)}\delta_{X^{(i)}}\left(x\right)\ \text{ where } W^{(i)}=\frac{w^{*}\left(X^{(i)}\right)}{\sum_{j=1}^{N}w^{*}\left(X^{(j)}\right)},$$

$$E_{\widehat{\pi}_{N}}\left(\varphi\left(X\right)\right)=\sum_{i=1}^{N}W^{(i)}\varphi\left(X^{(i)}\right).$$

- The estimates are a ratio of estimates.

3.4– Normalized Importance Sampling

- Contrary to standard IS, this estimate is biased but asymptotically unbiased by the LLN it is asymptotically consistent.

- Derivation of the asymptotic bias and variance based on the delta method.

3.5– Proof using the Delta Method

- Assume you have $Z = g(A, B)$ with $E(A) = \mu_A$ and $E(B) = \mu_B$ then a two-dimensional Taylor series gives around $\mu = (\mu_A, \mu_B)$

$$Z \simeq g(\mu) + (A - \mu_A) \frac{\partial g}{\partial a}(\mu) + (B - \mu_B) \frac{\partial g}{\partial b}(\mu).$$

It follows that

$$E(Z) \simeq g(\mu),$$

$$Var(Z) \simeq \sigma_A^2 \frac{\partial g}{\partial a}^2 (\mu) + \sigma_B^2 \frac{\partial g}{\partial b}^2 (\mu) + 2 \frac{\partial g}{\partial a}(\mu) \frac{\partial g}{\partial b}(\mu) \sigma_{A,B}.$$

- In our case

$$Z = E_{\widehat{\pi}_N}(\varphi(X)) = \frac{E_{q_N}(w^*(X) \varphi(X))}{E_{q_N}(w^*(X))} = \frac{A}{B}$$

- We have

$$\frac{\partial g}{\partial a}(\mu)\frac{\partial g}{\partial b}(\mu) = -\frac{\mu_A}{\mu_B^3},\ \frac{\partial g}{\partial a}^2(\mu) = \frac{1}{\mu_B^2},\ \frac{\partial g}{\partial b}^2(\mu) = \frac{\mu_A^2}{\mu_B^4},$$

$$\begin{aligned} \mu_A &= E_q\left(w^*\left(X\right)\varphi\left(X\right)\right),\ \mu_B = E_q\left(w^*\left(X\right)\right), \\ \sigma_A^2 &= \frac{var_q\left(w^*\left(X\right)\varphi\left(X\right)\right)}{N},\ \sigma_B^2 = \frac{var_q\left(w^*\left(X\right)\right)}{N} \end{aligned}$$

$$\sigma_{A,B} = \frac{E_q\left(w^*\left(X\right)^2\varphi\left(X\right)\right) - \mu_A.\mu_B}{N}.$$

- It follows that

$$
\begin{aligned}
Var\left(E_{\widehat{\pi}_N}\left(\varphi\left(X\right)\right)\right) &\simeq \sigma_A^2 \frac{\partial g}{\partial a}^2\left(\mu\right) + \sigma_B^2 \frac{\partial g}{\partial b}^2\left(\mu\right) + 2\frac{\partial g}{\partial a}\left(\mu\right)\frac{\partial g}{\partial b}\left(\mu\right)\sigma_{A,B} \\
&= \frac{\sigma_A^2}{\mu_B^2} + \frac{\sigma_B^2\mu_A^2}{\mu_B^4} - 2\frac{\mu_A\sigma_{A,B}}{\mu_B^3}
\end{aligned}
$$

- Asymptotically, we have a central limit theorem

$$
\sqrt{N}\left(E_{\widehat{\pi}_N}\left(\varphi\left(X\right)\right) - E_{\pi}\left(\varphi\left(X\right)\right)\right) \Rightarrow \mathcal{N}\left(0, \sigma_{IS}^2\left(\varphi\right)\right)
$$

where

$$
\sigma_{IS}^2\left(\varphi\right) = \int \frac{\pi^2\left(x\right)}{q\left(x\right)}\left(\varphi\left(x\right) - E_{\pi}\left(\varphi\right)\right)^2 dx
$$

- In practice, it is now necessary but highly recommended to select the proposal q such that

$$\sup_{x\in\mathcal{X}} w(x) < \infty \text{ or equivalently } \sup_{x\in\mathcal{X}} w^{*}(x) < \infty.$$

- There is some empirical evidence that Normalized IS performs better than standard IS in numerous cases.

- Using a second order Taylor expansion

$$
\begin{aligned}
Z \quad \simeq \quad & g(\mu) + (A-\mu_A)\frac{\partial g}{\partial a}(\mu) + (B-\mu_B)\frac{\partial g}{\partial b}(\mu) \\
& + \frac{1}{2}(A-\mu_A)^2 \frac{\partial^2 g}{\partial a^2}(\mu) + \frac{1}{2}(B-\mu_B)^2\frac{\partial^2 g}{\partial b^2}(\mu) + (A-\mu_A)(B-\mu_B)\frac{\partial^2 g}{\partial a \partial b}(\mu)
\end{aligned}
$$

gives

$$
E\left(E_{\widehat{\pi}_N}(\varphi(X))\right) \simeq g(\mu) + \frac{1}{2}\sigma_A^2 \frac{\partial^2 g}{\partial a^2}(\mu) + \frac{1}{2}\sigma_B^2 \frac{\partial^2 g}{\partial b^2}(\mu) + \sigma_{A,B}\frac{\partial^2 g}{\partial a \partial b}(\mu).
$$

- It follows that asymptotically we have

$$
N\left(E_{\widehat{\pi}_N}(\varphi(X)) - E_\pi(\varphi(X))\right) \to -\int \frac{\pi^2(x)}{q(x)}\left(\varphi(x) - E_\pi(\varphi)\right)dx.
$$

- We have $Bias^2$ of order $1/N^2$ and Variance of order $1/N$.

3.8– Optimal Importance Sampling

• For a given test function, one can minimize the normalized IS asymptotic variance using

$$q^{\text{opt}}(x) = \frac{\left|\varphi(x) - E_{\pi(\varphi)}\right| \pi(x)}{\int_{\mathcal{X}} \left|\varphi(x) - E_{\pi(\varphi)}\right| \pi(x)\, dx}$$

Proof:

$$\begin{aligned} \int q(x) \frac{\pi^2(x)}{q^2(x)} (\varphi(x) - E_\pi(\varphi))^2\, dx &\geq \left(\int q(x) \frac{\pi(x) \left|\varphi(x) - E_\pi(\varphi)\right|}{q(x)} dx \right)^2 \\ &= \left(\int \pi(x) \left|\varphi(x) - E_\pi(\varphi)\right| dx \right)^2 \end{aligned}$$

and this lower bound is attained for $q^{\text{opt}}(x)$.

• This result is practically useless because it requires knowing $E_\pi(\varphi)$ but it suggests approximations.

3.9– In practice...

• In statistics, we are usually not interested in a specific φ but in several functions and we prefer having $q(x)$ as close as possible to $\pi(x)$.

• For flat functions, one can approximate the variance by

$$var\left(E_{\widehat{\pi}_N}\left(\varphi\left(X\right)\right)\right) \simeq \left(1 + var_q\left(w\left(X\right)\right)\right) \frac{var\left(E_\pi\left(\varphi\left(X\right)\right)\right)}{N}.$$

• Simple interpretation: The N weighted samples are approximately equivalent to M unweighted samples from π where

$$M = \frac{N}{1 + var_q\left(w\left(X\right)\right)} \leq N.$$

• However, we are often interested in estimating the ratio of normalizing constants

$$\frac{\int \pi^{*}(x)\,dx}{\int q^{*}(x)\,dx} = \int w^{*}(x)\,q(x)\,dx = E_{q}\left[w^{*}(X)\right].$$

using

$$E_{q_N}\left[w^{*}(X)\right] = \frac{1}{N}\sum_{i=1}^{N} w^{*}\left(X^{(i)}\right)$$

which is unbiased and has variance

$$var\left[E_{q_N}\left[w^{*}(X)\right]\right] = \frac{var_q\left(w^{*}(X)\right)}{N}.$$

3.10– Open Question

- Clearly if you have $q(x) = \pi(x)$ then

$$var\left[E_{q_N}\left[w^*(X)\right]\right] = 0$$

- However if $q(x) = \pi(x)$ then the estimate is simply

$$E_{q_N}\left[w^*(X)\right] = \frac{\int \pi^*(x)\,dx}{\int q^*(x)\,dx}.$$

- **Open Question**: How could you come up with a good estimate of $\int \pi^*(x)\,dx$ based on samples of π.

3.11– Application to Bayesian Inference

- Consider a Bayesian model: prior $\pi(\theta)$ and likelihood $f(x|\theta)$.

- The posterior distribution is given by

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int_\Theta \pi(\theta) f(x|\theta)\, d\theta} \propto \pi^*(\theta|x) \text{ where } \pi^*(\theta|x) = \pi(\theta) f(x|\theta).$$

- We can use the prior distribution as a candidate distribution $q(\theta) = q^*(\theta) = \pi(\theta)$.

- We also get an estimate of the marginal likelihood

$$\int_\Theta \pi(\theta) f(x|\theta)\, d\theta.$$

- IS is more powerful than you think.

- Assume you have say to compute the importance weight

$$w\left(\theta^{(i)}\right) \propto \int f\left(x, z \mid \theta\right) dz;$$

i.e. the likelihood is very complex and might not admit a closed-form expression.

- You do NOT need to compute $w\left(\theta^{(i)}\right)$ exactly, an unbiased estimate of it is sufficient.

3.12– Importance sampling does not work well in high-dimension

- Consider the case where $\mathcal{X} = \mathbb{R}^n$

$$\pi(\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n \theta_i^2}{2}\right)$$

and

$$q_\sigma(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n \theta_i^2}{2\sigma^2}\right)$$

- We have for any $\sigma > 1$

$$w_\sigma(\theta) = \frac{\pi(\theta)}{q_\sigma(\theta)} = \sigma^n \exp\left(-\sum_{i=1}^n \frac{\theta_i^2}{2}\left(1 - \frac{1}{\sigma^2}\right)\right) \leq \sigma^n \text{ for any } \theta$$

and

$$var_{q_\sigma}\left(\frac{\pi(\theta)}{q_\sigma(\theta)}\right) = \sigma^n \sigma'^n - 1 \text{ with } \sigma'^2 = \frac{\sigma^2}{\sigma^2 - 1/2} > 1$$

- Despites having a very good proposal then the variance of the weights increases exponentially fast with the dimension of the problem.

3.13– Rejection Sampling versus Importance Sampling

- Given N samples from q, we estimate $E_\pi(\varphi(X))$ through IS

$$\widehat{E}_\pi^{IS}(\varphi(X)) = \frac{\sum_{i=1}^N w^*\left(X^{(i)}\right)\varphi\left(X^{(i)}\right)}{\sum_{i=1}^N w^*\left(X^{(i)}\right)}$$

or we "filter" the samples through rejection and propose instead

$$\widehat{E}_\pi^{RS}(\varphi(X)) = \frac{1}{K}\sum_{k=1}^K \varphi\left(X^{(i_k)}\right)$$

where K is a random variable.

- We want to know which strategy performs the best.

3.14– Rejection Sampling is a special case of Importance Sampling

- Define the artificial target $\overline{\pi}(x,y)$ on $\mathcal{X}\times[0,1]$ as

$$\overline{\pi}(x,y)=\begin{cases}\frac{Cq^{*}(x)}{\int\pi^{*}(x)dx}, & \text{for } x\in\mathcal{X},\ y\in\left[0,\frac{\pi^{*}(x)}{Cq^{*}(x)}\right]\\ \\ 0 & \text{otherwise}\end{cases}$$

then

$$\int\overline{\pi}(x,y)\,dy=\int_{0}^{\frac{\pi^{*}(x)}{Cq^{*}(x)}}\frac{Cq^{*}(x)}{\int\pi^{*}(x)\,dx}dy=\pi(x).$$

- Now let us consider the proposal distribution

$$q(x,y)=q(x)\,U_{[0,1]}(y)\ \text{ for }\ (x,y)\in\mathcal{X}\times[0,1].$$

3.14– Rejection Sampling is a special case of Importance Sampling

- Then rejection sampling is nothing but IS on $\mathcal{X} \times [0,1]$ where

$$w(x,y) = \frac{\overline{\pi}(x,y)}{q(x)\,U_{[0,1]}(y)} = \begin{cases} \frac{C\int q^*(x)dx}{\int \pi^*(x)dx} & \text{for } Y^{(i)} \in \left[0, \frac{\pi^*\left(X^{(i)}\right)}{Cq^*\left(X^{(i)}\right)}\right] \\ 0, & \text{otherwise.} \end{cases}$$

- We have

$$\widehat{E}_{\pi}^{RS}(\varphi(X)) = \frac{1}{K}\sum_{k=1}^{K}\varphi\left(X^{(i_k)}\right) = \frac{\sum_{i=1}^{N} w\left(X^{(i)},Y^{(i)}\right)\varphi\left(X^{(i)}\right)}{\sum_{i=1}^{N} w\left(X^{(i)},Y^{(i)}\right)}.$$

- Compared to standard IS, RS performs IS on an enlarged space.

3.14– Rejection Sampling is a special case of Importance Sampling

• The variance of the importance weights from RS is higher than for standard IS:

$$var_q\left[w\left(X,Y\right)\right] \geq var_q\left[w\left(X\right)\right].$$

More precisely, we have

$$\begin{aligned} var\left[w\left(X,Y\right)\right] &= var\left[E\left[\left.w\left(X,Y\right)\right|X\right]\right] + E\left[var\left[\left.w\left(X,Y\right)\right|X\right]\right] \\ &= var\left[w\left(X\right)\right] + E\left[var\left[\left.w\left(X,Y\right)\right|X\right]\right]. \end{aligned}$$

• To compute integrals, Rejection sampling is inefficient and you should simply use IS.

3.15– Discussion

• Like Rejection, IS is useful for small non-standard distributions but collapses for most "interesting" problems.

• In both cases, the problem is to be able to design "clever" proposal distributions.

• Towards the end of this course, we will present advanced dynamic method to address this problem.