CPSC 535
Importance Sampling Methods

AD

8th February 2007

- Importance Sampling

- Importance Sampling
- Normalized Importance Sampling.

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

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

- Let $\pi(x)$ be a probability density on $\mathcal{X}$.
- Monte Carlo approximation is given by

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

- 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 $\varphi : \mathcal{X} \rightarrow \mathbb{R}$

$$\mathbb{E}_{\widehat{\pi}_N}(\varphi(X)) = \frac{1}{N}\sum_{i=1}^{N}\varphi\left(X^{(i)}\right) \approx \mathbb{E}_{\pi}(\varphi(X))$$

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

$$\mathbb{E}_{\widehat{\pi}_N}(\varphi(X)) = \frac{1}{N}\sum_{i=1}^{N}\varphi\left(X^{(i)}\right) \approx \mathbb{E}_{\pi}(\varphi(X))$$

- More precisely, we have

$$\begin{aligned} \mathbb{E}_{\left\{X^{(i)}\right\}}\left[\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right] &= \mathbb{E}_{\pi}(\varphi(X)), \\ var_{\left\{X^{(i)}\right\}}\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right) &= \frac{var_{\pi}(\varphi(X))}{N}. \end{aligned}$$

Accept-Reject Procedure

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

Accept-Reject Procedure

- 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 $\mathcal{X}$ where $q \propto q^*$.

Accept-Reject Procedure

- 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 $\mathcal{X}$ where $q \propto q^*$.
- We need q to 'dominate' π; i.e.

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

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

Accept/Reject procedure

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

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.

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

- 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 $\mathcal{X}$.

- 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 $\mathcal{X}$.
- It seems you are wasting some information by rejecting samples.

- 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 $\mathcal{X}$.
- It seems you are wasting some information by rejecting samples.
- You need to wait a random time to obtain some samples from π.

- 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 $\mathcal{X}$.
- 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?

Importance Sampling

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

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

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

$$\pi(x) = w(x)\, q(x)$$

where the so-called Importance Weight is given by

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

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

$$\pi(x) = w(x)\, q(x)$$

where the so-called Importance Weight is given by

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

- It follows that

$$\begin{aligned}
\mathbb{E}_\pi(\varphi(X)) &= \int_{\mathcal{X}} \varphi(x)\pi(x)dx = \int_{\mathcal{X}} \varphi(x)\frac{\pi(x)}{q(x)}q(x)dx \\
&= \mathbb{E}_q(w(X)\varphi(X))
\end{aligned}$$

- Monte Carlo approximation of q is

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

- Monte Carlo approximation of q is

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

- It corresponds to the following approximation

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

- Monte Carlo approximation of q is

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

- It corresponds to the following approximation

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

- It follows that an estimate of $\mathbb{E}_\pi(\varphi(X)) = \mathbb{E}_q(w(X)\varphi(X))$ is

$$\mathbb{E}_{\widehat{q}_N}(w(X)\varphi(X)) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)})\varphi(X^{(i)}).$$

- We have

$$\mathbb{E}_{\{X^{(i)}\}}\left[\mathbb{E}_{\widehat{q}_N}\left(w(X)\varphi\left(X\right)\right)\right]=\mathbb{E}_{\pi}\left(\varphi\left(X\right)\right)$$

and

$$\begin{aligned} var_{\{X^{(i)}\}}\left(\mathbb{E}_{\widehat{q}_N}\left(\varphi\left(X\right)\right)\right) &= \frac{var_q\left(w(X)\varphi\left(X\right)\right)}{N} \\ &= \frac{\mathbb{E}_{\pi}\left(w(X)\varphi^2\left(X\right)\right)-\mathbb{E}_{\pi}^2\left(\varphi\left(X\right)\right)}{N} \end{aligned}$$

- We have
$$\mathbb{E}_{\{X^{(i)}\}}\left[\mathbb{E}_{\widehat{q}_N}\left(w(X)\varphi\left(X\right)\right)\right]=\mathbb{E}_{\pi}\left(\varphi\left(X\right)\right)$$
and
$$\begin{aligned} var_{\{X^{(i)}\}}\left(\mathbb{E}_{\widehat{q}_N}\left(\varphi\left(X\right)\right)\right) &= \frac{var_q\left(w(X)\varphi\left(X\right)\right)}{N} \\ &= \frac{\mathbb{E}_{\pi}\left(w(X)\varphi^2\left(X\right)\right)-\mathbb{E}_{\pi}^2\left(\varphi\left(X\right)\right)}{N} \end{aligned}$$

- In practice, it is recommended to ensure
$$\mathbb{E}_{\pi}\left(w(X)\right)=\int\frac{\pi^2\left(x\right)}{q\left(x\right)}dx<\infty.$$

- We have

$$\mathbb{E}_{\{X^{(i)}\}}\left[\mathbb{E}_{\widehat{q}_N}\left(w(X)\varphi\left(X\right)\right)\right]=\mathbb{E}_{\pi}\left(\varphi\left(X\right)\right)$$

and

$$\begin{aligned} var_{\{X^{(i)}\}}\left(\mathbb{E}_{\widehat{q}_N}\left(\varphi\left(X\right)\right)\right) &= \frac{var_q\left(w(X)\varphi\left(X\right)\right)}{N} \\ &= \frac{\mathbb{E}_{\pi}\left(w(X)\varphi^2\left(X\right)\right)-\mathbb{E}_{\pi}^2\left(\varphi\left(X\right)\right)}{N} \end{aligned}$$

- In practice, it is recommended to ensure

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

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

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

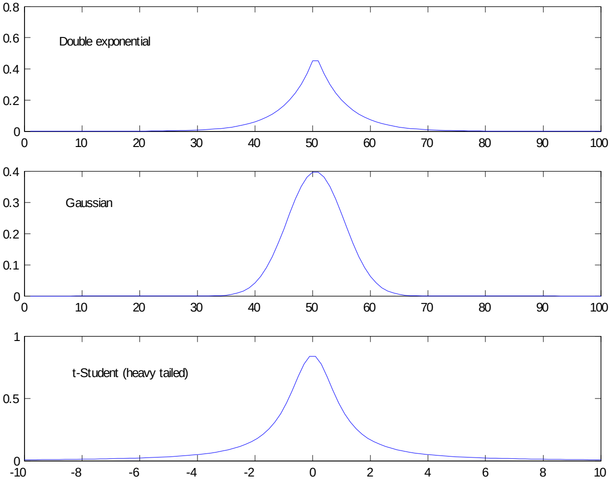

Figure: Target double exponential distributions and two IS distributions

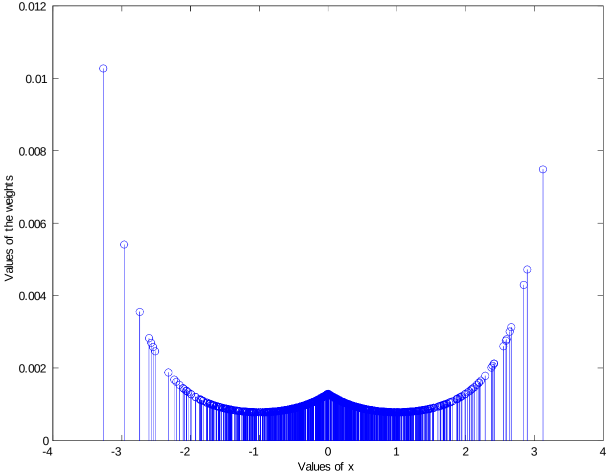

Figure: IS approximation obtained using a Gaussian IS distribution

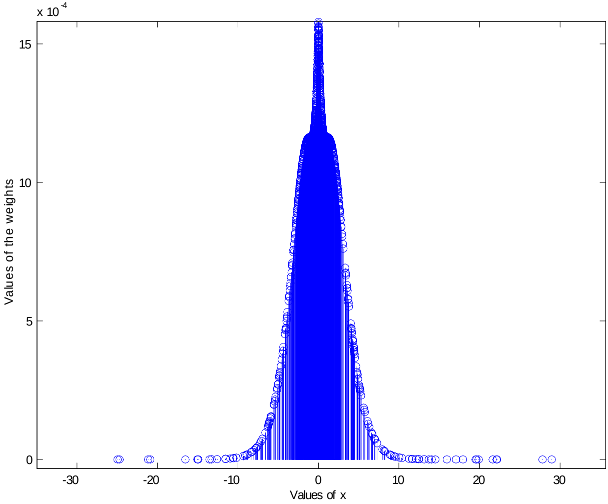

Figure: IS approximation obtained using a Student-t IS distribution

- We try to compute

$$\int \sqrt{\frac{x}{1-x}} \pi (x) \, dx$$

where

$$\pi (x) = \frac{\Gamma ((\nu + 1)/2)}{\sqrt{\nu \pi} \Gamma (\nu/2)} \left(1 + \frac{x}{\nu}\right)^{-(\nu+1)/2}$$

is a t-student distribution with $\nu > 1$ (you can sample from it by composition $\mathcal{N}(0, 1) / \mathcal{G}a(\nu/2, \nu/2)$) using Monte Carlo.

- We try to compute

$$\int \sqrt{\frac{x}{1-x}} \pi(x) \, dx$$

where

$$\pi(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x}{\nu}\right)^{-(\nu+1)/2}$$

is a t-student distribution with $\nu > 1$ (you can sample from it by composition $\mathcal{N}(0,1) / \mathcal{G}a(\nu/2, \nu/2)$) using Monte Carlo.

- We use $q_1(x) = \pi(x)$, $q_2(x) = \frac{\Gamma(1)}{\sqrt{\nu\pi}\Gamma(1/2)} \left(1 + \frac{x}{\nu\sigma}\right)^{-1}$ (Cauchy distribution) and $q_3(x) = \mathcal{N}\left(x; 0, \frac{\nu}{\nu-2}\right)$ (variance chosen to match the variance of $\pi(x)$)

- We try to compute

$$\int \sqrt{\frac{x}{1-x}} \pi(x)\, dx$$

where

$$\pi(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x}{\nu}\right)^{-(\nu+1)/2}$$

is a t-student distribution with $\nu > 1$ (you can sample from it by composition $\mathcal{N}(0,1) / \mathcal{G}a(\nu/2, \nu/2)$) using Monte Carlo.

- We use $q_1(x) = \pi(x)$, $q_2(x) = \frac{\Gamma(1)}{\sqrt{\nu\pi}\Gamma(1/2)} \left(1 + \frac{x}{\nu\sigma}\right)^{-1}$ (Cauchy distribution) and $q_3(x) = \mathcal{N}\left(x; 0, \frac{\nu}{\nu-2}\right)$ (variance chosen to match the variance of $\pi(x)$)
- It is easy to see that

$$\frac{\pi(x)}{q_2(x)} < \infty \text{ and } \int \frac{\pi(x)^2}{q_3(x)} dx = \infty, \ \frac{\pi(x)}{q_3(x)} \text{ is unbounded}$$

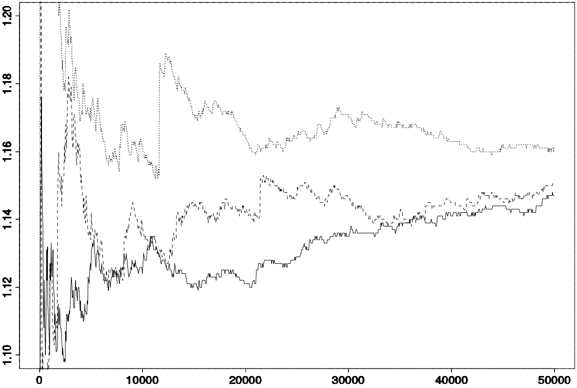

Figure: Performance for $\nu = 12$ with q_1 (solid line), q_2 (dashes) and q_3 (light dots). Final values 1.14, 1.14 and 1.16 vs true value 1.13

- We now try to compute

$$\int_{2.1}^{\infty} x^5 \pi(x)\, dx$$

- We now try to compute

$$\int_{2.1}^{\infty} x^5 \pi(x)\, dx$$

- We try to use the same importance distribution but also use the fact that using a change of variables $u = 1/x$, we have

$$\begin{aligned} \int_{2.1}^{\infty} x^5 \pi(x)\, dx &= \int_0^{1/2.1} u^{-7} \pi(1/u)\, du \\ &= \frac{1}{2.1} \int_0^{1/2.1} 2.1 u^{-7} \pi(1/u)\, du \end{aligned}$$

which is the expectation of $2.1u^{-7}\pi(1/u)$ with respect to $\mathcal{U}[0, 1/2.1]$.

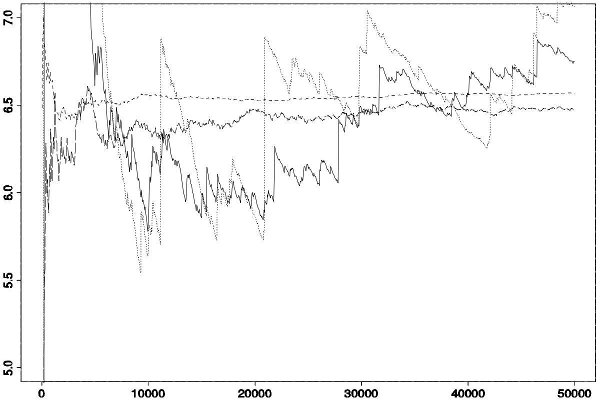

Figure: Performance for $\nu = 12$ with q_1 (solid), q_2 (short dashes), q_3 (dots), uniform (long dashes). Final values 6.75, 6.48, 7.06 and 6.48 vs true value 6.54

Optimal Importance function

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

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

Optimal Importance function

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

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

- *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

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

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

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 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

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

where

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

- For any test function φ, we can also write

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

- For any test function φ, we can also write

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

- 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

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

- For any test function φ, we can also write

$$\mathbb{E}_{\pi}(\varphi(X)) = \frac{\mathbb{E}_q(w^*(X)\varphi(X))}{\mathbb{E}_q(w^*(X))} = \frac{\mathbb{E}_q(w(X)\varphi(X))}{\mathbb{E}_q(w(X))}.$$

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

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

- The estimates are a ratio of estimates.

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

- Contrary to standard IS, this estimate is biased but by the LLN it is asymptotically consistent.
- Derivation of the asymptotic bias and variance based on the delta method.

Asymptotic Bias and Variance

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

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

  It follows that

$$\mathbb{E}(Z) \approx g(\mu),$$

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

Asymptotic Bias and Variance

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

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

It follows that

$$\mathbb{E}(Z) \approx g(\mu),$$

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

- In our case

$$Z = \mathbb{E}_{\widehat{\pi}_N}(\varphi(X)) = \frac{\mathbb{E}_{\widehat{q}_N}(w^*(X)\varphi(X))}{\mathbb{E}_{\widehat{q}_N}(w^*(X))} = \frac{A}{B}$$

- We have

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

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

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

- It follows that

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

- It follows that

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

- Asymptotically, we have a central limit theorem

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

where

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

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

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

- 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

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

gives

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right) \approx\ & g(\mu) + \tfrac{1}{2}\sigma_A^2\frac{\partial^2 g}{\partial a^2}(\mu) + \tfrac{1}{2}\sigma_B^2\frac{\partial^2 g}{\partial b^2}(\mu) \\
& + \sigma_{A,B}\frac{\partial^2 g}{\partial a \partial b}(\mu).
\end{aligned}$$

- Using a second order Taylor expansion

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

gives

$$
\begin{aligned}
\mathbb{E}\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right) \approx\ & g(\mu) + \frac{1}{2}\sigma_A^2\frac{\partial^2 g}{\partial a^2}(\mu) + \frac{1}{2}\sigma_B^2\frac{\partial^2 g}{\partial b^2}(\mu) \\
& + \sigma_{A,B}\frac{\partial^2 g}{\partial a \partial b}(\mu).
\end{aligned}
$$

- It follows that asymptotically we have

$$
N\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X)) - \mathbb{E}_{\pi}(\varphi(X))\right) \rightarrow -\int \frac{\pi^2(x)}{q(x)}(\varphi(x) - \mathbb{E}_{\pi}(\varphi))\,dx.
$$

- Using a second order Taylor expansion

$$Z \approx g(\mu) + (A - \mu_A) \frac{\partial g}{\partial a}(\mu) + (B - \mu_B) \frac{\partial g}{\partial b}(\mu)$$
$$+ \frac{1}{2}(A - \mu_A)^2 \frac{\partial^2 g}{\partial a^2}(\mu) + \frac{1}{2}(B - \mu_B)^2 \frac{\partial^2 g}{\partial b^2}(\mu)$$
$$+ (A - \mu_A)(B - \mu_B) \frac{\partial^2 g}{\partial a \partial b}(\mu)$$

gives

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

- It follows that asymptotically we have

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

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

- The asymptotic variance (and also the asymptotic bias) can be consistently estimated from the data using

$$\frac{\widehat{\sigma_{IS}^2(\varphi)}}{N} = \frac{\widehat{\sigma}_A^2}{\widehat{\mu}_B^2} + \frac{\widehat{\sigma}_B^2 \widehat{\mu}_A^2}{\widehat{\mu}_B^4} - 2\frac{\widehat{\mu}_A \widehat{\sigma}_{A,B}}{\widehat{\mu}_B^3}.$$

- The asymptotic variance (and also the asymptotic bias) can be consistently estimated from the data using

$$\frac{\widehat{\sigma_{IS}^{2}\left(\varphi\right)}}{N}=\frac{\widehat{\sigma}_{A}^{2}}{\widehat{\mu}_{B}^{2}}+\frac{\widehat{\sigma}_{B}^{2}\widehat{\mu}_{A}^{2}}{\widehat{\mu}_{B}^{4}}-2\frac{\widehat{\mu}_{A}\widehat{\sigma}_{A,B}}{\widehat{\mu}_{B}^{3}}.$$

- You can also compute the variance of the variance estimate; see Geweke (1989).

Application to Bayesian Statistics

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

Application to Bayesian Statistics

- Consider a Bayesian model: prior $\pi(\theta)$ and likelihood $f(x|\theta)$.
- The posterior distribution is given by

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

where $\pi^*(\theta|x) = \pi(\theta) f(x|\theta)$.

Application to Bayesian Statistics

- Consider a Bayesian model: prior $\pi(\theta)$ and likelihood $f(x|\theta)$.
- The posterior distribution is given by

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

where $\pi^*(\theta|x) = \pi(\theta) f(x|\theta)$.

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

Application to Bayesian Statistics

- 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

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

- *Example*: Application to Bayesian analysis of Markov chain. Consider a two state Markov chain with transition matrix F

$$\begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}$$

that is $\Pr\left(X_{t+1}=1\middle|\, X_t=1\right)=1-\Pr\left(X_{t+1}=2\middle|\, X_t=1\right)=p_1$ and $\Pr\left(X_{t+1}=2\middle|\, X_t=2\right)=1-\Pr\left(X_{t+1}=1\middle|\, X_t=2\right)=p_2$. Physical constraints tell us that $p_1+p_2<1$.

- *Example*: Application to Bayesian analysis of Markov chain. Consider a two state Markov chain with transition matrix F

$$\begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}$$

that is $\Pr(X_{t+1}=1|\, X_t=1) = 1-\Pr(X_{t+1}=2|\, X_t=1) = p_1$ and $\Pr(X_{t+1}=2|\, X_t=2) = 1-\Pr(X_{t+1}=1|\, X_t=2) = p_2$. Physical constraints tell us that $p_1+p_2<1$.

- Assume we observe $x_1, ..., x_m$ and the prior is

$$\pi(p_1, p_2) = 2\mathbb{I}_{p_1+p_2\le 1}$$

then the posterior is

$$\pi(p_1, p_2|\, x_{1:m}) \propto p_1^{m_{1,1}}(1-p_1)^{m_{1,2}}(1-p_2)^{m_{2,1}}\, p_2^{m_{2,2}}\mathbb{I}_{p_1+p_2\le 1}$$

where

$$m_{i,j} = \sum_{t=1}^{m-1}\mathbb{I}_{x_t=i}\mathbb{I}_{x_{t+1}=i}$$

- *Example*: Application to Bayesian analysis of Markov chain. Consider a two state Markov chain with transition matrix F

$$\begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}$$

that is $\Pr\left(X_{t+1}=1\middle|X_t=1\right)=1-\Pr\left(X_{t+1}=2\middle|X_t=1\right)=p_1$ and $\Pr\left(X_{t+1}=2\middle|X_t=2\right)=1-\Pr\left(X_{t+1}=1\middle|X_t=2\right)=p_2$. Physical constraints tell us that $p_1+p_2<1$.

- Assume we observe $x_1,...,x_m$ and the prior is

$$\pi\left(p_1,p_2\right)=2\mathbb{I}_{p_1+p_2\leq 1}$$

then the posterior is

$$\pi\left(p_1,p_2\middle|x_{1:m}\right)\propto p_1^{m_{1,1}}\left(1-p_1\right)^{m_{1,2}}\left(1-p_2\right)^{m_{2,1}}p_2^{m_{2,2}}\mathbb{I}_{p_1+p_2\leq 1}$$

where

$$m_{i,j}=\sum_{t=1}^{m-1}\mathbb{I}_{x_t=i}\mathbb{I}_{x_{t+1}=i}$$

- The posterior does not admit a standard expression and its normalizing constant is unnknown. We can sample from it using rejection sampling.

- We are interested in estimating $\mathbb{E}\left[\left.\varphi_i\left(p_1, p_2\right)\right| x_{1:m}\right]$ for $\varphi_1\left(p_1, p_2\right) = p_1$, $\varphi_2\left(p_1, p_2\right) = p_2$, $\varphi_3\left(p_1, p_2\right) = p_1 / \left(1 - p_1\right)$, $\varphi_4\left(p_1, p_2\right) = p_2 / \left(1 - p_2\right)$ and $\varphi_5\left(p_1, p_2\right) = \log \frac{p_1\left(1-p_2\right)}{p_2\left(1-p_1\right)}$ using Importance Sampling.

- We are interested in estimating $\mathbb{E}\left[\left.\varphi_i\left(p_1, p_2\right)\right| x_{1:m}\right]$ for $\varphi_1\left(p_1, p_2\right) = p_1$, $\varphi_2\left(p_1, p_2\right) = p_2$, $\varphi_3\left(p_1, p_2\right) = p_1 / \left(1 - p_1\right)$, $\varphi_4\left(p_1, p_2\right) = p_2 / \left(1 - p_2\right)$ and $\varphi_5\left(p_1, p_2\right) = \log \frac{p_1(1-p_2)}{p_2(1-p_1)}$ using Importance Sampling.
- If there was no on $p_1 + p_2 < 1$ and $\pi\left(p_1, p_2\right)$ was uniform on $[0, 1] \times [0, 1]$, then the posterior would be

$$\begin{aligned}\pi_0\left(\left.p_1, p_2\right| x_{1:m}\right) &= \mathcal{B}e\left(p_1; m_{1,1}+1, m_{1,2}+1\right) \\ &\quad \mathcal{B}e\left(p_2; m_{2,2}+1, m_{2,1}+1\right)\end{aligned}$$

but this is inefficient as for the given data $\left(m_{1,1}, m_{1,2}, m_{2,2}, m_{2,1}\right)$ we have $\pi_0\left(\left.p_1 + p_2 < 1\right| x_{1:m}\right) = 0.21$.

- We are interested in estimating $\mathbb{E}\left[\left.\varphi_i\left(p_1,p_2\right)\right| x_{1:m}\right]$ for $\varphi_1\left(p_1,p_2\right)=p_1$, $\varphi_2\left(p_1,p_2\right)=p_2$, $\varphi_3\left(p_1,p_2\right)=p_1/\left(1-p_1\right)$, $\varphi_4\left(p_1,p_2\right)=p_2/\left(1-p_2\right)$ and $\varphi_5\left(p_1,p_2\right)=\log\frac{p_1(1-p_2)}{p_2(1-p_1)}$ using Importance Sampling.
- If there was no on $p_1+p_2<1$ and $\pi\left(p_1,p_2\right)$ was uniform on $[0,1]\times[0,1]$, then the posterior would be

$$\begin{aligned}\pi_0\left(\left.p_1,p_2\right| x_{1:m}\right) &= \mathcal{B}e\left(p_1;m_{1,1}+1,m_{1,2}+1\right)\\ &\quad \mathcal{B}e\left(p_2;m_{2,2}+1,m_{2,1}+1\right)\end{aligned}$$

but this is inefficient as for the given data $\left(m_{1,1},m_{1,2},m_{2,2},m_{2,1}\right)$ we have $\pi_0\left(\left.p_1+p_2<1\right| x_{1:m}\right)=0.21$.
- The form of the posterior suggests using a Dirichlet distribution with density

$$\pi_1\left(\left.p_1,p_2\right| x_{1:m}\right)\propto p_1^{m_{1,1}}p_2^{m_{2,2}}\left(1-p_1-p_2\right)^{m_{1,2}+m_{2,1}}$$

but $\pi\left(\left.p_1,p_2\right| x_{1:m}\right)/\pi_1\left(\left.p_1,p_2\right| x_{1:m}\right)$ is unbounded.

- (Geweke, 1989) proposed using the normal approximation to the binomial distribution

$$\begin{aligned}\pi_2\left(p_1, p_2 \middle| x_{1:m}\right) &\propto \exp\left(-\left(m_{1,1}+m_{1,2}\right)\left(p_1-\widehat{p}_1\right)^2 / \left(2\widehat{p}_1\left(1-\widehat{p}_1\right)\right)\right) \\ &\quad \times \exp\left(-\left(m_{2,1}+m_{2,2}\right)\left(p_2-\widehat{p}_2\right)^2 / \left(2\widehat{p}_2\left(1-\widehat{p}_2\right)\right)\right)\end{aligned}$$

where $\widehat{p}_1 = m_{1,1} / \left(m_{1,1}+m_{1,2}\right)$, $\widehat{p}_1 = m_{2,2} / \left(m_{2,2}+m_{2,1}\right)$. Then to simulate from this distribution, we simulate first $\pi_2\left(p_1 \middle| x_{1:m}\right)$ and then $\pi_2\left(p_2 \middle| x_{1:m}, p_1\right)$ which are univariate truncated Gaussian distribution which can be sampled using the inverse cdf method. The ratio $\pi\left(p_1, p_2 \middle| x_{1:m}\right) / \pi_2\left(p_1, p_2 \middle| x_{1:m}\right)$ is upper bounded.

- (Geweke, 1989) proposed using the normal approximation to the binomial distribution

$$\begin{aligned}\pi_2\left(p_1, p_2 \middle| x_{1:m}\right) &\propto \exp\left(-\left(m_{1,1}+m_{1,2}\right)\left(p_1-\widehat{p}_1\right)^2 / \left(2\widehat{p}_1\left(1-\widehat{p}_1\right)\right)\right) \\ &\quad \times \exp\left(-\left(m_{2,1}+m_{2,2}\right)\left(p_2-\widehat{p}_2\right)^2 / \left(2\widehat{p}_2\left(1-\widehat{p}_2\right)\right)\right)\end{aligned}$$

where $\widehat{p}_1 = m_{1,1}/\left(m_{1,1}+m_{1,2}\right)$, $\widehat{p}_1 = m_{2,2}/\left(m_{2,2}+m_{2,1}\right)$. Then to simulate from this distribution, we simulate first $\pi_2\left(p_1 \middle| x_{1:m}\right)$ and then $\pi_2\left(p_2 \middle| x_{1:m}, p_1\right)$ which are univariate truncated Gaussian distribution which can be sampled using the inverse cdf method. The ratio $\pi\left(p_1, p_2 \middle| x_{1:m}\right)/\pi_2\left(p_1, p_2 \middle| x_{1:m}\right)$ is upper bounded.
- A final one consists of using

$$\pi_3\left(p_1, p_2 \middle| x_{1:m}\right) = \mathcal{B}e\left(p_1; m_{1,1}+1, m_{1,2}+1\right)\pi_3\left(p_2 \middle| x_{1:m}, p_1\right)$$

where $\pi\left(p_2 \middle| x_{1:m}, p_1\right) \propto \left(1-p_2\right)^{m_{2,1}} p_2^{m_{2,2}} \mathbb{I}_{p_2 \leq 1-p_1}$ is badly approximated through $\pi_3\left(p_2 \middle| x_{1:m}, p_1\right) = \frac{2}{\left(1-p_1\right)^2} p_2 \mathbb{I}_{p_2 \leq 1-p_1}$. It is straightforward to check that $\pi\left(p_1, p_2 \middle| x_{1:m}\right)/\pi_3\left(p_1, p_2 \middle| x_{1:m}\right) \propto \left(1-p_2\right)^{m_{2,1}} p_2^{m_{2,2}} / \frac{2}{\left(1-p_1\right)^2} p_2 < \infty$.

- Performance for $N = 10,000$

Distribution	φ_1	φ_2	φ_3	φ_4	φ_5
π_1	0.748	0.139	3.184	0.163	2.957
π_2	0.689	0.210	2.319	0.283	2.211
π_3	0.697	0.189	2.379	0.241	2.358
π	0.697	0.189	2.373	0.240	2.358

- Performance for $N = 10,000$

Distribution	φ_1	φ_2	φ_3	φ_4	φ_5
π_1	0.748	0.139	3.184	0.163	2.957
π_2	0.689	0.210	2.319	0.283	2.211
π_3	0.697	0.189	2.379	0.241	2.358
π	0.697	0.189	2.373	0.240	2.358

- Sampling from π using rejection sampling works well but is computationally expensive. π_3 is computationally much cheaper whereas π_1 does extremely poorly as expected.

Optimal Normalized Importance Sampling

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

$$q^{\text{opt}}(x) = \frac{|\varphi(x) - \mathbb{E}_{\pi}(\varphi)|\, \pi(x)}{\int_{\mathcal{X}} |\varphi(x) - \mathbb{E}_{\pi}(\varphi)|\, \pi(x)\, dx}$$

Optimal Normalized Importance Sampling

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

$$q^{\text{opt}}(x) = \frac{|\varphi(x) - \mathbb{E}_\pi(\varphi)|\,\pi(x)}{\int_{\mathcal{X}} |\varphi(x) - \mathbb{E}_\pi(\varphi)|\,\pi(x)\,dx}$$

- *Proof*:

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

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

Optimal Normalized Importance Sampling

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

$$q^{\text{opt}}(x) = \frac{|\varphi(x) - \mathbb{E}_\pi(\varphi)|\,\pi(x)}{\int_{\mathcal{X}} |\varphi(x) - \mathbb{E}_\pi(\varphi)|\,\pi(x)\,dx}$$

- *Proof*:

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

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

- This result is practically useless because it requires knowing $\mathbb{E}_\pi(\varphi)$ but it suggests approximations.

Effective Sample Size

- 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 $\pi(x)$.

Effective Sample Size

- 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

$$var\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right) \approx (1 + var_q(w(X)))\,\frac{var(\mathbb{E}_\pi(\varphi(X)))}{N}.$$

Effective Sample Size

- 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

$$var\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right) \approx (1 + var_q(w(X)))\frac{var(\mathbb{E}_\pi(\varphi(X)))}{N}.$$

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

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

Computing Ratio of Normalizing Constant

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

$$\frac{\int \pi^* (x)\, dx}{\int q^* (x)\, dx} = \int w^* (x)\, q (x)\, dx = \mathbb{E}_q [w^* (X)] .$$

using

$$\mathbb{E}_{\widehat{q}_N} [w^* (X)] = \frac{1}{N} \sum_{i=1}^{N} w^* \left(X^{(i)} \right)$$

Computing Ratio of Normalizing Constant

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

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

using

$$\mathbb{E}_{\widehat{q}_N}\left[w^{*}(X)\right] = \frac{1}{N}\sum_{i=1}^{N} w^{*}\left(X^{(i)}\right)$$

- It is unbiased and has variance

$$var\left[\mathbb{E}_{\widehat{q}_N}\left[w^{*}(X)\right]\right] = \frac{var_{q}\left(w^{*}(X)\right)}{N}.$$

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

$$var\left[\mathbb{E}_{\widehat{q}_N}\left[w^*(X)\right]\right] = 0$$

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

$$var\left[\mathbb{E}_{\widehat{q}_N}\left[w^*(X)\right]\right] = 0$$

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

$$\mathbb{E}_{\widehat{q}_N}\left[w^*(X)\right] = \frac{\int \pi^*(x)\,dx}{\int q^*(x)\,dx}.$$

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

$$var\left[\mathbb{E}_{\widehat{q}_N}\left[w^*(X)\right]\right] = 0$$

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

$$\mathbb{E}_{\widehat{q}_N}\left[w^*(X)\right] = \frac{\int \pi^*(x)\,dx}{\int q^*(x)\,dx}.$$

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

- IS is more powerful than you think.

- IS is more powerful than you think.
- Assume you have say to compute the importance weight

$$w(\theta) \propto \int f(x, z|\theta)\, dz$$

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

- IS is more powerful than you think.
- Assume you have say to compute the importance weight

$$w(\theta) \propto \int f(x, z|\,\theta)\, dz$$

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.

Limitations of Importance Sampling

- 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

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

Limitations of Importance Sampling

- 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

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

Limitations of Importance Sampling

- 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.

Normalized Importance Sampling versus Rejection Sampling

- Given N samples from q, we estimate $\mathbb{E}_{\pi}(\varphi(X))$ through IS

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

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

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

where $K \leq N$ is a random variable corresponding to the number of samples accepted.

Normalized Importance Sampling versus Rejection Sampling

- Given N samples from q, we estimate $\mathbb{E}_{\pi}(\varphi(X))$ through IS

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

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

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

  where $K \leq N$ is a random variable corresponding to the number of samples accepted.
- We want to know which strategy performs the best.

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

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

then

$$\int \overline{\pi}(x, y)\, dy = \int_0^{\frac{\pi^*(x)}{Cq^*(x)}} \frac{Cq^*(x)}{\int \pi^*(x)\, dx} dy = \pi(x).$$

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

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

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

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

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

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

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

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

- We have

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

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

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

- We have

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

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

- 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

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

- 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

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

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

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

- 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.

- 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 methods to address this problem.