# Advanced Simulation - Lecture 4

Patrick Rebeschini

January 24th, 2018

# Generic Sampling Methods

- For Monte Carlo methods, you need samples from distributions.

- Seen: inversion, transformation, composition, rejection.

- Today: importance sampling.

# Importance Sampling

- We want to compute

$$I = \mathbb{E}_{\pi}(\varphi(X)) = \int_{\mathbb{X}} \varphi(x)\, \pi(x)\, dx.$$

- We do not know how to sample from the target $\pi$ but have access to a proposal distribution of density $q$.

- We only require that

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

  i.e. the support of $q$ includes the support of $\pi$.

- $q$ is called the proposal, or importance, distribution.

# Importance Sampling

- We have the following identity

$$I = \mathbb{E}_\pi(\varphi(X)) = \mathbb{E}_q(\varphi(X) w\left(X\right)),$$

where $w : \mathbb{X} \to \mathbb{R}^+$ is the importance weight function

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

- Hence for $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} q$,

$$\widehat{I}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) w(X_i).$$

- It can be interpreted as performing the following approximation of $\pi$

$$\widehat{\pi}_n^{\text{IS}}\left(dx\right) = \frac{1}{n} \sum_{i=1}^{n} w(X_i) \delta_{X_i}\left(dx\right).$$

# Importance Sampling Properties

- **SLLN**: If $\mathbb{E}_q(|\varphi(X)|\, w\,(X)) < \infty$ then $\lim_{n\to\infty} \widehat{I}_n^{\mathrm{IS}} = I$.

- **Unbiased**: $\mathbb{E}_q\left(\widehat{I}_n^{\mathrm{IS}}\right) = I$.

- **Variance & CLT**: $\mathbb{V}_q\left(\widehat{I}_n^{\mathrm{IS}}\right) = \sigma_{\mathrm{IS}}^2/n$ where

$$\sigma_{\mathrm{IS}}^2 := \mathbb{V}_q\left(\varphi(X) w\,(X)\right)$$

and

$$\lim_{n\to\infty} \sqrt{n}\left(\widehat{I}_n^{\mathrm{IS}} - I\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \sigma_{\mathrm{IS}}^2\right).$$

# Importance Sampling: Practical Advices

- Consistency does not require $\sigma^2_{\text{IS}} < \infty$ but highly recommended in practice (!).

- **Sufficient condition**: If $\mathbb{E}_\pi \left(\varphi^2(X)\right) < \infty$ and $w\left(x\right) \leq M$ for all $x$ for some $M < \infty$, then $\sigma^2_{\text{IS}} < \infty$.

- In practice ensure $w\left(x\right) \leq M$ although it is neither necessary nor sufficient, as seen in the following example.

# Importance Sampling: Example

- $\pi(x) = \mathcal{N}(x; 0, 1)$, $q(x) = \mathcal{N}(x; 0, \sigma^2)$.

- For $\sigma^2 \geq 1$, $w(x) \leq M$ for all $x$,
  and for $\sigma^2 < 1$, $w(x) \to \infty$ as $|x| \to \infty$.

- For $\varphi(x) = x^2$, we have $\sigma^2_{\text{IS}} < \infty$ for all $\sigma^2 > 1/2$.

- For $\varphi(x) = \exp\left(\frac{\beta}{2} x^2\right)$, we have $I < \infty$ for $\beta < 1$
  but $\sigma^2_{\text{IS}} = \infty$ for $\beta > 1 - \frac{1}{2\sigma^2}$.

# Optimal Importance Distribution

- **Proposition:** The optimal proposal minimising $\mathbb{V}_q\left(\widehat{I}_n^{\mathrm{IS}}\right)$ is given by

$$q_{\mathrm{opt}}\left(x\right) = \frac{\left|\varphi(x)\right| \pi\left(x\right)}{\int_{\mathbb{X}} \left|\varphi(x)\right| \pi\left(x\right) dx}.$$

- **Proof**. We have indeed

$$\mathbb{V}_q\left(\varphi(X) w\left(X\right)\right) = \mathbb{E}_q\left(\varphi^2(X) w^2\left(X\right)\right) - I^2.$$

For $q = q_{\mathrm{opt}}$, we have

$$\begin{aligned}
\mathbb{E}_{q_{\mathrm{opt}}}\left(\varphi^2(X) w^2\left(X\right)\right) &= \int_{\mathbb{X}} \frac{\varphi^2(x)\pi^2\left(x\right)}{\left|\varphi(x)\right| \pi\left(x\right)} dx. \int_{\mathbb{X}} \left|\varphi(x)\right| \pi\left(x\right) dx \\
&= \left(\int_{\mathbb{X}} \left|\varphi(x)\right| \pi\left(x\right) dx\right)^2
\end{aligned}$$

We also have by Jensen's inequality for any $q$

$$\mathbb{E}_q\left(\varphi^2(X) w^2\left(X\right)\right) \geq \mathbb{E}_q^2\left(\left|\varphi(X)\right| w\left(X\right)\right) = \left(\int_{\mathbb{X}} \left|\varphi(x)\right| \pi\left(x\right) dx\right)^2.$$

# Optimal Importance Distribution

- $q_{\text{opt}}(x)$ can never be used in practice!
- For $\varphi(x) > 0$ we have $q_{\text{opt}}(x) = \varphi(x)\pi(x)/I$ and $\mathbb{V}_{q_{\text{opt}}}\left(\widehat{I}_n^{\text{IS}}\right) = 0$ but this is because

  $$\varphi(x)\, w(x) = \varphi(x)\, \frac{\pi(x)}{q_{\text{opt}}(x)} = I,$$

  it requires knowing $I$!
- This can be used as a guideline to select $q$; i.e. select $q(x)$ such that $q(x) \approx q_{\text{opt}}(x)$.
- Particularly interesting in rare event simulation, not quite in statistics.

# Normalised Importance Sampling

- Standard IS has limited applications in statistics as it requires knowing $\pi(x)$ and $q(x)$ exactly.

- Assume $\pi(x) = C_\pi \times \pi_u(x)$ and $q(x) = C_q \times q_u(x)$, $\pi(x) > 0 \Rightarrow q(x) > 0$ and and define

$$w_u(x) = \frac{\pi_u(x)}{q_u(x)}.$$

- An alternative identity is

$$I = \mathbb{E}_\pi(\varphi(X)) = \frac{\int_{\mathbb{X}} \varphi(x)\, w_u(x)\, q(x) dx}{\int_{\mathbb{X}} w_u(x) q(x) dx}.$$

- Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} q$ then

$$\widehat{I}_n^{\text{NIS}} = \frac{\sum_{i=1}^n \varphi(X_i) w_u(X_i)}{\sum_{i=1}^n w_u(X_i)}$$

is strongly consistent through the SLLN as long as $\mathbb{E}_q(|\varphi(X)|\, w(X)) < \infty$.

- Variance of IS:

$$\mathbb{V}\left(\widehat{I}_n^{\text{IS}}\right) = \frac{1}{n} \int \frac{(\varphi(x)\pi(x) - Iq(x))^2}{q(x)} dx$$

while variance of NIS (using the Delta method):

$$\mathbb{V}\left(\widehat{I}_n^{\text{NIS}}\right) = \frac{1}{n} \int \frac{\pi(x)^2 (\varphi(x) - I)^2}{q(x)} dx.$$

# Details on the Delta method

If

$$\sqrt{n}\left(\begin{pmatrix}X_n\\Y_n\end{pmatrix}-\begin{pmatrix}\mu_x\\\mu_y\end{pmatrix}\right)\xrightarrow{D}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\Sigma_{xx}&\Sigma_{xy}\\\Sigma_{yx}&\Sigma_{yy}\end{pmatrix}\right)$$

and $g:(x\ y)^T\mapsto g(x,y)$ then

$$\sqrt{n}\left(g\begin{pmatrix}X_n\\Y_n\end{pmatrix}-g\begin{pmatrix}\mu_x\\\mu_y\end{pmatrix}\right)\xrightarrow{D}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\nabla g_{|\mu}^T\begin{pmatrix}\Sigma_{xx}&\Sigma_{xy}\\\Sigma_{yx}&\Sigma_{yy}\end{pmatrix}\nabla g_{|\mu}\right).$$

With $g:(x\ y)^T\mapsto x/y$ we have

$$\nabla g:(x\ y)^T\mapsto\begin{pmatrix}\frac{1}{y}&-\frac{x}{y^2},\end{pmatrix}^T$$

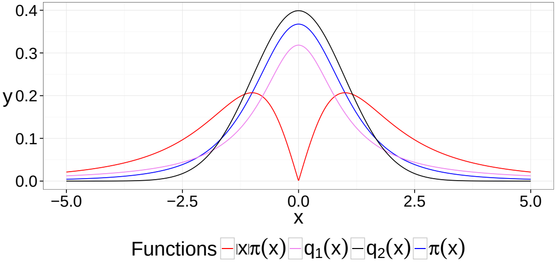
and thus

$$\nabla g_{|\mu}=\begin{pmatrix}\frac{1}{\mu_y}&-\frac{\mu_x}{\mu_y^2}.\end{pmatrix}^T$$

# Toy Example: t-distribution

- We want to compute $I = \mathbb{E}_\pi(|X|)$ where $\pi(x) \propto \left(1 + x^2/3\right)^{-2}$ ($t_3$-distribution).

1. Directly sample from $\pi$.
2. Use $q_1(x) = g_{t_1}(x) \propto \left(1 + x^2\right)^{-1}$ ($t_1$-distribution).
3. Use $q_2(x) \propto \exp\left(-x^2/2\right)$ (normal).



# Toy Example: t-distribution

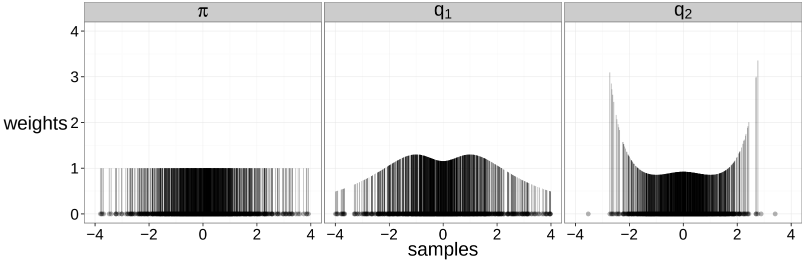



Figure: Sample weights obtained for 1000 realisations of $X_i$, from the different proposal distributions.

# Toy Example: t-distribution

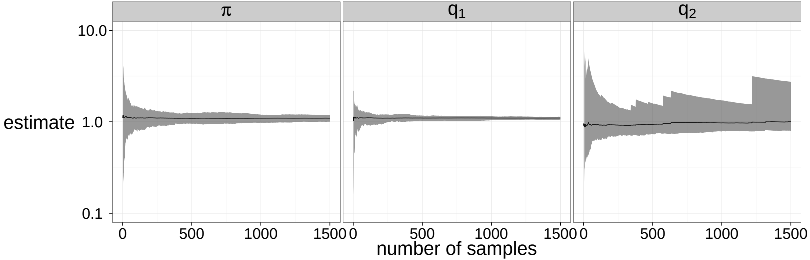



Figure: Estimates $\widehat{I}_n$ of $I$ obtained after 1 to 1500 samples. The grey shaded areas correpond to the range of 100 independent replications.

# Variance of importance sampling estimators

- Standard Importance Sampling: $X_1, \ldots, X_n \overset{iid}{\sim} q$,

$$\widehat{I}_n^{\mathrm{IS}} = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) w(X_i).$$

- Asymptotic Variance:

$$\begin{aligned}\mathbb{V}_{as}\left(\widehat{I}_n^{\mathrm{IS}}\right) &= \mathbb{E}_q\left[\left(\varphi(X)w(X) - \mathbb{E}_q\left(\varphi(X)w(X)\right)\right)^2\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i)w(X_i) - \widehat{I}_n^{\mathrm{IS}}\right)^2.\end{aligned}$$

- Thus the asymptotic variance can be estimated consistently with

$$\frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i)w(X_i) - \widehat{I}_n^{\mathrm{IS}}\right)^2.$$

# Variance of importance sampling estimators

- Normalised Importance Sampling: $X_1, \ldots, X_n \overset{iid}{\sim} q$,

$$\widehat{I}_n^{\text{NIS}} = \frac{\sum_{i=1}^n \varphi(X_i) w_u(X_i)}{\sum_{i=1}^n w_u(X_i)}.$$

- Asymptotic Variance:

$$\mathbb{V}_{as}\left(\widehat{I}_n^{\text{NIS}}\right) = \frac{\mathbb{E}_q\left[\left(\varphi(X) w(X) - I \times w(X)\right)^2\right]}{\mathbb{E}_q\left[w(X)\right]^2}.$$

- Thus the asymptotic variance can be estimated consistently with

$$\frac{\frac{1}{n}\sum_{i=1}^N w_u(X_i)^2\left(\varphi(X_i) - \widehat{I}_n^{\text{NIS}}\right)^2}{\left(\frac{1}{n}\sum_{i=1}^N w_u(X_i)\right)^2}.$$

# Diagnostics

- If only one weight, say $w_u(X_j)$, is significant compared to the others, then

$$\widehat{I}_n^{\mathrm{NIS}} = \frac{\sum_{i=1}^n \varphi(X_i) w_u(X_i)}{\sum_{i=1}^n w_u(X_i)} \approx \varphi(X_j).$$

  The "effective sample size" is one.

- To how many unweighted samples correspond our weighted samples of size $n$? Solve for $n_e$ in

$$\frac{1}{n}\mathbb{V}_{as}\left(\widehat{I}_n^{\mathrm{NIS}}\right) = \frac{\sigma^2}{n_e},$$

  where $\sigma^2/n_e$ corresponds to the variance of an unweighted sample of size $n_e$.

# Diagnostics

- We solve by matching $\varphi(X_i) - \widehat{I}^{\text{NIS}}$ with $\varphi(X_i) - I \approx \sigma$ as if they were i.i.d samples:

$$\frac{1}{n} \frac{\frac{1}{n}\sum_{i=1}^{N} w_u(X_i)^2 \left(\varphi(X_i) - \widehat{I}_n^{\text{NIS}}\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{N} w_u(X_i)\right)^2} \approx \frac{\sigma^2}{n_e}$$

$$\text{i.e.} \qquad \frac{1}{n} \frac{\frac{1}{n}\sum_{i=1}^{N} w_u(X_i)^2}{\left(\frac{1}{n}\sum_{i=1}^{N} w_u(X_i)\right)^2} = \frac{1}{n_e}.$$

- The solution is

$$n_e = \frac{\left(\sum_{i=1}^{n} w_u(X_i)\right)^2}{\sum_{i=1}^{n} w_u(X_i)^2},$$

and is called the effective sample size.

# Rejection and Importance Sampling in High Dimensions

- **Toy example:** Let $\mathbb{X} = \mathbb{R}^d$ and

$$\pi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\sum_{i=1}^{d} x_i^2}{2}\right)$$

and

$$q(x) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\sum_{i=1}^{d} x_i^2}{2\sigma^2}\right).$$

- How do Rejection sampling and Importance sampling scale in this context?

# Performance of Rejection Sampling

- We have

$$w(x) = \frac{\pi(x)}{q(x)} = \sigma^d \exp\left(-\frac{\sum_{i=1}^d x_i^2}{2}\left(1 - \frac{1}{\sigma^2}\right)\right) \leq \sigma^d$$

  for $\sigma > 1$.

- Acceptance probability is

$$\mathbb{P}(X \text{ accepted}) = \frac{1}{\sigma^d} \to 0 \text{ as } d \to \infty,$$

  i.e. exponential degradation of performance.

- For $d = 100$, $\sigma = 1.2$, we have

$$\mathbb{P}(X \text{ accepted}) \approx 1.2 \times 10^{-8}.$$

# Performance of Importance Sampling

- We have

$$w\left(x\right)=\sigma^{d}\exp\left(-\frac{\sum_{i=1}^{d}x_{i}^{2}}{2}\left(1-\frac{1}{\sigma^{2}}\right)\right).$$

- Variance of the weights:

$$\mathbb{V}_{q}\left[w\left(X\right)\right]=\left(\frac{\sigma^{4}}{2\sigma^{2}-1}\right)^{d/2}-1$$

where $\sigma^{4}/\left(2\sigma^{2}-1\right)>1$ for any $\sigma^{2}>1/2$.

- For $d=100$, $\sigma=1.2$, we have

$$\mathbb{V}_{q}\left[w\left(X\right)\right]\approx 1.8\times 10^{4}.$$

# Wait a minute. . .

Lecture 1:

- Simpson's rule for approximating integrals: error in $\mathcal{O}(n^{-1/d})$.

Lecture 2:

- Monte Carlo for approximating integrals: error in $\mathcal{O}(n^{-1/2})$ with rate independent of $d$.

And now:

- Importance Sampling standard deviation in the Gaussian example in $\exp(d)n^{-1/2}$.

The rate is indeed independent of $d$ but the "constant" (in $n$) explodes exponentially (in $d$).

# Markov chain Monte Carlo

- Revolutionary idea introduced by Metropolis et al., J. Chemical Physics, 1953.
- **Key idea**: Given a target distribution $\pi$, build a Markov chain $(X_t)_{t \geq 1}$ such that, as $t \to \infty$, $X_t \sim \pi$ and

$$\frac{1}{n} \sum_{t=1}^{n} \varphi(X_t) \to \int \varphi(x) \pi(x) \, dx$$

  when $n \to \infty$ e.g. almost surely.
- Central limit theorems with a rate in $1/\sqrt{n}$.
- In some cases the constant (in $n$) does not explode exponentially with the dimension $d$, but polynomially.

# Side Dish: Control Variates

- Variance reduction techniques, not always applicable but useful in some cases.
- Suppose that we want to compute

$$I = \int \varphi(x)\pi(x)dx$$

and that we know exactly

$$J = \int \psi(x)\pi(x)dx.$$

- Sample $X_1, \dots, X_n$ from $\pi$ and compute

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i) - \lambda(\psi(X_i) - J)\right).$$

- What is the benefit of $\widehat{I}_n$ over the standard Monte Carlo estimator?