Advanced Simulation - Lecture 3

George Deligiannidis

January 25th, 2016
Recall

**Algorithm (Rejection Sampling).** Given two densities $\pi, q$ with $\pi(x) \leq M q(x)$ for all $x$, we can generate a sample from $\pi$ by

1. Draw $X \sim q$, draw $U \sim U_{[0,1]}$.
2. Accept $X = x$ as a sample from $\pi$ if

$$U \leq \frac{\pi(x)}{M q(x)},$$

otherwise go to step 1.

**Proposition**

*The distribution of the samples accepted by rejection sampling is $\pi$.***
Rejection Sampling

- Often we only know $\pi$ and $q$ up to some normalising constants; i.e.
  \[ \pi = \tilde{\pi} / Z_{\pi} \quad \text{and} \quad q = \tilde{q} / Z_q \]
  where $\tilde{\pi}, \tilde{q}$ are known but $Z_{\pi}, Z_q$ are unknown. 
  You still need to be able to sample from $q(\cdot)$.

- If you can upper bound:
  \[ \tilde{\pi}(x) / \tilde{q}(x) \leq \tilde{M}, \]
  then using $\tilde{\pi}, \tilde{q}$ and $\tilde{M}$ in the algorithm is correct.

- Indeed we have
  \[ \frac{\tilde{\pi}(x)}{\tilde{q}(x)} \leq \tilde{M} \iff \frac{\pi(x)}{q(x)} \leq \tilde{M} \frac{Z_q}{Z_{\pi}} = M. \]
Rejection Sampling

Let $T$ denote the number of pairs $(X, U)$ that have to be generated until $X$ is accepted for the first time.

**Lemma**

*T is geometrically distributed with parameter $1/M$ and in particular $\mathbb{E}(T) = M$.*

In the unnormalised case, this yields

$$\mathbb{P}(X \text{ accepted}) = \frac{1}{M} = \frac{Z_\pi}{\tilde{Z} q},$$

$$\mathbb{E}(T) = M = \frac{Z_q \tilde{M}}{Z_\pi},$$

and it can be used to provide unbiased estimates of $Z_\pi/Z_q$ and $Z_q/Z_\pi$. 
Let $B \subset \mathbb{R}^p$, a bounded subset of $\mathbb{R}^p$:

$$\pi(x) \propto I_B(x).$$

Let $R$ be a rectangle containing $B \subset R$ and

$$q(x) \propto I_R(x).$$

Then we can use $\tilde{M} = 1$ and

$$\tilde{\pi}(x) / \left( \tilde{M}' \tilde{q}(x) \right) = I_B(x).$$

The probability of accepting a sample is then $Z_{\pi} / Z_q$. 
Example: Normal density

- Let \( \tilde{\pi}(x) = \exp\left(-\frac{1}{2} x^2\right) \) and \( \tilde{q}(x) = \frac{1}{1 + x^2} \). We have

\[
\frac{\tilde{\pi}(x)}{\tilde{q}(x)} = (1 + x^2) \exp\left(-\frac{1}{2} x^2\right) \leq \frac{2}{\sqrt{e}} = \tilde{M}
\]

which is attained at \( \pm 1 \).

- Let \( X \sim \tilde{q} \). The acceptance probability is

\[
\mathbb{P}\left(U \leq \frac{\tilde{\pi}(X)}{\tilde{M}\tilde{q}(X)}\right) = \frac{Z_\pi}{\tilde{M}Z_q} = \frac{\sqrt{2\pi}}{\frac{2}{\sqrt{e}} \pi} = \sqrt{\frac{e}{2\pi}} \approx 0.66,
\]

and the mean number of trials to success is approximately \( 1/0.66 \approx 1.52 \).
Examples: Genetic Linkage model

- We observe

\[ (Y_1, Y_2, Y_3, Y_4) \sim \mathcal{M} \left( n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4} (1 - \theta), \frac{1}{4} (1 - \theta), \frac{\theta}{4} \right) \]

where \( \mathcal{M} \) is the **multinomial distribution** and \( \theta \in (0, 1) \).

- The likelihood of the observations is thus

\[
p (y_1, ..., y_4; \theta) = \frac{n!}{y_1! y_2! y_3! y_4!} \left( \frac{1}{2} + \frac{\theta}{4} \right)^{y_1} \left( \frac{1}{4} (1 - \theta) \right)^{y_2+y_3} \left( \frac{\theta}{4} \right)^{y_4} \propto (2 + \theta)^{y_1} (1 - \theta)^{y_2+y_3} \theta^{y_4}.
\]

- Bayesian approach where we select \( p (\theta) = \mathbb{I}_{[0,1]} (\theta) \) and are interested in

\[
p (\theta | y_1, ..., y_4) \propto (2 + \theta)^{y_1} (1 - \theta)^{y_2+y_3} \theta^{y_4} \mathbb{I}_{[0,1]} (\theta).
\]
Examples: Genetic linkage model

- Rejection sampling using the prior as proposal
  \[ q(\theta) = \tilde{q}(\theta) = p(\theta) \]
  to sample from \[ p(\theta \mid y_1, ..., y_4) \].

- To use accept-reject, we need to upper bound
  \[ \frac{\tilde{\pi}(\theta)}{\tilde{q}(\theta)} = \tilde{\pi}(\theta) = (2 + \theta)^{y_1}(1 - \theta)^{y_2 + y_3} \theta^{y_4} \]

- Maximum of \( \tilde{\pi} \) can be found using standard optimization procedure to perform rejection sampling.

- For a realisation of \( (Y_1, Y_2, Y_3, Y_4) \) equal to \( (69, 9, 11, 11) \)
  obtained with \( n = 100 \) and \( \theta^* = 0.6 \), results shown in following figure.
Examples: Genetic linkage model

Figure: Histogram of 10,000 samples drawn from posterior obtained by rejection sampling (left); and histogram of waiting time distribution before acceptance (right).
Rejection sampling requires

1. Samples from some distribution $q$;

2. Evaluation of $\pi(\cdot)$ point-wise, or unnormalized $\tilde{\pi}$;

3. An upper bound $M$ on $\pi(x)/q(x)$, or $\tilde{\pi}/q$ and so on.

Sometimes the upper bound is not feasible.
Importance Sampling

- We want to compute

\[ I = \mathbb{E}_\pi(\varphi(X)) = \int_X \varphi(x) \pi(x) \, dx. \]

- 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

\[ \pi(x) > 0 \implies q(x) > 0; \]

i.e. the support of \( q \) includes the support of \( \pi \).

- \( q \) is called the proposal, or importance distribution.
We have the following identity

\[ I = \mathbb{E}_\pi(\varphi(X)) = \mathbb{E}_q(\varphi(X)w(X)), \]

where \( w : \mathbb{X} \rightarrow \mathbb{R}^+ \) is the importance weight function

\[ w(x) = \frac{\pi(x)}{q(x)}. \]

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

\[ \widehat{I}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)w(X_i). \]
Importance Sampling Properties

Proposition

(a) **Unbiased:** \( \mathbb{E}_q[\hat{I}^{IS}_n] = I; \)

(b) **Strongly consistent:** If \( \mathbb{E}_q(|\varphi(X)|w(X)) < \infty \) then

\[
\lim_{n \to \infty} \hat{I}^{IS}_n = I, \quad \text{a.s.}
\]

(c) **CLT:** \( \mathbb{V}_q(\hat{I}^{IS}_n) = \sigma_{IS}^2/n \) where

\[
\sigma_{IS}^2 := \mathbb{V}_q(\varphi(X)w(X))
\]

If \( \sigma_{IS}^2 < \infty \) then

\[
\lim_{n \to \infty} \sqrt{n} \left( \hat{I}^{IS}_n - I \right) \xrightarrow{D} \mathcal{N}(0, \sigma_{IS}^2).
\]
Consistency does not require $\sigma_{\text{IS}}^2 < \infty$ but highly recommended in practice (!).

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

In practice ensure $w(x) \leq M$ although it is neither necessary nor sufficient, as seen in the following example.
Importance Sampling: Example

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

\[
    w(x) = \frac{\pi(x)}{q(x)} \propto \exp \left[ -x^2 \left( 1 - \frac{1}{\sigma^2} \right) \right].
\]

- 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 \( \sigma_{IS}^2 < \infty \) 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 \( \sigma_{IS}^2 = \infty \) for \( \beta > 1 - \frac{1}{2\sigma^2} \).
Question

Is there a best proposal that minimizes the variance \( \sigma_{\text{IS}}^2 \)?

Proposition

The optimal proposal minimising \( \mathbb{V}_q \left( \hat{I}_n^{\text{IS}} \right) \) is given by

\[
q_{\text{opt}}(x) = \frac{|\varphi(x)| \pi(x)}{\int_X |\varphi(x)| \pi(x) \, dx}.
\]
Proof.

We have indeed

$$\sigma_{IS}^2 = \nabla_q (\phi(X)w(X)) = \mathbb{E}_q (\phi^2(X)w^2(X)) - I^2.$$

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

$$\mathbb{E}_q (\phi^2(X)w^2(X)) \geq \left( \int_X |\phi(x)| \pi(x) \, dx \right)^2.$$

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

$$\mathbb{E}_{q_{opt}} (\phi^2(X)w^2(X)) = \int_X \frac{\phi^2(x)\pi^2(x)}{|\phi(x)| \pi(x)} \, dx \times \int_X |\phi(x)| \pi(x) \, dx$$

$$= \left( \int_X |\phi(x)| \pi(x) \, dx \right)^2.$$
Optimal Importance Distribution

- $q_{\text{opt}}(x)$ can never be used in practice!

- For $\phi(x) > 0$ we have $q_{\text{opt}}(x) = \phi(x)\pi(x) / I$ and $\nabla_{q_{\text{opt}}} \left( \widehat{I}_{n}^{\text{IS}} \right) = 0$ but this is because

$$\phi(x) \, \omega(x) = \phi(x) \frac{\pi(x)}{q_{\text{opt}}(x)} = I,$$

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 $\pi(x) = \tilde{\pi}(x)/Z_\pi$ and $q(x) = \tilde{q}(x)/Z_q$, $\pi(x) > 0 \Rightarrow q(x) > 0$ and and define

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

- An alternative identity is

  $$I = E_\pi(\varphi(X)) = \frac{\int_X \varphi(x) \tilde{w}(x) q(x) dx}{\int_X \tilde{w}(x) q(x) dx}.$$
Proposition (SLLN for NIS)

Let \( X_1, ..., X_n \) \( i.i.d. \sim q \) and assume that \( \mathbb{E}_q(\lvert \varphi(X) \rvert w(X)) < \infty \). Then

\[
\hat{I}_n^{NIS} = \frac{\sum_{i=1}^{n} \varphi(X_i) \tilde{w}(X_i)}{\sum_{i=1}^{n} \tilde{w}(X_i)}
\]

is strongly consistent.

Proof.

Divide numerator and denominator by \( n \). Both converge almost surely by the strong law of large numbers.
Proposition

If $V_q(\varphi(X)w(X)) < \infty$ and $V_q(w(X)) < \infty$ then

$$\sqrt{n}(\hat{I}_n^\text{NIS} - I) \Rightarrow \mathcal{N}(0, \sigma^2_{\text{NIS}}),$$

where

$$\sigma^2_{\text{NIS}} := V_q\left(\left[\varphi(X)w(X)\right] - Iw(X)\right)$$

$$= \int \frac{\pi(x)^2 (\varphi(x) - I)^2}{q(x)} \, dx.$$
Proof.

First notice that with $X_1, \ldots, X_n$ i.i.d. $\sim q$

$$\sqrt{n}(\hat{I}_{n}^{\text{NIS}} - I) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{w}(X_i) [\varphi(X_i) - I]$$

where since $\tilde{w}(x) = \tilde{\pi} / \tilde{q}$

$$\mathbb{E}_q \left[ \tilde{w}(X_n)(\varphi(X_i) - I) \right] = 0.$$

Since $\mathbb{V}_q(\varphi(X)w(X)) < \infty$ by standard CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{w}(X_i) [\varphi(X_i) - I] \Rightarrow \mathcal{N}\left(0, \mathbb{V}_q \left( \tilde{w}(X_1)[\varphi(X_1) - I] \right) \right).$$
Proof.

The strong law of large numbers applied to the denominator

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{w}(X_i) \to \mathbb{E}_q[\tilde{w}(X_1)] = Z_{\pi} / Z_q, \quad \text{a.s.}$$

By Slutsky’s theorem, combining the two

$$\sqrt{n}(\hat{I}_n^{\text{NIS}} - I) \Rightarrow \mathcal{N}\left(0, \mathbb{V}_q(\tilde{w}(X_1) [\varphi(X_1) - I]) \frac{Z_q^2}{Z_{\pi}^2}\right)$$

$$\sim \mathcal{N}\left(0, \sigma_{\text{NIS}}^2 \right).$$

Alternatively, use Delta method.
We want to compute $I = \mathbb{E}_\pi(|X|)$ where $\pi(x) \propto (1 + x^2/3)^{-2}$ (t$_3$-distribution).

1. Directly sample from $\pi$.
2. Use $q_1(x) = g_{t_1}(x) \propto (1 + x^2)^{-1}$ (t$_1$-distribution).
3. Use $q_2(x) \propto \exp(-x^2/2)$ (normal).
Toy Example: t-distribution

Figure: Sample weights obtained for 1000 realisations of $X_i$, from the different proposal distributions.
Figure: Estimates $\hat{I}_n$ of $I$ obtained after 1 to 1500 samples. The grey shaded areas correspond to the range of 100 independent replications.
Variance of importance sampling estimators

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

\[
\hat{I}_n^{IS} = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)w(X_i). 
\]

- **Asymptotic Variance**:

\[
\text{Var}_{as} \left( \hat{I}_n^{IS} \right) = \mathbb{E}_q \left[ \left( \phi(X)w(X) - \mathbb{E}_q (\phi(X)w(X)) \right)^2 \right] 
\approx \frac{1}{n} \sum_{i=1}^{n} \left( \phi(X_i)w(X_i) - \hat{I}_n^{IS} \right)^2. 
\]

- Thus the asymptotic variance can be estimated consistently with

\[
\frac{1}{n} \sum_{i=1}^{n} \left( \phi(X_i)w(X_i) - \hat{I}_n^{IS} \right)^2. 
\]
Variance of importance sampling estimators

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

\[
\hat{I}_{NIS}^n = \frac{\sum_{i=1}^{n} \varphi(X_i) \tilde{w}(X_i)}{\sum_{i=1}^{n} \tilde{w}(X_i)}.
\]

- **Asymptotic Variance:**

\[
\text{Var}_{as} \left( \hat{I}_{n}^{NIS} \right) = \frac{\mathbb{E}_q \left[ (\varphi(X)w(X) - I \times w(X))^2 \right]}{\mathbb{E}_q [w(X)]^2}.
\]

- Thus the asymptotic variance can be estimated consistently with

\[
\frac{1}{n} \sum_{i=1}^{N} \tilde{w}(X_i)^2 \left( \varphi(X_i) - \hat{I}_{n}^{NIS} \right)^2
\]

\[
\left( \frac{1}{n} \sum_{i=1}^{N} \tilde{w}(X_i) \right)^2.
\]
Diagnostics

- Importance sampling works well when all weights roughly equal.
- If dominated by one $\tilde{w}(X_j)$,

$$\hat{I}_{n}^{\text{NIS}} = \frac{\sum_{i=1}^{n} \varphi(X_i)\tilde{w}(X_i)}{\sum_{i=1}^{n} \tilde{w}(X_i)} \approx \tilde{w}(X_j)\varphi(X_j).$$

The “effective sample size” is one.

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

$$\frac{1}{n} \text{Var}_{\text{as}}\left(\hat{I}_{n}^{\text{NIS}}\right) = \frac{\sigma^2}{n_e},$$

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

Diagnostics

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

$$
\frac{1}{n} \frac{1}{n} \sum_{i=1}^{N} \tilde{w}(X_i)^2 \left( \varphi(X_i) - \hat{I}_n \right)^2 \\
\approx \frac{\sigma^2}{n_e}
$$

i.e.

$$
\frac{1}{n} \frac{1}{n} \sum_{i=1}^{n} \tilde{w}(X_i)^2 \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{w}(X_i) \right)^2 = \frac{1}{n_e}.
$$

- The solution is

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

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

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

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

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

- Variance of the weights:

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

where \( \sigma^4 / (2\sigma^2 - 1) > 1 \) for any \( \sigma^2 > 1/2 \).

- For \( d = 100, \sigma = 1.2 \), we have

\[ \mathbb{V}_q[w(X)] \approx 1.8 \times 10^4. \]
Lecture 1:
- Simpson’s rule for approximating integrals: error in $O(n^{-1/d})$.

Lecture 2:
- Monte Carlo for approximating integrals: error in $O(n^{-1/2})$ 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 \(X_1, \ldots, X_n\) from \(\pi\) and compute
  \[ \hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} (\varphi(X_i) - \lambda(\psi(X_i) - J)) . \]
- What is the benefit of \(\hat{I}_n\) over the standard Monte Carlo estimator?