Monte Carlo Methods: Lecture 3 : Importance Sampling

Nick Whiteley

Overview of this lecture

What we have seen ...

Rejection sampling.

This lecture will cover ...

Importance sampling.

- Basic importance sampling
- Importance sampling using self-normalised weights
- Finite variance estimates
- Optimal proposals
- Example

Recall rejection sampling

Algorithm 2.1: Rejection sampling

Given two densities f, g with $f(x) < M \cdot g(x)$ for all x, we can generate a sample from f by

1. Draw $X \sim g$.
2. Accept X as a sample from f with probability

$$\frac{f(X)}{M \cdot g(X)},$$

otherwise go back to step 1.

Drawbacks:

- We need that $f(x) < M \cdot g(x)$
- On average we need to repeat the first step M times before we can accept a value proposed by g.

2.3 Importance sampling

The fundamental identities behind importance sampling (1)

Assume that $g(x) > 0$ for (almost) all x with $f(x) > 0$. Then for a measurable set A:

$$\mathbb{P}(X \in A) = \int_A f(x)\, dx = \int_A g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} dx = \int_A g(x) w(x)\, dx$$

For some integrable function h, assume that $g(x) > 0$ for (almost) all x with $f(x) \cdot h(x) \neq 0$

$$\mathbb{E}_f(h(X)) = \int f(x) h(x)\, dx = \int g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} h(x)\, dx$$

$$= \int g(x) w(x) h(x)\, dx = \mathbb{E}_g(w(X) \cdot h(X)),$$

The fundamental identities behind importance sampling (2)

- How can we make use of $\mathbb{E}_f(h(X)) = \mathbb{E}_g(w(X) \cdot h(X))$?
- Consider $X_1, \ldots, X_n \sim g$ and $\mathbb{E}_g|w(X) \cdot h(X)| < +\infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_g(w(X) \cdot h(X))$$

(law of large numbers), which implies

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_f(h(X)).$$

- Thus we can estimate $\mu := \mathbb{E}_f(h(X))$ by
  1. Sample $X_1, \ldots, X_n \sim g$
  2. $\tilde{\mu} := \frac{1}{n}\sum_{i=1}^{n} w(X_i)h(X_i)$

The importance sampling algorithm

Algorithm 2.1a: Importance Sampling

Choose g such that $\text{supp}(g) \supset \text{supp}(f \cdot h)$.

1. For $i = 1, \ldots, n$:
   i. Generate $X_i \sim g$.
   ii. Set $w(X_i) = \frac{f(X_i)}{g(X_i)}$.
2. Return

$$\tilde{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{n}$$

as an estimate of $\mathbb{E}_f(h(X))$.

- Contrary to rejection sampling, importance sampling does not yield realisations from f, but a *weighted sample* (X_i, W_i).
- The weighted sample can be used for estimating expectations $\mathbb{E}_f(h(X))$ (and thus probabilities, etc.)

Basic properties of the importance sampling estimate

- We have already seen that $\tilde{\mu}$ is consistent if $\mathrm{supp}(g) \supset \mathrm{supp}(f \cdot h)$ and $\mathbb{E}_g|w(X) \cdot h(X)| < +\infty$, as

$$\tilde{\mu} := \frac{1}{n} \sum_{i=1}^{n} w(X_i)h(X_i) \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_f(h(X))$$

- The expected value of the weights is $\mathbb{E}_g(w(X)) = 1$.
- $\tilde{\mu}$ is unbiased (see theorem below)

Theorem 2.2: Bias and Variance of Importance Sampling

$$\begin{aligned} \mathbb{E}_g(\tilde{\mu}) &= \mu \\ \mathrm{Var}_g(\tilde{\mu}) &= \frac{\mathrm{Var}_g(w(X) \cdot h(X))}{n} \end{aligned}$$

What if f is known only up to a multiplicative constant?

- Assume $f(x) = C\pi(x)$. Then

$$\tilde{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{C\pi(X_i)}{g(X_i)} h(X_i)$$

- Idea: Estimate $1/C$ as well. Consider the estimator

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{\sum_{i=1}^{n} w(X_i)}$$

- Now we have that

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{\sum_{i=1}^{n} w(X_i)} = \frac{\sum_{i=1}^{n} \frac{\pi(X_i)}{g(X_i)} h(X_i)}{\sum_{i=1}^{n} \frac{\pi(X_i)}{g(X_i)}},$$

$\rightsquigarrow$ $\hat{\mu}$ does not depend on C

The importance sampling algorithm (2)

Algorithm 2.1b: Importance Sampling using self-normalised weights

Choose g such that $\text{supp}(g) \supset \text{supp}(f \cdot h)$.

1. For $i = 1, \ldots, n$:
   i. Generate $X_i \sim g$.
   ii. Set $w(X_i) = \frac{f(X_i)}{g(X_i)}$.
2. Return

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{\sum_{i=1}^{n} w(X_i)}$$

as an estimate of $\mathbb{E}_f(h(X))$.

Basic properties of the self-normalised estimate

- $\hat{\mu}$ is consistent as

$$\hat{\mu} = \underbrace{\frac{\sum_{i=1}^n w(X_i)h(X_i)}{n}}_{=\tilde{\mu}\longrightarrow\mathbb{E}_f(h(X))} \underbrace{\frac{n}{\sum_{i=1}^n w(X_i)}}_{\longrightarrow 1} \overset{a.s.}{\underset{n\to\infty}{\longrightarrow}} \mathbb{E}_f(h(X)),$$

(provided $\text{supp}(g) \supset \text{supp}(f \cdot h)$ and $\mathbb{E}_g|w(X) \cdot h(X)| < +\infty$)

- $\hat{\mu}$ is biased, but asymptotically unbiased (see theorem below)

Theorem 2.2: Bias and Variance (ctd.)

$$\begin{aligned}
\mathbb{E}_g(\hat{\mu}) &= \mu + \frac{\mu \text{Var}_g(w(X)) - \text{Cov}_g(w(X), w(X) \cdot h(X))}{n} + O(n^{-2}) \\
\text{Var}_g(\hat{\mu}) &= \frac{\text{Var}_g(w(X) \cdot h(X)) - 2\mu \text{Cov}_g(w(X), w(X) \cdot h(X))}{n} \\
&\quad + \frac{\mu^2 \text{Var}_g(w(X))}{n} + O(n^{-2})
\end{aligned}$$

Finite variance estimators

- Importance sampling estimate consistent for large choice of g. (only need that ...)
- More important in practice: *finite variance estimators*, i.e.

$$\mathrm{Var}(\tilde{\mu}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} w(X_i)h(X_i)}{n}\right) < +\infty$$

- Sufficient conditions for finite variance of $\tilde{\mu}$:
  - $f(x) < M \cdot g(x)$ and $\mathrm{Var}_f(h(X)) < \infty$, or
  - E is compact, f is bounded above on E, and g is bounded below on E.
- Note: If f has heavier tails then g, then the weights will have infinite variance!

Optimal proposals

Theorem 2.3: Optimal proposal

The proposal distribution g that minimises the variance of $\tilde{\mu}$ is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(t)|f(t)\, dt}.$$

- Theorem of little practical use: the optimal proposal involves $\int |h(t)|f(t)\, dt$, which is the integral we want to estimate!
- Practical relevance of theorem 2.3:
  Choose g such that it is close to $|h(x)| \cdot f(x)$

Super-efficiency of importance sampling

- For the optimal g^* we have that

$$\mathrm{Var}_f\left(\frac{h(X_1)+\ldots+h(X_n)}{n}\right) > \mathrm{Var}_{g^\star}(\tilde{\mu}),$$

if h is not almost surely constant.

Superefficiency of importance sampling

The variance of the importance sampling estimate can be *less* than the variance obtained when sampling directly from the target f.

- Intuition: Importance sampling allows us to choose g such that we focus on areas which contribute most to the integral $\int h(x)f(x)\,dx$.
- Even sub-optimal proposals can be super-efficient.

Example 2.5: Setup

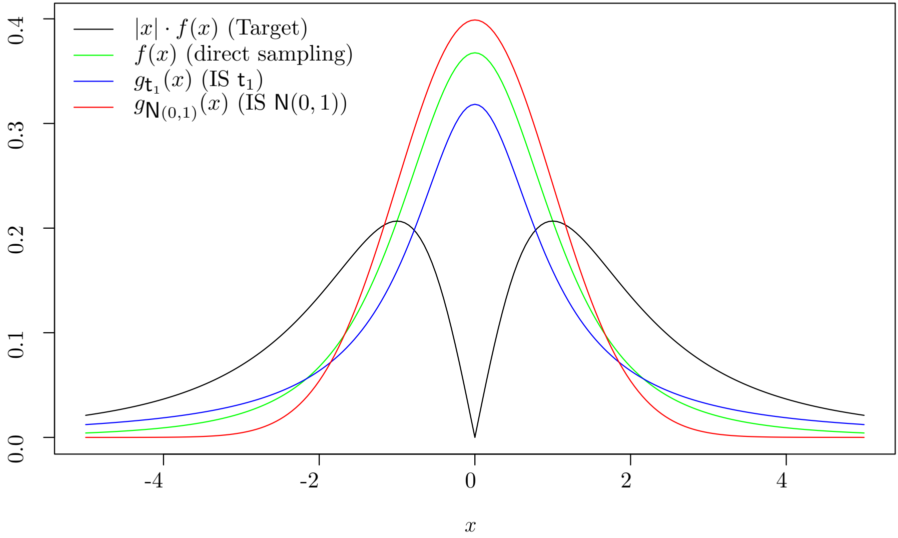
Compute $\mathbb{E}_f|X|$ for $X \sim \mathrm{t}_3$ by ...

(a) sampling directly from t_3.

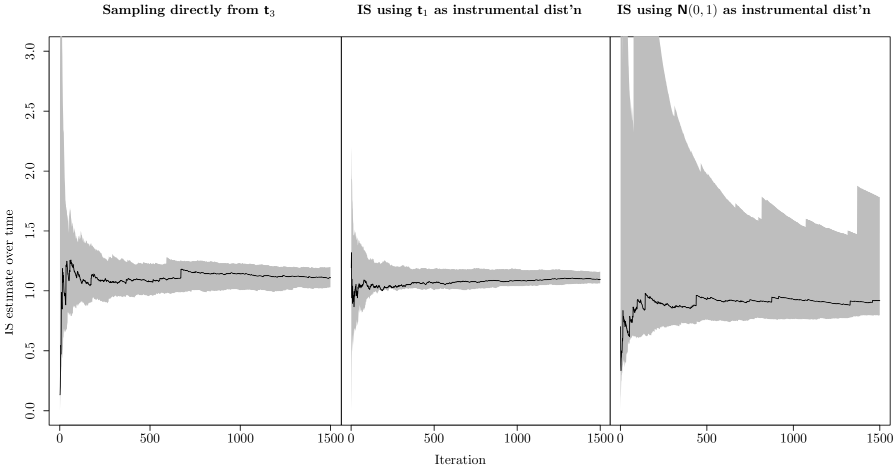
(b) using a t_1 distribution as instrumental distribution.

(c) using a $\mathrm{N}(0, 1)$ distribution as instrumental distribution.

Example 2.5: Densities



Example 2.5: Estimates obtained



Example 2.5: Weights