

# Importance sampling. [RC]

## Chapter 3.

Anatoli Iambartsev

IME-USP

**Importance sampling.** [RC, Chapter 3.3] “The method we now study is called importance sampling because it relies on so-called importance functions, which are instrumental distributions, in lieu of the original distributions. In fact, an evaluation of

$$m := \mathbb{E}_f(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx, \quad (1)$$

based on simulations from  $f$  is almost never optimal in the sense that using alternative distributions can improve the variance of the resulting estimator of (1).”

**Importance sampling.** [RC] Importance sampling is based on an alternative representation of the integral (1). Given an arbitrary density  $g$  that is strictly positive when  $h \cdot f$  is different from zero, i.e.

$$\text{supp}(g) \supseteq \text{supp}(h \cdot f),$$

we can indeed rewrite (1) as

$$\mathbb{E}_f(h(X)) = \int_{\text{supp}(g)} h(x) \frac{f(x)}{g(x)} g(x) dx = \mathbb{E}_g \left[ \frac{h(X)f(X)}{g(X)} \right]$$

and for any measurable set  $A$

$$\mathbb{P}(X \in A) = \int_A f(x) dx = \int_A \frac{f(x)}{g(x)} g(x) dx$$

$$\mathbb{E}_f(h(X)) = \int_{\text{supp}(g)} h(x) \frac{f(x)}{g(x)} g(x) dx = \mathbb{E}_g \left[ \frac{h(X)f(X)}{g(X)} \right]$$

[RC] This *importance sampling fundamental identity* justifies the use of the estimator

$$m_n^{IS} = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)} h(X_i) \rightarrow \mathbb{E}_f(h(X)),$$

based on a sample  $X_1, \dots, X_n$  generated from  $g$  (not from  $f$ !). By strong law of large numbers, if

$$\mathbb{E}_g \left| \frac{h(X)f(X)}{g(X)} \right| < \infty$$

then the convergence is almost sure.

**Bias and variance of importance sampling.**

$$\begin{aligned}\mathbb{E}_g(m_n^{IS}) &= m \\ \mathbb{V}ar_g(m_n^{IS}) &= \frac{\mathbb{V}ar_g(w(X) \cdot h(X))}{n}\end{aligned}$$

where

$$w(X) = \frac{f(X)}{g(X)}.$$

Moreover, the expected value of the weights is

$$\mathbb{E}_g(w(X)) = 1.$$

**Example 3.5 [RC].**

For example, if  $Z \sim N(0, 1)$  and we are interested in the probability  $\mathbb{P}(Z > 4.5)$ , which is very small,

```
> pnorm(-4.5,log=T)
```

```
[1] -12.59242
```

```
> pnorm(-4.5)
```

```
[1] 3.397673e-06
```

simulating  $Z_i \sim N(0, 1)$  only produces a hit once in about 3 million iterations!

**Exercise 3.5 [RC].**

Thanks to importance sampling, we can greatly improve our accuracy and thus bring down the number of simulations by several orders of magnitude.

For instance, if we consider a distribution with support restricted to  $(4.5, \infty)$ , the additional and unnecessary variation of the Monte Carlo estimator due to simulating zeros (i.e., when  $x < 4.5$ ) disappears. A natural choice is to take  $g$  as the density of the exponential distribution  $Exp(1)$  truncated at 4.5,

$$g(y) = \frac{e^{-y}}{\int_{4.5}^{\infty} e^{-x} dx} = e^{-(y-4.5)}$$

**Exercise 3.5 [RC].** A natural choice is to take  $g$  as the density of the exponential distribution  $Exp(1)$  truncated at 4.5,

$$g(y) = \frac{e^{-y}}{\int_{4.5}^{\infty} e^{-x} dx} = e^{-(y-4.5)}$$

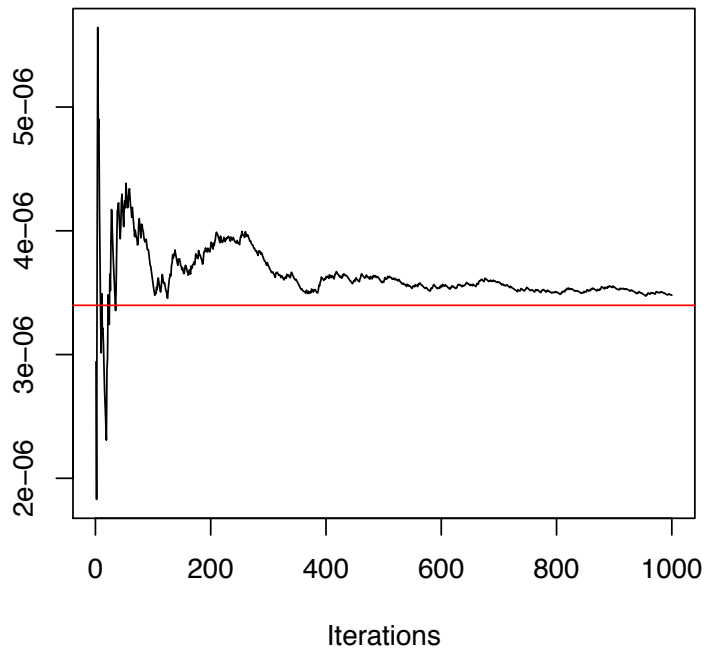
and the corresponding importance sampling estimator of the tail probability is

$$\frac{1}{n} \sum_{i=1}^n \frac{f(Y_i)}{g(Y_i)} = \frac{1}{n} \sum_{i=1}^n \frac{e^{-Y_i^2/2+Y_i-4.5}}{\sqrt{2\pi}}$$

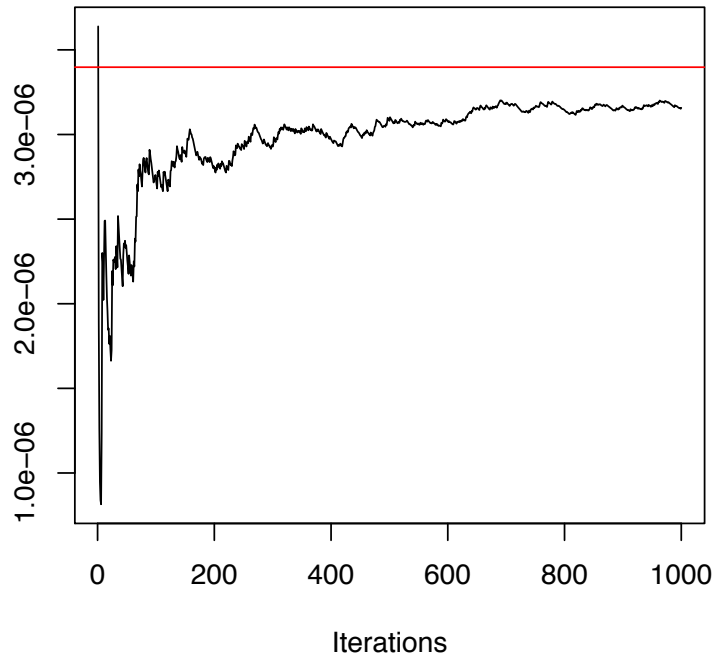
where  $Y_i$ 's are i.i.d. generations from truncated at 4.5 exponential  $Exp(1)$  distribution.



**Exercise 3.5 [RC].** The true value is  $\mathbb{P}(Z > 4.5) \cong 3.398 \cdot 10^{-6}$  (red line). The estimated value is  $3.480 \cdot 10^{-6}$



**Exercise 3.5 [RC].** The true value is  $\mathbb{P}(Z > 4.5) \cong 3.398 \cdot 10^{-6}$  (red line). The estimated value is  $3.158 \cdot 10^{-06}$



**Exercise 3.5 [RC]. Code p.71.**

```
> Nsim=10^3
```

```
> y=rexp(Nsim)+4.5
```

```
> weit=dnorm(y)/dexp(y-4.5)
```

```
> plot(cumsum(weit)/1:Nsim,type="l",ylab="",xlab="Iterations")
```

```
> abline(a=pnorm(-4.5),b=0,col="red")
```

```
> est=sum(weit)/Nsim
```

**Example 3.5 [RC].**

The accuracy of the approximation is remarkable, especially when compared with the original size requirements imposed by a normal simulation.

See Anexo 1 for detailed discussion about rare-event probability estimation.

**Abstract theory.** Let  $Z := h(X)$

$Z : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B}, \mathbb{P}_Z)$  measurable, i.e.  $Z$  is real-valued random variable s.t. for all Borel sets  $B \in \mathcal{B}$

$$\mathbb{P}_Z(B) := \mathbb{P}(Z^{-1}(B)) = \mathbb{P}(\omega : Z(\omega) \in B).$$

Consider another probability  $\mathbb{Q}$  s.t.  $\mathbb{Q} \ll \mathbb{P}$  (we say that  $\mathbb{Q}$  is absolute continuous with respect to  $\mathbb{P}$ ): i.e. for any  $A \in \mathcal{F}$

$$\mathbb{Q}(A) = 0 \Rightarrow \mathbb{P}(A) = 0.$$

**Abstract theory.**

Consider the probability  $\mathbb{Q}$  s.t.  $\mathbb{Q} \ll \mathbb{P}$ . Then there exists a *likelihood ratio* (or Radon-Nikodym derivative)  $L : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B})$  s.t.

$$d\mathbb{P} = Ld\mathbb{Q}.$$

Equivalently

$$\mathbb{E}(Z) = \mathbb{E}_{\mathbb{Q}}(LZ).$$

It suffices a probability  $\mathbb{Q}$  s.t. for all  $\omega \in \Omega$

$$\mathbb{1}(Z(\omega) \neq 0)d\mathbb{P}(\omega) = \mathbb{1}(Z(\omega) \neq 0)Ld\mathbb{Q}(\omega).$$

**Abstract theory.**

In the case of absolute continuous random variables

$$\mathbb{Q} \ll \mathbb{P} \quad \Leftrightarrow \quad \text{supp}(g) \supseteq \text{supp}(f)$$

and

$$d\mathbb{P} = Ld\mathbb{Q} \quad \Leftrightarrow \quad f(x)dx = \frac{f(x)}{g(x)}g(x)dx$$

**Variance of importance sampling.** Recall

$$m_n^{IS} = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)} h(X_i) \rightarrow \mathbb{E}_f(h(X)),$$

The variance of estimator

$$\text{Var}_g(m_n^{IS}) = \frac{\text{Var}_g(w(X) \cdot h(X))}{n}$$

where  $w(X) = f(X)/g(X)$ . Consider  $\text{Var}_g(w(X)h(X))$

$$\text{Var}_g(w(X)h(X)) = \int h^2(x) \frac{f^2(x)}{g(x)} dx - m^2.$$

More important condition on choosing  $g$  is finiteness of variance estimator, i.e.

$$\text{Var}_g(w(X)h(X)) < \infty.$$

Note that if  $f$  has heavier tails than  $g$ , then the weights will have infinite variance.



**Variance of importance sampling.** [RC1,pp.94-95] An alternative to

$$m_n^{IS} = \frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)} h(X_i) \approx \mathbb{E}_f(h(X)), \quad (2)$$

which addresses the finite variance issue, and generally yields a more stable estimator, is to use

$$\frac{\sum_{i=1}^n \frac{f(x_i)}{g(x_i)} h(x_i)}{\sum_{i=1}^n \frac{f(x_i)}{g(x_i)}}, \quad (3)$$

where we replaced  $n$  with the sum of the weights. Since

$$\frac{1}{n} \sum_{i=1}^n \frac{f(x_i)}{g(x_i)} \rightarrow 1, \text{ as } n \rightarrow \infty,$$

this estimator also converges to  $\mathbb{E}_f(h(X))$  by the Strong Law of Large Numbers. Although this estimator is biased, the bias is small, and the improvement in variance makes it a preferred alternative to (2).

**Variance. Optimal proposal. Anexo 2.**

**Optimality Theorem.** The distribution  $g$  that minimizes the variances of  $m_n^{IS}$  (for fixed  $n$ ) is

$$\hat{g} = \frac{|h(x)|f(x)}{\int |h(t)|f(t)dt} \propto |h(x)|f(x)$$

- 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: choose  $g$  such that it is close to  $|h(x)|f(x)$ , i.e. looking for such distributions  $g$  for which  $\frac{|h|f}{g}$  is almost constant with finite variance.

[RC ] It is important to note that although the finite variance constraint is not necessary for the convergence, importance sampling performs quite poorly when

$$\int \frac{f^2(x)}{g(x)} dx = \infty.$$

**Variance. Optimal proposal. Proof.**

$$\begin{aligned}\text{Var}_{\hat{g}}(\hat{w}(X)h(X)) &= \int h^2(x) \frac{f^2(x)}{\hat{g}(x)} dx - m^2 \\ &= \int \frac{h^2(x)f^2(x)}{|h(x)|f(x)} \left( \int |h(t)|f(t) dt \right) dx - m^2 \\ &= \left( \int |h(x)|f(x) dx \right)^2 - m^2.\end{aligned}$$

Cauchy-Swartz inequality shows that any other choice of  $g$  increase the variance:

$$\begin{aligned}\left( \int |h(x)|f(x) dx \right)^2 &= \left( \int |h(x)|f(x) \frac{\sqrt{g(x)}}{\sqrt{g(x)}} dx \right)^2 \\ &\leq \int \frac{h^2(x)f^2(x)}{g(x)} dx \int g(x) dx \leq \int \frac{h^2(x)f^2(x)}{g(x)} dx\end{aligned}$$

□

**Self-normalized estimator. Anexo 2.**

Importance sampling can be useful when the density  $f$  is known up to some unknown constant

$$f(x) \propto \pi(x), \text{ or } f(x) = C\pi(x).$$

Applying directly the importance sampling estimator provides

$$m_n^{IS} = \frac{1}{n} \sum_{i=1}^n \frac{C\pi(X_i)}{g(X_i)} h(X_i) = \frac{1}{n} \sum_{i=1}^n w(X_i) h(X_i).$$

What about the unknown constant  $C$ ?

**Self-normalized estimator. Anexo 2.**

$$m_n^{IS} = \frac{1}{n} \sum_{i=1}^n \frac{C\pi(X_i)}{g(X_i)} h(X_i) = \frac{1}{n} \sum_{i=1}^n w(X_i) h(X_i).$$

What about the unknown constant  $C$ ? The idea: we should estimate  $C$  as well, and use normalized weights.

$$\tilde{w}(X_i) = \frac{w(X_i)}{\sum_{k=1}^n w(X_k)},$$

which gives us an idea to importance sampling simulation of distribution  $f$ . Consider the estimator

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

**Self-normalized estimator. Anexo 2.** Observe that the *self-normalized importance sampling estimator* does not depend on  $C$ :

$$m_n^{SNIS} = \frac{\sum_{i=1}^n w(X_i)h(X_i)}{\sum_{i=1}^n w(X_i)} = \frac{\sum_{i=1}^n \frac{\pi(X_i)}{g(X_i)}h(X_i)}{\sum_{i=1}^n \frac{\pi(X_i)}{g(X_i)}}$$

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

1. For  $i = 1, \dots, n$  :

(a) Generate  $X_i \sim g$ .

(b) Set  $w(X_i) = \frac{\pi(X_i)}{g(X_i)}$ .

2. Return

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

**Self-normalized estimator. Properties. Anexo 2.**

$m_n^{SNIS}$  is consistent estimator. Indeed,

$$m_n^{SNIS} = \frac{\sum_{i=1}^n w(X_i)h(X_i)}{\sum_{i=1}^n w(X_i)} = \frac{\sum_{i=1}^n w(X_i)h(X_i)}{n} \times \frac{n}{\sum_{i=1}^n w(X_i)} \times \frac{\sum_{i=1}^n w(X_i)}{\sum_{i=1}^n w(X_i)}$$

$$\begin{array}{ccc} \downarrow & & \downarrow \\ \mathbb{E}_f(h(X)) & & \mathbb{E}_g\left(\frac{f(X)}{g(X)}\right) = 1 \end{array}$$

where the convergence is almost sure, requiring

$$\text{supp}(g) \supset \text{supp}(f \cdot h)$$

and  $\mathbb{E}_g|w(X)h(X)| < \infty$ .

**Self-normalized estimator. Properties. Anexo 2.**

$m_n^{SNIS} = \frac{\sum_{i=1}^n w(X_i)h(X_i)}{\sum_{i=1}^n w(X_i)}$  is biased, but asymptotically unbiased.

**Theorem.** Bias and variance of  $m_n^{SNIS}$ :

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



### Optimality Theorem and SNIS.

[RC1, p.95] A practical alternative taking advantage of The Optimality Theorem is to use the *self-normalized importance sampling estimator* (SNIS). Where instead of  $f$  we know  $g$  up to an unknown constant  $g \propto |h|f$ . Observe that the SNIS can be defined in the same way. In this case SNIS is

$$m_n = \frac{\sum_{i=1}^n \frac{f(x_i)}{|h(x_i)|f(x_i)} h(x_i)}{\sum_{i=1}^n \frac{f(x_i)}{|h(x_i)|f(x_i)}} = \frac{\sum_{i=1}^n \frac{h(x_i)}{|h(x_i)|}}{\sum_{i=1}^n \frac{1}{|h(x_i)|}}, \quad (4)$$

where  $x_i \sim g \propto |h|f$ . “Note that the numerator is the number of times  $h(x_i)$  is positive minus the number of times when it is negative. In particular, when  $h$  is positive, (4) is the *harmonic mean*. Unfortunately, the optimality of Theorem does not transfer to (4), which is biased and may exhibit severe instability.”

The problem to use (4) in practice is the generation  $x_i \sim g \propto |h|f$ .

**References:**

- [RC ] Cristian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Series “Use R!”. Springer.
- [RC1 ] Cristian P. Robert and George Casella. *Monte Carlo Statistical Methods*. “Springer Texts in Statistics Series”. Springer, Second Edition, 2004.
- [Ross ] Ross, S. *Simulation*. 2012.

Anexo 1, Anexo 2