For example, suppose $X \sim p$ and we want to estimate

$$\mathbb{P}(X \geq x_0) = \mathbb{E}_p[\mathbf{1}\,(X \geq x_0)]$$

with x_0 in the extreme upper tail of $p(x)$. If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$, we may not get any samples $X_i \geq x_0$ and the usual Monte Carlo estimate

$$I_p^{(n)}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\,(X_i \geq x_0)$$

is zero with high probability. We can take an proposal density q that puts more probability at large Y, and then reweight to get expectations in X. By using IS, we can reduce the variance significantly.

Example 4.16. Say $p(x)$ is the standard normal density and we want to estimate $\theta = \mathbb{P}(X \geq x_0)$ for some $x_0 \geq 3$.

Take q as the shifted exponential,

$$q(y) := r \exp\big(-r(y - x_0)\big)\mathbf{1}\,(y \geq x_0)\,.$$

Let us determine r so that q approximates the optimal distribution (the conditional tail of p) locally: $(\log p)' = (\log q)'$ at x_0, that is,

$$r = g'(x_0), \qquad g(x) = -\log p(x) = \frac{x^2}{2} \qquad \Longrightarrow \qquad r = x_0.$$

The weights are, for $y \geq x_0$,

$$\begin{aligned} w(y) &= \frac{p(y)}{q(y)} \\ &= \frac{1}{r\sqrt{2\pi}} \exp\bigg(-\frac{y^2}{2} + r(y - x_0)\bigg) \end{aligned}$$

and the IS estimator of θ is $\frac{1}{n}\sum_{i=1}^{n} w(Y_i)\mathbf{1}\,(Y_i \geq x_0)$; See Figure 8.

4.3 Self-normalised importance sampling

The rejection sampling algorithm is straightforward to apply in case of unknown normalising constants, that is, when only the unnormalised densities $p_u(x) \propto p(x)$ and $q_u(x) \propto q(x)$ are available.

In importance sampling, this means that we can access the *unnormalised* importance weights

$$w_u(x) := \frac{p_u(x)}{q_u(x)} = \frac{Z_p}{Z_q} w(x), \qquad q(x) > 0,$$

and $w_u(x) := 0$ when $q(x) = 0$. In order to apply (unbiased) importance sampling, we would need w. We can get around by *simultaneously* estimating the ratio Z_p/Z_q, with a cost of introducing a bias (which is asymptotically vanishing).

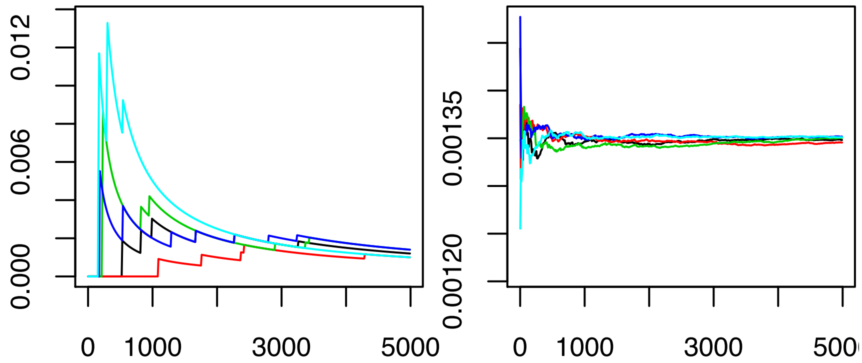

Figure 8: Five trajectories of classical Monte Carlo (left) and IS of Example 4.16 (right) with $x_0 = 3$. Number of samples in x-axis and value of estimate in y-axis.

Definition 4.17 (Self-normalised importance sampling)**.** Suppose p and q are p.d.f.s or p.m.f.s, such that

$$\boxed{\text{Assumption: } q(x) = 0 \implies p(x) = 0.} \tag{10}$$

Then, if $Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} q$,

$$\hat{I}_{p,q}^{(n)}(f) := \sum_{k=1}^{n} f(Y_k) W_k^{(n)}, \tag{11}$$

$$\text{where} \quad W_k^{(n)} := \begin{cases} \frac{w_u(Y_k)}{\sum_{j=1}^{n} w_u(Y_j)}, & \text{if } w_u(Y_j) > 0 \text{ for some } 1 \le j \le n \\ \mathbf{1}\,(k=1)\,, & \text{otherwise} \end{cases}$$

is the *self-normalised* (or *rescaled*) IS approximation of $\mathbb{E}_p[f(X)]$.

Remark 4.18. Note that

(a) $\beta = \mathbb{P}_q(w_u(Y_j) > 0) = \mathbb{P}_q(p(Y_j) > 0) > 0$, and therefore

$$\mathbb{P}_q\big(w_u(Y_j) > 0 \text{ for some } 1 \le j \le n\big) = 1 - (1-\beta)^n \xrightarrow{n\to\infty} 1.$$

(b) We always have $\sum_{k=1}^{n} W_k^{(n)} = 1$.

The drawback of the self-normalised IS is that the estimator $\hat{I}_{p,q}^{(n)}(f)$ is generally biased for finite n. However, the estimator is (strongly) consistent.

Theorem 4.19. *Suppose* (10) *holds. Then,* $\hat{I}_{p,q}^{(n)}(f) \xrightarrow{n\to\infty} \mathbb{E}_p[f(X)]$ *(almost surely).*

Proof. Because $w_u(Y_j) > 0$ for some $1 \le j \le n$ eventually (almost surely; cf. Remark 4.18), we may consider only such n.

$$\hat{I}_{p,q}^{(n)}(f) = \frac{\sum_{k=1}^{n} f(Y_k) w_u(Y_k)}{\sum_{k=1}^{n} w_u(Y_k)} = \frac{\frac{1}{n}\sum_{k=1}^{n} f(Y_k) w(Y_k)}{\frac{1}{n}\sum_{k=1}^{n} w(Y_k)} = \frac{I_{p,q}^{(n)}(f)}{I_{p,q}^{(n)}(1)}.$$

Theorem 4.3 (b) implies that $I_{p,q}^{(n)}(f) \xrightarrow{n\to\infty} \mathbb{E}_p[f(X)]$ almost surely and $I_{p,q}^{(n)}(1) \xrightarrow{n\to\infty} \mathbb{E}_p[1] = 1$ almost surely. □

Remark 4.20. In the proof of Theorem 4.19, we need the condition $q(x) = 0 \implies p(x) = 0$ in order to ensure $I_{p,q}^{(n)}(1) \to 1$. This is more stringent than with unbiased IS, where we only need $q(x) = 0 \implies p(x)f(x) = 0$ which ensures $I_{p,q}^{(n)}(f) \to \mathbb{E}_p[f(X)]$.

Remark 4.21. Note that

$$\mathbb{E}_q[w_u(Y)] = \frac{Z_p}{Z_q}\mathbb{E}_q[w(Y)] = \frac{Z_p}{Z_q},$$

so the mean of unnormalised SNIS weights is unbiased and (strongly) consistent estimator of the ratio of normalising constants,

$$\frac{1}{n}\sum_{k=1}^{n} w_u(Y_k) \xrightarrow{n\to\infty} \frac{Z_p}{Z_q} \qquad \text{(almost surely)}.$$

This is important in certain applications.

Example 4.22. We saw in Example 4.5 that if $Y_i \sim \Gamma(a, b)$ and

$$w(y) = \frac{\Gamma(a)\beta^\alpha}{\Gamma(\alpha)b^a} y^{\alpha-a} \exp(-(\beta - b)y)$$

then

$$I_{p,q}^{(n)}(f) = \frac{1}{n}\sum_{i=1}^{n} f(Y_i)w(Y_i)$$

is unbiased and consistent estimator of $\mathbb{E}_p[f(X)]$ with $p = \Gamma(\alpha, \beta)$.

To avoid calculating $\Gamma(a)/\Gamma(\alpha)$, we can use

$$w_u(y) = y^{\alpha-a} \exp(-(\beta - b)y)$$

and then the self-normalised IS estimator

$$\hat{I}_{p,q}^{(n)}(f) := \frac{\sum_{i=1}^{n} f(Y_i)w_u(Y_i)}{\sum_{i=1}^{n} w_u(Y_i)}$$

is a consistent estimator of $\mathbb{E}_p[f(X)]$.

```
function snis_gamma(n, alpha, beta, f)
   y = -log.(rand(n))                          # y ~ Exp(1) = Gamma(1,1)
   w_u = y.^(alpha-1) .* exp.(-(beta-1)*y)  # Unnormalised w
   w = w_u/sum(w_u)                            # Normalised w
   sum(f.(y) .* w)                             # SNIS estimate
end
# Use the function f(x)=x to estimate mean:
snis_gamma(1000, 2, 4, x -> x)
```

The self-normalised IS satisfies a CLT with same variance as the unbiased IS for zero mean functions, in which case they are asymptotically equally efficient. A consistent confidence interval can also be easily constructed.

Theorem 4.23. *Suppose* (10) *holds and* $\bar\sigma^2_{p,q} := \mathbb{E}_p[w(X)\bar f^2(X)] < \infty$, *where* $\bar f(x) = f(x) - \mathbb{E}_p[f(X)]$.

(i) $\sqrt{n}\big(\hat I^{(n)}_{p,q}(f) - \mathbb{E}_p[f(X)]\big) \xrightarrow{n\to\infty} N(0,\bar\sigma^2_{p,q})$ *in distribution.*

(ii) If also $\mathbb{E}_p[w(X)] < \infty$, *then the following hold:*

- $n v^{(n)}_{p,q} \xrightarrow{n\to\infty} \bar\sigma^2_{p,q}$ *(a.s.), where* $v^{(n)}_{p,q} := \sum_{k=1}^n (W_k^{(n)})^2\big[f(Y_k) - \hat I^{(n)}_{p,q}(f)\big]^2$, *and*
- $\mathbb{P}\Big(\mathbb{E}_p[f(X)] \in \Big[\hat I^{(n)}_{p,q}(f) \pm \alpha\sqrt{v^{(n)}_{p,q}}\Big]\Big) \to 1 - 2\Phi(-\alpha)$ *for any* $\alpha \in (0,\infty)$.

Proof. (i) Because $\sum_{k=1}^n W_k^{(n)} = 1$, $\hat I^{(n)}_{p,q}(f) - \mathbb{E}_p[f(X)] = \hat I^{(n)}_{p,q}(\bar f)$. Now, as in the proof of Theorem 4.19, $\sqrt{n}\hat I^{(n)}_{p,q}(\bar f) = \sqrt{n} I^{(n)}_{p,q}(\bar f)/I^{(n)}_{p,q}(1)$. Corollary 4.8 implies that the numerator converges in distribution to $N(0,\bar\sigma^2_{p,q})$ and the denominator converges to 1 almost surely. Slutsky's theorem (Lemma 1.14) concludes the proof. The first part of (ii), that is, $n v^{(n)}_{p,q} \to \bar\sigma^2_{p,q}$ is an exercise, and the second claim follows from (i), as in the proof of Proposition 1.13 (iii). □

Remark 4.24 (*). The quantity $n_{\text{eff}} = \big(\sum_{k=1}^n (W_k^{(n)})^2\big)^{-1} \in [1,n]$ is widely known as the *effective sample size* of (self-normalised) IS.

This may be (loosely) justified when the function is of the form $f(x) := c\mathbf{1}(x \in A)$ with $c > 0$ and A such that $\mathbb{E}_p[\mathbf{1}(X \in A)] = 1/2$. In this case, $\bar f(x) \equiv \frac{c}{2}$, and standard Monte Carlo estimator $I^{(n)}_p(f)$ would have variance $\mathrm{Var}_p(f(X))/n = (c/2)^2/n$, but the corresponding limiting CLT variance of the SNIS estimator is $\mathbb{E}_p[w(X)\bar f^2(X)]/n$. It is not hard to see (cf. the proof of Theorem 4.23 (ii)) that then

$$\frac{n}{n_{\text{eff}}} \xrightarrow{n\to\infty} \mathbb{E}_p[w(X)],$$

so $\mathbb{E}_p[w(X)]/n \approx \mathrm{Var}_p(f(X))/n_{\text{eff}}$ for large n. Therefore, the self-normalised IS with n samples may be (loosely) thought of as having n_{eff} 'effective independent samples'.

Remark 4.25 (*). It is sometimes useful to consider the SNIS as an empirical approximation of the distribution p. That is,

$$\hat\mu^{(n)}_{p,q}(A) := \sum_{k=1}^n W_k^{(n)}\mathbf{1}(Y_k \in A) \approx \mathbb{P}(X \in A), \qquad A \subset \mathbb{X},$$

where $X \sim p$. The approximation is consistent assuming (10), in the following sense:

$$\hat\mu^{(n)}_{p,q}(A) \xrightarrow{n\to\infty} \mathbb{P}(X \in A) \qquad \text{almost surely},$$

for any (measurable) $A \subset \mathbb{X}$.

With unbiased IS, we have

$$\mu^{(n)}_{p,q}(A) := \frac{1}{n}\sum_{k=1}^n w(Y_k)\mathbf{1}(Y_k \in A).$$

Given (10) this is consistent and also unbiased $\mathbb{E}[\mu^{(n)}_{p,q}(A)] = \mathbb{P}(X \in A)$, but unlike self-normalised IS and plain MC, $\mu^{(n)}_{p,q}$ is not a probability distribution, because $\mu^{(n)}_{p,q}(\mathbb{X}) \neq 1$ in general.