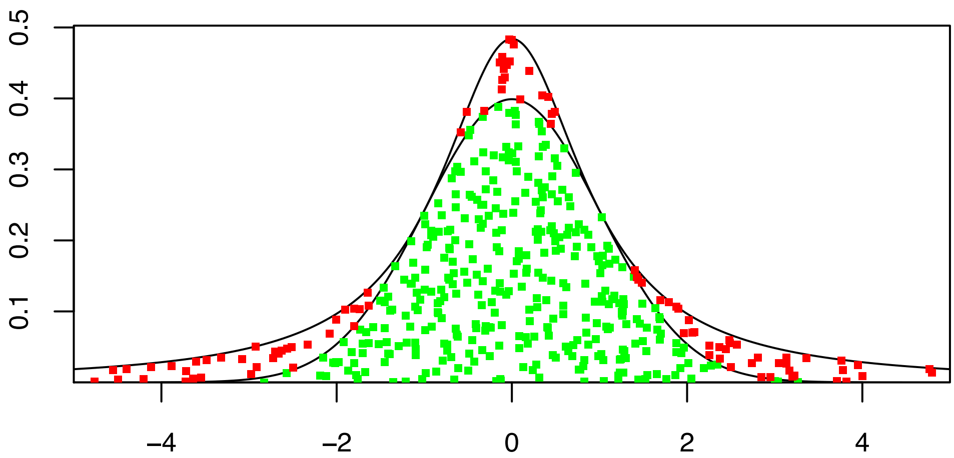

Figure 4: All points $(Y_k, U_k M q(Y_k))$ simulated in the rejection algorithm of Example 3.8 are *distributed uniformly* in beteeen the x-axis and the function $Mq(x)$ (the upper curve). The points that fall below the curve $p(x)$ are accepted (green), others are rejected (red).

3.2 Unnormalised distributions and rejection sampling

A p.d.f. $p(x)$ on $\mathbb{X}$ (resp. p.m.f. $p(x)$ on $\mathbb{X}$) must satisfy

$$\int_{\mathbb{X}} p(x)\mathrm{d}x = 1 \qquad \left(\text{resp.} \quad \sum_{x\in\mathbb{X}} p(x) = 1\right).$$

We can specify a p.d.f (resp. p.m.f.) by just giving a non-negative function $p_u(x)$, which is propotional to $p(x)$. More specifically, if

$$p(x) \propto p_u(x) \qquad \text{then} \qquad p(x) = \frac{p_u(x)}{Z_p},$$

with the *normalising constant*

$$Z_p := \int_{\mathbb{X}} p_u(x)\mathrm{d}x. \qquad \left(\text{resp.} \quad Z_p = \sum_{x\in\mathbb{X}} p_u(x)\right).$$

The distribution $p(x)$ is fully determined by $p_u(x)$, even though we could not calculate values of $p(x)$. (Of course, we must have $Z_p \in (0, \infty)$.)

Example 3.9. Suppose we know $p(x)$, the density of random variable X, and we are interested in the conditional density of X given $X \geq t$, of the following form:

$$p_t(x) = \frac{p(x)\mathbf{1}\,(x \geq t)}{\int_t^\infty p(t)\mathrm{d}t} \propto p(x)\mathbf{1}\,(x \geq t)\,.$$

It is clear that we could sample from $(X_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} p$, and accept only those for which $X_k \geq t$, which would be samples from p_t.

Example 3.10. In Bayesian inference, we are interested in a conditional distribution (the posterior)

$$p(x) = f_{X|Y}(x \mid y^*) = \frac{f_{Y|X}(y^* \mid x) f_X(x)}{\int_{\mathbb{X}} f_{Y|X}(y^* \mid \hat{x}) f_X(\hat{x}) \mathrm{d}\hat{x}} \propto f_{Y|X}(y^* \mid x) f_X(x),$$

where y^* stands for the observed value of random variable Y and random variable X is the unknown. (Above, $p(x)$ is the conditional density of $X \mid (Y = y^*)$ and $f_{X|Y}$ stands for the conditional density of X given Y.) We can only calculate $p_u(x) = f_{Y|X}(y^* \mid x) f_X(x)$.

We would like an algorithm to simulate $X \sim p$ and use only the unnormalised density $p_u(x)$, without need to calculate $p(x)$. The rejection algorithm can be used in such a case.

Algorithm 3.11 (Rejection sampling with unnormalised distributions)**.** Suppose q and p are p.d.f.s (or p.m.f.s) such that $q \propto q_u$ and $p \propto p_u$, with

$$\boxed{\text{Assumption: } \frac{p_u(x)}{q_u(x)} \le M \qquad \text{for all } x \in \mathbb{X},} \tag{5}$$

and that $(Y_k)_{k\ge 1} \overset{\text{i.i.d.}}{\sim} q$ independent of $(U_k)_{k\ge1} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$. Set $T = 1$ and

(A) If $U_T \le \dfrac{p_u(Y_T)}{M q_u(Y_T)}$, then output $X = Y_T$.

(B) Otherwise, increment $T = T + 1$ and retry (A).

Algorithm 3.11 is valid by the proof of Theorem 3.4, with minor adjustments. Namely,

$$\mathbb{P}(Y_t = x, B_t = 1) = \frac{1}{M} q(x) \frac{p_u(x)}{q_u(x)} = \left(\frac{1}{M} \cdot \frac{Z_p}{Z_q}\right) p(x),$$

from which we notice also that $T \sim \text{Geometric}(1/\hat{M})$ where $\hat{M} = M Z_q / Z_p$.

(In fact, $\frac{p_u(y)}{M q_u(y)} = \frac{p(y)}{\hat{M} q(y)}$, so Algorithm 3.11 coincides with Algorithm 3.2 with $\hat{M} = M$.)

Example 3.12. Consider the probability density

$$p(x) \propto p_u(x) := \frac{\sin^2(x)}{x^2} \mathbf{1}\left(x \neq 0\right), \qquad -\infty < x < \infty, \ (x \neq 0)$$

and the standard Cauchy distribution $q(x) \propto q_u(x) = (1 + x^2)^{-1}$, which can be simulated with the inverse c.d.f. method (exercise). We have

$$\frac{p_u(x)}{q_u(x)} = \frac{\sin^2 x(1 + x^2)}{x^2} \le \min\left\{\frac{1 + x^2}{x^2}, 1 + x^2\right\} \le 2,$$

because $|\sin x| \le \min\{1, x\}$. (Optimal bound is slightly less than 1.5.)

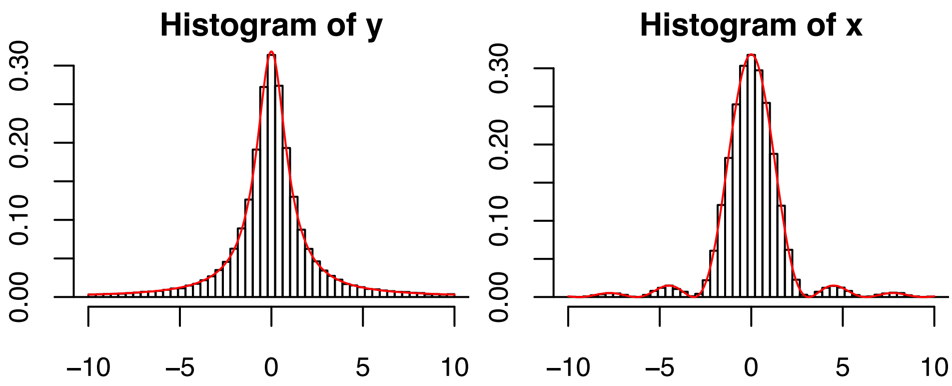

Figure 5: Simulated samples from the standard Cauchy distribution (left) and samples from $p(x) \propto \sin^2(x)/x^2$ (right) with corresponding densities.

```
using Distributions
max_n = 100_000; x = zeros(0) # Empty (zero-length) vector
for k = 1:max_n
    y = rand(Cauchy())
    ratio_pu_qu_M = sin(y)^2*(1+y^2) / (2y^2)
    if rand() <= ratio_pu_qu_M
        push!(x, y)  # Append y to the end of vector x
    end
end
```

4 Importance sampling

All methods up to this point have aimed at simulating i.i.d. random variables $(X_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} p$. It is possible to use an auxiliary distribution q for Monte Carlo integration similar to rejection sampling, but without an explicit accept-reject mechanism.

This can be of interest from different reasons, for instance:

- Being less wasteful by ‘recycling’ samples that would be rejected in rejection sampling.
- Reducing Monte Carlo variance.
- Use when M in (4) or (5) is unknown, or even when no such finite M exists.

4.1 Unbiased importance sampling

Definition 4.1 (Importance sampling)**.** Suppose p and q are two p.d.f.s or p.m.f.s on $\mathbb{X}$ and $f : \mathbb{X} \to \mathbb{R}$.

$$\boxed{\text{Assumption: } q(x) = 0 \implies p(x)f(x) = 0.} \tag{6}$$

Define

$$w(x) := \begin{cases} \frac{p(x)}{q(x)}, & \text{if } q(x) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

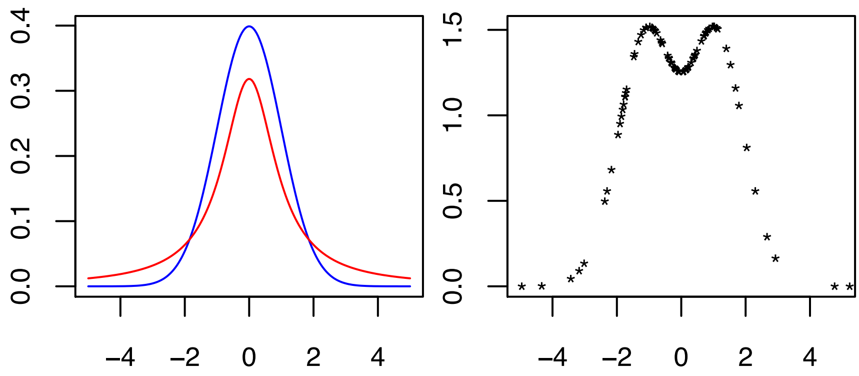

Figure 6: Importance sampling with p standard Normal (blue) and q Cauchy (red), as in Example 3.8). The importance weights $w(Y_k)$ are shown on the right.

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

$$I_{p,q}^{(n)}(f) := \frac{1}{n}\sum_{k=1}^{n} f(Y_k)w(Y_k) \tag{7}$$

is the (unbiased) importance sampling (IS) approximation of $\mathbb{E}_p[f(X)]$.

Remark 4.2. The distribution q is called the *proposal distribution* (sometimes also *importance* or *instrumental*). The term $w(Y_k)$ is called the *(importance) weight* related to the sample Y_k.

Theorem 4.3. *Assuming* (6) *holds, then the IS estimator is*

(a) *Unbiased:* $\mathbb{E}[I_{p,q}^{(n)}(f)] = \mathbb{E}_p[f(X)]$, *for all* $n \in \mathbb{N}$

(b) *Consistent:* $I_{p,q}^{(n)}(f) \xrightarrow{n\to\infty} \mathbb{E}_p[f(X)]$ *(almost surely).*

Proof. Because the random variables $Z_k := f(Y_k)w(Y_k)$ are i.i.d., it is sufficient for (a) to check that $\mathbb{E}[Z_1] = \mathbb{E}_p[f(X)]$. In the discrete case,

$$\mathbb{E}[Z_1] = \sum_{y\in\mathbb{X}\,:\,q(y)>0} f(y)\frac{p(y)}{q(y)}q(y) = \sum_{y\in\mathbb{X}} f(y)p(y)\mathrm{d}y = \mathbb{E}_p[f(X)],$$

and similarly in the continuous case, changing the sum to an integral. The almost sure convergence (b) follows from the strong law of large numbers. □

Remark 4.4 (*). In terms of general probability, importance sampling is a *change of measure*, and the function w is the related *Radon-Nikodym derivative*.

Example 4.5 (Gamma distribution). Example 2.14 showed how to simulate $Y \sim \Gamma(a, b)$ for $a \in \mathbb{N}_+$ and $b > 0$ by summing exponentials.

Suppose we have simulated $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \Gamma(a, b)$, but want to estimate the expectation of $f(X)$ where $X \sim \Gamma(\alpha, \beta)$, with some other parameters $\alpha, \beta > 0$.

Recall that the density of $\Gamma(\alpha, \beta)$ is

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)\mathbf{1}\,(x > 0)$$

so the importance weights are given as

$$w(y) = \frac{p(y)}{q(y)} = \frac{\Gamma(a)\beta^\alpha}{\Gamma(\alpha)b^a} y^{\alpha-a} \exp\big(-(\beta-b)y\big), \qquad \text{for } y > 0.$$

The importance sampling estimator is (NB: $\mathbb{P}(q(Y_i)=0)=0$!)

$$\begin{aligned} I_{p,q}^{(n)}(f) &= \frac{1}{n}\sum_{i=1}^n f(Y_i)w(Y_i) \\ &= \frac{\Gamma(a)\beta^\alpha}{\Gamma(\alpha)b^a} \cdot \frac{1}{n}\sum_{i=1}^n f(Y_i)Y_i^{\alpha-a}\exp\big(-(\beta-b)Y_i\big). \end{aligned}$$

We know that this is unbiased and (strongly) consistent estimator of $\mathbb{E}_p[f(X)]$.

Remark 4.6. In fact, we can 'recycle' the samples the $Y_1,\ldots,Y_n \overset{\text{i.i.d.}}{\sim} q$ in Example 4.5 to obtain estimates of $\mathbb{E}_{p_{\alpha,\beta}}[f(X)]$ with $p_{\alpha,\beta}$ corresponding to $\Gamma(\alpha,\beta)$, for a *range* of values α and $\beta\ldots$

Theorem 4.3 showed that IS is consistent with minimal conditions. How about the variance of IS?

Proposition 4.7. *Suppose that* (6) *holds. Then, the variance of the IS estimator can be given as*

$$\operatorname{Var}\big(I_{p,q}^{(n)}(f)\big) = \frac{\sigma_{p,q}^2}{n} \qquad \textit{where} \qquad \sigma_{p,q}^2 := \mathbb{E}_p[f^2(X)w(X)] - \mathbb{E}_p[f(X)]^2.$$

Note that this permits the case $\sigma_{p,q}^2 = \infty \implies \operatorname{Var}\big(I_{p,q}^{(n)}(f)\big) = \infty\ \forall n \in \mathbb{N}$.

Proof. Denote $Z_k := f(Y_k)w(Y_k)$, then in the discrete case

$$\mathbb{E}[Z_1^2] = \sum_{y\in\mathbb{X}\,:\,q(y)>0} f^2(y)\frac{p^2(y)}{q^2(y)}q(y) = \sum_{y\in\mathbb{X}} f^2(y)w(y)p(y)\mathrm{d}y = \mathbb{E}_p[f^2(X)w(X)].$$

Now, $\sigma_{p,q}^2 = \operatorname{Var}(Z_1) = \mathbb{E}Z_1^2 - (\mathbb{E}Z_1)^2$ and $\mathbb{E}Z_1 = \mathbb{E}_p[f(X)]$, and as (Z_k) are i.i.d., $\operatorname{Var}(I_{p,q}^{(n)(f)}) = \sigma_{p,q}^2/n$. The continuous case follows similarly. □

Because $I_{p,q}^{(n)}(f)$ is a sum of i.i.d. random variables, the proof of Proposition 4.7 implies the following:

Corollary 4.8. *Suppose* (6) *holds and*

$$\mathbb{E}_p[f^2(X)w(X)] < \infty. \tag{8}$$

Then, $\sqrt{n}[I_{p,q}^{(n)}(f) - \mathbb{E}_p[f(X)]] \xrightarrow{n\to\infty} N(0,\sigma_{p,q}^2)$ *in distribution.*

Remark 4.9. Because IS is just usual Monte Carlo approximating $\mathbb{E}_q[g(X)]$ with $g(x) = f(x)w(x)$, Proposition 1.13 holds, and gives confidence intervals also for the IS estimator.

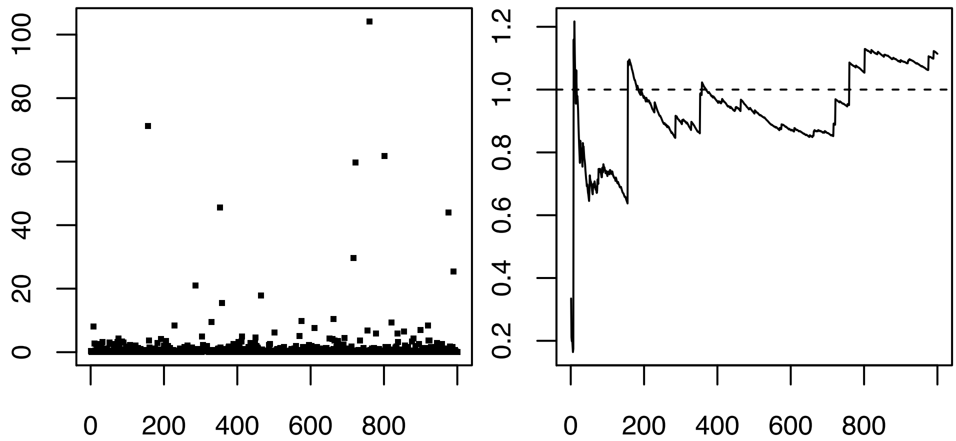

Figure 7: Example 4.10 with $a = 2$, $b = 2$ and $\beta = 2.5$, $\alpha = 0.5$ (NB $\alpha < a$) and $f(x) \equiv 1$. Values of the weights $w(Y_n)$ (left) and the sequence of estimates $I_{p,q}^{(n)}(f)$ (right) for $n = 1, 2, \ldots, 1000$.

Example 4.10 (Gamma distribution (cont.)). Let us consider the variance of the IS estimator for the Gamma distributions in Example 4.5. We may write

$$w(x)f^2(x) = \frac{\Gamma(a)\beta^\alpha}{\Gamma(\alpha)b^a} x^{\alpha-a} \exp(-(\beta - b)x) f^2(x),$$

so

$$\mathbb{E}_p[w(X)f^2(X)] = c_{a,b,\alpha,\beta}\mathbb{E}_p\big[X^{\alpha-a} \exp(-(\beta - b)X)f^2(X)\big].$$

If $\alpha \geq a$ and $\beta > b$, then

$$\sup_{x>0} \big[x^{\alpha-a} \exp(-(\beta - b)x)\big] < \infty.$$

In this case $\mathbb{E}_p[w(X)f^2(X)] \leq c\mathbb{E}_p[f^2(X)]$, so if also $\mathrm{Var}_p(f(X)) < \infty \iff \mathbb{E}_p[f^2(X)] < \infty$, then we have $\mathbb{E}_p[w(X)f^2(X)] < \infty$ and the importance sampling estimator is guaranteed to have finite variance.

Figure 7 shows an example simulation where $\mathrm{Var}(I_{p,q}^{(n)}(f)) = \infty$. Exercise: What would happen if we used $f(x) = x$ instead?

We formalise the sufficient condition found in the Gamma example above.

Proposition 4.11. *Suppose* (6) *holds and*

$$M := \sup_x w(x) = \sup_x \frac{p(x)}{q(x)} < \infty, \tag{9}$$

where the supremum is taken over all $x \in \mathbb{X}$ such that $p(x)f(x) > 0$. Then, if $\mathrm{Var}_p(f(X)) < \infty$*, the variance of the IS estimator is finite, and can be upper bounded by*

$$\begin{aligned}\sigma_{p,q}^2 &\leq M\mathbb{E}_p[f^2(X)] - \mathbb{E}_p[f(X)]^2 \\ &= M\mathrm{Var}_p(f(X)) + (M - 1)\mathbb{E}_p[f(X)]^2.\end{aligned}$$

Remark 4.12. If $\mathbb{E}_p[f(X)]^2 \ll \mathbb{E}_p[f^2(X)]$, Proposition 4.11 indicates that the IS estimator is (roughly) at most M times worse than the classical Monte Carlo estimate. How does this result relate with using rejection sampling instead of IS?

> Rule of thumb: Try to make sure that (9) holds (unless you have a specific f in mind).

What is the best possible proposal density q for a specific f?

Proposition 4.13. *Suppose that $f : \mathbb{X} \to \mathbb{R}$ satisfies $\mathbb{E}_p[|f(X)|] > 0$. Then, the proposal distribution*

$$q_*(x) := \frac{p(x)|f(x)|}{\mathbb{E}_p[|f(X)|]} \propto p(x)|f(x)|$$

admits the minimum variance among all distributions q satisfying (6).

Proof. In the discrete case, we have with $w_*(x) = p(x)/q_*(x)$,

$$\mathbb{E}_p[f^2(X)w_*(X)] = \sum_{x \in \mathbb{X}\,:\, q_*(x)>0} f^2(x)\frac{p^2(x)}{q_*(x)} = \big(\mathbb{E}_p[|f(X)|]\big)^2$$

On the other hand, for any q satisfying (6),

$$\big(\mathbb{E}_p[|f(X)|]\big)^2 = \Big(\mathbb{E}_q[|f(X)|w(X)]\Big)^2 \leq \mathbb{E}_q[f^2(X)w^2(X)] = \mathbb{E}_p[f^2(X)w(X)],$$

by Jensen's inequality. This implies $\sigma^2_{p,q_*} \leq \sigma^2_{p,q}$ by Proposition 4.7. □

Remark 4.14. The result of Proposition 4.13 is, of course, mostly theoretical, but leads to:

> Rule of thumb: Try to find q that is approximately proportional to $p(x)|f(x)|$.

In particular, if f is zero (or has very small absolute values) in some regions of the space, we avoid putting any (or put less) mass of q to such regions.

Remark 4.15. Note in particular that IS can have, in fact, a (significantly) smaller variance than the classical Monte Carlo estimate. We restate the main reasons to use IS rather than classical Monte Carlo:

- Use IS when we cannot sample (efficiently) from p.
- Use IS to reduce variance over the classical Monte Carlo estimator.
- Rejection sampling is not applicable (because we do not know $M < \infty$, or $M = \infty$)

4.2 Application: Rare event estimation

One important class of applications of IS as variance reduction is problems in which we estimate the probability of a rare event. In such scenarios, we may be able to sample from p directly but this leads to high variance.