

Chapter 3

Fundamental Concepts: Transformation, Rejection, and Reweighting

3.1 Transformation methods

In section 2.4 we have seen how to create (pseudo-)random numbers from the uniform distribution $U[0, 1]$. One of the simplest methods of generating random samples from a distribution with cumulative distribution function (c.d.f.) $F(x) = \mathbb{P}(X \leq x)$ is based on the inverse of the c.d.f..

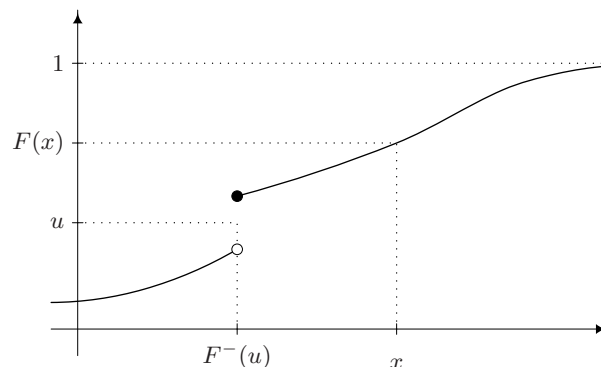


Figure 3.1. Illustration of the definition of the generalised inverse F^- of a c.d.f. F

The c.d.f. is an increasing function, however it is not necessarily continuous. Thus we define the *generalised inverse* $F^-(u) = \inf\{x : F(x) \geq u\}$. Figure 3.1 illustrates its definition. If F is continuous, then $F^-(u) = F^{-1}(u)$.

Theorem 3.1 (Inversion Method). Let $U \sim U[0, 1]$ and F be a c.d.f.. Then $F^-(U)$ has the c.d.f. F .

Proof. It is easy to see (e.g. in figure 3.1) that $F^-(u) \leq x$ is equivalent to $u \leq F(x)$. Thus for $U \sim U[0, 1]$

$$\mathbb{P}(F^-(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

thus F is the c.d.f. of $X = F^-(U)$. □

Example 3.1 (Exponential Distribution). The exponential distribution with rate $\lambda > 0$ has the c.d.f. $F_\lambda(x) = 1 - \exp(-\lambda x)$ for $x \geq 0$. Thus $F_\lambda^-(u) = F_\lambda^{-1}(u) = -\log(1 - u)/\lambda$. Thus we can generate random samples from $\text{Exp}(\lambda)$ by applying the transformation $-\log(1 - U)/\lambda$ to a uniform $U[0, 1]$ random variable U . As U and $1 - U$, of course, have the same distribution we can use $-\log(U)/\lambda$ as well. ◁

The Inversion Method is a very efficient tool for generating random numbers. However very few distributions possess a c.d.f. whose (generalised) inverse can be evaluated efficiently. Take the example of the Gaussian distribution, whose c.d.f. is not even available in closed form.

Note however that the generalised inverse of the c.d.f. is just one possible transformation and that there might be other transformations that yield the desired distribution. An example of such a method is the Box-Muller method for generating Gaussian random variables.

Example 3.2 (Box-Muller Method for Sampling from Gaussians). When sampling from the normal distribution, one faces the problem that neither the c.d.f. $\Phi(\cdot)$, nor its inverse has a closed-form expression. Thus we cannot use the inversion method.

It turns out however, that if we consider a pair $X_1, X_2 \stackrel{\text{i.i.d.}}{\sim} \text{N}(0, 1)$, as a point (X_1, X_2) in the plane, then its polar coordinates (R, θ) are again independent and have distributions we can easily sample from: $\theta \sim \text{U}[0, 2\pi]$, and $R^2 \sim \text{Expo}(1/2)$.

This can be shown as follows. Assume that $\theta \sim \text{U}[0, 2\pi]$ and $R^2 \sim \text{Expo}(1/2)$. Then the joint density of (θ, r^2) is

$$f_{(\theta, r^2)}(\theta, r^2) = \frac{1}{2\pi} \mathbf{1}_{[0, 2\pi]}(\theta) \cdot \frac{1}{2} \exp\left(-\frac{1}{2}r^2\right) = \frac{1}{4\pi} \exp\left(-\frac{1}{2}r^2\right) \cdot \mathbf{1}_{[0, 2\pi]}(\theta)$$

To obtain the probability density function of

$$X_1 = \sqrt{R^2} \cdot \cos(\theta), \quad X_2 = \sqrt{R^2} \cdot \sin(\theta)$$

we need to use the transformation of densities formula.

$$\begin{aligned} f_{(X_1, X_2)}(x_1, x_2) &= f_{(\theta, r^2)}(\theta(x_1, x_2), r^2(x_1, x_2)) \cdot \left| \begin{array}{cc} \frac{\partial x_1}{\partial \theta} & \frac{\partial x_1}{\partial r^2} \\ \frac{\partial x_2}{\partial \theta} & \frac{\partial x_2}{\partial r^2} \end{array} \right|^{-1} = \frac{1}{4\pi} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right) \cdot 2 \\ &= \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_1^2\right)\right) \cdot \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_2^2\right)\right) \end{aligned}$$

as

$$\left| \begin{array}{cc} \frac{\partial x_1}{\partial \theta} & \frac{\partial x_1}{\partial r^2} \\ \frac{\partial x_2}{\partial \theta} & \frac{\partial x_2}{\partial r^2} \end{array} \right| = \left| \begin{array}{cc} -r \sin(\theta) & \frac{\cos(\theta)}{2r} \\ r \cos(\theta) & \frac{\sin(\theta)}{2r} \end{array} \right| = \left| -\frac{r \sin(\theta)^2}{2r} - \frac{r \cos(\theta)^2}{2r} \right| = \frac{1}{2}$$

Thus $X_1, X_2 \sim \text{N}(0, 1)$. As their joint density factorises, X_1 and X_2 are independent, as required.

Thus we only need to generate $\theta \sim \text{U}[0, 2\pi]$, and $R^2 \sim \text{Expo}(1/2)$. Using $U_1, U_2 \stackrel{\text{i.i.d.}}{\sim} \text{U}[0, 1]$ and example 3.1 we can generate $R = \sqrt{R^2}$ and θ by

$$R = \sqrt{-2 \log(U_1)}, \quad \theta = 2\pi U_2,$$

and thus

$$X_1 = \sqrt{-2 \log(U_1)} \cdot \cos(2\pi U_2), \quad X_2 = \sqrt{-2 \log(U_1)} \cdot \sin(2\pi U_2)$$

are two independent realisations from a $\text{N}(0, 1)$ distribution. ◁

The idea of transformation methods like the Inversion Method was to generate random samples from a distribution other than the target distribution and to transform them such that they come from the desired target distribution. In many situations, we cannot find such a transformation in closed form. In these cases we have to find other ways of correcting for the fact that we sample from the “wrong” distribution. The next two sections present two such ideas: rejection sampling and importance sampling.

3.2 Rejection sampling

The basic idea of rejection sampling is to sample from an *instrumental distribution*¹ and reject samples that are “unlikely” under the target distribution.

Assume that we want to sample from a target distribution whose density f is known to us. The simple idea underlying rejection sampling (and other Monte Carlo algorithms) is the rather trivial identity

$$f(x) = \int_0^{f(x)} 1 \, du = \int_0^1 \underbrace{1_{0 < u < f(x)}}_{=f(x,u)} \, du$$

Thus $f(x)$ can be interpreted as the marginal density of a uniform distribution on the area under the density $f(x)$

$$\{(x, u) : 0 \leq u \leq f(x)\}.$$

Figure 3.2 illustrates this idea. This suggests that we can generate a sample from f by sampling from the area under the curve.

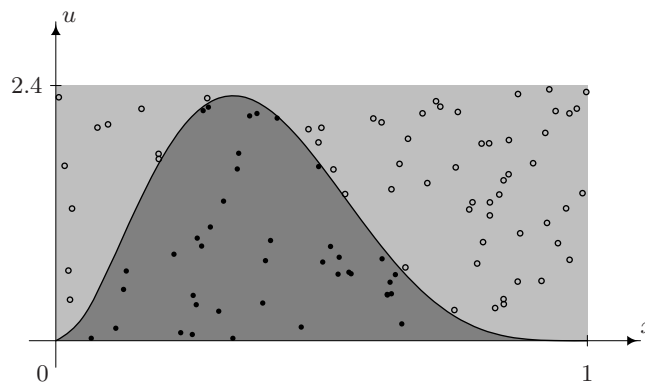


Figure 3.2. Illustration of example 3.3. Sampling from the area under the curve (dark grey) corresponds to sampling from the Beta(3, 5) density. In example 3.3 we use a uniform distribution of the light grey rectangle as proposal distribution. Empty circles denote rejected values, filled circles denote accepted values.

Example 3.3 (Sampling from a Beta distribution). The Beta(a, b) distribution ($a, b \geq 0$) has the density

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad \text{for } 0 < x < 1,$$

where $\Gamma(a) = \int_0^{+\infty} t^{a-1} \exp(-t) \, dt$ is the Gamma function. For $a, b > 1$ the Beta(a, b) density is unimodal with mode $(a-1)/(a+b-2)$. Figure 3.2 shows the density of a Beta(3, 5) distribution. It attains its maximum of $1680/729 \approx 2.305$ at $x = 1/3$.

Using the above identity we can draw from Beta(3, 5) by drawing from a uniform distribution on the area under the density $\{(x, u) : 0 < u < f(x)\}$ (the area shaded in dark gray in figure 3.2).

In order to sample from the area under the density, we will use a similar trick as in examples 2.1 and 2.2. We will sample from the light grey rectangle and only keep the samples that fall in the area under the curve. Figure 3.2 illustrates this idea.

Mathematically speaking, we sample independently $X \sim U[0, 1]$ and $U \sim U[0, 2.4]$. We keep the pair (X, U) if $U < f(X)$, otherwise we reject it.

The conditional probability that a pair (X, U) is kept if $X = x$ is

$$\mathbb{P}(U < f(X) | X = x) = \mathbb{P}(U < f(x)) = f(x)/2.4$$

¹ The instrumental distribution is sometimes referred to as *proposal distribution*.

As X and U were drawn independently we can rewrite our algorithm as: Draw X from $U[0, 1]$ and accept X with probability $f(X)/2.4$, otherwise reject X . \triangleleft

The method proposed in example 3.3 is based on bounding the density of the Beta distribution by a box. Whilst this is a powerful idea, it cannot be directly applied to other distributions, as the density might be unbounded or have infinite support. However we might be able to bound the density of $f(x)$ by $M \cdot g(x)$, where $g(x)$ is a density that we can easily sample from.

Algorithm 3.1 (Rejection sampling). Given two densities f, g with $f(x) < M \cdot g(x)$ for all x , we can generate a sample from f as follows:

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.

Proof. We have

$$\mathbb{P}(X \in \mathcal{X} \text{ and is accepted}) = \int_{\mathcal{X}} g(x) \underbrace{\frac{f(x)}{M \cdot g(x)}}_{=\mathbb{P}(X \text{ is accepted} | X=x)} dx = \frac{\int_{\mathcal{X}} f(x) dx}{M}, \quad (3.1)$$

and thus²

$$\mathbb{P}(X \text{ is accepted}) = \mathbb{P}(X \in S \text{ and is accepted}) = \frac{1}{M}, \quad (3.2)$$

yielding

$$\mathbb{P}(x \in \mathcal{X} | X \text{ is accepted}) = \frac{\mathbb{P}(X \in \mathcal{X} \text{ and is accepted})}{\mathbb{P}(X \text{ is accepted})} = \frac{\int_{\mathcal{X}} f(x) dx / M}{1/M} = \int_{\mathcal{X}} f(x) dx. \quad (3.3)$$

Thus the density of the values accepted by the algorithm is $f(\cdot)$. \square

Remark 3.2. If we know f only up to a multiplicative constant, i.e. if we only know $\pi(x)$, where $f(x) = C \cdot \pi(x)$, we can carry out rejection sampling using

$$\frac{\pi(X)}{M \cdot g(X)}$$

as probability of rejecting X , provided $\pi(x) < M \cdot g(x)$ for all x . Then by analogy with (3.1) - (3.3) we have

$$\mathbb{P}(X \in \mathcal{X} \text{ and is accepted}) = \int_{\mathcal{X}} g(x) \frac{\pi(x)}{M \cdot g(x)} dx = \frac{\int_{\mathcal{X}} \pi(x) dx}{M} = \frac{\int_{\mathcal{X}} f(x) dx}{C \cdot M},$$

$\mathbb{P}(X \text{ is accepted}) = 1/(C \cdot M)$, and thus

$$\mathbb{P}(x \in \mathcal{X} | X \text{ is accepted}) = \frac{\int_{\mathcal{X}} f(x) dx / (C \cdot M)}{1/(C \cdot M)} = \int_{\mathcal{X}} f(x) dx$$

Example 3.4 (Rejection sampling from the $N(0, 1)$ distribution using a Cauchy proposal). Assume we want to sample from the $N(0, 1)$ distribution with density

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

using a Cauchy distribution with density

$$g(x) = \frac{1}{\pi(1+x^2)}$$

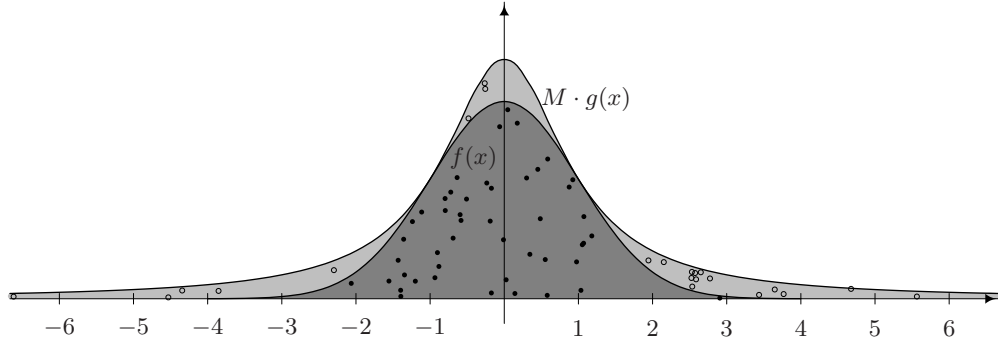


Figure 3.3. Illustration of example 3.3. Sampling from the area under the density $f(x)$ (dark grey) corresponds to sampling from the $N(0, 1)$ density. The proposal $g(x)$ is a $Cauchy(0, 1)$.

as instrumental distribution.³ The smallest M we can choose such that $f(x) \leq M g(x)$ is $M = \sqrt{2\pi} \cdot \exp(-1/2)$. Figure 3.3 illustrates the results. As before, filled circles correspond to accepted values whereas open circles correspond to rejected values.

Note that it is impossible to do rejection sampling from a Cauchy distribution using a $N(0, 1)$ distribution as instrumental distribution: there is no $M \in \mathbb{R}$ such that

$$\frac{1}{\pi(1+x^2)} < M \cdot \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{x^2}{2}\right);$$

the Cauchy distribution has heavier tails than the Gaussian distribution. ◁

3.3 Importance sampling

In rejection sampling we have compensated for the fact that we sampled from the instrumental distribution $g(x)$ instead of $f(x)$ by rejecting some of the values proposed by $g(x)$. Importance sampling is based on the idea of using weights to correct for the fact that we sample from the instrumental distribution $g(x)$ instead of the target distribution $f(x)$.

Importance sampling is based on the identity

$$\mathbb{P}(X \in A) = \int_A f(x) dx = \int_A g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} dx = \int_A g(x)w(x) dx \tag{3.4}$$

for all $g(\cdot)$, such that $g(x) > 0$ for (almost) all x with $f(x) > 0$. We can generalise this identity by considering the expectation $\mathbb{E}_f(h(X))$ of a measurable function h :

$$\mathbb{E}_f(h(X)) = \int_S f(x)h(x) dx = \int_S g(x) \underbrace{\frac{f(x)}{g(x)}}_{=:w(x)} h(x) dx = \int_S g(x)w(x)h(x) dx = \mathbb{E}_g(w(X) \cdot h(X)), \tag{3.5}$$

if $g(x) > 0$ for (almost) all x with $f(x) \cdot h(x) \neq 0$.

Assume we have a sample $X_1, \dots, X_n \sim g$. Then, provided $\mathbb{E}_g|w(X) \cdot h(X)|$ exists,

$$\frac{1}{n} \sum_{i=1}^n w(X_i)h(X_i) \xrightarrow[n \rightarrow \infty]{a.s.} \mathbb{E}_g(w(X) \cdot h(X))$$

(by the Law of Large Numbers) and thus by (3.5)

$$\frac{1}{n} \sum_{i=1}^n w(X_i)h(X_i) \xrightarrow[n \rightarrow \infty]{a.s.} \mathbb{E}_f(h(X)).$$

² We denote by S the set of all possible values X can take, i.e., $\int_S f(x)dx = 1$.

³ There is not much point in using this method is practise. The Box-Muller method is more efficient.

In other words, we can estimate $\mu = \mathbb{E}_f(h(X))$ by

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

Note that whilst $\mathbb{E}_g(w(X)) = \int_S \frac{f(x)}{g(x)}g(x) dx = \int_S f(x) = 1$, the weights $w_1(X), \dots, w_n(X)$ do not necessarily sum up to n , so one might want to consider the *self-normalised* version

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

This gives rise to the following algorithm:

Algorithm 3.2 (Importance Sampling). Choose g such that $\text{supp}(g) \supset \text{supp}(f \cdot h)$.

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

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

or

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

The following theorem gives the bias and the variance of importance sampling.

Theorem 3.3 (Bias and Variance of Importance Sampling). (a) $\mathbb{E}_g(\tilde{\mu}) = \mu$

$$(b) \text{Var}_g(\tilde{\mu}) = \frac{\text{Var}_g(w(X) \cdot h(X))}{n}$$

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

$$(d) \text{Var}_g(\hat{\mu}) = \frac{\text{Var}_g(w(X) \cdot h(X)) - 2\mu \text{Cov}_g(w(X), w(X) \cdot h(X)) + \mu^2 \text{Var}_g(w(X))}{n} + O(n^{-2})$$

$$\text{Proof. (a) } \mathbb{E}_g \left(\frac{1}{n} \sum_{i=1}^n w(X_i)h(X_i) \right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_g(w(X_i)h(X_i)) = \mathbb{E}_f(h(X))$$

$$(b) \text{Var}_g \left(\frac{1}{n} \sum_{i=1}^n w(X_i)h(X_i) \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}_g(w(X_i)h(X_i)) = \frac{\text{Var}_g(w(X)h(X))}{n}$$

(c) and (d) see (Liu, 2001, p. 35) □

Note that the theorem implies that in contrast to $\tilde{\mu}$ the self-normalised estimator $\hat{\mu}$ is biased. The self-normalised estimator $\hat{\mu}$ however might have a lower variance. In addition, it has another advantage: we only need to know the density up to a multiplicative constant, as it is often the case in hierarchical Bayesian modelling. Assume $f(x) = C \cdot \pi(x)$, then

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

i.e. the self-normalised estimator $\hat{\mu}$ does not depend on the normalisation constant C .⁴ On the other hand, as we have seen in the proof of theorem 3.3 it is a lot harder to analyse the theoretical properties of the self-normalised estimator $\hat{\mu}$.

Although the above equations (3.4) and (3.5) hold for every g with $\text{supp}(g) \supset \text{supp}(f \cdot h)$ and the importance sampling algorithm converges for a large choice of such g , one typically only considers choices of g that lead to *finite variance estimators*. The following two conditions are each sufficient (albeit rather restrictive) for a finite variance of $\tilde{\mu}$:

⁴ By complete analogy one can show that is enough to know g up to a multiplicative constant.

- $f(x) < M \cdot g(x)$ and $\text{Var}_f(h(X)) < +\infty$.
- S is compact, f is bounded above on S , and g is bounded below on S .

Note that under the first condition rejection sampling can also be used to sample from f .

So far we have only studied whether a distribution g leads to a finite-variance estimator. This leads to the question which instrumental distribution is *optimal*, i.e. for which choice $\text{Var}(\tilde{\mu})$ is minimal. The following theorem answers this question:

Theorem 3.4 (Optimal proposal). *The proposal distribution g that minimises the variance of $\tilde{\mu}$ is*

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

Proof. We have from theorem 3.3 (b) that

$$n \cdot \text{Var}_g(\tilde{\mu}) = \text{Var}_g(w(X) \cdot h(X)) = \text{Var}_g\left(\frac{h(X) \cdot f(X)}{g(X)}\right) = \mathbb{E}_g\left(\left(\frac{h(X) \cdot f(X)}{g(X)}\right)^2\right) - \underbrace{\left(\mathbb{E}_g\left(\frac{h(X) \cdot f(X)}{g(X)}\right)\right)^2}_{=\mathbb{E}_g(\tilde{\mu})=\mu}.$$

Thus we only have to minimise $\mathbb{E}_g\left(\left(\frac{h(X) \cdot f(X)}{g(X)}\right)^2\right)$. When plugging in g^* we obtain:

$$\begin{aligned} \mathbb{E}_{g^*}\left(\left(\frac{h(X) \cdot f(X)}{g^*(X)}\right)^2\right) &= \int_S \frac{h(x)^2 \cdot f(x)^2}{g^*(x)} dx = \left(\int_S \frac{h(x)^2 \cdot f(x)^2}{|h(x)|f(x)} dx\right) \cdot \left(\int_S |h(t)|f(t) dt\right) \\ &= \left(\int_S |h(x)|f(x) dx\right)^2 \end{aligned}$$

On the other hand, we can apply the Jensen inequality⁵ to $\mathbb{E}_g\left(\left(\frac{h(X) \cdot f(X)}{g(X)}\right)^2\right)$ yielding

$$\mathbb{E}_g\left(\left(\frac{h(X) \cdot f(X)}{g(X)}\right)^2\right) \geq \left(\mathbb{E}_g\left(\frac{|h(X)| \cdot f(X)}{g(X)}\right)\right)^2 = \left(\int_S |h(x)|f(x) dx\right)^2.$$

□

An important corollary of theorem 3.4 is that importance sampling can be *super-efficient*, i.e. when using the optimal g^* from theorem 3.4 the variance of $\tilde{\mu}$ is less than the variance obtained when sampling directly from f :

$$\begin{aligned} n \cdot \text{Var}_f\left(\frac{h(X_1) + \dots + h(X_n)}{n}\right) &= \mathbb{E}_f(h(X)^2) - \mu^2 \\ &\geq (\mathbb{E}_f|h(X)|)^2 - \mu^2 = \left(\int_S |h(x)|f(x) dx\right)^2 - \mu^2 = n \cdot \text{Var}_{g^*}(\tilde{\mu}) \end{aligned}$$

by Jensen's inequality. Unless $h(X)$ is (almost surely) constant the inequality is strict. There is an intuitive explanation to the super-efficiency of importance sampling. Using g^* instead of f causes us to focus on regions of high probability where $|h|$ is large, which contribute most to the integral $\mathbb{E}_f(h(X))$.

Theorem 3.4 is, however, a rather formal optimality result. When using $\tilde{\mu}$ we need to know the normalisation constant of g^* , which is exactly the integral we are looking for. Further we need to be able to draw samples from g^* efficiently. The practically important corollary of theorem 3.4 is that we should choose an instrumental distribution g whose shape is close to the one of $f \cdot |h|$.

Example 3.5 (Computing $\mathbb{E}_f|X|$ for $X \sim t_3$). Assume we want to compute $\mathbb{E}_f|X|$ for X from a t-distribution with 3 degrees of freedom (t_3) using a Monte Carlo method. Three different schemes are considered

⁵ If X is real-valued random variable, and ψ a convex function, then $\psi(\mathbb{E}(X)) \leq \mathbb{E}(\psi(X))$.

– Sampling X_1, \dots, X_n directly from t_3 and estimating $\mathbb{E}_f|X|$ by

$$\frac{1}{n} \sum_{i=1}^n |X_i|.$$

- Alternatively we could use importance sampling using a t_1 (which is nothing other than a Cauchy distribution) as instrumental distribution. The idea behind this choice is that the density $g_{t_1}(x)$ of a t_1 distribution is closer to $f(x)|x|$, where $f(x)$ is the density of a t_3 distribution, as figure 3.4 shows.
- Third, we will consider importance sampling using a $N(0, 1)$ distribution as instrumental distribution.

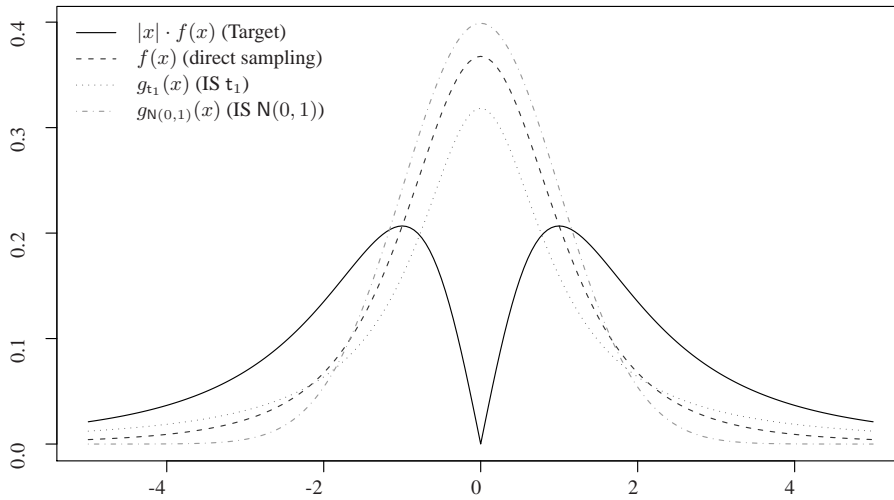


Figure 3.4. Illustration of the different instrumental distributions in example 3.5.

Note that the third choice yields weights of infinite variance, as the instrumental distribution $(N(0, 1))$ has lighter tails than the distribution we want to sample from (t_3). The right-hand panel of figure 3.5 illustrates that this choice yields a very poor estimate of the integral $\int |x|f(x) dx$.

Sampling directly from the t_3 distribution can be seen as importance sampling with all weights $w_i \equiv 1$, this choice clearly minimises the variance of the weights. This however does not imply that this yields an estimate of the integral $\int |x|f(x) dx$ of minimal variance. Indeed, after 1500 iterations the empirical standard deviation (over 100 realisations) of the direct estimate is 0.0345, which is larger than the empirical standard deviation of $\tilde{\mu}$ when using a t_1 distribution as instrumental distribution, which is 0.0182. So using a t_1 distribution as instrumental distribution is super-efficient (see figure 3.5).

Figure 3.6 somewhat explains why the t_1 distribution is a far better choice than the $N(0, 1)$ distribution. As the $N(0, 1)$ distribution does not have heavy enough tails, the weight tends to infinity as $|x| \rightarrow +\infty$. Thus large $|x|$ get large weights, causing the jumps of the estimate $\tilde{\mu}$ shown in figure 3.5. The t_1 distribution has heavy enough tails, so the weights are small for large values of $|x|$, explaining the small variance of the estimate $\tilde{\mu}$ when using a t_1 distribution as instrumental distribution. ◁

Example 3.6 (Partially labelled data). Suppose that we are given count data from observations in two groups, such that

$$\begin{aligned} Y_i &\sim \text{Poi}(\lambda_1) && \text{if the } i\text{-th observation is from group 1} \\ Y_i &\sim \text{Poi}(\lambda_2) && \text{if the } i\text{-th observation is from group 2} \end{aligned}$$

The data is given in the table 3.1. Note that only the first ten observations are labelled, the group label is missing for the remaining ten observations.

We will use a $\text{Gamma}(\alpha, \beta)$ distribution as (conjugate) prior distribution for λ_j , i.e. the prior density of λ_j is

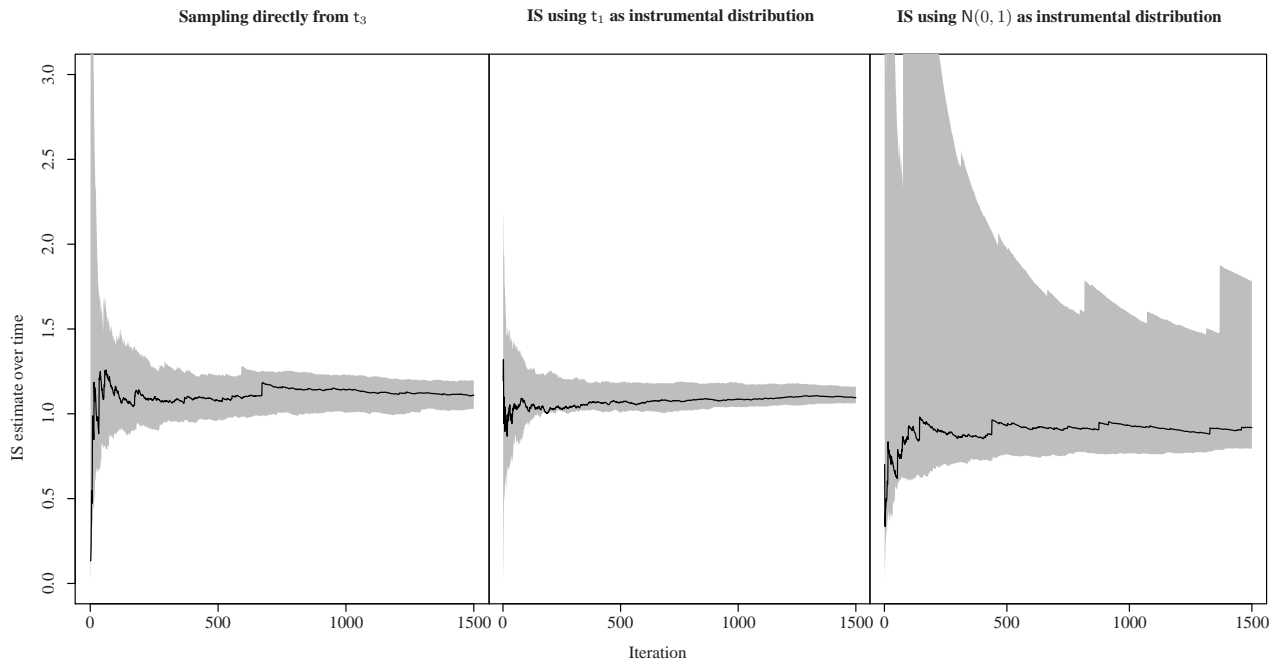


Figure 3.5. Estimates of $E|X|$ for $X \sim t_3$ obtained after 1 to 1500 iterations. The three panels correspond to the three different sampling schemes used. The areas shaded in grey correspond to the range of 100 replications.

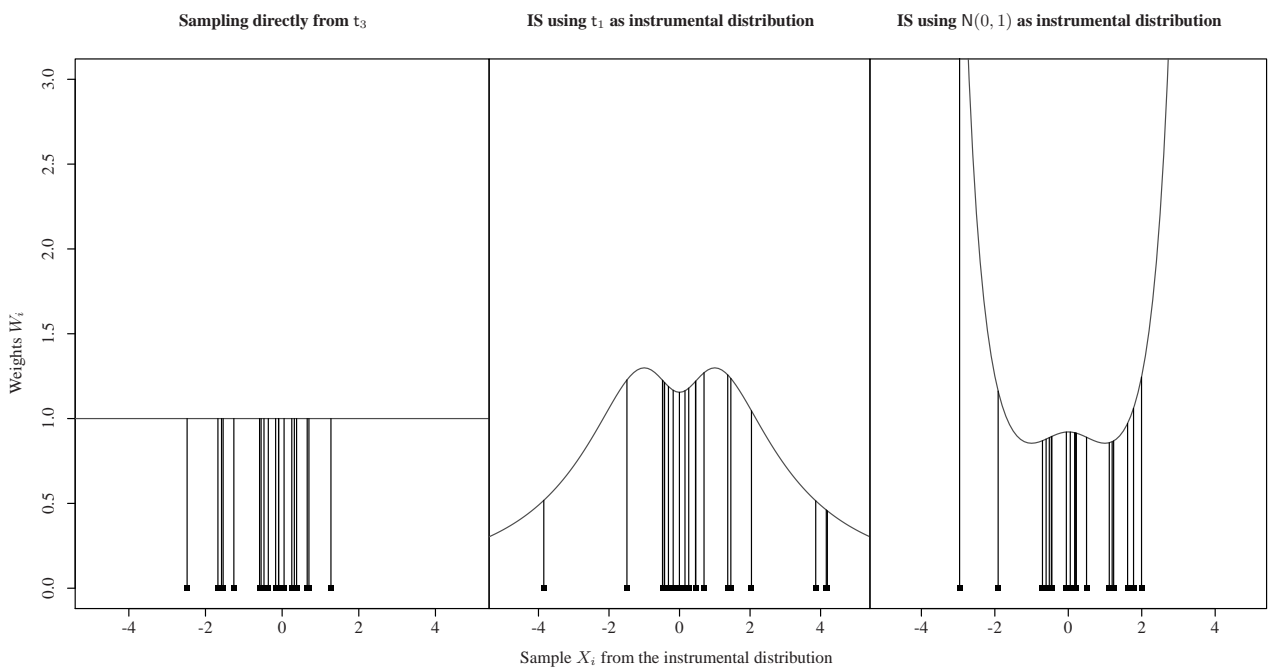


Figure 3.6. Weights W_i obtained for 20 realisations X_i from the different instrumental distributions.

Group	Count Y_i	Group	Count Y_i	Group	Count Y_i	Group	Count Y_i
1	3	2	14	*	15	*	21
1	6	2	12	*	4	*	11
1	3	2	11	*	1	*	3
1	5	2	19	*	6	*	7
1	9	2	18	*	11	*	18

Table 3.1. Data of example 3.6.

$$f(\lambda_j) = \frac{1}{\Gamma(\alpha)} \lambda_j^{\alpha-1} \beta_j^\alpha \exp(-\beta \lambda_j).$$

Furthermore, we believe that a priori each observation is equally likely to stem from group 1 or group 2.

We start with analysing the labelled data only, ignoring the 10 unlabelled observations. In this case, we can analyse the two groups separately. In group 1 we have that the joint distribution of $Y_1, \dots, Y_5, \lambda_1$ is given by

$$\begin{aligned} f(y_1, \dots, y_5, \lambda_1) &= f(y_1, \dots, y_5 | \lambda_1) f(\lambda_1) = \left(\prod_{i=1}^5 \frac{\exp(-\lambda_1) \lambda_1^{y_i}}{y_i!} \right) \cdot \frac{1}{\Gamma(\alpha)} \lambda_1^{\alpha-1} \beta^\alpha \exp(-\beta \lambda_1) \\ &= \frac{1}{\prod_{i=1}^5 y_i!} \cdot \frac{1}{\Gamma(\alpha)} \lambda_1^{\alpha + \sum_{i=1}^5 y_i} \beta^\alpha \exp(-(\beta + 5) \lambda_1) \propto \lambda_1^{\alpha + \sum_{i=1}^5 y_i} \exp(-(\beta + 5) \lambda_1) \end{aligned}$$

The posterior distribution of λ_1 given the data from group 1 is

$$\begin{aligned} f(\lambda_1 | y_1, \dots, y_5) &= \frac{f(y_1, \dots, y_5, \lambda_1)}{\int_{\lambda} f(y_1, \dots, y_5, \lambda) d\lambda} \propto f(y_1, \dots, y_5, \lambda_1) \\ &\propto \lambda_1^{\alpha + \sum_{i=1}^5 y_i} \exp(-(\beta + 5) \lambda_1) \end{aligned}$$

Comparing this to the density of the Gamma distribution we obtain that

$$\lambda_1 | Y_1, \dots, Y_5 \sim \text{Gamma} \left(\alpha + \sum_{i=1}^5 y_i, \beta + 5 \right),$$

and similarly

$$\lambda_2 | Y_6, \dots, Y_{10} \sim \text{Gamma} \left(\alpha + \sum_{i=6}^{10} y_i, \beta + 5 \right).$$

Thus, when only using the labelled data, we do not have to resort to Monte Carlo methods for finding the posterior distribution.

This however is not the case any more once we also want to include the unlabelled data. The conditional density of $Y_i | \lambda_1, \lambda_2$ for an unlabelled observation ($i > 10$) is

$$f(y_i | \lambda_1, \lambda_2) = \frac{1}{2} \frac{\exp(-\lambda_1) \lambda_1^{y_i}}{y_i!} + \frac{1}{2} \frac{\exp(-\lambda_2) \lambda_2^{y_i}}{y_i!}$$

The posterior density for the entire sample (using both labelled and unlabelled data) is

$$\begin{aligned} f(\lambda_1, \lambda_2 | y_1, \dots, y_{20}) &\propto \underbrace{f(\lambda_1) f(y_1, \dots, y_5 | \lambda_1)}_{\propto f(\lambda_1 | y_1, \dots, y_5)} \underbrace{f(\lambda_2) f(y_6, \dots, y_{10} | \lambda_2)}_{\propto f(\lambda_2 | y_6, \dots, y_{10})} \cdot \underbrace{f(y_{11}, \dots, y_{20} | \lambda_1, \lambda_2)}_{= \prod_{i=11}^{20} f(y_i | \lambda_1, \lambda_2)} \\ &\propto f(\lambda_1 | y_1, \dots, y_5) f(\lambda_2 | y_6, \dots, y_{10}) \prod_{i=11}^{20} f(y_i | \lambda_1, \lambda_2) \end{aligned}$$

This suggests using importance sampling with the product of the distributions of $\lambda_1 | Y_1, \dots, Y_5$ and $\lambda_2 | Y_6, \dots, Y_{10}$ as instrumental distributions, i. e. use

$$g(\lambda_1, \lambda_2) = f(\lambda_1 | y_1, \dots, y_5) f(\lambda_2 | y_6, \dots, y_{10}).$$

The target distribution is $f(\lambda_1, \lambda_2 | y_1, \dots, y_{20})$, thus the weights are

$$\begin{aligned}
w(\lambda_1, \lambda_2) &= \frac{f(\lambda_1, \lambda_2 | y_1, \dots, y_{20})}{g(\lambda_1, \lambda_2)} \\
&\propto \frac{f(\lambda_1 | y_1, \dots, y_5) f(\lambda_2 | y_6, \dots, y_{10}) \prod_{i=11}^{20} f(y_i | \lambda_1, \lambda_2)}{f(\lambda_1 | y_1, \dots, y_5) f(\lambda_2 | y_6, \dots, y_{10})} \\
&= \prod_{i=11}^{20} f(y_i | \lambda_1, \lambda_2) = \prod_{i=11}^{20} \left(\frac{1}{2} \frac{\exp(-\lambda_1) \lambda_1^{y_i}}{y_i!} + \frac{1}{2} \frac{\exp(-\lambda_2) \lambda_2^{y_i}}{y_i!} \right)
\end{aligned} \tag{3.6}$$

Thus we can draw a weighted sample of size n from the distribution of $f(\lambda_1, \lambda_2 | y_1, \dots, y_{20})$ by repeating the three steps below n times:

1. Draw $\lambda_1 \sim \text{Gamma}(\alpha + \sum_{i=1}^5 y_i, \beta + 5)$
2. Draw $\lambda_2 \sim \text{Gamma}(\alpha + \sum_{i=6}^{10} y_i, \beta + 5)$
3. Compute the weight $w(\lambda_1, \lambda_2)$ using equation (3.6).

From a simulation with $n = 50,000$ I obtained 4.4604 as posterior mean of λ_1 and 14.5294 as posterior mean of λ_2 . The posterior densities are shown in figure 3.7. ◀

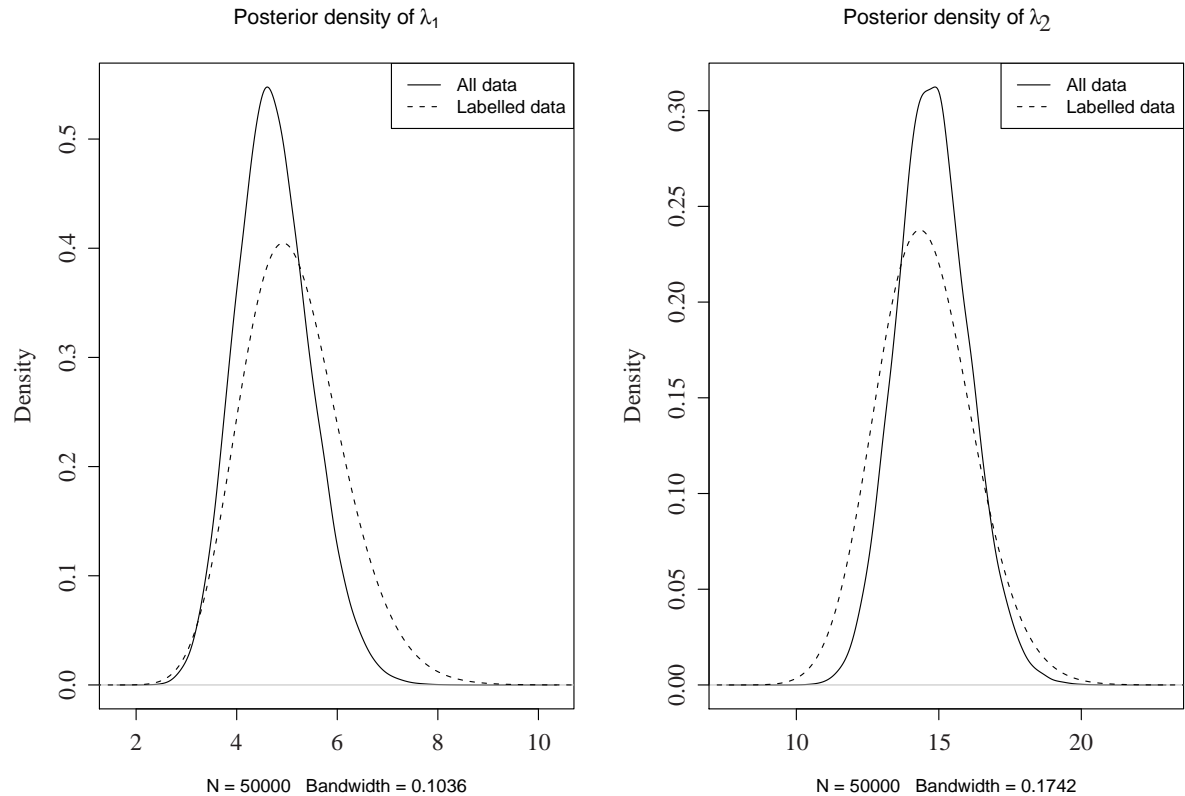


Figure 3.7. Posterior distributions of λ_1 and λ_2 in example 3.6. The dashed line is the posterior density obtained only from the labelled data.

