Exercise 1. *Monte Carlo integration and importance sampling*

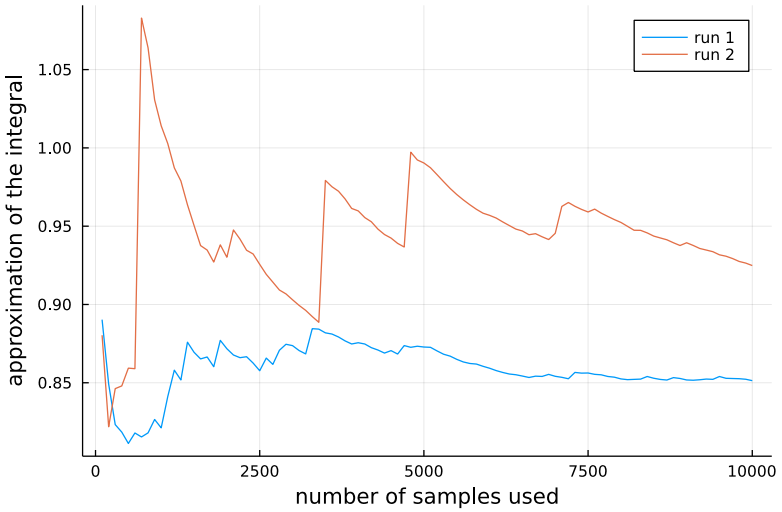
A standard Cauchy distribution has the density function (pdf)

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

with $x \in \mathbb{R}$. A friend would like to verify that $\int p(x)dx = 1$ but doesn't quite know how to solve the integral analytically. They thus use importance sampling and approximate the integral as

$$\int p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)} \qquad x_i \sim q \tag{2}$$

where q is the density of the auxiliary/importance distribution. Your friend chooses a standard normal density for q and produces the following figure:



The figure shows two independent runs. In each run, your friend computes the approximation with different sample sizes by subsequently including more and more x_i in the approximation, so that, for example, the approximation with $n = 2000$ shares the first 1000 samples with the approximation that uses $n = 1000$.

Your friend is puzzled that the two runs give rather different results (which are not equal to one), and also that within each run, the estimate very much depends on the sample size. Explain these findings, both mathematically and intuitively.

Solution. While the estimate $\hat{I}_n$

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)} \tag{S.1}$$

is unbiased by construction, we have to check whether its second moment is finite. Otherwise, we have an invalid estimator that behaves erratically in practice. The ratio $w(x)$ between $p(x)$

and $q(x)$ equals

$$w(x) = \frac{p(x)}{q(x)} \tag{S.2}$$

$$= \frac{\frac{1}{\pi}\frac{1}{1+x^2}}{\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)} \tag{S.3}$$

which can be simplified to

$$w(x) = \frac{\sqrt{2\pi}\exp(x^2/2)}{\pi(1+x^2)}. \tag{S.4}$$

The second moment of $w(x)$ under $q(x)$ thus is

$$\mathbb{E}_{q(x)}\left[w(x)^2\right] = \int_{-\infty}^{\infty} \frac{2\pi}{\pi^2}\frac{\exp(x^2)}{(1+x^2)^2} q(x)dx \tag{S.5}$$

$$= \int_{-\infty}^{\infty} \frac{2\pi}{\pi^2}\frac{\exp(x^2)}{(1+x^2)^2}\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)dx \tag{S.6}$$

$$\propto \int_{-\infty}^{\infty} \frac{\exp(x^2/2)}{(1+x^2)^2}dx \tag{S.7}$$

The exponential function grows more quickly than any polynomial so that the integral becomes arbitrarily large. Hence, the second moment (and the variance) of $\hat{I}_n$ is unbounded, which explains the erratic behaviour of the curves in the plot.

A less formal but quicker way to see that, for this problem, a standard normal is a poor choice of an importance distribution is to note that its density decays more quickly than the Cauchy pdf in (1), which means that the standard normal pdf is "small" when the Cauchy pdf is still "large" (see Figure 1). This leads to large variance of the estimate. The overall conclusion is that the integral $\int p(x)dx$ should not be approximated with importance sampling with a Gaussian importance distribution.

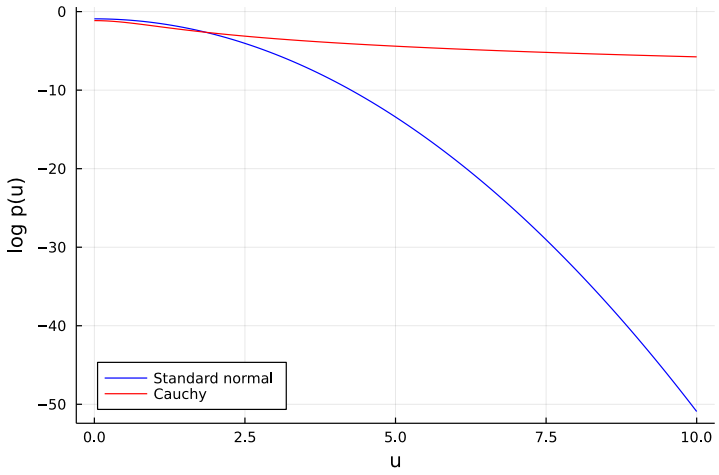


Figure 1: Exercise 1. Comparison of the log pdf of a standard normal (blue) and the Cauchy random variable (red) for positive inputs. The Cauchy pdf has much heavier tails than a Gaussian so that the Gaussian pdf is already "small" when the Cauchy pdf is still "large".

Exercise 2. ***Inverse transform sampling for*** $\mathcal{B}(x; 2, 1)$

We here use inverse transform sampling to sample from $\mathcal{B}(x; 2, 1)$.

(a) What is the density and cumulative distribution function (cdf) of $\mathcal{B}(x; 2, 1)$*?*

Solution. We have seen that $\mathcal{B}(x; \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}, x \in [0, 1]$ and hence

$$\mathcal{B}(x; 2, 1) \propto x, \quad x \in [0, 1] \tag{S.8}$$

We compute the normalising constant as

$$\begin{aligned} Z &= \int_0^1 x \mathrm{d}x && \text{(S.9)} \\ &= \frac{x^2}{2}\Big|_0^1 && \text{(S.10)} \\ &= \frac{1}{2} && \text{(S.11)} \end{aligned}$$

Hence the density is

$$\mathcal{B}(x; 2, 1) = 2x, \quad x \in [0, 1] \tag{S.12}$$

For the cdf, we compute a similar integral:

$$\begin{aligned} F_x(\alpha) &= \int_0^\alpha \mathcal{B}(x; 2, 1) \mathrm{d}x && \text{(S.13)} \\ &= \int_0^\alpha 2x \mathrm{d}x && \text{(S.14)} \\ &= 2\frac{x^2}{2}\Big|_0^\alpha && \text{(S.15)} \\ &= \alpha^2 && \text{(S.16)} \end{aligned}$$

with $\alpha \in [0, 1]$.

(b) Derive an explicit formula to generate samples $x \sim \mathcal{B}(x; 2, 1)$ *from samples* $u \sim \mathcal{U}(0, 1)$.

Solution. From the theory of inverse transform sampling, we know that

$$x = F_x^{-1}(u), \qquad u \sim \mathcal{U}(0, 1) \tag{S.17}$$

follows $\mathcal{B}(x; 2, 1)$ if F_x is the corresponding cdf.

All we need to do is thus to compute the inverse cdf (quantile function) F_x^{-1}. Setting $y = F_x(\alpha) = \alpha^2$, we solve for $\alpha \in [0, 1]$, which gives $\alpha = \sqrt{y}$. Hence

$$x = \sqrt{u}, \qquad u \sim \mathcal{U}(0, 1) \tag{S.18}$$

generates samples from $\mathcal{B}(x; 2, 1)$.

Exercise 3. ***Sampling from a restricted Boltzmann machine***

The restricted Boltzmann machine (RBM) is a model for binary variables $\mathbf{v} = (v_1, \ldots, v_n)^\top$ *and* $\mathbf{h} = (h_1, \ldots, h_m)^\top$ *which asserts that the joint distribution of* $(\mathbf{v}, \mathbf{h})$ *can be described by the probability mass function*

$$p(\mathbf{v}, \mathbf{h}) \propto \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right), \tag{3}$$

where $\mathbf{W}$ *is a* $n \times m$ *matrix, and* $\mathbf{a}$ *and* $\mathbf{b}$ *vectors of size* n *and* m*, respectively. Both the* v_i *and* h_i *take values in* $\{0, 1\}$*. The* v_i *are called the "visibles" variables since they are assumed to be observed while the* h_i *are the hidden variables since it is assumed that we cannot measure them.*

Use Gibbs sampling to generate samples from the marginal $p(\mathbf{v})$*,*

$$p(\mathbf{v}) = \frac{\sum_{\mathbf{h}} \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right)}{\sum_{\mathbf{h},\mathbf{v}} \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right)}, \tag{4}$$

for any given values of $\mathbf{W}$*,* $\mathbf{a}$*, and* $\mathbf{b}$*.*

Hint: *You may use that*

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{m} p(h_i|\mathbf{v}), \qquad p(h_i = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-\sum_j v_j W_{ji} - b_i\right)}, \tag{5}$$

$$p(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{n} p(v_i|\mathbf{h}), \qquad p(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp\left(-\sum_j W_{ij} h_j - a_i\right)}. \tag{6}$$

Solution. In order to generate samples $\mathbf{v}^{(k)}$ from $p(\mathbf{v})$ we generate samples $(\mathbf{v}^{(k)}, \mathbf{h}^{(k)})$ from $p(\mathbf{v}, \mathbf{h})$ and then ignore the $\mathbf{h}^{(k)}$.

Gibbs sampling is a MCMC method to produce a sequence of samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$ that follow a pdf/pmf $p(\mathbf{x})$ (if the chain is run long enough). Assuming that $\mathbf{x}$ is d-dimensional, we generate the next sample $\mathbf{x}^{(k+1)}$ in the sequence from the previous sample $\mathbf{x}^{(k)}$ by:

1. picking (randomly) an index $i \in \{1, \ldots, d\}$
2. sampling $x_i^{(k+1)}$ from $p(x_i \mid \mathbf{x}_{\setminus i}^{(k)})$ where $\mathbf{x}_{\setminus i}^{(k)}$ is vector $\mathbf{x}$ with x_i removed, i.e. $\mathbf{x}_{\setminus i}^{(k)} = (x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_{i+1}^{(k)}, \ldots, x_d^{(k)})$
3. setting $\mathbf{x}^{(k+1)} = (x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_d^{(k)})$.

For the RBM, the tuple $(\mathbf{h}, \mathbf{v})$ corresponds to $\mathbf{x}$ so that a x_i in the above steps can either be a hidden variable or a visible. Hence

$$p(x_i \mid \mathbf{x}_{\setminus i}) = \begin{cases} p(h_i \mid \mathbf{h}_{\setminus i}, \mathbf{v}) & \text{if } x_i \text{ is a hidden variable } h_i \\ p(v_i \mid \mathbf{v}_{\setminus i}, \mathbf{h}) & \text{if } x_i \text{ is a visible variable } v_i \end{cases} \tag{S.19}$$

($\mathbf{h}_{\setminus i}$ denotes the vector $\mathbf{h}$ with element h_i removed, and equivalently for $\mathbf{v}_{\setminus i}$)

To compute the conditionals on the right hand side, we use the hint:

$$p(\mathbf{h}|\mathbf{v}) = \prod_{i=1}^{m} p(h_i|\mathbf{v}), \qquad p(h_i = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-\sum_j v_j W_{ji} - b_i\right)}, \tag{S.20}$$

$$p(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{n} p(v_i|\mathbf{h}), \qquad p(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp\left(-\sum_j W_{ij} h_j - a_i\right)}. \tag{S.21}$$

Given the independencies between the hiddens given the visibles and vice versa, we have

$$p(h_i \mid \mathbf{h}_{\setminus i}, \mathbf{v}) = p(h_i \mid \mathbf{v}) \qquad\qquad p(v_i \mid \mathbf{v}_{\setminus i}, \mathbf{h}) = p(v_i \mid \mathbf{h}) \tag{S.22}$$

so that the expressions for $p(h_i = 1|\mathbf{v})$ and $p(v_i = 1|\mathbf{h})$ allow us to implement the Gibbs sampler.

Given the independencies, it makes further sense to sample the $\mathbf{h}$ and $\mathbf{v}$ variables in blocks: first we sample all the h_i given $\mathbf{v}$, and then all the v_i given the $\mathbf{h}$ (or vice versa). This is known as block Gibbs sampling.

In summary, given a sample $(\mathbf{h}^{(k)}, \mathbf{v}^{(k)})$, we thus generate the next sample $(\mathbf{h}^{(k+1)}, \mathbf{v}^{(k+1)})$ in the sequence as follows:

- For all h_i, $i = 1, \ldots, m$:
  - compute $p_i^h = p(h_i = 1|\mathbf{v}^{(k)})$
  - sample u_i from a uniform distribution on $[0, 1]$ and set $h_i^{(k+1)}$ to 1 if $u_i \leq p_i^h$.
- For all v_i, $i = 1, \ldots, n$:
  - compute $p_i^v = p(v_i = 1|\mathbf{h}^{(k+1)})$
  - sample u_i from a uniform distribution on $[0, 1]$ and set $v_i^{(k+1)}$ to 1 if $u_i \leq p_i^v$.

As final step, after sampling S pairs $(\mathbf{h}^{(k)}, \mathbf{v}^{(k)})$, $k = 1, \ldots, S$, the set of visibles $\mathbf{v}^{(k)}$ form samples from the marginal $p(\mathbf{v})$.