# Solutions for Exercise Round 7

## Exercise 1. (Optimal importance distributions)

Recall that in order to compute $\int f(x)q(x)\mathrm{d}x = \int f(x)\frac{q(x)}{\pi(x)}\pi(x)\mathrm{d}x$ you can use the empirical approximation $\int f(x)\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})\frac{q(x^{(i)})}{\pi(x^{(i)})}$, where the $x^{(i)}$'s are i.i.d. sampled from $\pi$.

(a) Suppose that $f(x) > 0$ for all $x$. Find the distribution $\pi$ that minimises the variance of $\frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})\frac{q(x^{(i)})}{\pi(x^{(i)})}$.

*Proof.* We have

$$\begin{aligned}\mathbb{V}\left[\frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})\frac{q(x^{(i)})}{\pi(x^{(i)})}\right] &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}\left[f(x^{(i)})\frac{q(x^{(i)})}{\pi(x^{(i)})}\right] \\ &= \frac{N}{N^2}\mathbb{V}_{\pi}\left[f(x)\frac{q(x)}{\pi(x)}\right]\end{aligned}$$

The first equality being by independence and the second one by the fact that they are all samples from $\pi$. So that the variance of the estimate is minimised when the variance of the single term $\mathbb{E}\left[f(x)\frac{q(x)}{\pi(x)}\right]$ is. Now, the variance is a non-negative quantity, so that if we find something so that it's 0, we have found the optimal $\pi$. Taking $\pi(x) \propto f(x)q(x)$, that is $\pi(x) = \frac{f(x)q(x)}{\int f(x)q(x)\mathrm{d}x}$ will make the term inside the variance constant, so of null variance. □

(b) Derive the optimal $\pi$ in the general case when $f$ can be positive or negative.

*Proof.* We now need to take $\pi(x) \propto |f(x)|q(x)$. This can be seen by splitting $f$ into its positive and negative parts $f^+ = \frac{f+|f|}{2}$ and negative part $f^- = \frac{|f|-f}{2}$, with $f = f^+ - f^-$ and repeating the analysis above. The critical argument is that changin the value of $\pi$ on the domain where $f$ is negative does not affect the variance of the positive part and vice versa. □

(c) Is the goal of designing such an importance density realistic? What would you do in practice?

*Proof.* Computing $\int |f(x)|q(x)\mathrm{d}x$ appearing in the expression of $\pi$ is generally as complicated as computing the original problem $\int f(x)q(x)\mathrm{d}x$ for which we resorted to importance sampling in the first place... Instead this gives us a rule of thumb for designing $\pi$: we need to take it to be as close as possible to the target that we want to integrate. There are many solutions that you will see in the next lectures: in particular using the Gaussian approximations given by extended and unscented Kalman filtering methods is a typical go-to solution (also think Laplace estimates, or neural nets, etc). □

## Exercise 2. (Rejection Sampling for Beta distribution)

See the notebook.

## Exercise 3. (Rejection Sampling Failure)

Consider the model

$$\begin{aligned} p(x) &= \mathrm{N}(x \mid 0, 1), \\ p(y \mid x) &= \mathrm{N}(y \mid x, 1). \end{aligned} \tag{1}$$

(a) Compute the posterior distribution $p(x \mid y)$.

*Proof.* The posterior distribution can be computed using, for example, the Kalman filter update and it is $p(x \mid y) = \mathrm{N}(x \mid y/2, 1/2)$. □

(b) Let us now assume that we use a distribution $\pi(x) = \mathrm{N}(x \mid 0, v)$ with some variance $v$ as the proposal distribution for rejection sampling. Show that there exists no bound $M$ such that $p(x \mid y)/\pi(x) \le M$. *Hint:* What is the maximum of $-(x - y/2)^2 + \frac{1}{2v}x^2$ over $x$ and $y$?

*Proof.* The required bound (which should be less than a constant $M$) now has the form

$$\frac{p(x \mid y)}{\pi(x)} = \frac{\mathrm{N}(x \mid y/2, 1/2)}{\mathrm{N}(x \mid 0, v)} = \frac{\sqrt{2\pi\, v}}{\sqrt{\pi}} \exp\left(-(x - y/2)^2 + \frac{1}{2v}x^2\right). \tag{2}$$

The maximum of this w.r.t. $y$ is obtained when $y/2 = x$ which gives

$$\frac{\sqrt{2\pi\, v}}{\sqrt{\pi}} \exp\left(\frac{1}{2v}x^2\right), \tag{3}$$

which is unbounded in $x$. □

(c) What does this result imply on the weights of importance sampling if we use $\pi(x)$ as the importance distribution?

*Proof.* The result implies that the importance weights will also be unbounded which further implies that the variance of the weights can become large. $\square$