MATS442 Stochastic simulation — Problems 2, 24.1.2020

1. Suppose $F_1(x), F_2(x), \ldots, F_n(x)$ are continuous and strictly increasing c.d.f.s on $\mathbb{R}$.

(a) Suppose $\alpha \in (0,1)$. Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$, and let $U \sim U(0,1)$ be independent of Y_1 and Y_2. Define X as follows

$$X := \begin{cases} Y_1, & \text{if } U \le \alpha \\ Y_2, & \text{if } U > \alpha. \end{cases}$$

Determine the c.d.f. of X.

(b) Suppose $w_1, \ldots, w_n \ge 0$ with $\sum_i w_i = 1$. Describe an algorithm to simulate from c.d.f.

$$F(x) = \sum_{i=1}^{n} w_i F_i(x),$$

using random variables $Y_1 \sim F_1, \ldots, Y_n \sim F_n$ and an independent $U \sim U(0,1)$.

2. Suppose you have a sample $X \sim p$, where p is a density on $\mathbb{R}$, and $a, b \in \mathbb{R}$ are constants with $a \neq 0$. Find a method to transform X to $Y \sim q$, with the density

$$q(x) = \frac{1}{|a|} p\Big(\frac{x-b}{a}\Big).$$

3. Suppose we are interested in simulating a distribution with density

$$p(x) = \begin{cases} \sqrt{\frac{2}{\pi}} \exp\Big(-\frac{1}{2}x^2\Big), & x \ge 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) How can you use rejection sampling to simulate $X \sim p$ with proposal density $N(0,1)$?

(b) How can you transform $Z \sim N(0,1)$ directly to $X \sim p$?

4.C Suppose that we are interested in the probability

$$\mathbb{P}(Z \ge 6), \qquad \text{where} \qquad Z \sim N(0,1).$$

Try to estimate this ‘rare event probability’ by using $N = 1000000$ samples with

(a) Classical Monte Carlo. (Hint: You can draw standard normal random variables with Julia function `randn`)

(b) Importance sampling with importance distribution

$$q(x) = \begin{cases} \sqrt{\frac{2}{\pi}} \exp\Big(-\frac{1}{2}(x-6)^2\Big), & x \ge 6 \\ 0, & \text{otherwise.} \end{cases}$$

(Hint: Problems 2 and 3.)

Repeat the tests 100 times and compare the results. Which method seems more reliable?

5.C Consider the rare-event simulation example in the lectures, but a bit more generally.

- Suppose $p(x)$ is the normal density with $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$.
- Suppose $q(y) := r \exp\big(-r(y - x_0)\big)\mathbf{1}(y \geq x_0)$ is a shifted exponential.
- We want to estimate $\int_{x_0}^{\infty} p(y)\mathrm{d}y$ with $x_0 > \mu$.

Write a R function which takes μ, σ, x_0 and number of samples n as arguments, and returns the corresponding importance sampling estimate.

- Determine r so that $(\log p)'(x_0) = (\log q)'(x_0)$.
- Use only *uniform* random variables (`rand()`) in your function.

6.C Suppose we are interested in Bayesian logistic regression based on the model $Y_i \sim \mathrm{Bernoulli}(\pi_i)$ where $\log(\frac{\pi_i}{1-\pi_i}) = \alpha + \beta X_i$, with a Gaussian prior. That is, consider a bivariate density

$$p(\alpha, \beta) \propto \exp\Big(-\frac{1}{8}(\alpha^2 + \beta^2)\Big)\prod_{i=1}^{N} \pi_i^{\mathbf{1}(y_i=1)}(1-\pi_i)^{\mathbf{1}(y_i=0)} \quad \text{with} \quad \pi_i = \mathrm{logit}^{-1}(\alpha + \beta x_i),$$

and where $\mathrm{logit}^{-1}(x) = e^x/(1+e^x)$. The following code generates "data" $(x_1, y_1), \ldots, (x_{100}, y_{100})$ using the model with $\alpha = -2$ and $\beta = 2$, and defines a function returning values $-\log p_u(\theta)$ where $\theta = (\alpha, \beta)^T$:

```
using Random; Random.seed!(123)
N = 100; alpha = -2; beta = 2; x = sort(rand(N)*3)
plogis(x) = exp(x)/(1 + exp(x)) # Inverse logit
pi_ = plogis.(alpha .+ beta*x); y = rand(N).<pi_
function nlog_p(theta)
  a = theta[1]; b = theta[2]; L = 0.125*(a^2+b^2)
  for k = 1:length(y)
     L -= y[k] ? log(plogis(a + b*x[k])) : log(1.0-plogis(a + b*x[k]))
  end
  L
end
```

The *Laplace approximation* of a density p is based on a (second-order) Taylor expansion of $\log p_u$ around its mode:

```
using Optim # Install by: Using Pkg; Pkg.add("Optim")
func = TwiceDifferentiable(nlog_p, zeros(2))
o = optimize(func, zeros(2), NelderMead()); m = o.minimizer
S = inv(Optim.hessian!(func, m)); S = (S+S')/2
```

The Laplace approximation $q \approx p$ is then $q(\alpha, \beta) = N\big((\alpha, \beta); m, S\big)$.

(i) Use q as a proposal for self-normalised importance sampling, with 10000 samples, and calculate an estimate of

$$\mathbb{P}\big(\alpha \in [-2 \pm 0.5], \beta \in [2 \pm 0.5] \mid y_1, \ldots, y_{40}\big) = \int_{-2.5}^{-1.5}\Big[\int_{1.5}^{2.5} p(\alpha, \beta)\mathrm{d}\beta\Big]\mathrm{d}\alpha.$$

(ii) Construct a 95% confidence interval for your estimate.

(iii) Calculate also a Monte Carlo estimate of $\int_{-2.5}^{-1.5}\big[\int_{1.5}^{2.5} q(\alpha, \beta)\mathrm{d}\beta\big]\mathrm{d}\alpha$.[1]

(iv) Inspect the weights (e.g. looking at the histogram and using 3-dimensional `scatter`; how do the weights look like near `m`?)

1. This may be regarded as a direct approximation of the integral of interest, relying on $q \approx p$...