MA 589 — Computational Statistics

Project 4

(Due: Tuesday, 11/8/16)

1. One simple way of sampling from a standard normal distribution is the *polar method*. It exploits the fact that if $X, Y \overset{\text{iid}}{\sim} N(0,1)$ then $R = \sqrt{X^2 + Y^2} \sim F$ with cdf[1] $F(r) = 1 - \exp\{-r^2/2\}$ and $\theta = \arctan(Y/X) \sim U(0, 2\pi)$. Thus, if we sample R and θ, we can set $X = R\cos\theta$ and $Y = R\sin\theta$ to obtain a pair of independent standard normal deviates; this is known as the *Box-Muller* method.

   (a) Write a function that implements Box and Muller's polar method: use the inverse cdf method to sample R, then sample θ, and then return X and Y as above[2].

   (b) Marsaglia proposed a more efficient method[3]:

   Step 1. Sample (X, Y) uniformly from the unit circle

   Step 2. Set $S = X^2 + Y^2$ and return $X\sqrt{-2\log S/S}$ and $Y\sqrt{-2\log S/S}$ as a pair of independent normal deviates[4].

   Write a function that implements Marsaglia's polar method.

   (c) Now verify your functions: sample 1,000 normal deviates using Box-Muller and Marsaglia methods and make a QQplot of the two sample sets. Describe how the two sample sets agree (you might also want to use `qqnorm` to compare directly to standard normal quantiles.)

2. You want to sample a random variable X that has the following cdf:

$$\mathbb{P}(X \le x) \doteq \mathbb{P}(Y \le x \,|\, Y > 0),$$

   where $Y \sim N(-3, 1)$. In this case X is said to have a *truncated* normal distribution. Note that the density f of X is proportional to the exponential of a quadratic term:

$$f(x) \propto \exp\left\{ -\frac{(x+3)^2}{2} \right\} I(x > 0).$$

   (a) Write a function `rs.supp` to sample from X using rejection "by support": sample from Y until $Y > 0$ and take that as a sample from X. Explain why your function works; for instance, what is your envelope function?

   (b) Write a function `rs.g` to rejection sample from X using $g(x) \propto \exp\{-3x\}$ as an envelope. Specify how you sample according to g, and what is your criterion for accepting a sample.

[1] In case you're curious: R follows a standard *Rayleigh* distribution, which is a χ (not a χ^2!) distribution with 2 degrees of freedom.

[2] Can you come up with a vectorized version?

[3] Mostly for avoiding the "expensive" trigonometric functions.

[4] Can you show that $S \sim U(0,1)$? This fact can give you one way of sampling uniformly from the unit circle (other than rejection sampling.)

(c) How much more likely are you to accept a sample from `rs.g` than from `rs.supp`? It is OK to estimate—or even confirm—your answer by simulation: run the rejection samplers many times and compare the proportion of accepted samples in each function.

3. Given the function

$$g(\mathbf{x}) = \left| \sin\left(2\pi x_1 \sum_{i=1}^{100} x_i\right) \right| \left\{ \cos\left(3\pi x_2 \sum_{i=1}^{100} x_i^2\right) \right\}^2,$$

defined for any $\mathbf{x} = (x_1, \ldots, x_{100}) \in [0,1]^{100}$, you want to evaluate the integral of g in its domain, that is, you want

$$I = \int_{[0,1]^{100}} g(\mathbf{x}) d\mathbf{x}.$$

(a) Obtain a Monte Carlo estimate of I in the form of a 95% confidence interval. Your interval should be no larger than 0.015.

(b) $*$ After looking at g more carefully, you realize that I should be somehow close to

$$I^* = \int_0^1 \int_0^1 h(x_1, x_2) dx_1 dx_2,$$

where $h(x_1, x_2) = |\sin(100\pi x_1)| \cos(100\pi x_2)^2$.

Give an explanation for why that would be the case. The true value[5] of I^* is $1/\pi$; now use $h(x_1, x_2)$ as a *control variate* to obtain a new estimate for I based on 10,000 samples. Report the standard error of your estimate and compare it with the MC estimate above. Has it improved? Why or why not?

4. A BU student leaves a party close to Agganis arena and two groups of friends invite him to the Paradise Club (PC) and to the BU Pub (BP). Since he is undecided, he decides to play the following game[6]: he divides the path between PC and BP in *twenty* segments and labels the positions from 0, starting at PC, to 20, ending at BP. He's currently at position 1. At each round of the game he flips a *fair* coin; if it's tails he walks west and so decrements his current position; if it's heads, he walks east and increments his position. The game ends if he reaches either PC or BP.

The following `R` function simulates his *random walk*; parameter `p` is the probability of moving east.

```
rwalk <- function (p) {
  j <- 1 # start
  walk <- c() # store movements
  repeat {
```

[5] Can you show this? Wolfram Alpha, for instance, times out on I^*.

[6] No, he's not drunk; he is just overly enthusiastic about randomness.

```
    j <- j + (2 * rbinom(1, 1, p) - 1) # move
    walk <- append(walk, j)
    if (j == 0 || j == 20) return(walk)
  }
}
```

(a) What is the probability that he ends up at BP? Run 100,000 simulations and obtain a Monte Carlo estimate. Report a 95% confidence interval.

(b) What is the shape of the distribution of the *length* of a walk? Plot a histogram, and provide an estimate for the probability that the student takes more than 200 "steps"[7].

(c) What is the shape of the distribution of the length of a walk *given* that the student reached BP? Plot a histogram and comment on the shape when compared to the distribution in the previous item. (*Hint*: filter your MC samples to only consider walks where the last position is 20.)

(d) The Dugout is at position 18. Estimate the expected number of times that the student will be in front of this pub.

Even with a large number of samples you still feel that the confidence interval for the probability that the student goes to BP is too large. Now you wish to devise an importance sampling (IS) scheme to make it smaller, and so you adopt a (slightly, but "importantly") loaded coin with probability of heads `p` = 0.55.

(e) Write a function that computes the importance sampling weight of a random walk.

(f) Re-estimate the probability of ending up at BP. What is the ratio of the MC standard deviation to the IS standard deviation?

[7] That is, that he spends the night playing the game...