MA 589 — Computational Statistics

Project 4

(Due: Friday, 11/9/18)

1. Write quantile functions and then use the inverse cdf method to write samplers for the following distributions:

   (a) *Gumbel* distribution: if $X \sim \mathsf{Gumbel}(\mu, \sigma)$, $\sigma > 0$, then X has cdf

   $$\mathbb{P}(X \le x) = \exp\left\{ -\exp\left\{ -\frac{x-\mu}{\sigma} \right\} \right\}.$$

   Draw 1,000 samples from a $\mathsf{Gumbel}(-\gamma, 1)$, where $\gamma = -\psi(1)$, with ψ the digamma function, is the Euler-Mascheroni constant. Now, using the quantile function you wrote, make a QQplot to verify that your sampler is working properly.

   (b) *Truncated* exponential distribution: if $X \sim \mathsf{TruncExp}(\lambda, a, b)$, $\lambda > 0$, $0 \le a < b \le \infty$, then X has cdf

   $$\mathbb{P}(X \le x) = \mathbb{P}(Y \le x \,|\, a \le Y < b) I(x \ge a),$$

   where $Y \sim \mathsf{Exp}(\lambda)$. Draw 1,000 samples from a $\mathsf{TruncExp}(1, 0.5, 1.5)$ and check that your sampler is working by making a QQplot. Now repeat using samples from a $\mathsf{TruncExp}(2, 1.5, \infty)$.

2. If X follows a K-*categorical* distribution with weights $w = (w_1, \ldots, w_K)$ then $\mathbb{P}(X = i) \propto w_i$ for $i = 1, \ldots, K$. Sampling from a categorical distribution is similar to sampling from a Bernoulli: we normalize the weights to probabilities $p_i := w_i / \sum_{k=1}^K w_k$, compute cumulative probabilities $c_i := \sum_{k=1}^i p_k$, then draw from a standard uniform and just check to which cumulative probability stratum the sample belongs: if $U \sim U(0,1)$ and $c_{i-1} \le U < c_i$ then we take i as our sample from the categorical. Here is a function that draws `n` samples from a categorical distribution with weights `w`:

   ```
   rcat <- function (n, w) findInterval(runif(n), cumsum(w / sum(w))) + 1
   ```

   Now suppose you have *log* weights $l_i = \log(w_i)$ and still want to sample from a categorical with weights w but cannot retrieve the weights with $\exp(l_i)$ to avoid underflows and overflows.

   (a) Write a new sampler for the categorical that takes log weights as arguments. Remember: do not compute $\exp(l_i)$.

   (b) If $G_i \overset{\text{iid}}{\sim} \mathsf{Gumbel}(-\gamma, 1)$ for $i = 1, \ldots, K$, then $\arg\max_{i=1,\ldots,K}\{G_i + l_i\}$ follows a categorical distribution with weights w. This is known as the *Gumbel max trick*[1]; use it to build another sampler for the categorical using log weights. Note that, surprisingly, it does not require any normalization of the weights or log weights!

[1] Can you show why it works?

3. The family of exponential envelopes

$$e_z(x) := \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{z^2}{2} - zx\right\}$$

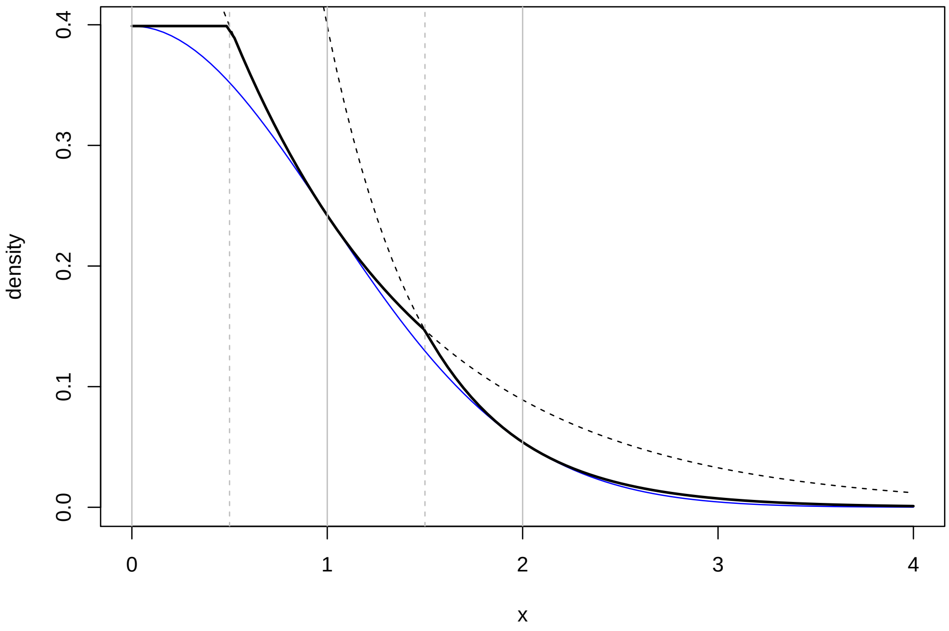
can be used to build an efficient rejection sampler for the standard normal.

(a) Show that $e_z(x) \geq \phi(x)$ for $x, z \in \mathbb{R}^+$, where ϕ is the standard normal density. Where is $e_z(x) = \phi(x)$?

(b) You decide to use a three-piece envelope for higher efficiency:

$$Mg(x) := e_0(x)I(0 \leq x < x_0^*) + e_1(x)I(x_0^* \leq x < x_1^*) + e_2(x)I(x \geq x_1^*),$$

where x_0^* is such that $e_0(x_0^*) = e_1(x_0^*)$ and x_1^* is such that $e_1(x_1^*) = e_2(x_1^*)$. In the figure below we can see the three pieces of the envelope; the blue curve is the target ϕ, the solid black curve is Mg.



Find x_0^* and x_1^* and write g as a *mixture* of truncated exponentials. Compute M, the mixture weights, and the truncation limits.

(c) Now write a rejection sampler based on this envelope for the standard normal. Remember that since the envelope only covers the positive part of the domain you still have to sample from the negative part. What is the efficiency of the sampler, that is, what is the expected proportion of rejections?

(d) Verify your sampler: draw 1,000 samples from it and make a QQplot.

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[2]: 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 {
    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"[3].

   (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?

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

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