HOMEWORK 2
COMPUTATIONAL METHODS FOR DATA SCIENCE
FALL SEMESTER 2023

1. **Generating Random Variables for Common Distributions.** *(24 points)* Assume we can generate a random variable U that follows Uniform$(0, 1)$. Suggest a method of generating random variables that follow each of the following distributions:

   (a) Normal(μ, σ^2).

   (b) Exponential(λ).

   (c) Poisson(λ).

   (d) Chi-Square(df=k).

   (e) $F_{k,m}$.

   (f) Binomial(n,p).

   (g) Negative Binomial(r, p).

   (h) Dirichlet(α_1,. . . ,α_k).

2. **Sampling Problem.** *(10 points)* Consider finding $\sigma^2 = E(X^2)$ where X has the density that is proportional to $q(x) = e^{-|x|^3/3}$.

   (a) Estimate σ^2 using importance sampling with standardized weights.

   (b) Repeat the estimation using rejection sampling.

3. π **Estimation** *(26 points)* Consider a disk D of radius 1 inscribed within a square of perimeter 8 centered at the origin. Then the ratio of the area of the disk to that of the square is $\pi/4$. Let f represent the uniform distribution on the square. Then for a sample of points $(X_i, Y_i) \sim f(x, y)$ for $i = 1, \ldots, n$, $\hat{\pi} = (4/n)\Sigma_{i=1}^{n} 1_{(X_i,Y_i)\in D}$ is an estimator of π, where 1_A is 1 if A is true, and 0 otherwise.
   Consider the strategy for estimating π. We start with $(x^{(0)}, y^{(0)}) = (0, 0)$. Thereafter, generate candidates as follows: First, generate both $\epsilon_x^{(0)}$ and $\epsilon_y^{(0)}$ follow Uniform$(-h, h)$. If $(x^{(i)} + \epsilon_x^{(i)}, y^{(i)} + \epsilon_y^{(i)})$ falls outside the square, regenerate $\epsilon_x^{(0)}$ and $\epsilon_y^{(0)}$ until the step taken remains within the square. Let $(X^{i+1}, Y^{i+1}) = (x^{(i)} + \epsilon_x^{(i)}, y^{(i)} + \epsilon_y^{(i)})$. Increment t. This generates a sample of points over the square. When $t = n$, stop and calculate $\hat{\pi}$ as given above.

   (a) Implement this method for $h = 1$ and $n = 20000$. Compute $\hat{\pi}$. What is the effect of increasing n? What is the effect of increasing and decreasing h?

   (b) Explain why this method is flawed. Using the same method to generate candidates, develop the correct approach by referring to the Metropolis-Hastings ratio. Prove that your sampling appraoch has a stationary distribution that is uniform on the square.

   (c) Implement your approach from Part (b) and caluclate $\hat{\pi}$. Experiment again with π and h. Comment on your result.

4. **Coal-Mining Disaster Analysis***(40 points)* In this problem, we consider a famous coal-mining disaster annual data between 1851 and 1962. The rate of accidents per year appears to decrease around 1900, so we consider a change-point model for these data. Let $j = 1$ in 1851,

and index each year therafter, so $j = 112$ in 1962. Let X_j be the number of accidents in year j, with $X_1, \ldots, X_\theta$ follows i.i.d. Poisson(λ_1) and $X_{\theta+1}, \ldots, X_{112}$ follows i.i.d. Poisson(λ_2). Thus the change-point occurs after the θth year in the series, where $\theta \in \{1, \ldots, 111\}$. This model has parameters θ, λ_1, and λ_2. For the standard setup, we assume a discrete uniform prior for θ on $\{1, 2, \ldots, 111\}$, and priors $\lambda_i | a_i \sim$ Gamma$(3, a_i)$, and $a_i \sim$ Gamma$(10, 10)$ independently for $i = 1, 2$.

(a) Estimate the posterior mean and provide a histogram each for θ, λ_1 and λ_2.

(b) Consider a modification that $\lambda_2 = \alpha\lambda_1$. Assume the same discrete uniform prior for θ and the same prior $\lambda_1 | a \sim$ Gamma$(3, a)$, $a \sim$ Gamma$(10, 10)$. In addition, assume $log\alpha \sim$ Uniform$(log1/8, log2)$. Estimate the posterior mean and provide a histogram each for θ, λ_1 and λ_2.

(c) Consider a modification that priors $\lambda_i | a_i \sim$ Gamma$(3, a_i)$, and $a_i \sim$ Uniform$(0, 100)$ independently for $i = 1, 2$ instead. Assume the same discrete uniform prior for θ. Estimate the posterior mean and provide a histogram each for θ, λ_1 and λ_2.

(d) Using the standard setup, derive the conditional distributions necessary to carry out Gibbs sampling for the change-point model.

(e) Implement the Gibbs sampler. Use a suite of convergence diagnostics to evaluate the convergence and mixing of your sampler.

(f) Construct density histogram and a table of summary statistics for the approximate posterior distributions of θ, λ_1 and λ_2.

B. **Bootstrapping Practice** *(Bonus 10 points)* Suppose $\theta = g(\mu)$, where g is a smooth function and μ is the mean of the distribution from which the data arise. Consider bootstrapping $R(X, F) = g(\bar{X}) - g(\mu)$.

(a) Show that $E^*(X^*) = \bar{x}$ and $var^*(X^*) = \hat{\mu^2}/n$, where $\hat{mu}_k$, where $\hat{\mu}_k = \Sigma_{i=1}^n (x_i - \bar{x})^k$.

(b) By Taylor Series Expansion, find $E^*(R(X^*, \hat{F}))$ and $var^*(R(X^*, \hat{F}))$. You can show the first two terms only of the expansion only.