STATS 406 F15: Lab 12
Final Review Questions

1. **R coding.**

   (a) Suppose we have a vector x of length n stored in memory, write a few lines of R code that uses vectorization to perform the following transformation $T(x_i)$ on each element of x:

$$T(u) = \begin{cases} -1 & \text{if } u < -1 \\ 3u + 1 & \text{if } -1 \leq u < 1 \\ 1 & \text{if } u \geq 1 \end{cases}$$

   (b) Suppose we have an integer vector:

```
y = c(1,2,3,5,1);
```

   What are the outputs of the following commands:

   - `y[c(1,2)]`
   - `y[y==1]`
   - `y[-c(1,3)]`
   - `y + c(1,2)`
   - `which.max(y)`
   - `as.list(y)[1]`

2. **Random number generation.**

   (a) Recall that the standard Cauchy distribution has PDF and CDF as follows:

$$f(x) = \frac{1}{\pi(1 + x^2)}$$
$$F(x) = \frac{1}{\pi} \arctan(x) + \frac{1}{2}$$

   If we are only allowed access to a Uniform $(0, 1)$ random number generator, how to sample from the standard Cauchy distribution?

   (b) Write a block of R code to sample from the Geometric(p) distribution, whose PMF is:

$$\mathbb{E}(X = k) = p(1 - p)^{k-1}$$

   for $k = 1, 2, 3, ...$

(c) Given the following CDF of a random variable supported on $[0, 1]$:

$$F(x) = \begin{cases} \sin(\pi x) & \text{if } 0 \le x \le \frac{1}{2} \\ 0 & \text{if } x < 0 \\ 1 & \text{if } x > \frac{1}{2} \end{cases}$$

Describe the steps of a rejection sampling to sample from $F(x)$. Notice here you are given a CDF.

3. **Monte-Carlo integration.**

(a) Describe the steps to compute

$$I := \int_1^{+\infty} \frac{1}{x^2} \mathrm{d}x$$

Using Monte-Carlo integration.

(b) Describe the steps to compute $\mathbb{E}[Y]$, where

$$Y = \begin{cases} X^3 & \text{if } X > 0 \\ 0 & \text{otherwise} \end{cases}$$

and $X \sim N(0, 1)$, using importance sampling with an exponential distribution as the auxiliary distribution. Choose a proper λ parameter for your exponential distribution.

(c) Suppose the random variable $X > 0$ follows a distribution with PDF as follows:

$$f_X(x) \propto \exp\left(-x^{3/2}\right)$$

How to use a proper form of importance sampling to compute $\mathbb{E}[X]$?

4. **Monte-Carlo performance evaluation and bootstrap.** Suppose we collect a sample $X := \{X_1, \ldots, X_n\}$ from $N(\mu, 1)$, where the true $\mu = 1$. Consider the following estimator for μ^2:

$$\widehat{\mu^2} = \left(\bar{X}\right)^2$$

(a) What is the mathematical definition of $\mathrm{MSE}(\widehat{\mu^2})$?

(b) If we know the true $\mu = 1$, how to compute $\mathrm{MSE}(\widehat{\mu^2})$?

(c) If we do not know the true μ, how to estimate $\mathrm{MSE}(\widehat{\mu^2})$?

5. **SQL.** The database contains the following three tables:

Table 1: Composers

ID	Composer	Era	Country
1	JS Bach	Baroque	Germany
2	WA Mozart	Classical	Austria
3	LV Beethoven	Classical, Romantic	Germany
4	FF Chopin	Romantic	Poland

Table 2: Works

Work ID	Title	Composer	Pianist
001	Well-Temptered Clavier	JS Bach	Glenn Gould
002	Goldberg Variations	JS Bach	Glenn Gould
003	Italian Concerto	JS Bach	Sviatoslav Richter
004	Complete Solo Piano Works	WA Mozart	Walter Gieseking
005	Piano Sonata No. 18	WA Mozart	Claudio Arrau
006	Piano Concerto No. 20	WA Mozart	Friedrich Gulda
007	Piano Sontat No. 12	LV Beethoven	Emil Gilels
008	Piano Sonata No. 57	LV Beethoven	Friedrich Gulda
009	Waltzes	FF Chopin	Claudio Arrau
010	Etudes	FF Chopin	Sviatoslav Richter

Table 3: Pianists

ID	Pianist	Country
1	Glenn Gould	Canada
2	Sviatoslav Richter	Soviet
3	Walter Gieseking	Germany
4	Claudio Arrau	Chile
5	Friedrich Gulda	Austria
6	Emil Gilels	Soviet

Write lines of SQL commands that complete each of the following tasks:

(a) Query all *pianists* from *Soviet*. Only report *pianist* and *country*.

(b) Query all *works* performed by *Glenn Gould*. Only report *title* and *composer*.

(c) Combine tables *Composers* and *Works* and query works played by European(including Soviet) pianists. Only report *title*, *era*, *composer* and *pianist*.

(d) Combine all three tables and query works composed by *Germany composers* and performed by *Soviet pianists*. Only report *title*, *composer* and *pianist*.

6. **XML.** Rewrite the following entry, transforming the attributes into children:

```
<book Title="The Return of The Native" Author="Thomas Hardy" Published year="1878" />
```

Consider the rewritten version: write an R command (assume the package "XML" is loaded) to query the content of the *Author* tag.

7. **Optimization.**

   (a) We want to minimize the following function:

   $$f(x, y) := x^2 + 4(y - 1)^2$$

   Start at the initial guess $(x, y) = (2, 3)$.

   i. Describe a gradient method to find the optimum. Calculate the first two iterations. Use a fixed step size of 0.5.

   ii. Describe a Newton's method to find the optimum. Calculate the first two iterations.

   (b) Consider a mixture of Poisson distributions. There are 3 Poisson distributions, with $\lambda = \lambda_1, \lambda_2, \lambda_3$ respectively, where λ_i's are mutually different. The mixing probabilities (π_1, π_2, π_3) are strictly positive and unknown. We have a set of collected data $X := \{X_1, \ldots, X_n\}$, and recall the PMF of Poisson distribution with rate λ:

   $$\mathbb{P}(X = k) = \frac{\lambda_k}{k!} e^{-\lambda}$$

   i. Write down the (incomplete) log-likelihood function.

   ii. Describe (without explicitly formulating any step) how to obtain the MLE for the parameters $\Theta := (\lambda_1, \lambda_2, \lambda_3, \pi_1, \pi_2, \pi_3)$ directly using optimization tools.

8. **EM algorithm.** Continuing Question 7b, describe the EM algorithm for estimating the set of parameters Θ.