Practise Problems for Computational Statistics Final Exam

Raazesh Sainudiin

Problem 1 If a RV X is equally likely to be either positive or negative, then the *Laplace*(λ) RV, with the rate parameter $\lambda > 0, \lambda \in \mathbb{R}$, may be used to model it. The density function for the *Laplace*(λ) RV given by $f(x;\lambda)$ is **proportional** to $\exp(-\lambda|x|)$, as follows:

$$f(x;\lambda) \propto e^{-\lambda|x|}, \quad where, \quad x \in \mathbb{R} \equiv (-\infty, \infty) \ .$$

We say x is proportional to y and write '$x \propto y$' when we mean $x = cy$ for some constant c. By '$|x|$' we mean the absolute value of x.

Problems marked by $\star$'s are for extra credit and more $\star$'s means more credits. The comments enclosed in square brackets ('[...]') at the end of each question is the problem handing-in protocol. Problems 1 through 8 make 10% of your final grade and there are 3 problems for extra credit.

$\star$. Assume the above proportionality and show step by step that the pdf of the *Laplace*(λ) RV X is

$$f(x;\lambda) = \frac{\lambda}{2} e^{-\lambda|x|}$$

1. Write a `Matlab function` that will take as **input** (1) the parameter $\lambda > 0$ and (2) a value $x \in \mathbb{R}$ and **output or return** $f(x;\lambda)$. [The M-file 'Laplacepdf.m' should start with '`function fx = Laplacepdf(x,lambda);`' and you **have** to comment the M-file as we did for the functions written in Lab 2 for full credit. Hand-in the M-file.]

$\star\star$. Show that the DF

$$F(x;\lambda) = \int_{-\infty}^{x} f(y;\lambda)dy = \frac{1}{2}\left(1 + sign(x)\left(1 - e^{-\lambda|x|}\right)\right), \quad where, \quad sign(x) = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0 \end{cases}$$

2. Write a `Matlab function` with **input** (1) the parameter $\lambda > 0$ and (2) a value $x \in \mathbb{R}$ and **output** $F(x;\lambda)$. [The M-file 'LaplaceCDF.m' should start with '`function Fx = LaplaceCDF(x,lambda);`' and be commented. Hand-in the M-file. You may need to use a vector multiplier '`.*`' instead of the scalar multiplier '`*`' in the function to pass x as a vector argument for 3. below.]

3. Plot $f(x;\lambda = 1)$, $f(x;\lambda = 0.1)$ and $f(x;\lambda = 10)$ on the same **figure** and explain in **words** the behavior of the probability as λ changes, in terms of the pdf. [Follow the plotting exercises in Lab 2 and recall the `hold` command and the various options to `plot`, including `Color` and `axes`. Choose appropriate ranges to capture the qualitative aspects of the plot. Submit the **figure** with the **words** below it (not the M-file for the plot).]

4. Plot $F(x;\lambda = 1)$, $F(x;\lambda = 0.1)$ and $F(x;\lambda = 10)$ on the same **figure** and explain in **words** the behavior of the probability as λ changes, in terms of the CDF. [same handing-in protocol as in 3. above.]

$\star\star\star$. Show that the inverse CDF for the *Laplace*(λ) RV X is

$$F^{-1}(u;\lambda) = -\frac{1}{\lambda}\ sign\left(u - \frac{1}{2}\right)\log\left(1 - 2\left|u - \frac{1}{2}\right|\right) \ , \ u \in [0,1]$$

5. Using the Inverse-CDF method (or Inversion Sampler) write a `Matlab function` that will take as **input** (1) the parameter $\lambda > 0$ and (2) a value $u \in [0,1]$ and **output** $F^{-1}(u;\lambda) \in \mathbb{R}$. [The M-file 'LaplaceInvCDF.m' should start with '`function x = LaplaceInvCDF(u,lambda);`' and be commented. Hand-in the M-file. Comments about '`.*`' made in 2. above may apply in this case too.]

6. Plot $F^{-1}(u;\lambda = 1)$, $F^{-1}(u;\lambda = 0.1)$ and $F^{-1}(u;\lambda = 10)$ on the same **figure** and interpret in **words** the behavior of the inverse CDF as λ changes. [same handing-in protocol as in 3. above. Note that the domain or 'x-axis' for this plot is $[0,1]$.]

7. Using the `rand` function in `Matlab` that generates samples from $Uniform[0,1]$ RV, generate $n = 10^3$ realizations of the $Laplace(\lambda = 1)$ RV X and store it in an array called $x1000$. [use the Mersenne twister algorithm with default seed by the command '`rand('twister', 5489);`', as in Lab 2 (Labwork 4), prior to generating samples from $Laplace(\lambda)$ RV via the function `LaplaceInvCDF(u,lambda);`' you wrote in 5. above. Hand-in the `Matlab` script (not necessarily an M-file) for creating the vector $x1000$.]

8. Using the `Matlab` function given below, plot the empirical CDF, as done in Lab 2 (LabWork 5), based on $x1000$ and the CDF $F(x;\lambda = 1)$ on the same **figure**. Are you convinced from the plot that you have simulated from the intended target $Laplace(\lambda = 1)$ RV ? Explain in **words**. [Turn in the **figure** with **words** below it. You need to call the function below or write your own empirical CDF plotter using the definition given in Lab 2 (LabWork 5; Equation (1)).]

```
function [x y] = LaplaceECDF(R,lambda);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% return the x and y values of empirical CDF
% based on R samples from the Laplace(lambda) RV X
%
% File Dates : Created  07/27/07
% Author(s)  : Raaz
%
% Call Syntax: [x y] = LaplaceECDF(R,lambda);
%              or LaplaceECDF(R,lambda);
%
% Input      : lambda = rate parameter,
%              R = number of samples from Laplace(lambda) RV
% Output     : [x y] & empirical CDF Plot
%
% Requires   : availability of the function LaplaceInvCDF(u,lambda)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
CDFPlotPoints=100; % large enough for a smooth CDF plot ?
x=[]; y=[];        % initialize x and y to null vectors
for i=1:1:R        % loop to append to x and y axis values of plot
x=[x LaplaceInvCDF(rand,lambda)]; % append samples from Laplace(lambda) RV to x
y=[y i/R];                   % append equi-increments of 1/R to y
end                % end of for loop
x=sort(x)          % sorting the sample values
x=[x(1) x x(R)];   % padding x for emp CDF to start at min(x) and end at max(x)
y=[0 y 1];         % padding y so emp CDF start at y=0 and end at y=1
stairs(x,y,'LineWidth',2) % stairs makes a stair-like plot with x and y pairs
hold                      % hold the stairs plot of emp CDF
x1=linspace(x(1),x(R),CDFPlotPoints); % get CDFPlotPoints many points in domain
plot(x1,LaplaceCDF(x1,lambda),'Color',[0.8 0 0]) %CDF plot of Laplace(lambda) RV
```

Problem 2 1. Recall the $Laplace(\lambda)$ RV, with the rate parameter $\lambda > 0, \lambda \in \mathbb{R}$ with density function

$$f(x;\lambda) = \frac{\lambda}{2} e^{-\lambda|x|} \ ,$$

DF

$$F(x;\lambda) = \int_{-\infty}^{x} f(y;\lambda)dy = \frac{1}{2}\left(1 + sign(x)\left(1 - e^{-\lambda|x|}\right)\right), \quad where, \quad sign(x) = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0 \end{cases},$$

and inverse DF

$$F^{-1}(u;\lambda) = -\frac{1}{\lambda}\ sign\left(u - \frac{1}{2}\right)\log\left(1 - 2\left|u - \frac{1}{2}\right|\right)\ ,\ u \in [0,1]\ .$$

You should be intimate with this function from the previous assignment. Let us use these RVs to make a toy model for pollen dispersal from a source at the origin along a line. In other words, we want a model that would simulate the 'pollen landing points along a line or transect of interest'. Note that the pollen may land on either side of the source that is located at the origin ($x = 0$) in our line of interest. When there is only one type of pollen being produced (by a plant) at the source, we can model the dispersal of the first pollen using the *Laplace*($\lambda = 1.0$) RV X_1. Let the RV X_i model the dispersal of the i-th pollen from the source. To specify the RV X_i, let us make further simplifying assumptions in our model of dispersal of n pollens. Under the **two assumptions** that (1) the dispersal of each pollen is **independent** of the dispersal of any other pollen and that (2) the dispersal of every pollen is **identical** in distribution to that of any other pollen, our model for the dispersal of n pollens from the source is given by the RVs:

$$X_1, X_2, \ldots, X_n \sim F(x;\lambda = 1.0)$$

(a) Write a `Matlab` **script** (use the `diary` command and edit it to make the final script with generous comments) that would simulate the dispersal of 10 consecutive pollens. The output of this simulation should be 10 real numbers that are stored in a vector named '`x10`'. Using **words** interpret the meaning of these numbers in terms of our model for pollen dispersal. You are required to initialize the fundamental sampler prior to this simulation using '`rand('twister',007)`' . [Submit the **script** file with **words** below it – if you use the functions in M-files you created during a previous lab or assignment then this M-file has to be included in your report for full credit].

(b) Reset the fundamental sampler using '`rand('twister',007)`' and simulate the dispersal of 100 pollens by modifying the above script and store the samples in a vector named '`x100`'. Report the five statistics; mean, median, sample standard deviation, minimum and maximum, for the 100 samples and their absolute values:

i. 100 samples stored in $x100$,

ii. $|x100|$, ie. the absolute values of the 100 samples in $x100$ ('`abs(x100)`').

using the built-in `Matlab` commands `mean`, `median`, `std`, `max`, and `min`, respectively. Interpret in **words**, these five statistics for $x100$ and $|x100|$, in terms of our pollen dispersal model. [Submit the final `Matlab` script (use the `diary` command and edit it to make the final script with generous comments) used to carry out the above task of statistical summarization with **words** below it. The two sets of the five summaries should unambiguously appear in the submitted final script. You don't need to submit the modified simulation script from 1(a)].

2. So far we have assumed that there is only one plant, say PLANT located at the origin ($x = 0$). Suppose we have two new plants FAT-PLANT and SKINNY-PLANT that are also located at the origin. Suppose FAT-PLANT produces fatter or heavier pollen and SKINNY-PLANT produces skinnier or lighter pollen when compared with the pollens produced by PLANT. Since a lighter pollen, that is dispersed by wind, would travel farther than a heavier pollen, we may model the dispersal of a pollen from FAT-PLANT according to *Laplace*($\lambda = 10$) RV and that of a pollen from SKINNY-PLANT according to *Laplace*($\lambda = 0.1$) RV. We still assume that each of the three plants independently and

identically disperse their pollens. Finally, we assume that there is no differences in the rate of pollen production between the three plants. Thus, the i-th pollen (of any kind) is equally likely to have come from any of the three plants. Such a model of dispersal for a pollen, that may be any one of the three kinds, is called a finite equi-weighted mixture model. Therefore, our model for the dispersal of n pollens (of any kind) from the source is given by the independent and identically distributed RVs:

$$Y_1, Y_2, \ldots, Y_n \sim F(y; \lambda_1 = 1.0, \lambda_2 = 10, \lambda_3 = 0.1)$$

with DF

$$F(y; \lambda_1 = 1.0, \lambda_2 = 10, \lambda_3 = 0.1) = \frac{1}{3}F(x; \lambda_1 = 1.0) + \frac{1}{3}F(x; \lambda_2 = 10) + \frac{1}{3}F(x; \lambda_3 = 0.1)$$

(a) Write a `Matlab` **script** (use the `diary` command and edit it to make the final script with generous comments) that would simulate the dispersal of 10 consecutive pollens of any kind according to the equi-weighted finite mixture model. In other words, draw 10 samples from the RV Y above. The output of this simulation should be 10 real numbers that are stored in a vector named '`y10`'. Using **words** interpret the meaning of these numbers in terms of our model for pollen dispersal. You are required to initialize the fundamental sampler prior to this simulation using '`rand('twister',007)`' . [Submit the **script** file with **words** below it – if you use the functions in M-files you created during a previous lab or assignment then this M-file has to be included in your report for full credit].

(b) Reset the fundamental sampler using '`rand('twister',007)`' and simulate the dispersal of 100 pollens of any kind by modifying the script of 2(a), ie. draw 100 samples from the RV Y above, and store the samples in a vector named '`y100`'. Report the five statistics; mean, median, sample standard deviation, minimum and maximum, for the 100 samples and their absolute values:

i. 100 samples stored in $y100$,

ii. $|y100|$.

using the built-in `Matlab` commands `mean`, `median`, `std`, `max`, and `min`, respectively. Interpret in **words**, these five statistics for $y100$ and $|y100|$, in terms of our dispersal model for three kinds of pollen from the same source. Compare and contrast these statistics with their counterparts in (b) above and try to interpret them in **words**. Feel free to increase the sample sizes when computing the five statistics to sharpen your intuitions. [Submit the final `Matlab` script (use the `diary` command and edit it to make the final script with generous comments) used to carry out the above task of statistical summarization with **words** below it. The two sets of the five summaries should unambiguously appear in the submitted final script. You don't need to submit the modified simulation script from 2(a)].

Problem 3 You have to generate your own polynomial for this problem:

$$g_c(x) = e^{-c_1 x - c_2 x^2 - c_3 x^3 - c_4 x^4} ,$$

where the following code generates the array $c = (c_1, c_2, c_3, c_4)$ that is specific to your Student ID. If your student ID is 35598968, then the following steps generate your coefficient vector c :

```
>> MyID = 35598968; % write your student ID in the variable MyID
rand('twister',MyID); % Set the seed for rand with MyID
c=rand(1,4)*2.0+2.0 % store the c's in the array called c

c =

    2.9276    3.5945    3.9367    3.0110
```

From hereon we assume that such an array `c` that is uniquely tied to your student ID has been declared and available in memory. Next write an M-file called `MyIDPloy4.m` to encode the c-specific $g_c(x)$ as follows:

MyIDPoly4.m

```
function y = MyIDPoly4(x,c);
% Return: the exponential of negative of a degree 4 polynomial with coefficients
%         specified by the array c
% Input : c is a vector of 4 numbers
% Input : x is a real number at which the polynomial is evaluated
% Output: y = exp( - (c_1 x + c_2 x^2 + ... + c_4 x^4) )
%                                                 using .^ for vectorized powers
y = exp( - c(1) * x - c(2) * x .^2 - c(3) * x .^3 - c(4) * x .^4 );
```

1. Using `QuadMR` compute an approximation $\overset{mr}{\mathbb{I}}_n(g_c)$ of $\int_{-10}^{10} g_c(x)dx$ as n, the number of calls to the integrand, ranges in $\{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$. [Report the 6 approximations $\overset{mr}{\mathbb{I}}_n(g_c)$ (in `format long`) to the integral returned by `QuadMR` as n varies from 10 to 10^6.] For hints, use `QuadMRplot` to visualize the `QuadMR` (you don't need to turn in any plots).

2. Compute the approximation to the integral $\int_{-10}^{10} g_c(x)dx$ using `MATLAB`'s built-in quadrature rule `quad` [Report two numbers: (1) the estimate of the integral and (2) the number of function calls needed by `quad`].

3. Using IID samples $x_1, x_2, \ldots, x_n$ drawn from the $Uniform(-10, 10)$ RV X with density:

$$f(x) = \begin{cases} \frac{1}{20} & if -10 \leq x \leq 10 \\ 0 & otherwise \end{cases}$$

   compute a Monte Carlo point estimate of the integral $\int_{-10}^{10} g_c(x)dx = \int_{-10}^{10} w_c(x)f(x)dx$, with $w_c(x) = 20g_c(x)$ from the sample mean:

$$\widehat{g_c}_n = \frac{1}{n}\sum_{i=1}^{n} w_c(x_i)$$

   as n ranges in $\{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$. [Report the six point-estimates $\widehat{g_c}_{10}$, $\widehat{g_c}_{10^2}$, $\widehat{g_c}_{10^3}$, $\widehat{g_c}_{10^4}$, $\widehat{g_c}_{10^5}$ and $\widehat{g_c}_{10^6}$ after initializing the the fundamental sampler by "`rand('twister',666)`" for each n in the set $\{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$]

4. Compute the standard error $\widehat{se}_n$ for the 6 point estimates $\widehat{g_c}_{10}$, $\widehat{g_c}_{10^2}$, $\widehat{g_c}_{10^3}$, $\widehat{g_c}_{10^4}$, $\widehat{g_c}_{10^5}$ and $\widehat{g_c}_{10^6}$ computed in 1(c), where

$$\widehat{se}_n = \frac{S_n}{\sqrt{n}}, \qquad S_n^2 = \frac{\sum_{i=1}^{n} \left(w_c(x_i) - \widehat{g_c}_n\right)^2}{n-1}$$

   [Report 6 numbers $\widehat{se}_{10}$, $\widehat{se}_{10^2}$, $\widehat{se}_{10^3}$, $\widehat{se}_{10^4}$, $\widehat{se}_{10^5}$ and $\widehat{se}_{10^6}$]

Problem 4 Let Q_{g_c} denote the quadrature approximation of $\int_{-10}^{10} g_c(x)dx$ from Problem 1(b). Now, let us turn $g_c(x)$ into a density function $f_c(x)$ by normalizing it with Q_{g_c} :

$$f_c(x) \equiv \frac{1}{Q_{g_c}} g_c(x) = \frac{1}{Q_{g_c}} e^{-c_1 x - c_2 x^2 - c_3 x^3 - c_4 x^4},$$

where the array c is generated as in Problem 1. Use `MATLAB`'s built-in function `fminbnd` to find the minimum of $-f_c(x)$ for $x \in [-10, 10]$.

```
[IQ Calls] = quad(@(x)MyIDPoly4(x,c),-10,10);
[Xm Fmin] = fminbnd(@(x)-(1/IQ)*MyIDPoly4(x,c),-10,10);
Fmax=-Fmin;
```

With the aid of `Fmax` we can build a rejection sampler to draw samples from the RV X with density $f_c(x)$ using samples proposed from a RV Y $\sim Uniform(-10, 10)$ as follows:

```
SamplesNeeded=100; % number of samples need
NumberSampled=0; % number of samples obtained so far
NumberProposed=0; % variable to track the number of porposed samples
Samples=zeros(1,SamplesNeeded);% initialize an array for storing samples
rand('twister',666); % initialize rand
while NumberSampled < SamplesNeeded, % keep going until
    proposed = (-10 + rand * 20); % propose a sample from Uniform(-10,10) RV
    NumberProposed = NumberProposed + 1; % increment number proposed
    height = (rand*Fmax); % propose a random height from Uniform(0,Fmax) RV
    if  height <= (1/IQ)*MyIDPoly4(proposed, c) % accept the proposed sample
        NumberSampled = NumberSampled + 1; % increment number sampled
        Samples(NumberSampled)=proposed; % save the accepted proposals as samples
    end
end
```

Denote the 100 samples from X stored in the array `Samples` by $(x_1, x_2, \ldots, x_{100})$ after executing the above code. This will be thought of as our data.

1. Using the Dvoretzky-Kiefer-Wolfowitz Inequality obtain a nonparametric 95% confidence interval for the DF $F_c(y) = \int_{-10}^{y} f_c(x)dx$ at $y = 0$. [Report 2 numbers that define the confidence interval]
2. Compute the plug-in estimate of the mean or $E(X)$ and a Normal-based 95% confidence interval for it. [Report the three numbers]
3. Compute the plug-in estimate for the median and a nonparametric bootstrap-based 95% confidence interval for it. [Report the three numbers. Initialize the rand by `rand('twister',777999)` prior to generating 1000 nonparametric bootstrapped data sets from our original data `Samples`.]
4. (*extra credit only) Make a more efficient sampler to draw samples from $X \sim F_c$.

Problem 5 Weblog data analysis.

1. Plot the nonparametric estimate of the DF for web login times starting October 2nd.
2. Plot the 95% confidence band around the above estimate of the DF.
3. Do a permutation test of the null hypothesis that the DF of the web login times starting October 1st is the same as that for the web login times starting October 2nd.

Problem 6 Let $A = \{a, b, c, d, e\}$ be the set of offspring or children of the set of parents $B = \{\alpha, \beta, \gamma, \delta\}$. The offspring-parent relation can be summarised by the following set of ordered pairs:

$$\{(a, \alpha), (b, \alpha), (c, \beta), (e, \gamma), (d, \delta)\} \ .$$

1. Draw a picture of the function or map $f : A \to B$ depicting the above offspring-parent relations (you need to make a picture with two sets A and B and 5 arrows between them).
2. What is $(\{f(a)\} \cap \{f(b)\}) \cup \{f(e)\}$?
3. What is $f^{[-1]}(\alpha) \cup f^{[-1]}(\gamma)$?
4. Will the following set of ordered pairs:

$$\{(a, \alpha), (b, \alpha), (c, \beta), (e, \gamma), (a, \delta), (d, \delta)\}$$

constitute a function from A to B? Explain your answer.

5. Will the following set of ordered pairs:

$$\{(a,\alpha),(b,\beta),(c,\gamma),(d,\delta)\}$$

constitute a function from A to B? Explain your answer.

Problem 7 Suppose we model the NZ Lotto as follows:

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} de\ Moivre(1/40, 1/40, \ldots, 1/40)$$

Show that the probability of any subset A of the sample space $\{1, 2, \ldots, 40\}$ is $|A|/40$. This implies that $P(\{2, 4, 6, \ldots, 38, 40) = 20/40 = 1/2$, for instance. You have to explain each step in your general derivation that $P(A) = |A|/40$ using the basic laws of probability. Recall that $|A|$ is the number of elements in A.

Problem 8 Suppose we roll a fair dice twice in an independent and identical manner. Let A be the event that the sum is 5 and B be the event that the first die is 2. What is $P(B|A)$, the conditional probability that the first die is 2 given that the sum is 5? Show each step in your computation.

Problem 9 Consider the following random variable X that is a discrete mixture of two Uniform(θ_1, θ_2) RVs. The PDF of X can be expressed as follows:

$$f(x) = \frac{1}{2}3\,\mathbb{1}_{[0,1/3]}(x) + \frac{1}{2}3\,\mathbb{1}_{[2/3,1]}(x)$$

Show each step in your computation of the following quantities.

1. Draw the graph of the PDF $f(x)$.
2. Find the CDF $F(x) := \int_{-\infty}^{x} f(y)dy$ of X.
3. Draw the graph of the CDF $F(x)$.
4. Find the Expectation $E(X) := \int_{-\infty}^{\infty} xf(x)dx$.
5. What is $P(X \in [2/3, 3/4])$?

Problem 10

The RV X in **Q.4** can be constructed from the following process. I flip a fair coin and if I get heads then I choose a real number uniformly at random from $[0, 1/3]$ and if I get tails then I choose a real number uniformly at random from $[2/3, 1]$. Using the inversion sampler write the pseudocode to simulate a sample x from X. Note that you have to transform IID samples from Uniform$(0, 1)$ RV to one sample from X.

Problem 11 For a given ordered pair of parameters $(k, \lambda) \in (0, \infty)^2$, the RV X is said to be Weibull(k, λ) distributed if its DF is:

$$F(x; k, \lambda) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^k\right), \qquad x \in [0, \infty)\ .$$

Devise an algorithm that can transform IID samples from Uniform$(0, 1)$ into IID samples from $X \sim$ Weibull(k, λ). Present this Algorithm as pseudocode.

Problem 12 Suppose you observe the following five data points from some product experiment:

$$0, 2, 1, 0, 3.$$

1. compute the sample mean.
2. compute the sample variance.

3. compute the order statistics.
4. draw a graph of the empirical mass function.
5. draw a graph of the empirical distribution function.

Problem 13 Recall that a linear congruential generator, $LCG(m, a, c, x_0, n)$ given by:

$$x_i \leftarrow (ax_{i-1} + c) \mod m, \quad i = 1, 2, \ldots, n$$

can have the full period m if and only if:

1. c and m are relatively prime, i.e., the greatest common divisor of c and m is 1.
2. $a - 1$ is divisible by all prime factors of m,
3. $a - 1$ is a multiple of 4 if m is a multiple of 4
4. Show that the LCG with $(m, a, c, x_0, n) = (256, 137, 123, 13, 256)$ has a full period of 256.
5. Find the period length of the LCG with $(m, a, c, x_0, n) = (256, 138, 3, 0, 256)$ and explain why it does not have the maximal period length of 256.

Problem 14 Given a real parameter $\lambda > 0$, the discrete RV X is said to be Poisson(λ) distributed if X has PDF:

$$f(x; \lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{if } x \in \mathbb{Z}_+ := \{0, 1, 2, \ldots\} \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Now suppose we make n observations and model these as a product Poisson(λ^*) experiment:

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda^*) \ .$$

1. Set up the log-likelihood function $\ell(\lambda)$ for the observed data $(x_1, x_2, \ldots, x_n)$ and simplify the expression as much as you can.
2. Find the derivative of $\ell(\lambda)$ with respect to λ.
3. Solve for λ in the equation: $\frac{\partial \ell(\lambda)}{\partial \lambda} = 0$.
4. What is the maximum likelihood estimator $\widehat{\Lambda}_n$ of the unknown parameter λ^*?
5. If you observed 5 data points $(x_1, x_2, x_3, x_4, x_5) = (0, 2, 1, 0, 3)$, what is the maximum likelihood estimate $\widehat{\lambda}_5$ of λ^*?

Problem 15 Suppose I drop a ball into the Quincunx and only observe that it ended up at the bucket labelled 7. The buckets are labelled by the number of right turns made by the ball along its journey down the 21 levels of nails.

1. How many left turns did the ball make under the assumption that it behaved well, in terms of only going right or left at each nail it hit, along its journey?
2. What is the probability of this observed outcome under a binomial model with 21 trials and equal probability of taking a right versus left turn, i.e. with parameters $n = 21$ and $\theta = 1/2$?
3. What is the likelihood function of the parameter $\theta \in [0, 1]$ for this observation under the binomial model?

4. If you had to choose the most likely parameter from the set

$$\{1/10, 1/2, 9/10\}$$

based on your single observation of 7 right turns, then what would it be? Justify your answer.

5. If you dropped another ball and it fell into the bucket labelled 0 then what is your most likely estimate of θ from the same set above on the basis of this new ball's outcome alone? Justify your answer.

Problem 16 Inference and Simulation for the binomial model.

1. Derive the maximum likelihood estimator $\widehat{\Theta}_m$ from the log-likelihood function $\ell(\theta)$ for the product binomial experiment from m IID observations with a known n:

$$X_i, X_2, \ldots, X_m \overset{IID}{\sim} \text{Binomial}(n, \theta)$$

2. Write an algorithm in pseudocode to simulate samples from the Binomial(n, θ) RV X with sample space $\mathbb{X} = \{0, 1, 2, \ldots, n\}$. Explain the steps in detail. You only have pseudo-random numbers to mimic IID samples from Uniform$(0, 1)$ at you disposal. The algorithm should take n, the number of trials, and θ, the "success probability," as input and return x, a sample from X as output.

Problem 17 Suppose you observe that two sampled females of a particular species of Lemurs weigh 2.2 and 1.6 units and two sampled males weigh 1.9 and 4.1 units. All measurements are made in units of Standard Lemur Kilograms. Set up the hypothesis testing problem of attempting to reject the null hypothesis that the males and females have the same underlying distribution for their weights. Conduct a non-parametric permutation test based on the test statistic of the absolute difference of the gender-specific sample means. Tabulate the probabilities in the sample space under the null hypothesis, compute the p-value. Did you reject or fail to reject the null hypothesis? Explain your steps in detail.

Problem 18 This question has four parts.

(a.) Show that the following algorithm produces independent and identically distributed (IID) samples from the Exponential(λ) random variable (RV) X with DF $F(x; \lambda) = 1 - \exp(-\lambda x)$, with $x \in (0, \infty)$ and the parameter $\lambda \in (0, \infty)$:

$$u \leftarrow \text{Uniform}(0, 1)$$
$$x \leftarrow -\lambda^{-1} \log(u)$$

(b.) For a given parameter $\alpha > 0$, we say $X \sim \text{Rayleigh}(\alpha)$ if the DF of X is:

$$F(x; \alpha) = 1 - \exp\left(\frac{-x^2}{2\alpha^2}\right), \qquad x \in [0, \infty) \ .$$

Devise an algorithm that can transform IID samples from $U \sim \text{Uniform}(0, 1)$ into IID samples from $X \sim \text{Rayleigh}(\alpha)$.

(c.) For a given ordered pair of parameters $(k, \lambda) \in (0, \infty)^2$, the RV X is said to be Weibull(k, λ) distributed if its DF is:

$$F(x; k, \lambda) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^k\right), \qquad x \in [0, \infty) \ .$$

Devise an algorithm that can transform IID samples from $U \sim \text{Uniform}(0, 1)$ into IID samples from $X \sim \text{Weibull}(k, \lambda)$.

(d.) For a given ordered triple of parameters $(\mu, \sigma, \xi) \in (-\infty, \infty) \times (0, \infty) \times (0, \infty)$, the RV X is said to have the generalised extreme value (GEV) distribution and denoted $X \sim \text{GEV}(\mu, \sigma, \xi)$, if its DF is:

$$F(x; \mu, \sigma, \xi) = \exp\left(-\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right), \qquad 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0 \ .$$

Devise an algorithm that can transform IID samples from $U \sim \text{Uniform}(0, 1)$ into IID samples from $X \sim \text{GEV}(\mu, \sigma, \xi)$.

Problem 19 This question has three parts. Let us recall the von Neumann Rejection Sampler. Suppose,

1. we can generate random variables with PDF g;
2. the support of g contains the support of f,
3. a constant $\tilde{a} > 0$ exists, such that:
$$\tilde{f}(x) \leq \tilde{a} g(x), \tag{1}$$
for any x in the support of X. Then x can be generated from Algorithm 1.

Algorithm 1 Rejection Sampler (RS) of von Neumann – target shape

input:

(1) shape of a target density $\tilde{f}(x) = \left(\int \tilde{f}(x)dx\right) f(x)$,

(2) a proposal density $g(x)$ satisfying (a), (b) and (c) above.

output: a sample x from RV X with density f
repeat
  Generate $y \sim g$ and $u \sim \text{Uniform}(0, 1)$
until $u \leq \frac{\tilde{f}(y)}{\tilde{a} g(y)}$
return: $x \leftarrow y$

The probability of accepting a proposed sample is the integral ratio: $\int \tilde{f}(x)dx / \int \tilde{a} g(y)dy$.

(a.) Let $\theta \geq 0$ be a parameter. Consider the following θ-specific target shape:

$$\widetilde{f}(x; \theta) = x^{\theta}(1-x)^{\theta}, \qquad x \in [0, 1] \ .$$

Using IID proposals from Uniform$(0, 1)$, devise a Rejection Sampling Algorithm to draw IID samples from the normalised target shape :

$$f(x; \theta) = \frac{\widetilde{f}(x; \theta)}{\int_0^1 \widetilde{f}(x; \theta)dx}, \qquad x \in [0, 1] \ .$$

Show that you have satisfied all the conditions of the Rejection Sampler and produce the θ-specific constant $\widetilde{a}$ when the proposal density g is the PDF of Uniform$(0, 1)$.

(b.) For what value of the parameter θ does this Rejection Sampler become optimal?

(c.) How does the acceptance probability of this sampler behave as a function of the parameter θ?

Problem 20 This question has two parts. Recall the Algorithm in pseudo-code for basic Monte Carlo integral estimation:

Recall the importance sampling estimator of of the integral $\vartheta^* = \int h(x)f(x)dx$ is

$$\widehat{\Theta}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{h(X_i)f(X_i)}{g(X_i)}, \quad X_1, \ldots, X_n \overset{IID}{\sim} g, \quad \text{and} \quad V_g(\widehat{\Theta}_n) = \frac{1}{n} V_g\left(\frac{h(X_1)f(X_1)}{g(X_1)}\right) \ .$$

Algorithm 2 Basic Monte Carlo Integral Estimation for $\vartheta^* = \int_{[\underline{a}_1,\overline{a}_1]\times\cdots\times[\underline{a}_k,\overline{a}_k]} h(x)dx$

1: *input:*

1. $n \leftarrow$ the number of samples.
2. $h(x) \leftarrow$ the integrand function over $\mathbb{R}$
3. $[\underline{a}_j, \overline{a}_j] \leftarrow$ lower and upper bounds of integration for each $j = 1, 2, \ldots, k$
4. capability to draw nk IID samples from Uniform$(0, 1)$ RV

2: *output:* a point estimate $\widehat{\vartheta}_n$ of ϑ^* and the estimated standard error $\widehat{\mathsf{se}}_n$

3: *initialize:* $y \leftarrow (0, 0, \ldots, 0)$, initialize y as a zero vector of length n

4: **while** $i \leq n$ **do**

5:
1. $i \leftarrow i + 1$,
2. $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,k})$, with $x_{i,j} \leftarrow u_j$, $u_j \sim$ Uniform$(\underline{a}_j, \overline{a}_j)$, for $j = 1, 2, \ldots, k$,
3. $y_i \leftarrow w(x_i) = h(x_i) \prod_{j=1}^{k} (\overline{a}_j - \underline{a}_j)$

6: **end while**

1. $\widehat{\vartheta}_n \leftarrow \overline{y}_n$, the sample mean of $y = (y_1, y_2, \ldots, y_n)$
2. $\widehat{\mathsf{se}}_n = s_n(y)/\sqrt{n}$, where $s_n(y)$ is the sample standard deviation of y

7: *return:* $\widehat{\vartheta}_n$ and $\widehat{\mathsf{se}}_n$

(a.) Imagine a rigid ball $B_3 := \{(x_1, x_2, x_3) : \sqrt{x_1^2 + x_2^2 + x_3^2} \leq 1/2\}$ perfectly enclosed by a hollow cube C_3 so that the ball is immobilized. Describe in words **and** in pseudo-code how you would use Monte Carlo integral estimation to compute the volume of air inside C_3, i.e. the volume of C_3 that is not occupied by B_3.

(b.) List the pseudo-code to compute the point estimate, standard error and a Normal-based 95% confidence interval for ϑ^*. The listing should have details similar to Algorithm 3 above.

Problem 21 This question has ten parts. You may need the following three facts.

1. Recall that for integer values of parameter k,

$$\text{If} \quad X_1, X_2, \ldots, X_k \overset{IID}{\sim} \text{Exponential}(v) \quad \text{then} \quad Y \equiv \sum_{i=1}^{k} X_i \sim \text{Gamma}(v, k)$$

2. The probability density function for the $Y \sim \text{Gamma}(v, k)$ is:

$$f_Y(y; k, v) = \frac{v(vy)^{k-1}}{\Gamma(k)} e^{-vy}, \quad y \geq 0 \ .$$

3. For a Normal-based 95% confidence interval $\widehat{\theta}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}_n$, for an unknown θ^*, $z_{\alpha/2} = 1.96$.

(a.) Suppose that the parameter k^* is known but the parameter v^* is unknown. Let

$$Y_1, Y_2, \ldots, Y_n \overset{IID}{\sim} \text{Gamma}(v^*, k^*)$$

Based on n independent and identically distributed observations from Gamma(v^*, k^*), with known k^*, show that the maximum likelihood estimate $\widehat{v}_n$ of v^* is:

$$\widehat{v}_n = \frac{k^*}{\overline{y}_n} \ .$$

(b.) Obtain the estimated standard error $\widehat{\mathsf{se}}_n$ for the ML estimator of v^* using the Fisher Information:

$$\widehat{\mathsf{se}}_n(\widehat{V}_n) = \frac{1}{\sqrt{-nE_{k^*,\widehat{v}_n}\left(\frac{\partial^2}{\partial v^2}\log(f_Y(y;k^*,\widehat{v}_n))\right)}}$$

(c.) Using the Delta method find the maximum likelihood estimate of $\psi^* = g(v^*) = 1/{v^*}^2$ and the standard error of the maximum likelihood estimate of ψ^*. Recall

$$\widehat{\mathsf{se}}_n(\widehat{\Psi}_n) = |g'(\widehat{v}_n)|\widehat{\mathsf{se}}_n(\widehat{V}_n)$$

For problems (d)-(f), we suppose that the inter-arrival time between any two consecutive buses arriving at a bus-stop is independent and identically distributed according to $X_i \sim \text{Exponential}(v^*)$. Suppose we observe the inter-arrival time of every 7-th bus at a bus-stop and record these values to the nearest minute, as follows:

$$45,\ 89,\ 61,\ 63,\ 102,\ 98,\ 86,\ 55,\ 52,\ 99\ .$$

(d.) Find the maximum likelihood estimate for the unknown parameter v^* and ψ^*.

(e.) Obtain the 95% confidence interval for the unknown parameter v^* using the approximation:

$$\frac{\widehat{v}_n - v^*}{\widehat{\mathsf{se}}_n(\widehat{V}_n)} \rightsquigarrow \text{Normal}(0,1)\ .$$

(f.) Obtain the 95% confidence interval for the unknown quantity ψ^* using the approximation:

$$\frac{\widehat{\psi}_n - \psi^*}{\widehat{\mathsf{se}}_n(\widehat{\Psi}_n)} \rightsquigarrow \text{Normal}(0,1)\ .$$

For problems (g)-(j), we suppose that the inter-arrival time between any two consecutive buses arriving at a bus-stop is independent and identically distributed according to $X_i \sim F^* \in \{\text{all DFs}\}$. Suppose we observe the inter-arrival time of every 7-th bus at a bus-stop and record these values to the nearest minute, as follows:

$$45,\ 89,\ 61,\ 63,\ 102,\ 98,\ 86,\ 55,\ 52,\ 99\ .$$

(g.) Plot the empirical CDF $\widehat{F}_n$.

(h.) Compute the plug-in estimate of the median.

(i.) List the pseudo-code for the non-parametric bootstrap algorithm to obtain a 95% confidence interval for the plug-in estimate of the median. Discuss the limitations of the procedure for this data set.

(j.) List the pseudo-code for a non-parametric hypothesis testing algorithm to possibly reject the null hypothesis that the mean inter-arrival time of every 7-th bus is less than 60 minutes. Discuss the limitations of the procedure for this data set.

Problem 22 For a given ordered triple of parameters $(\mu, \sigma, \xi) \in (-\infty, \infty) \times (0, \infty) \times (0, \infty)$, the RV X is said to have the generalised extreme value (GEV) distribution and denoted $X \sim \text{GEV}(\mu, \sigma, \xi)$, if its DF is:

$$F(x;\mu,\sigma,\xi) = \exp\left(-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right), \qquad 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\ .$$

Devise an algorithm that can transform IID samples from $U \sim \text{Uniform}(0,1)$ into IID samples from

$$X \sim \frac{1}{2}\,\text{GEV}(\mu_1,\sigma_1,\xi_1) + \frac{1}{2}\,\text{GEV}(\mu_2,\sigma_2,\xi_2)\ .$$

Note that X is a RV whose distribution is obtained from an equi-probable mixture of two independent and possibly non-identical GEV distributions. Give all details of your algorithm and show all steps in the derivation of any expressions used in the algorithm.

Problem 23 This question has two parts.

(a.) Write an Algorithm in pseudocode or in Matlab syntax to produce 100 samples from a random vector Y with sample space:

$$\mathbb{Y} = \{(y_1, y_2, y_3) \in \mathbb{Z}_+^3 : \sum_{i=1}^{3} y_i = 1973\} \ ,$$

and probability mass function:

$$P(Y = (y_1, y_2, y_3)) = \mathbb{1}_{\mathbb{Y}}((y_1, y_2, y_3)) \frac{1973!}{y_1! y_2! y_3!} \left(\frac{1}{3}\right)^{\sum_{i=1}^{3} y_i} \ .$$

Recall that $\mathbb{Z}_+^3$ is the set of all non-negative integer vectors in three dimensions. You need to explain each step in your algorithm.

(b.) Explain the additional algorithmic step(s) needed to obtain samples from the random vector $W := (y_1, y_2, y_3)/1973$.

(c.) Formally describe the sample space $\mathbb{W}$

(d.) What is the probability mass function of our transformed random vector W?

Problem 24 This question has two parts. Recall the Algorithm in pseudo-code for basic Monte Carlo integral estimation:

Algorithm 3 Basic Monte Carlo Integral Estimation for $\vartheta^* = \int_{[\underline{a}_1, \overline{a}_1] \times \cdots \times [\underline{a}_k, \overline{a}_k]} h(x) dx$

1: *input:*

1. $n \leftarrow$ the number of samples.
2. $h(x) \leftarrow$ the integrand function over $\mathbb{R}$
3. $[\underline{a}_j, \overline{a}_j] \leftarrow$ lower and upper bounds of integration for each $j = 1, 2, \ldots, k$
4. capability to draw nk IID samples from Uniform$(0, 1)$ RV

2: *output:* a point estimate $\widehat{\vartheta}_n$ of ϑ^* and the estimated standard error $\widehat{\mathsf{se}}_n$

3: *initialize:* $y \leftarrow (0, 0, \ldots, 0)$, initialize y as a zero vector of length n

4: **while** $i \leq n$ **do**

5:
1. $i \leftarrow i + 1$,
2. $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,k})$, with $x_{i,j} \leftarrow u_j$, $u_j \sim$ Uniform$(\underline{a}_j, \overline{a}_j)$, for $j = 1, 2, \ldots, k$,
3. $y_i \leftarrow w(x_i) = h(x_i) \prod_{j=1}^{k} (\overline{a}_j - \underline{a}_j)$

6: **end while**

1. $\widehat{\vartheta}_n \leftarrow \overline{y}_n$, the sample mean of $y = (y_1, y_2, \ldots, y_n)$
2. $\widehat{\mathsf{se}}_n = s_n(y)/\sqrt{n}$, where $s_n(y)$ is the sample standard deviation of y

7: *return:* $\widehat{\vartheta}_n$ and $\widehat{\mathsf{se}}_n$

(a.) Imagine a rigid chunk of swiss-cheese S that is perfectly enclosed by a hollow cube C_3. Suppose you have at your disposal the indicator function $\mathbb{1}_S((x_1, x_2, x_3))$ which returns 1 if the point (x_1, x_2, x_3) is in S and 0 otherwise. Describe in words **and** in pseudo-code how you would use Monte Carlo integral estimation to compute ϑ^*, the volume of cheese in S.

(b.) List the pseudo-code to compute the point estimate, standard error and a Normal-based 95% confidence interval for ϑ^*. The listing should have details similar to Algorithm 3 above.