CS215 Assignment 3 Report

Atharva Bendale (22B0901)
Vishal Bysani (22B1061)

September 2023

Contents

Introduction

In this report, we have written down our solutions for the problems in the Assigment 3 of CS 215 course on Data Analysis and Interpretation.

1 Problem 1

Consider a shelf containing n books, each one with a distinct color. Let us suppose that you pick a book uniformly at random with replacement (i.e. you put the book back on the shelf after picking it) and independently of what was picked earlier. Let $X^{(n)}$ be the number of times you would need to pick a book in this fashion, such that you have chosen a book of each color at least once. We can write that $X^{(n)} = X_1 + X_2 + ... + X_n$ where X_i denotes the additional number of times you have to pick a book such that you move from having picked books of $i-1$ distinct colors to i distinct colors. We wish to determine $E(X)$ and $Var(X)$. To this end, do as follows:

1. What is X_1? When books with $i-1$ distinct types of colors have been collected, what is the probability of picking a book with a different color (i.e. different from the previous $i-1$ colors)? [3 points]
2. Due to independence, X_i is a geometric random variable. What is its parameter? Let Z be a random variable for the trial number for the first head obtained in a sequence of independent Bernoulli trials with head probability p. Then $P(Z = k) = (1-p)^{k-1}p$ where $k = 1, 2, 3, ...$, and Z is said to be a geometric random variable with parameter p. [3 points]
3. Show that the expected value of a geometric random variable with parameter p is $1/p$. Derive the variance of a geometric random variable. [4+4=8 points]
4. Hence derive $E(X^{(n)})$ for this problem. [3 points]
5. Hence derive an upper bound on $Var(X^{(n)})$ for this problem. You will need the inequality that the sum of reciprocals of squares of positive integers is upper bounded by $\pi^2/6$. [3 points]
6. Plot a graph of $E(X^{(n)})$ versus n for different n. If $E(X^{(n)}) = \Theta(f(n))$, what is $f(n)$? [3+2=5 points]

Solution :

1. X_1 denotes the additional number of times you have to pick a book such that you move from having picked books of 0 distinct colors to 1 distinct color. Since picking up only 1 book enables us to move to 1 distinct color, $X_1 = 1$.
   When books with $i-1$ distinct types of colors have been collected, the probability of picking a book with a different color is $\dfrac{n-i+1}{n}$.
2. X_i is a geometric random variable with parameter $p_i = \dfrac{n-i+1}{n}$. X_i basically denotes the number of trials needed for the first occurrence of success, with the probability of success being p_i.
3. Let Y be the geometric random variable with parameter p. The expected value of Y is given by

$$E(Y) = \textstyle\sum_{i=1}^{\infty} i \cdot (1-p)^{i-1} \cdot p$$

This is an arithmetic-geometric progression. It can be solved as follows (by shifting the sequence by 1, multiplying with $(1-p)$ and subtracting from the above sequence)

$$E(Y) = 1 \cdot p + 2(1-p)p + 3(1-p)^2 p + 4(1-p)^3 p \dots \tag{1}$$

$$E(Y)(1-p) = 1(1-p)p + 2(1-p)^2p + 3(1-p)^3p \dots \tag{2}$$

On subtracting 2 from 1,

$$E(Y) \cdot p = 1 \cdot p + (1-p)p + (1-p)^2p + (1-p)^3p \dots$$

This is a geometric progression with common ratio $(1-p)$.

$$\therefore E(Y) \cdot p = \frac{p}{1-(1-p)}$$
$$\implies E(Y) = \frac{1}{p}$$

Now the variance of Y is given by ,

$$\text{Var}(Y) = E[(Y - E(Y))^2] = E(Y^2) - (E(Y))^2$$

In order to find $E(Y^2)$, we follow the same procedure as above, but applied twice,

$$E(Y^2) = 1^2 \cdot p + 2^2(1-p)p + 3^2(1-p)^2p + 4^2(1-p)^3p \dots \tag{3}$$
$$E(Y^2) \cdot (1-p) = 1^2(1-p)p + 2^2(1-p)^2p + 3^2(1-p)^3p \dots \tag{4}$$

On subtracting 4 from 3,

$$E(Y^2) \cdot p = 1 \cdot p + 3(1-p)p + 5(1-p)^2p + 7(1-p)^3p \dots \tag{5}$$
$$E(Y^2) \cdot p \cdot (1-p) = 1p(1-p) + 3(1-p)^2p + 5(1-p)^3p \dots \tag{6}$$

On subtracting 6 from 5,

$$E(Y^2) \cdot p^2 = 1 \cdot p + 2(1-p)p + 2(1-p)^2p + 2(1-p)^3p \dots$$
$$\implies E(Y^2) \cdot p^2 = p + \frac{2(1-p)p}{1-(1-p)}$$
$$\implies E(Y^2) \cdot p^2 = 2 - p$$
$$\therefore E(Y^2) = \frac{2}{p^2} - \frac{1}{p}$$

Now to find variance of Y,

$$\text{Var}(Y) = \frac{2}{p^2} - \frac{1}{p} - \left(\frac{1}{p}\right)^2$$
$$\implies \text{Var}(Y) = \frac{1}{p^2} - \frac{1}{p}$$

4. $E(X_i) = \dfrac{1}{p_i}$

[$\because X_i$ is a geometric random variable and we showed that $E(X) = 1/p$ if X is a geometric random variable]

$E(X^{(n)}) = E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n \dfrac{1}{p_i} = \sum_{i=1}^n \dfrac{n}{n-i+1} = \sum_{j=1}^n \dfrac{n}{j}$

[Using change of variable $j = n - i + 1$]

5. X_i and X_j are independent for $i \neq j$.

$$\therefore \text{Var}(X^{(n)}) = \text{Var}(\textstyle\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i) = \sum_{i=1}^n \frac{1}{p_i^2} - \frac{1}{p_i} = \sum_{i=1}^n \frac{n^2}{(n-i+1)^2} - \frac{n}{n-i+1}$$

$$= \sum_{j=1}^n \frac{n^2}{j^2} - \frac{n}{j} = \sum_{j=1}^n \frac{n(n-j)}{j^2} \quad \leq \quad \sum_{j=1}^n \frac{n(n-1)}{j^2} \quad \leq \quad \sum_{j=1}^\infty \frac{n(n-1)}{j^2} \leq \frac{n(n-1)\pi^2}{6}$$

Hence an upper bound on $\text{Var}(X^{(n)})$ is $\dfrac{n(n-1)\pi^2}{6}$.

6. Below are the plots for $E(X^{(n)})$,$n\log(n)$ and $n(\log(n-1)+1)$

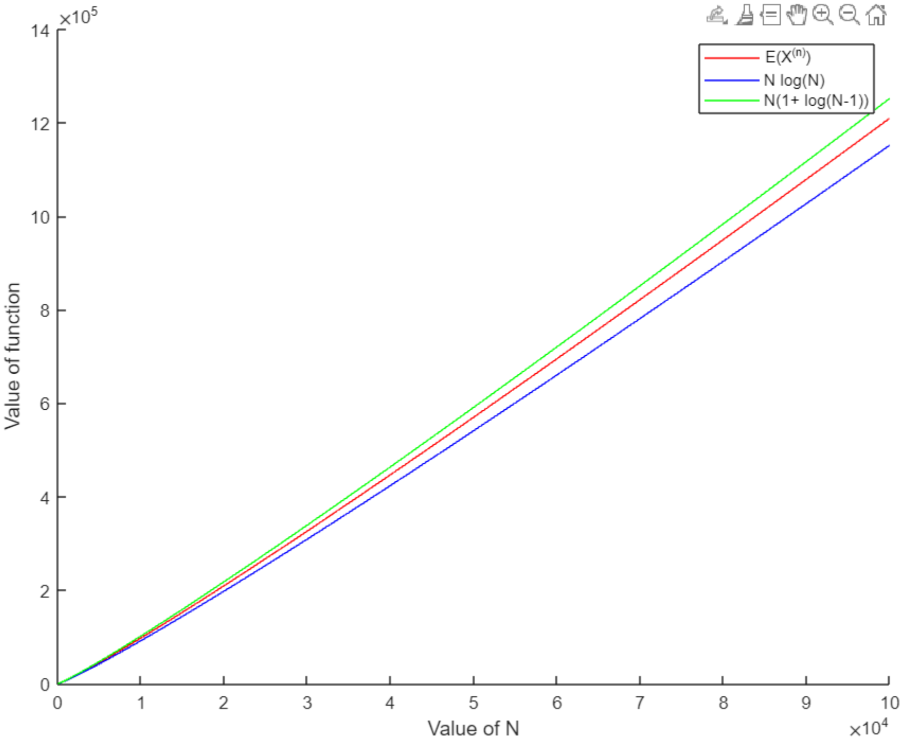

Figure 1: Plots of $E(X^{(n)})$ and its bounds

Now we need to find the function $f(n)$ which asymptotically binds $E(X^{(n)})$ from above and below. We can see graphically that,

$$\int_1^n \frac{1}{x}dx \leq \sum_{j=1}^n \frac{1}{j} \leq 1 + \int_2^n \frac{1}{x-1}dx$$

$$\implies \log(n) \leq \sum_{j=1}^n \frac{1}{j} \leq \log(n-1) + 1$$

$$\implies n\log(n) \leq \sum_{j=1}^n \frac{n}{j} \leq n \cdot (\log(n-1) + 1)$$

$$\implies n\log(n) \leq E(X^{(n)}) \leq n \cdot (\log(n-1) + 1)$$

Also,

$$n \cdot (\log(n-1)+1) \leq 2n\log(n)$$

Therefore,

$$n\log(n) \leq E(X^{(n)}) \leq 2n\log(n)$$

Hence, if $E(X^{(n)}) = \Theta(f(n))$, then $f(n) = n\log(n)$ i.e,

The code for generating the above plot is under the **Problem1** folder, saved under the name `Plot1.m`.

2 Problem 2

1. A student is trying to design a procedure to generate a sample from a distribution function F, where F is invertible. For this, (s)he generates a sample u_i from a $[0,1]$ uniform distribution using the 'rand' function of MATLAB, and computes $v_i = F^{-1}(u_i)$. This is repeated n times for $i = 1...n$. Prove that the values $\{v_i\}_{i=1}^n$ follow the distribution F. [6 points]
2. Let $Y_1, Y_2, ..., Y_n$ represent data from a continuous distribution F. The empirical distribution function F_e of these data is defined as $F_e(x) = \frac{\sum_{i=1}^n \mathbf{1}(Y_i \leq x)}{n}$ where $\mathbf{1}(z) = 1$ if the predicate z is true and 0 otherwise. Now define $D = \max_x|F_e(x) - F(x)|$. Also define $E = \max_{0\leq y \leq 1}\left|\frac{\sum_{i=1}^n \mathbf{1}(U_i \leq y)}{n} - y\right|$ where $U_1, U_2, ..., U_n$ represent data from a $[0,1]$ uniform distribution. Now prove that $P(E \geq d) = P(D \geq d)$. *Briefly* explain what you think is the practical significance of this result in statistics. [6+5=11 points]

Solution :

1. Let U be the uniform random variable over $[0,1]$ $\quad\therefore U \sim \text{Unif}[0,1]$
   We know that for a uniform distribution U on $[0,1]$, $P(U \leq u_i) = u_i$

   $$P(U \leq u_i) = u_i$$
   $$\implies P(F^{-1}(U) \leq F^{-1}(u_i)) = u_i$$
   [$\because$ F is invertible and since it is a distribution function, F is monotonically increasing as well]

   $$\implies P(F^{-1}(U) \leq v_i) = F(v_i) \quad [\because v_i = F^{-1}(u_i)]$$

   Hence, if we consider the random variable $Y = F^{-1}(U)$, then

   $$P(Y \leq v_i) = F(v_i)$$

   $\therefore$ F is the distribution of the random variable Y. Hence the values $\{v_i\}_{i=1}^n$ follow the distribution F.

2. Here,

   $$F_e(x) = \frac{\sum_{i=1}^n \mathbf{1}(Y_i \leq x)}{n}$$

   and

   $$D = max_x|F_e(x) - F(x)|$$

   Now

   $$D = max_x|\frac{\sum_{i=1}^n \mathbf{1}(Y_i \leq x)}{n} - F(x)| = max_{F(x)}|\frac{\sum_{i=1}^n \mathbf{1}(F(Y_i) \leq F(x))}{n} - F(x)| \qquad \textbf{as F is increasing}$$

Replacing $F(x)$ by y we can see that $D = E$ as $0 \leq F(x) \leq 1$ and $F(Y_i) = U_i$
Since $Y_i = F^{-1}(U_i)$ from part a) because a distribution form F can be witten in terms of a distribution from a **uniform** distribution
Hence their distribution function will be same or $P|D < d| = P|E < d|$
Now

$$P|D \geq d| = 1 - P|D < d| = 1 - P|E < d| = P|E \geq d|$$

Hence proved.

The observation is that as $P(D \geq d) = P(E \geq d)$, $P(D \geq d)$ is proved to be independent of the particular distribution function $F(x)$. This can be used to check whether the given data indeed belong to a pre specified distribution F. If the data indeed belong to F, then the value of D will likely not exceed the corresponding difference between the empirical CDF computed from random variables belonging to Uniform(0, 1) and the true uniform distribution (i.e. Uniform(0, 1)). This forms the motivation for a very famous statistical test called the Kolmogorov-Smirnov Test.

3 Problem 3

1. In this exercise, we will perform maximum likelihood based plane fitting. Let the equation of the plane be $z = ax + by + c$. Let us suppose we have access to accurate X and Y coordinates of some N points lying on the plane. We also have access to the Z coordinates of these points, but those have been corrupted independently by noise from $\mathcal{N}(0, \sigma^2)$. Write down the log-likelihood function $\mathcal{L}$ to be maximized in order to determine a, b, c. Write down three linear equations corresponding to setting partial derivatives of $\mathcal{L}$ w.r.t. a, b, c (respectively) to 0. Express these equations in matrix and vector form. [3+4=7 points]

2. Repeat the previous part if z had the form $z = a_1x^2 + a_2y^2 + a_3xy + a_4x + a_5y + a_6$. Again, let us suppose we have access to accurate X and Y coordinates of some N points lying on the plane. We also have access to the Z coordinates of these points, but those have been corrupted independently by noise from $\mathcal{N}(0, \sigma^2)$. Write down the log-likelihood function $\mathcal{L}$ to be maximized in order to determine $a_1, a_2, ..., a_6$. Write down linear equations corresponding to setting partial derivatives of $\mathcal{L}$ w.r.t. $a_1, a_2, ..., a_6$ (respectively) to 0. Express these equations in matrix and vector form. [4+4=8 points]

3. Now write MATLAB code to solve this linear system for data consisting of XYZ coordinates of $N =$ 2000 points, stored in the file 'XYZ.txt' in the homework folder. Read the data using the MATLAB function 'dlmwrite'. The data consist of N rows, each containing the X,Y,Z coordinates of a point (in that order). What is the predicted equation of the plane? What is the predicted noise variance? State these in your report, and print them out via your code. [10 points]

Solution :

1. Let (x_i, y_i, z_i) be the available coordinates of the i^{th} point. As the z coordinate of the points are corrupted by Gaussian noise $\mathcal{N}(0, \sigma^2)$, we can write :

$$z_i = ax_i + by_i + c + \epsilon_0$$

where $\epsilon_0 \in \mathcal{N}(0, \sigma^2)$.
The values of $\{x_i, y_i\}$ are known accurately, so the $\{z_i\}$ estimates are inaccurate.

$$\therefore z_i \in \mathcal{N}(ax_i + by_i + c, \sigma^2)$$

$$p(z_i \mid \{x_i, y_i\}, a, b, c) = \frac{e^{-\left(\frac{z_i - (ax_i + by_i + c)}{\sigma\sqrt{2}}\right)^2}}{\sigma\sqrt{2\pi}}$$

The log likelihood function $\mathcal{L}$ is :

$$\mathcal{L}(a, b, c) = \Sigma_i \log p(z_i \mid \{x_i, y_i\}, a, b, c)$$

$$\mathcal{L}(a, b, c) = -\sum_{i=1}^{n} \left[\frac{z_i - (ax_i + by_i + c)}{\sigma\sqrt{2}}\right]^2 - n\log\left(\sigma\sqrt{2\pi}\right)$$

This is the log likelihood function to be maximized for determining a, b, c. The three linear equations corresponding to partial derivative of $\mathcal{L}$ are :

$$\frac{\partial \mathcal{L}}{\partial a} = \sum_{i=1}^{n} x_i \cdot \left[\frac{z_i - (ax_i + by_i + c)}{\sigma\sqrt{2}}\right]$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} y_i \cdot \left[\frac{z_i - (ax_i + by_i + c)}{\sigma\sqrt{2}}\right]$$

$$\frac{\partial \mathcal{L}}{\partial c} = \sum_{i=1}^{n} \left[\frac{z_i - (ax_i + by_i + c)}{\sigma\sqrt{2}}\right]$$

For maximizing $\mathcal{L}$ it's partial derivatives with a, b & c should be zero, applying this condition yields these three linear equation in a, b, c :

$$a\left(\Sigma x_i^2\right) + b\left(\Sigma x_i y_i\right) + c\left(\Sigma x_i\right) = \Sigma x_i z_i$$

$$a\left(\Sigma x_i y_i\right) + b\left(\Sigma y_i^2\right) + c\left(\Sigma y_i\right) = \Sigma y_i z_i$$

$$a\left(\Sigma x_i\right) + b\left(\Sigma y_i\right) + (n)c = \Sigma z_i$$

Here all the summations are from $i = 1$ to n.
The matrix form of these linear equations is :

$$\begin{bmatrix} \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \\ \Sigma x_i y_i & \Sigma y_i^2 & \Sigma y_i \\ \Sigma x_i & \Sigma y_i & n \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \Sigma x_i z_i \\ \Sigma y_i z_i \\ \Sigma z_i \end{bmatrix}$$

In vector form :

$$a\begin{bmatrix} \Sigma x_i^2 \\ \Sigma x_i y_i \\ \Sigma x_i \end{bmatrix} + b\begin{bmatrix} \Sigma x_i y_i \\ \Sigma y_i^2 \\ \Sigma y_i \end{bmatrix} + c\begin{bmatrix} \Sigma x_i \\ \Sigma y_i \\ n \end{bmatrix} = \begin{bmatrix} \Sigma x_i z_i \\ \Sigma y_i z_i \\ \Sigma z_i \end{bmatrix}$$

2. Similar to the previous part we can write :

$$z_i = a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6 + \epsilon_0$$

where $\epsilon_0 \in \mathcal{N}(0, \sigma^2)$.
Similarly $\{x_i, y_i\}$ are known accurately, so the $\{z_i\}$ estimates are inaccurate.

$$\therefore z_i \in \mathcal{N}\left(a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6,\ \sigma^2\right)$$

$$p(z_i \mid \{x_i, y_i\}, a_1 \ldots a_6) = \frac{e^{-\left(\frac{z_i - (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6)}{\sigma\sqrt{2}}\right)^2}}{\sigma\sqrt{2\pi}}$$

The log likelihood function $\mathcal{L}$ is :

$$\mathcal{L}(a, b, c) = \sum_{i=1}^{n} \log p(z_i \mid \{x_i, y_i\}, a_1 \ldots a_6)$$

$$\mathcal{L}(a, b, c) = -\sum_{i=1}^{n} \left[\frac{z_i - (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6)}{\sigma\sqrt{2}}\right]^2 - n\log\left(\sigma\sqrt{2\pi}\right)$$

The partial derivatives of $\mathcal{L}$ wrt $\{a_1, \ldots a_6\}$ are :

$$\frac{\partial \mathcal{L}}{\partial a_1} = \sum_{i=1}^{n} x_i^2 \cdot \left[\frac{z_i - (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6)}{\sigma\sqrt{2}}\right]$$

$$\frac{\partial \mathcal{L}}{\partial a_2} = \sum_{i=1}^{n} y_i^2 \cdot \left[\frac{z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)}{\sigma\sqrt{2}}\right]$$

$$\frac{\partial \mathcal{L}}{\partial a_3} = \sum_{i=1}^{n} x_iy_i \cdot \left[\frac{z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)}{\sigma\sqrt{2}}\right]$$

$$\frac{\partial \mathcal{L}}{\partial a_4} = \sum_{i=1}^{n} x_i \cdot \left[\frac{z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)}{\sigma\sqrt{2}}\right]$$

$$\frac{\partial \mathcal{L}}{\partial a_5} = \sum_{i=1}^{n} y_i \cdot \left[\frac{z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)}{\sigma\sqrt{2}}\right]$$

$$\frac{\partial \mathcal{L}}{\partial a_6} = \sum_{i=1}^{n} \left[\frac{z_i - (a_1x_i^2 + a_2y_i^2 + a_3x_iy_i + a_4x_i + a_5y_i + a_6)}{\sigma\sqrt{2}}\right]$$

For maximizing $\mathcal{L}$ all these partial derivatives should be zeros. Applying this condition yields these 6 linear equations :

$$a_1(\Sigma x_i^4) + a_2(\Sigma x_i^2y_i^2) + a_3(\Sigma x_i^3y_i) + a_4(\Sigma x_i^3) + a_5(\Sigma x_i^2y_i) + a_6(\Sigma x_i^2) = \Sigma x_i^2 z_i$$

$$a_1(\Sigma x_i^2y_i^2) + a_2(\Sigma y_i^4) + a_3(\Sigma x_iy_i^3) + a_4(\Sigma x_iy_i^2) + a_5(\Sigma y_i^3) + a_6(\Sigma y_i^2) = \Sigma y_i^2 z_i$$

$$a_1(\Sigma x_i^3y_i) + a_2(\Sigma x_iy_i^3) + a_3(\Sigma x_i^2y_i^2) + a_4(\Sigma x_i^2y_i) + a_5(\Sigma x_iy_i^2) + a_6(\Sigma x_iy_i) = \Sigma x_iy_iz_i$$

$$a_1(\Sigma x_i^3) + a_2(\Sigma x_iy_i^2) + a_3(\Sigma x_i^2y_i) + a_4(\Sigma x_i^2) + a_5(\Sigma x_iy_i) + a_6(\Sigma x_i) = \Sigma x_iz_i$$

$$a_1(\Sigma x_i^2y_i) + a_2(\Sigma y_i^3) + a_3(\Sigma x_iy_i^2) + a_4(\Sigma x_iy_i) + a_5(\Sigma y_i^2) + a_6(\Sigma y_i) = \Sigma y_iz_i$$

$$a_1(\Sigma x_i^2) + a_2(\Sigma y_i^2) + a_3(\Sigma x_iy_i) + a_4(\Sigma x_i) + a_5(\Sigma y_i) + (n)a_6 = \Sigma z_i$$

Here all the summations are from $i = 1$ to n.
The matrix form of these equations is :

$$\begin{bmatrix} \Sigma x_i^4 & \Sigma x_i^2y_i^2 & \Sigma x_i^3y_i & \Sigma x_i^3 & \Sigma x_i^2y_i & \Sigma x_i^2 \\ \Sigma x_i^2y_i^2 & \Sigma y_i^4 & \Sigma x_iy_i^3 & \Sigma x_iy_i^2 & \Sigma y_i^3 & \Sigma y_i^2 \\ \Sigma x_i^3y_i & \Sigma x_iy_i^3 & \Sigma x_i^2y_i^2 & \Sigma x_i^2y_i & \Sigma x_iy_i^2 & \Sigma x_iy_i \\ \Sigma x_i^3 & \Sigma x_iy_i^2 & \Sigma x_i^2y_i & \Sigma x_i^2 & \Sigma x_iy_i & \Sigma x_i \\ \Sigma x_i^2y_i & \Sigma y_i^3 & \Sigma x_iy_i^2 & \Sigma x_iy_i & \Sigma y_i^2 & \Sigma y_i \\ \Sigma x_i^2 & \Sigma y_i^2 & \Sigma x_iy_i & \Sigma x_i & \Sigma y_i & n \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \begin{bmatrix} \Sigma x_i^2z_i \\ \Sigma y_i^2z_i \\ \Sigma x_iy_iz_i \\ \Sigma x_iz_i \\ \Sigma y_iz_i \\ \Sigma z_i \end{bmatrix}$$

In vector form :

$$a_1\cdot\begin{bmatrix} \Sigma x_i^4 \\ \Sigma x_i^2y_i^2 \\ \Sigma x_i^3y_i \\ \Sigma x_i^3 \\ \Sigma x_i^2y_i \\ \Sigma x_i^2 \end{bmatrix} + a_2\cdot\begin{bmatrix} \Sigma x_i^2y_i^2 \\ \Sigma y_i^4 \\ \Sigma x_iy_i^3 \\ \Sigma x_iy_i^2 \\ \Sigma y_i^3 \\ \Sigma y_i^2 \end{bmatrix} + a_3\cdot\begin{bmatrix} \Sigma x_i^3y_i \\ \Sigma x_iy_i^3 \\ \Sigma x_i^2y_i^2 \\ \Sigma x_i^2y_i \\ \Sigma x_iy_i^2 \\ \Sigma x_iy_i \end{bmatrix} + a_4\cdot\begin{bmatrix} \Sigma x_i^3 \\ \Sigma x_iy_i^2 \\ \Sigma x_i^2y_i \\ \Sigma x_i^2 \\ \Sigma x_iy_i \\ \Sigma x_i \end{bmatrix} + a_5\cdot\begin{bmatrix} \Sigma x_i^2y_i \\ \Sigma y_i^3 \\ \Sigma x_iy_i^2 \\ \Sigma x_iy_i \\ \Sigma y_i^2 \\ \Sigma y_i \end{bmatrix} + a_6\cdot\begin{bmatrix} \Sigma x_i^2 \\ \Sigma y_i^2 \\ \Sigma x_iy_i \\ \Sigma x_i \\ \Sigma y_i \\ n \end{bmatrix} = \begin{bmatrix} \Sigma x_i^2z_i \\ \Sigma y_i^2z_i \\ \Sigma x_iy_iz_i \\ \Sigma x_iz_i \\ \Sigma y_iz_i \\ \Sigma z_i \end{bmatrix}$$

3. The predicted equation of plane is :

$$z = (10.002)x + (19.998)y + 29.951$$

The predicted noise variance is **23.057**, every value is truncated to 3 decimal digits.
To run the code just use command `run("Q3.m")`, this will print all the above values. The script `Q3.m` is present inside the **Problem3** folder.

4 Problem 4

We have extensively seen parametric PDF estimation in class via maximum likelihood. In many situations, the family of the PDF is however unknown. Estimation under such a scenario is called nonparametric density estimation. We have studied one such technique in class, namely histogramming, and we also analyzed its rate of convergence. There is another popular technique for nonparametric density estimation. It is called KDE or Kernel density esitmation, the formula for which is given as $\hat{p}_n(x;\sigma) = \frac{\sum_{i=1}^{n} \exp\{(-(x-x_i)^2/(2\sigma^2))\}}{n\sigma\sqrt{2\pi}}$. Here $\hat{p_n}(x)$ is an estimate of the underlying probability density at value x, $\{x_i\}_{i=1}^{n}$ are the n samples values, from which the unknown PDF is being estimated, and σ is a bandwidth parameter (similar to a histogram bin-width parameter). The choice of the appropriate σ is not very straightforward. We will implement one possible procedure to choose σ - called cross-validation. For this, do as follows:

1. Use MATLAB to draw $n = 1000$ independent samples from $\mathcal{N}(0, 16)$. We will use a random subset of 750 samples (set T) for building the PDF, and the remaining 250 as the validation set V. Note that T and V must be disjoint sets.

2. In your report, write down an expression for the joint likelihood of the samples in V, based on the estimate of the PDF built from T with bandwidth parameter σ. [3 points]

3. For different values of σ from the set $\{0.001, 0.1, 0.2, 0.9, 1, 2, 3, 5, 10, 20, 100\}$, write MATLAB code to evaluate the log of the joint likelihood LL of the samples in V, based on the estimate of the PDF built from T. Plot of a graph of LL versus $\log\sigma$ and include it in your report. In the report, state which value of σ yielded the best LL value, and print it via your code as well. This procedure is called cross-validation. For this best sigma, plot a graph of $\hat{p}_n(x;\sigma)$ for $x \in [-8 : 0.1 : 8]$ and overlay the graph of the true density on it, for the same values of x. Include this plot in your report. [7 points]

4. In this experiment, we know the ground truth pdf which we shall denote as $p(x)$. So we can peek into it, in order to choose the best σ. This is impractical in actual experiments, but for now it will serve as a method of comparison. For each σ, write MATLAB code to evaluate $D = \sum_{x_i \in V}(p(x_i) - \hat{p}_n(x_i;\sigma))^2$. Plot of a graph of D versus $\log\sigma$ and include it in the report. In the report, state which value of σ yielded the best D value, and also what was the D value for the σ parameter which yielded the best LL. For this best sigma, plot a graph of $\hat{p}_n(x;\sigma)$ for $x \in [-8 : 0.1 : 8]$ and overlay the graph of the true density on it, for the same values of x. Include this plot in your report. [7 points]

5. Now, suppose the set T and V were equal to each other. What happens to the cross-validation procedure, and why? Explain in the report. [4+4=8 points]

Solution :

1. The code to generate T and V sets is in the script named `Script1.m` under the folder **Problem4**. On running the command `run("Script1.m")`, the two disjoint sets T and V with 750 and 250 samples are generated.

2. Suppose $y_1, y_2, ..., y_{250}$ are the independent(as mentioned in question) samples present in the validation set V and $x_1, x_2, ..., x_{750}$ be the samples .
   The PDF built from T with bandwidth parameter σ is given by,

   $$\hat{p}_{750}(x;\sigma) = \frac{\sum_{i=1}^{750} \exp\{(-(x-x_i)^2/(2\sigma^2))\}}{750\sigma\sqrt{2\pi}}$$

   The joint likelihood for the samples in V is,

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_{250} = y_{250}) = P(Y_1 = y_1)P(Y_2 = y_2)...P(Y_{250} = y_{250})$$
$$[\because Y_1, Y_2, ..., Y_{250} \text{ are independent}]$$

$$= \hat{p}_{750}(y_1; \sigma) \cdot \hat{p}_{750}(y_2; \sigma) \cdots \hat{p}_{750}(y_{250}; \sigma) = \prod_{j=1}^{250} \frac{\sum_{i=1}^{750} \exp\{(-(y_j - x_i)^2/(2\sigma^2))\}}{750\sigma\sqrt{2\pi}}$$

$$= \frac{1}{(750\sigma\sqrt{2\pi})^{250}} \prod_{j=1}^{250} \left(\sum_{i=1}^{750} \exp\{(-(y_j - x_i)^2/(2\sigma^2))\} \right)$$

3. The code for plotting the graph of LL versus $\log\sigma$ is present in the script `Script1.m`. On running the file, it generates the image of the plot, which is saved under the name `LL_log(sigma).png`. The code prints the value of the best LL value as well as the corresponding σ value. The script `EstimatedPDF.m` under **Problem4** folder is used to calculate the estimated PDF value built from $T : \hat{p}_{750}(x; \sigma)$ for any σ.

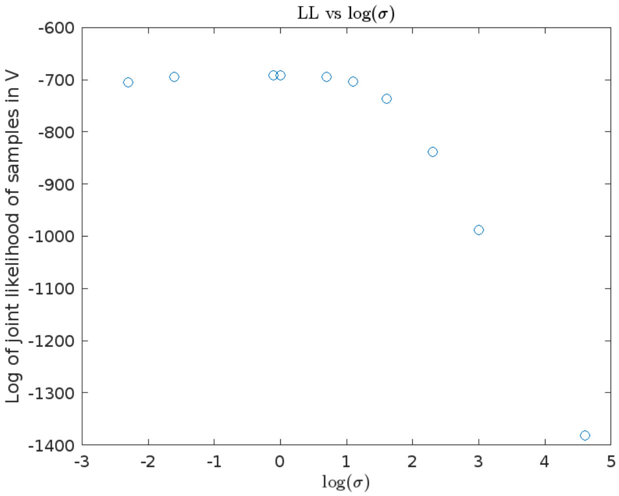

Figure 2: Plots of LL versus $\log\sigma$

The value of σ which generated the best LL value was mostly 1 or 0.9. Rarely it took the value 2. In the above image maxima occurs at $\log\sigma \approx 0 \implies \sigma = 0.9$.

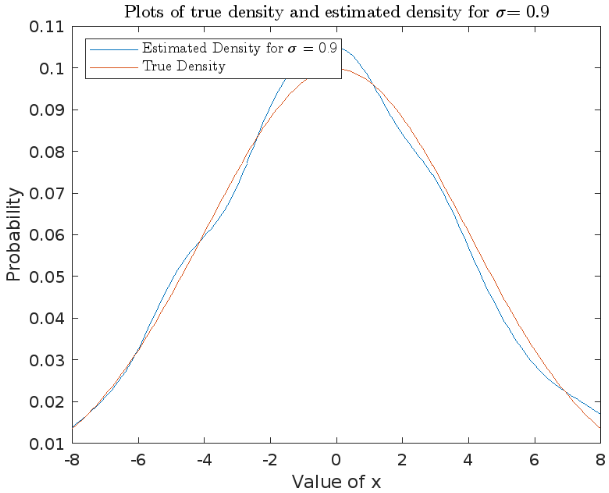

Figure 3: Plots of True density, Estimated density for $\sigma = 0.9$(Cross Validation)

The second plot is also obtained on running the command `run("Script1.m")` and is saved under the name `True_EstimatedPDF.png` in the folder **Problem4**.

4. The plot of D versus logσ is obtained on running the command `run("Script1.m")` and is saved under the name `D_vs_log(sigma).png` in the folder **Problem4**.

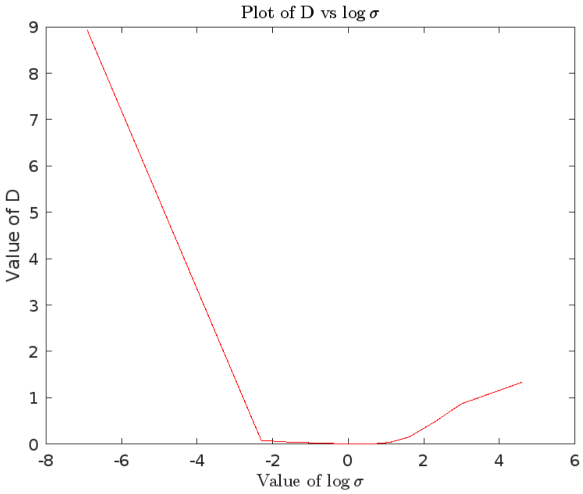

Figure 4: Plots of D versus logσ

$\sigma = 1$ yielded the best D value (0.002323). The D value for the σ parameter which yielded the best LL value (i.e $\sigma = 0.9$) is 0.0030902.

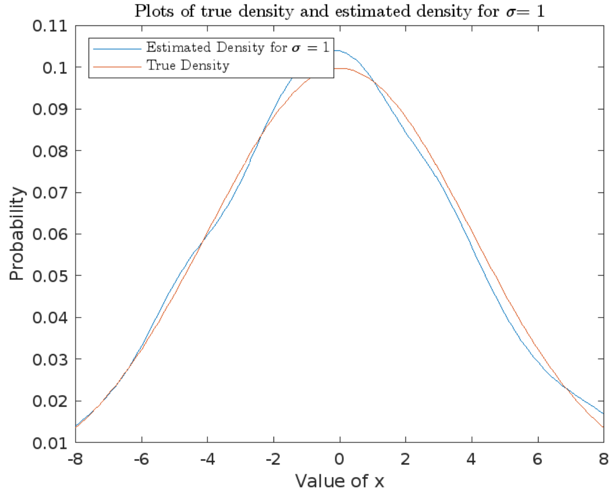

Figure 5: Plots of True density , Estimated density for $\sigma = 1$

The fourth plot is also obtained on running the command `run("Script1.m")` and is saved under the name `True_EstimatedPDF2.png` in the folder **Problem4**.

5. If the sets T and V were equal to each other,

- It would lead to over-fitting. It will choose the σ which would be the best for the given set, but may not give the correct PDF estimate for other unseen data.
- We cannot assess how well our estimated PDF is working with unseen data, as we have been validating it with the same data.

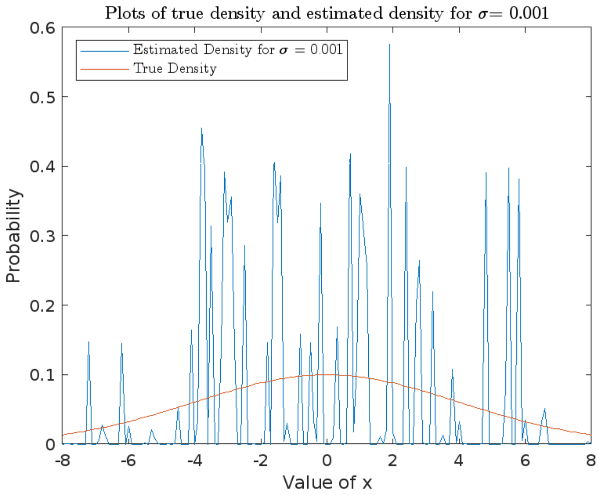

Figure 6: Large variation between the PDF of unseen data ($x \in [-8 : 0.1 : 8]$ and the true PDF

From Figure 6 we can that the estimated PDF gives a lot of incorrect values. This implies that σ which gave the best LL value in this case is incorrect and hence gives bad results. The above plot was obtained by equating both the T and V sets.

- Another observation is that the smallest value of σ is chosen. The LL of T wrt T does not have a maxima. This is because $LL \rightarrow \infty$ as $\sigma \rightarrow 0$. This is the reason why $\sigma = 0.001$ was chosen in this case.
  $\forall x, \exists k < 750 \ni x - x_k = 0 \quad [\because T = V]$
  $\therefore \exp\big\{(-(x - x_i)^2/(2\sigma^2))\big\} = 1 \quad$ for $i = k$

  $\exp\big\{(-(x - x_i)^2/(2\sigma^2))\big\} \rightarrow 0 \quad$ as $\sigma \rightarrow 0$

  $\therefore \hat{p}_{750}(x; \sigma) \rightarrow \infty \quad$ as $\sigma \rightarrow 0$ for $i \neq k$

  $\implies LL \rightarrow \infty$

  Hence there is no maximum likelihood value for LL when $T = V$, due to which the cross validation procedure fails and gives the wrong σ value at which the LL is not defined.

5 Problem 5

Let X be a real-valued random variable whose values lie from a to b always, where $a < b$. Then consider an intermediate result (called IR) that $E[e^{s(X-E[X])}] \leq e^{s^2(b-a)^2/8}$ where $s > 0$. Now, let $X_1, X_2, ..., X_n$ be independent random variables such for every i, we have X_i always lies in $[a_i, b_i]$ where $a_i < b_i$. Let $S_n = \sum_{i=1}^{n} X_i$. Derive an upper bound on $P(S_n - E[S_n] > t)$ in terms of a_i, b_i, t using Markov's inequality and IR, and upon suitable elimination of s. Notice that IR is an upper bound on the moment generating function of random variable X with bounded values. We will now proceed to prove IR as follows:

1. Without loss of generality, we consider $E(X) = 0$, because X can be replaced by $X - E(X)$ anyways. Hence, we consider $a \leq 0 \leq b$. The function e^{sx} is a convex function of x, and hence a line segment joining two distinct points of the graph always lies above the graph of the function between the two points. Hence $e^{sx} \leq \dfrac{(b-x)e^{sa}}{b-a} + \dfrac{(x-a)e^{sb}}{b-a}$.

2. Taking, expectation on both sides, prove that $E(e^{sx}) \leq e^{L(s(b-a))}$ where $L(h) = \dfrac{ha}{b-a} + \log\big(1 + (a - ae^h)/(b-a)\big)$.

3. Using Taylor's expansion and the result that $(x+y)/2 \geq \sqrt{xy}$ for real-valued x, y, prove that $L''(h) \leq 1/4$ for all real-valued h.

4. Hence, conclude the proof (write the final, now somewhat obvious step).

[5 + (1+3+1) = 10 points]

Solution :

2. We know from statement (a) that :

$$E(x) = 0$$

$$e^{sx} \leq \frac{(b-x)e^{sa}}{b-a} + \frac{(x-a)e^{sb}}{b-a} = e^{sa}\left[\frac{b - ae^{s(b-a)}}{b-a} + \frac{x \cdot (e^{s(b-a)} - 1)}{b-a}\right]$$

Taking expectation on both sides :

$$E(e^{sx}) \leq E\left(e^{sa}\left[\frac{b - ae^{s(b-a)}}{b-a} + \frac{x \cdot (e^{s(b-a)} - 1)}{b-a}\right]\right)$$

$$E(e^{sx}) \leq e^{sa}\left[\frac{b - ae^{s(b-a)}}{b-a} + \frac{E(x) \cdot (e^{s(b-a)} - 1)}{b-a}\right]$$

As $E(x) = 0$:

$$E(e^{sx}) \leq e^{sa}\left[1 + \frac{a - ae^{s(b-a)}}{b-a}\right]$$

$$E(e^{sx}) \leq e^{\frac{s(b-a)a}{b-a} + \log\left(1 + (a - ae^{s(b-a)})/(b-a)\right)}$$

$$E(e^{sx}) \leq e^{L(s(b-a))}$$

where $L(h) = \dfrac{ha}{b-a} + \log\big(1 + (a - ae^h)/(b-a)\big)$

3. We have to prove that $L''(h) \leq 1/4$.

$$L'(h) = \frac{a}{b-a} + \frac{ae^h}{ae^h - b}$$

$$L''(h) = \frac{b \cdot (-ae^h)}{(b + (-ae^h))^2}$$

We know that $(x+y)/2 \geq \sqrt{xy}$ for $x \geq 0$ and $y \geq 0$. Also from part (a), $a < 0$ and $b > 0$, also $e^h > 0 \quad \forall h$. So $(-ae^h) > 0$ and $b > 0$,

$$\frac{b + (-ae^h)}{2} \geq \sqrt{b \cdot (-ae^h)}$$

$$\frac{1}{2} \geq \frac{\sqrt{b \cdot (-ae^h)}}{b + (-ae^h)}$$

$$\frac{b \cdot (-ae^h)}{(b + (-ae^h))^2} \leq 1/4$$

$$L''(h) \leq 1/4$$

4. For any h, as $L''(h) \leq 1/4 \; \forall h$.

For any $h \geq 0$ we can say that :

$$L'(h) = \int_0^h \big(L''(t)\big)\,dt + L'(0) \leq \int_0^h \left(\frac{1}{4}\right) dt + 0 = \frac{h}{4} \qquad \forall h \geq 0$$

$$L'(h) \leq \frac{h}{4} \qquad \forall h \geq 0$$

$$L(h) = \int_0^h \big(L'(t)\big)\,dt + L(0) \leq \int_0^h \left(\frac{t}{4}\right) dt + 0 = \frac{h^2}{8} \qquad \forall h \geq 0$$

$$L(h) \leq \frac{h^2}{8} \qquad \forall h \geq 0$$

And for $h < 0$,

$$L'(0) = \int_h^0 \big(L''(t)\big)\,dt + L'(h) \leq \int_h^0 \left(\frac{1}{4}\right) dt + L'(h) = L'(h) - \frac{h}{4} \qquad \forall h < 0$$

$$L'(h) \geq \frac{h}{4} \qquad \forall h < 0$$

$$L(0) = \int_h^0 L'(t)dt + L(h) \geq \int_h^0 \left(\frac{t}{4}\right) dt + L(h) = L(h) - \frac{h^2}{8} \qquad \forall h < 0$$

$$L(h) \leq \frac{h^2}{8} \qquad \forall h < 0$$

Therefore we can say that,

$$\therefore L(h) \leq \frac{h^2}{8} \qquad \forall h$$

$$\therefore e^{L(h)} \leq e^{\frac{h^2}{8}}$$

$$e^{L(s(b-a))} \leq e^{\frac{s^2(b-a)^2}{8}}$$

As $E(e^{sx}) \leq e^{L(s(b-a))}$,

$$E(e^{sx}) \leq e^{\frac{s^2(b-a)^2}{8}}$$

Now we know $S_n = \Sigma_{i=1}^{n} X_i$ and $a_i \leq X_i \leq b_i$,

$$\therefore \Sigma_{i=1}^{n} a_i \leq S_n \leq \Sigma_{i=1}^{n} b_i$$

Let $u = (\Sigma_{i=1}^{n} b_i - \Sigma_{i=1}^{n} a_i)^2$ Also if S_n is a random variable then $e^{s(S_n - E[S_n])}$ is also a random variable (where s is a constant), then by Markov's inequality and IR :

$$P(S_n - E[S_n] > t) = P(e^{s(S_n - E[S_n])} > e^{st}) < \frac{E[e^{s(S_n - E[S_n])}]}{e^{st}} \leq \frac{e^{s^2(\Sigma_{i=1}^{n} b_i - \Sigma_{i=1}^{n} a_i)^2/8}}{e^{st}}$$

$$P(S_n - E[S_n] > t) < e^{(s^2u/8) - st}$$

For the tightest possible bound we will minimize the expression on the right w.r.t. s,

$$\frac{\partial e^{(s^2u/8) - st}}{\partial s} = 0$$

$$su/4 - t = 0$$

$$s = 4t/u$$

The tightest bound is $e^{-2t^2/u}$ at $s = 4t/u$.

$$\therefore P(S_n - E[S_n] > t) < e^{\frac{-2t^2}{(\Sigma(b_i - a_i))^2}}$$

where the summation is from $i = 1$ to n.