(Q.1) ~~X_1 is~~

(a) X_1 is the number of times you have to pick a book so that you move from 0 distinct books to 1 distinct book.

Hence, The first book you pick will be distinct.

So, $X_1 = 1$.

Now, everytime you pick a book with replacement, it is independent of earlier events.

$$\text{Hence, } P_{req} = \frac{\text{no. of books with colors distinct from chosen ones}}{\text{Total no. of books}}$$

$$= \frac{n-(i-1)}{n} = \frac{n+1-i}{n}$$

$$\Rightarrow \boxed{P_{req} = \frac{n+1-i}{n}}$$

(b) The parameter of a geometric random variable is the probability of success in a trial.

~~Here~~, let p_i be the parameter of random variable X_i.

So, p_i = probability of success of X_i

= probability that you choose a book of color different from the $i-1$ distinct colors ~~o~~ books already chosen

$= \frac{n+1-i}{n}$.

$$\Rightarrow \boxed{p_i = \frac{n+1-i}{n}}$$

$$\text{So, } P(X_i = k) = (1-p_i)^{k-1} p_i = \left(1 - \frac{n+1-i}{n}\right)^{k-1} \frac{n+1-i}{n}$$

$$= \left(\frac{i-1}{n}\right)^{k-1} \left(\frac{n+1-i}{n}\right)$$

(c) Let Z be a geometric random variable with parameter p.

So, $P(Z=k) = (1-p)^{k-1}p$; $k = 1, 2, 3, \dots$

Hence, $E(Z) = \sum_{k=1}^{\infty} (\text{value of } Z=k) \times (1-p)^{k-1} \cdot p = \sum_{k=1}^{\infty} k \cdot P(Z=k)$

$= \sum_{k=1}^{\infty} k(1-p)^{k-1} p = S_p$ (say)

(Note that $Z=k \Rightarrow$ first success comes on k^{th} trial for Z)

Now, $\sum_{k=1}^{\infty} k(1-p)^{k-1}p = p + 2(1-p)p + 3(1-p)^2 p + \dots$

is an arithmetic - geometric progression

$S_p = p + 2p(1-p) + 3p(1-p)^2 + \dots$

$S_p(1-p) = \quad p(1-p) + 2p(1-p)^2 + \dots$

$\Rightarrow S_p - S_p(1-p) = p + p(1-p) + p(1-p)^2 + \dots - \infty(1-p)^{\infty} p$ → is 0 ∵ (The value of the expression $\lim_{m\to\infty} m \times r^m$; $r<1$

$= \lim_{m\to\infty} \frac{m}{s^m}$; $s = \frac{1}{r} > 1$

$= \lim_{m\to\infty} \frac{1}{m s^{m-1}} = 0$)

(Applied L'Hospital rule here as denominator → ∞)

$S_p \times p = \frac{p}{1-(1-p)} = \frac{p}{p} = 1$

$\Rightarrow \boxed{S_p = \frac{1}{p}} \Rightarrow \boxed{E(Z) = \frac{1}{p}}$

Now, $Var(Z) = E(Z^2) - (E(Z))^2 = E(Z^2) - \frac{1}{p^2}$ — ①

Now, $E(Z^2) = \sum_{k=1}^{\infty} k^2 (1-p)^{k-1} p = \sum_{k=1}^{\infty} k^2 \cdot P(Z=k)$

$= F_p$

$\Rightarrow F_p = p + 4p(1-p) + 9p(1-p)^2 + 16p(1-p)^3 + \dots$

$F_p(1-p) = \quad p(1-p) + 4p(1-p)^2 + 9p(1-p)^3 + \dots$

$\Rightarrow F_p - F_p(1-p) = p + 3p(1-p) + 5p(1-p)^2 + 7p(1-p)^3 + \dots$

$\Rightarrow F_p \times p = p + 3p(1-p) + 5p(1-p)^2 + 7p(1-p)^3 + \dots$

$= G_p$ — ②

$$\Rightarrow G_p = p + 3p(1-p) + 5p(1-p)^2 + 7p(1-p)^3 + \dots$$

$$G_p(1-p) = p(1-p) + 3p(1-p)^2 + 5p(1-p)^3 + \dots$$

$$\Rightarrow G_p - G_p(1-p) = p + 2p(1-p) + 2p(1-p)^2 + 2p(1-p)^3 + \dots$$

$$pG_p = p + \frac{2p(1-p)}{1-(1-p)} = p + \frac{2\cancel{p}(1-p)}{\cancel{p}} = 2-p$$

$$\Rightarrow G_p = \frac{2-p}{p} \quad - (3)$$

Using eq^n (3) in eq^n (2),

$$pF_p = G_p = \frac{2-p}{p} \Rightarrow F_p = \frac{2-p}{p^2} \Rightarrow E(Z^2) = \frac{2-p}{p^2}$$

$$\Rightarrow Var(Z) = \frac{2-p}{p^2} - \frac{1}{p^2} \Rightarrow \boxed{Var(Z) = \frac{1-p}{p^2}}$$

(d) $E(X^{(n)}) = E(X_1 + X_2 + \dots + X_{n-1} + X_n)$

$= E(X_1) + E(X_2) + \dots + E(X_{n-1}) + E(X_n)$ (always applicable)

$= \frac{1}{p_1} + \frac{1}{p_2} + \dots + \frac{1}{p_n}$

$= \sum_{i=1}^{n} \frac{1}{p_i} = \sum_{i=1}^{n} \frac{1}{\left(\frac{n+1-i}{n}\right)} = \sum_{i=1}^{n} \frac{n}{n+1-i}$

$$\Rightarrow E(X^{(n)}) = n\left(\sum_{i=1}^{n} \frac{1}{n+1-i}\right) = n\left(\frac{1}{n} + \frac{1}{n-1} + \dots + \frac{1}{2} + \frac{1}{1}\right)$$

$$\Rightarrow \boxed{E(X^{(n)}) = n\left(1 + \frac{1}{2} + \dots + \frac{1}{n-1} + \frac{1}{n}\right)}$$

(e) $Var(X^{(n)}) = Var(X_1 + X_2 + \dots + X_{n-1} + X_n)$

$= Var(X_1) + Var(X_2) + \dots + Var(X_n)$ (I)

($\because$ The value of X_i, i.e. no of times, a book is picked to change from $i-1$ distinct colors to i distinct colors does not have any effect on value of X_j which has similar analogy, for $i \neq j$, X_i and X_j are independent $\forall\ i \neq j$).

$$\Rightarrow Var(X^{(n)}) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \frac{1-p_i}{p_i^2}$$

$$= \sum_{i=1}^{n} \frac{1-\left(\frac{n+1-i}{n}\right)}{\frac{(n+1-i)^2}{n^2}} = \sum_{i=1}^{n} \frac{(n-(n+1-i))n}{(n+1-i)^2} = \sum_{i=1}^{n} \frac{(i-1)n}{(n+1-i)^2}$$

$$\Rightarrow Var(X^{(n)}) = n\left(\sum_{i=1}^{n} \frac{(i-1)}{(n+1-i)^2}\right)$$

$$= n\left(\frac{0}{(n)^2} + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + \dots + \frac{n-1}{(1)^2}\right)$$

$$\leq n\left(\frac{n}{n^2} + \frac{n}{(n-1)^2} + \frac{n}{(n-2)^2} + \dots + \frac{n}{(1)^2}\right)$$

$$= n \times n\left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + \frac{1}{(n-2)^2} + \dots + \frac{1}{(1)^2}\right)$$

$$= n^2\left(\frac{1}{1^2} + \frac{1}{2^2} + \dots + \frac{1}{(n-1)^2} + \frac{1}{n^2}\right)$$

$$\leq n^2 \times \frac{\pi^2}{6} \left(\because \sum_{j=1}^{n} \frac{1}{j^2} \leq \frac{\pi^2}{6}\right)$$

$$\Rightarrow Var(X^{(n)}) < \frac{n^2\pi^2}{6}.$$

(f) Here, we know that $\int_1^n \frac{1}{x}\,dx = \ln(n)$.

we also see that :- $n \sum_{x=1}^{n} \frac{1}{x+1} \leq n \int_1^n \frac{1}{x}\,dx = n\ln(n) \leq n \sum_{x=1}^{n} \frac{1}{x}$

$$\Rightarrow E(X^{(n)}) = n\sum_{x=1}^{n}\frac{1}{x} \Rightarrow \boxed{n\ln(n) \leq E(X^{(n)})} \Rightarrow E(X^{(n)}) = \Omega(n\ln n)$$

Now, $n\sum_{x=1}^{n}\frac{1}{x+1} = n\left(\frac{E(X^{(n)})}{n} - 1 + \frac{1}{n+1}\right) \leq n\ln(n)$

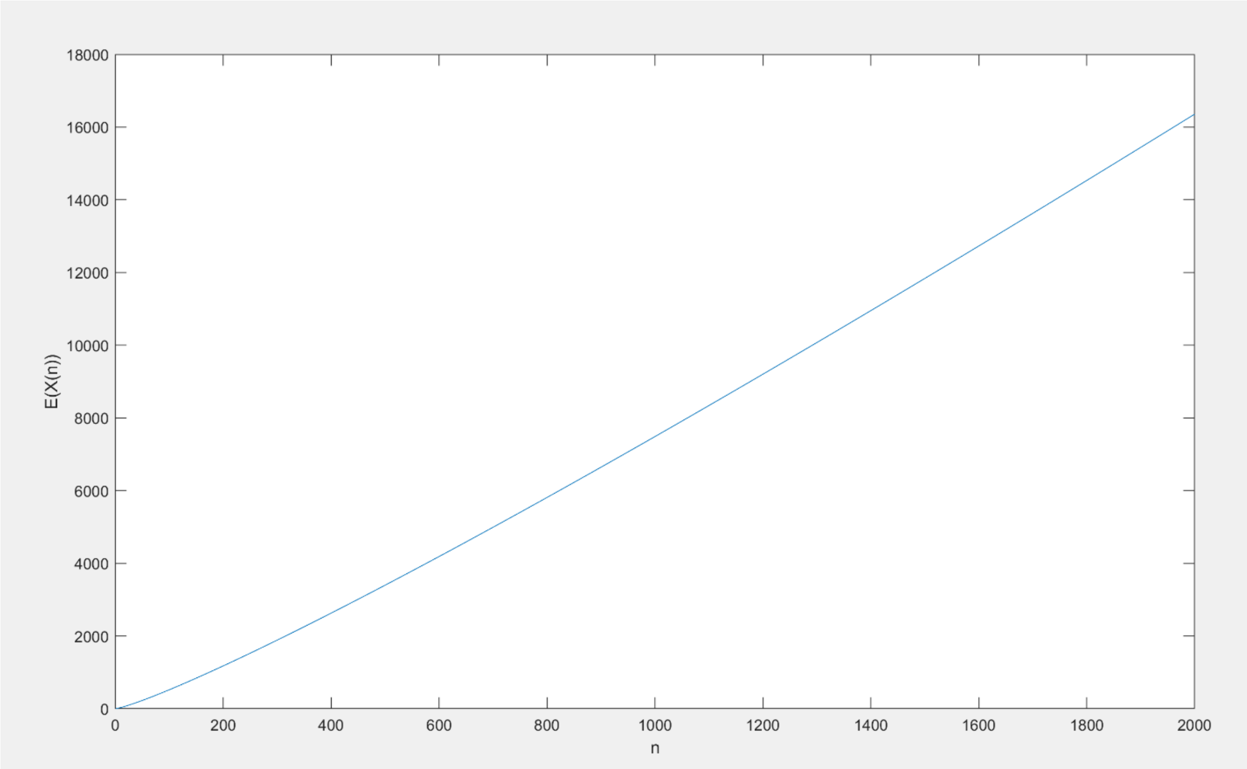
$$\Rightarrow E(X^{(n)}) - n + \frac{n}{n+1} \leq n\ln(n) \Rightarrow E(X^{(n)}) \leq n\ln(n) + \frac{n^2}{n+1}$$

$$\Rightarrow E(X^{(n)}) = O(n\ln n)$$

$\Rightarrow$ as $E(X^{(n)}) = \Omega(n\ln n)$ & $E(X^{(n)}) = O(n\ln n)$, $E(X^{(n)}) = \Theta(n\ln(n))$

$$\Rightarrow \boxed{f(n) = n\ln(n)}$$

Plot for Question 1(f)



Karan Agarwalla-180050045 Parth Laturia-180050071

Question 2

(Q.2)

(a) Let $v_i = F^{-1}(u_i)$

$\Rightarrow F_{v_i}(\alpha) = P(v_i \le \alpha)$

CDF of v_i

$= P(F^{-1}(u_i) \le \alpha)$

$= P(u_i \le F(\alpha))$ (∵ F is an increasing func always)

$= \int_0^{F(\alpha)} dt$ (∵ u_i is chosen from $[0,1]$ uniform distribution)

$= F(\alpha)$

$\Rightarrow F_{v_i}(\alpha) = F(\alpha)$

$\Rightarrow$ v follows the prob. distribution $F(x)$

$\Rightarrow$ CDF of $v = F(x)$

$\Rightarrow$ QED

(b) From the previous part, to obtain the distribution $Y_1, Y_2, \ldots, Y_{n-1}, Y_n$, we can choose the values $b_1, b_2, \ldots, b_n$ from $(0,1)$ uniformly and apply $F^{-1}(x)$ on each of v_i. (i.e. replace u_i & v_i of previous part by b_i and Y_i respectively)

Hence, there must exist $\{b_i\}_{i=1}^{n}$ such that

$F^{-1}(b_i) = y_i$

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

$$= \max_x \left| \sum_{i=1}^{n} \frac{1(Y_i \le x)}{n} - F(x) \right|$$

$$\cancel{= \max_x \left| \sum_{i=1}^{n} \frac{1[F^{-1}(b_i) \le x]}{n} - F(x) \right|}$$

$$= \max_x \left\{ \sum_{i=1}^{n} \frac{1[F^{-1}(b_i) \le x]}{n} - F(x) \right\}$$

$$= \max_x \left| \frac{\sum_{i=1}^{n} 1(b_i \le F(x))}{n} - F(x) \right|$$

let $F(x) = y$

The expression then becomes :-

$$D = \max_y \left| \sum_{i=1}^{n} \frac{1(b_i \le y)}{n} - y \right| \qquad y \in [0,1]$$

Note that D is a random variable as $\{b_i\}_{i=1}^{n}$ is a random variable. Since each of b_i and U_i have same distributions we conclude

$$(\text{CDF of } D)(x) = (\text{CDF of } E)(x)$$

$$\Rightarrow p(D \ge d) = p(E \ge d)$$

$$\Rightarrow P(E \ge d) = p(D \ge d)$$

With this result, we can predict the precision of D. We know that as we take larger and larger values of n, $\sum_{i=1}^{n} \frac{1(U_i \le y)}{n}$ approaches closer and closer to y

$$\because E\left(\sum_{i=1}^{n} \frac{1(U_i \le y)}{n}\right) = y.$$

Therefore, we observe that E has a greater tendency to be 0. So, both $P(E \ge d)$ and $P(D \ge d)$ become smaller with increasing n.

Here, the given result signifies the maximum error in our empirical cdf telling what is the maximum deviation from correct value.
Also, $P(D \geq d)$ signifies the max. deviation of D to exceed d.
By above arguement, as n increases, eventually this probability $P(D \geq d)$ will decrease towards 0 for a fixed d and hence our empirical CDF becomes even more accurate.

Question 3

Question 3

$$z_i = ax_i + by_i + c + u_i$$

where $u_i \sim N(0, \sigma^2)$

estimation of a, b, c, σ^2 given x_i, y_i (accurate) and z_i (inaccurate)

$$z_i \sim N(ax_i + by_i + c, \sigma^2)$$

$$f_Z(z; a, b, c, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\mu)^2/2\sigma^2}$$

$$\mu = ax + by + c$$

Taking joint pdf of $Z_1 = z_1, Z_2 = z_2, \ldots$ will be

$$f_{Z_1, Z_2, Z_3, \ldots, Z_n}(z_1, z_2, \ldots, z_n; a, b, c, \sigma) = \prod_{i=1}^{n} \frac{\exp\left(-\frac{(z_i - (ax_i + by_i + c))^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}$$

$$LL = \sum_{i=1}^{n} -\log(\sqrt{2\pi}) - \log\sigma - \frac{(z_i - (ax_i + by_i + c))^2}{2\sigma^2}$$

$$\frac{\partial LL}{\partial a} = \sum_{i=1}^{n} \frac{-2(z_i - (ax_i + by_i + c))}{2\sigma^2}(-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i\left[(z_i - (ax_i + by_i + c))\right] = 0$$

$$\Rightarrow a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i y_i + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i z_i \quad \text{— (1)}$$

$$\frac{\partial LL}{\partial b} = \sum_{i=1}^{n} -\frac{2(z_i-(ax_i+by_i+c))}{2\sigma^2}(-y_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (z_i-(ax_i+by_i+c))\, y_i = 0$$

$$\Rightarrow a\sum_{i=1}^{n} x_i y_i + b\sum_{i=1}^{n} y_i^2 + c\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i z_i \quad \text{— (2)}$$

$$\frac{\partial LL}{\partial c} = \sum_{i=1}^{n} -\frac{2(z_i-(ax_i+by_i+c))(-1)}{2\sigma^2} = 0$$

$$\Rightarrow a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i + cn = \sum_{i=1}^{n} z_i \quad \text{— (3)}$$

Expressing (1), (2) and (3) in matrix form

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i & n \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i z_i \\ \sum_{i=1}^{n} y_i z_i \\ \sum_{i=1}^{n} z_i \end{pmatrix} \quad -(4)$$

$$AX = b \Rightarrow X = A^{-1}B.$$

The matrix equation is solved in Q3.m yielding the expected equation of the plane is $ax+by+c=z$ where

$a = 10.0022$

$b = 19.9980$

$c =$ ~~29.9516~~ 29.9516

The predicted eq. of plane $z = 10.0022x + 19.9980y + 29.9516$.

$$\frac{\partial LL}{\partial \sigma} = \sum_{i=1}^{n} \left[-\frac{1}{\sigma} + \frac{(z_i - (ax_i + by_i + c))^2}{2\sigma^3} 2 \right] = 0$$

$$\Rightarrow -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (z_i - (ax_i + by_i + c))^2}{\sigma^3} = 0$$

$$\Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n} (z_i - (ax_i + by_i + c))^2}{n}$$

From our code, we get

$$\sigma^2 = 23.0570$$

Using the fact that x_i and y_i are known exactly and the linearity of expectation, we can write ~~equation~~ matrix (4) in the following form,

$$\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum x_i y_i & \sum x_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i \\ \sum x_i & \sum y_i & n \end{pmatrix} \begin{pmatrix} E(a) \\ E(b) \\ E(c) \end{pmatrix} = \begin{pmatrix} E(\sum x_i z_i) \\ E(\sum y_i z_i) \\ E(\sum z_i) \end{pmatrix}$$

$$= \begin{pmatrix} E(\sum x_i (ax_i + by_i + c)) \\ E(\sum y_i (ax_i + by_i + c)) \\ E(\sum (ax_i + by_i + c)) \end{pmatrix}$$

$$= \begin{pmatrix} a\sum x_i^2 + b\sum x_i y_i + c\sum x_i \\ a\sum x_i y_i + b\sum y_i^2 + c\sum y_i \\ a\sum x_i + b\sum y_i + cn \end{pmatrix}$$

$$\therefore \; A\begin{pmatrix} E(a) \\ E(b) \\ E(c) \end{pmatrix} = A\begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Given A is invertible,

$$E(a) = a, \; E(b) = b, \; E(c) = c$$

$$Var(z_i) = Var(ax_i + by_i + c + n_i)$$
$$= Var(n_i) = \sigma^2$$

Now ~~using~~ variance,

$$\begin{pmatrix} \left(\sum_{i=1}^{n} x_i^2\right)^2 & \left(\sum_{i=1}^{n} x_i y_i\right)^2 & \left(\sum_{i=1}^{n} x_i\right)^2 \\ \left(\sum_{i=1}^{n} x_i y_i\right)^2 & \left(\sum_{i=1}^{n} y_i^2\right)^2 & \left(\sum_{i=1}^{n} y_i\right)^2 \\ \left(\sum_{i=1}^{n} x_i\right)^2 & \left(\sum_{i=1}^{n} y_i\right)^2 & n^2 \end{pmatrix} \begin{pmatrix} Var(a) \\ Var(b) \\ Var(c) \end{pmatrix}$$

$$= \begin{pmatrix} Var\left(\sum_{i=1}^{n} x_i z_i\right) \\ Var\left(\sum_{i=1}^{n} y_i z_i\right) \\ Var\left(\sum_{i=1}^{n} z_i\right) \end{pmatrix} = \begin{pmatrix} \sigma^2 \sum_{i=1}^{n} x_i^2 \\ \sigma^2 \sum_{i=1}^{n} y_i^2 \\ n\sigma^2 \end{pmatrix}$$

$$\left(\text{Since } Var(ax) = a^2 Var(x)\right)$$

Now, $$M\begin{pmatrix} Var(a) \\ Var(b) \\ Var(c) \end{pmatrix} = \begin{pmatrix} \sigma^2 \sum x_i^2 \\ \sigma^2 \sum y_i^2 \\ n\sigma^2 \end{pmatrix}$$

$$\begin{pmatrix} Var(a) \\ Var(b) \\ Var(c) \end{pmatrix} = M^{-1} \begin{pmatrix} \sigma^2 \sum_{i=1}^{n} x_i^2 \\ \sigma^2 \sum_{i=1}^{n} y_i^2 \\ n\sigma^2 \end{pmatrix}$$

Question 4

Question 4

(b) given $\hat{P}_n(x;\sigma) = \dfrac{\sum_{i=1}^{n} \exp\left(-(x_i - x)^2/2\sigma^2\right)}{n\sigma\sqrt{2\pi}}$

where $x \in V$ and $x_i \in T$

the joint likelihood of samples in V is given by

$$JL = \prod_{x_j \in V} \left[\frac{\sum_{i=1}^{n} \exp\left(-(x_j - x_i)^2/(2\sigma^2)\right)}{n\sigma\sqrt{2\pi}} \right]$$

where $x_i \in T$

$$LL = \sum_{x_j \in V} \log \left[\frac{\sum_{i=1}^{n} \exp\left(-(x_j - x_i)^2/(2\sigma^2)\right)}{n\sigma\sqrt{2\pi}} \right]$$

(c) best LL value is for minimum $\mathcal{LL}$ value

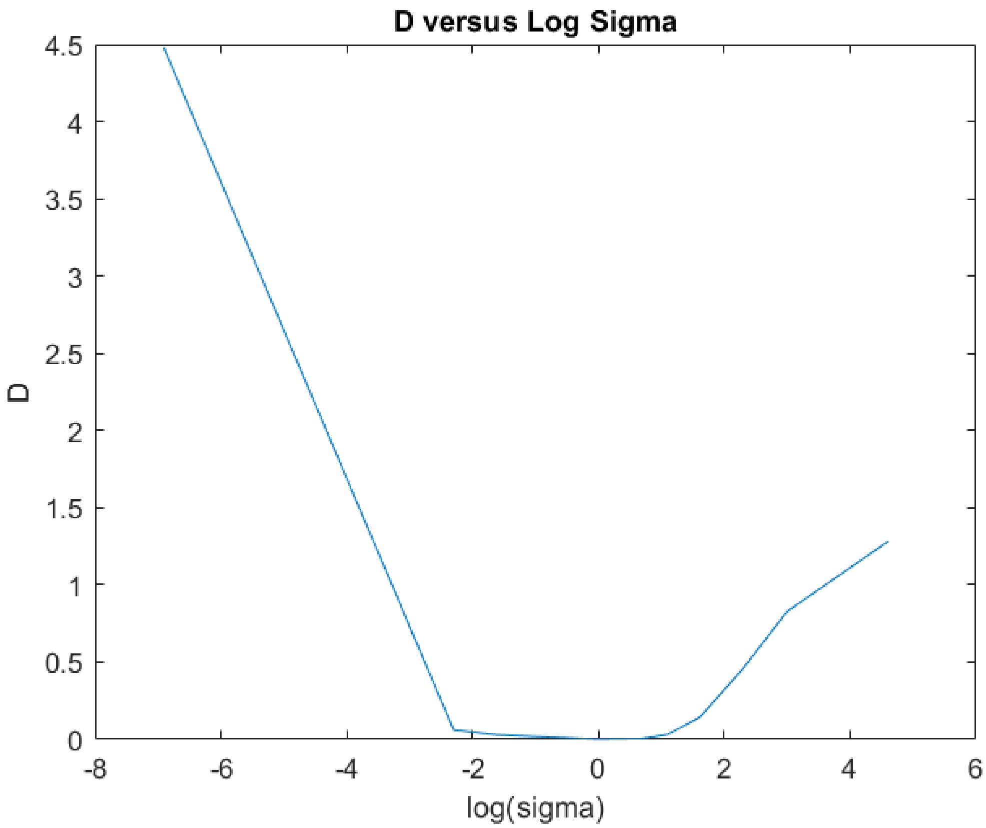
we get maximum LL value at $\sigma = 2$

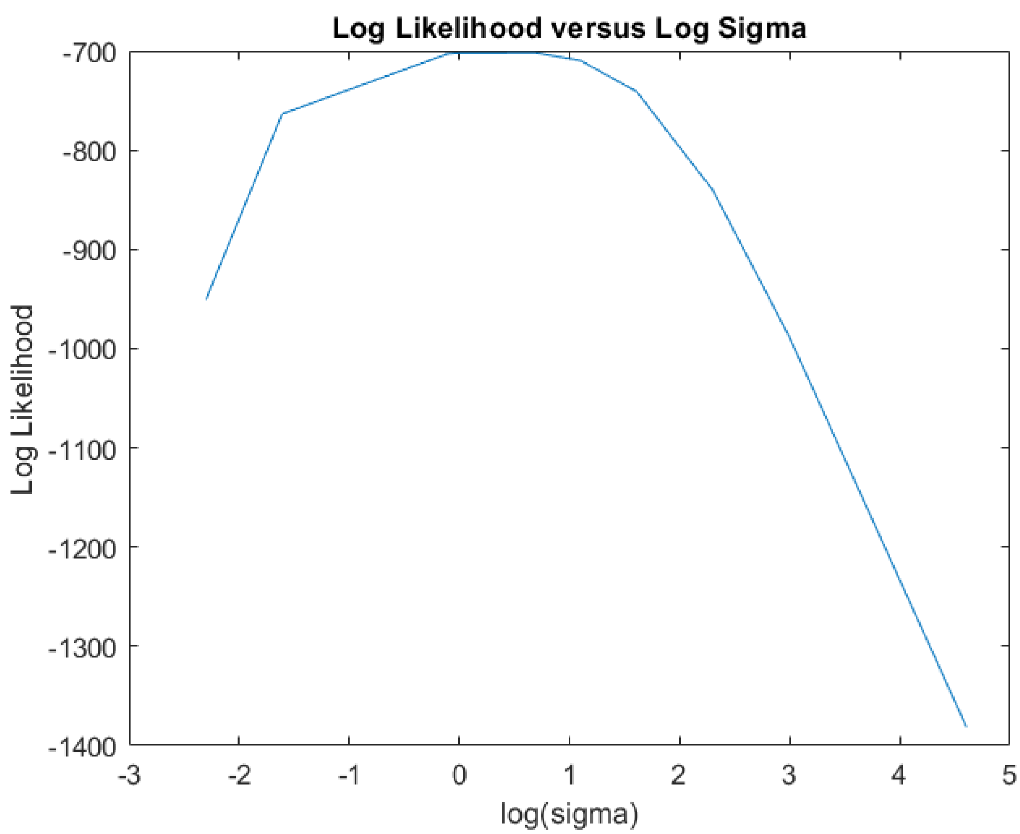
which is equal to -702.0721

(d) best value of D = minimum value of D

We get minimum value of D at $\sigma = 1$ which is 0.0026.

(e) If our validation set is same as training set, cross validation breaks down and may not yield correct answer. This is because the cross-validation procedure results in a value of σ that gives joint likelihood for the set T only, i.e., the chosen value of σ could yield poor joint likelihood s for some other sample set from the same distribution. This can be ~~attributed to the~~ seen as an equivalence of overfitting. In fact values of σ calculated from such a distribution will vary a lot or in other words has a very high variance. The $\hat{p}_n(x)$ calculated for such sample sets would be high since there are such $x = x_i$ where $e^{-[(x-x_i)^2/2\sigma^2]}$ (high value) This will predict our estimation formula to be much more accurate than it actually is and hence cross-validation will give erroneous answers. Hence ~~our~~ one must choose an independent sample set for cross-validation.


D versus Log Sigma
D
log(sigma)
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-8
-6
-4
-2
0
2
4
6


Log Likelihood versus Log Sigma
Log Likelihood
log(sigma)
-700
-800
-900
-1000
-1100
-1200
-1300
-1400
-3
-2
-1
0
1
2
3
4
5