

CS215: Data Analysis and Interpretation

Assignment 3

Anilesh Bansal, 22b0928
Arihant Vashista, 22b0958
Sanskar Shaurya, 22b0985

10 September 2023

Contents

1	Question 1	2
2	Question 2	4
3	Question 3	5
4	Question 4	6
5	Question 5	8

1 Question 1

(a)

Since no book has been picked yet, any book taken will be a new book. Hence $X_1 = 1$. When $(i - 1)$ distinct books have picked, the next book to be picked should be among the $(n - i + 1)$ books. Since there are n books to be picked from, the probability is $\frac{n-i+1}{n}$

(b)

The head probability of the Bernoulli trial corresponding to the geometric trial is the probability whether we select a distinct book other than the $(i - 1)$ books selected. This probability as calculated above is $\frac{n-i+1}{n}$. Hence the parameter $p = \frac{n-i+1}{n}$

(c)

Suppose $p - 1 = q$

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{i=1}^{\infty} iP(Z = i) \\ &= p + 2pq + 3pq^2 + 4pq^3 + \dots\end{aligned}\tag{1}$$

$$q\mathbb{E}[Z] = pq + 2pq^2 + 3pq^3 + \dots\tag{2}$$

Subtracting (2) from (1), we get

$$\begin{aligned}(1 - q)\mathbb{E}[Z] &= p + pq + pq^2 + pq^3 + \dots \\ &= \frac{p}{1 - q} \quad \text{Sum of GP and } q < 1 \\ \implies \mathbb{E}[Z] &= \frac{p}{(1 - q)^2} = \frac{1}{p}\end{aligned}\tag{3}$$

Now, for finding the variance,

$$\begin{aligned}\mathbb{E}[Z^2] &= \sum_{i=1}^{\infty} i^2 P(Z = i) \\ &= p + 4pq + 9pq^2 + 16pq^3 + 25pq^4 \dots\end{aligned}\tag{4}$$

$$q\mathbb{E}[Z^2] = pq + 4pq^2 + 9pq^3 + 16pq^4 \dots\tag{5}$$

Subtracting (5) from (4), we get

$$(1 - q)\mathbb{E}[Z^2] = p + 3pq + 5pq^2 + 7pq^3 + \dots\tag{6}$$

$$q(1 - q)\mathbb{E}[Z^2] = pq + 3pq^2 + 5pq^3 + \dots\tag{7}$$

Subtracting (7) from (6), we get

$$\begin{aligned}(1 - q)^2\mathbb{E}[Z^2] &= p + 2pq + 2pq^2 + 2pq^3 + \dots \\ &= \frac{2p}{1 - q} - p \quad \text{Sum of GP and } q < 1 \\ \implies \mathbb{E}[Z^2] &= \frac{2p}{(1 - q)^3} - \frac{p}{(1 - q)^2} = \frac{2}{p^2} - \frac{1}{p} \\ \text{Var}(Z) &= \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \\ &= \frac{1 - p}{p^2}\end{aligned}\tag{8}$$

(d)

From part (b) and 3 we can conclude ,

$$\mathbb{E} \left[X^{(n)} \right] = \mathbb{E} \left[\sum_{i=1}^n X_i \right] = \sum_{i=1}^n \mathbb{E} [X_i] = \sum_{i=1}^n \frac{n}{n-i+1} = n \sum_{i=1}^n \frac{1}{i}$$

(e)

For finding the variance we first need $Var(X_i)$. Again lets call $1-p=q$

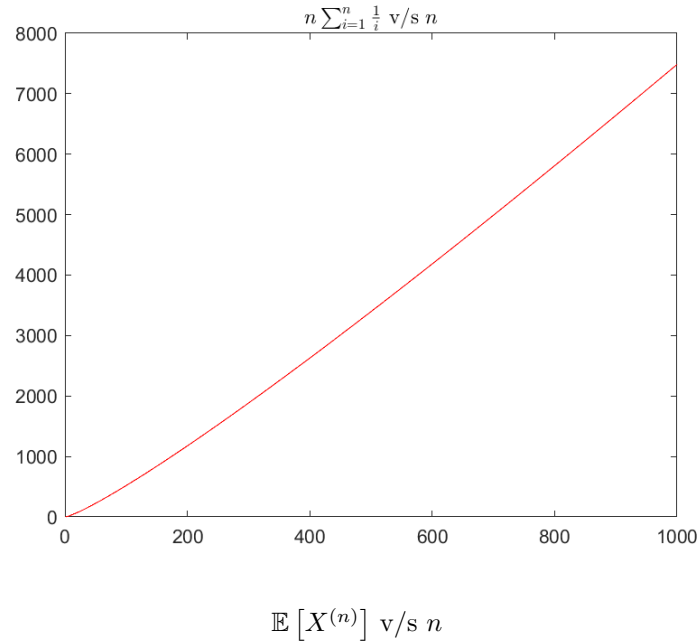
$$\begin{aligned} Var(X^{(n)}) &= Var \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n Var(X_i) \quad X_i' \text{ s are independent} \\ &= \sum_{i=1}^n \mathbb{E} [X_i^2] - \mathbb{E} [X_i]^2 \\ &= \sum_{i=1}^n \frac{1}{p_i^2} - \frac{1}{p_i} \quad \text{From 8} \\ &= \sum_{i=1}^n \left(\frac{n}{n-i+1} \right)^2 - \left(\frac{n}{n-i+1} \right) \\ &= \sum_{i=1}^n \left(\frac{n}{i} \right)^2 - \left(\frac{n}{i} \right) \\ &< \sum_{i=1}^n \left(\frac{n}{i} \right)^2 < \frac{n^2 \pi^2}{6} \end{aligned} \tag{9}$$

(f)

We know for a fact that , $\exists c > 0, N_0 \in \mathbb{N}$, such that

$$n \sum_{i=1}^n \frac{1}{i} < n(1 + \log(n)) < c \cdot n \log(n) \quad \forall n > N_0$$

Hence we can say $\mathbb{E} [X^{(n)}] \in \Theta(n \log(n))$ or $f(n) = n \log(n)$



2 Question 2

(a)

Suppose U is a uniform random variable. We know that for a uniform distribution, $P(U \leq x) = x$. Consider the random variable $V = F^{-1}(U)$. We are concerned with finding the CDF of this random variable V . (Since v_i are samples taken from V). This can be found by finding:

$$\begin{aligned} P(V \leq x) &= P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) \quad \text{Because } F \text{ is monotonically increasing, Apply } F \text{ on both sides} \\ &= F(x) \end{aligned}$$

(b)

Claim: For the samples $Y_i \sim F$, for some distribution F , $F(Y_i) \sim U[0, 1]$, where $U[0, 1]$ represents a uniform distribution.

$$\begin{aligned} P(F(Y_i) \leq y) &= P(Y_i \leq F^{-1}(y)) \\ &= F(F^{-1}(y)) = y \quad \text{From definition of CDF of } F \end{aligned}$$

This is identical to the uniform distribution CDF which says $P(U_i \leq u) = u$. Hence, our claim is proven.

Now, for the question, Proving $P(D < d) = P(E < d)$ is equivalent, Now

$$P(D < d) = P(\max_x |F_e(x) - F(x)| < d) = P(\forall x |F_e(x) - F(x)| < d)$$

. Now, let us substitute $y = F(x)$. As x varies over all possible values, y will vary from 0 to 1.

Also $x = F^{-1}(y)$.

$$\begin{aligned}
P(D < d) &= P(\forall y \in [0,1] | F_e(F^{-1}(x)) - y| < d) \\
&= P(\forall y \in [0,1] | \frac{1}{n} \sum \mathbf{1}(Y_i \leq F^{-1}(y)) - y| < d) \\
&= P(\forall y \in [0,1] | \frac{1}{n} \sum \mathbf{1}(F(Y_i) \leq y) - y| < d) \\
&= P(\forall y \in [0,1] | \frac{1}{n} \sum \mathbf{1}(U_i \leq y) - y| < d) \\
&= P(\max_{y \in [0,1]} E < d) = P(E < d)
\end{aligned}$$

Last step is due to the claim proven above.

This means that $P(D \geq d)$ is independent of the distribution F we are taking. Hence, this can be used to check whether a given set of samples Y_i belongs to a particular distribution F or not, as if it does belong to F , then the probability of D being higher than a particular value should be almost similar to probability that the max difference between the empirical CDF and the true uniform distribution ($\text{Uniform}(0,1)$) corresponding to the same number of samples is greater than that value. That is given a large set of values, the distribution of D tends to the distribution of E . Hence, we can use this property to check whether a distribution of samples belong to a particular distribution or not.

3 Question 3

(a)

First, observe that every coordinate z_i is a sample taken from the distribution $Z_i \sim \mathcal{N}(ax_i + by_i + c, \sigma^2)$, therefore, the likelihood is.

$$\begin{aligned}
f(z_1, z_2, z_3 \dots z_n; \theta) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{z_i - ax_i - by_i - c}{\sigma}\right)^2\right) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(\sum_{i=1}^n -\frac{1}{2} \left(\frac{z_i - ax_i - by_i - c}{\sigma}\right)^2\right)
\end{aligned}$$

Now ,

$$\mathcal{L} = \log(f(z_1, z_2, z_3 \dots z_n; \theta)) = -\sum_{i=1}^n \frac{1}{2} \left(\frac{z_i - ax_i - by_i - c}{\sigma}\right)^2 - n \log(\sigma\sqrt{2\pi})$$

We can get the required equations by setting :

$$\frac{\partial \mathcal{L}}{\partial a} = 0 \quad \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \frac{\partial \mathcal{L}}{\partial c} = 0$$

We get the following 3 equations:

$$\sum x_i(ax_i + by_i + c) = \sum z_i x_i \tag{10}$$

$$\sum y_i(ax_i + by_i + c) = \sum z_i y_i \tag{11}$$

$$\sum (ax_i + by_i + c) = \sum z_i \tag{12}$$

In matrix form :

$$\begin{bmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i \\ \sum x_i & \sum y_i & \sum 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum z_i x_i \\ \sum z_i y_i \\ \sum z_i \end{bmatrix}$$

(b)

Using similar logic as above we can get

$$\mathcal{L} = \log(f(z_1, z_2, z_3 \dots z_n; \theta)) = - \sum_{i=1}^n \frac{1}{2} \left(\frac{z_i - a_1 x_i^2 - a_2 y_i^2 - a_3 x_i y_i - a_4 x_i - a_5 y_i - a_6}{\sigma} \right)^2 - n \log(\sigma \sqrt{2\pi})$$

The required eqations can be obtained by setting :

$$\frac{\partial \mathcal{L}}{\partial a_i} = 0 \quad i \in \{1, 2, 3, 4, 5, 6\}$$

We get the following equations:

$$\sum x_i^2 (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6) = \sum x_i^2 z_i \quad (13)$$

$$\sum y_i^2 (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6) = \sum y_i^2 z_i \quad (14)$$

$$\sum x_i y_i (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6) = \sum x_i y_i z_i \quad (15)$$

$$\sum x_i (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6) = \sum x_i z_i \quad (16)$$

$$\sum y_i (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6) = \sum y_i z_i \quad (17)$$

$$\sum (a_1 x_i^2 + a_2 y_i^2 + a_3 x_i y_i + a_4 x_i + a_5 y_i + a_6) = \sum z_i \quad (18)$$

In matrix form:

$$\begin{bmatrix} \sum x_i^4 & \sum x_i^2 y_i^2 & \sum x_i^3 y_i & \sum x_i^3 & \sum x_i^2 y_i & \sum x_i^2 \\ \sum x_i^2 y_i^2 & \sum y_i^4 & \sum x_i y_i^3 & \sum x_i y_i^2 & \sum y_i^3 & \sum y_i^2 \\ \sum x_i^3 y_i & \sum x_i y_i^3 & \sum x_i^2 y_i^2 & \sum x_i^2 y_i & \sum x_i y_i^2 & \sum x_i y_i \\ \sum x_i^3 & \sum x_i y_i^2 & \sum x_i^2 y_i & \sum x_i^2 & \sum x_i y_i & \sum x_i \\ \sum x_i^2 y_i & \sum y_i^3 & \sum x_i y_i^2 & \sum x_i y_i & \sum y_i^2 & \sum y_i \\ \sum x_i^2 & \sum y_i^2 & \sum x_i y_i & \sum x_i & \sum y_i & \sum 1 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{bmatrix} = \begin{bmatrix} \sum x_i^2 z_i \\ \sum y_i^2 z_i \\ \sum x_i y_i z_i \\ \sum x_i z_i \\ \sum y_i z_i \\ \sum z_i \end{bmatrix}$$

(c)

On solving the equation on MATLAB, the predicted equation of the plane is :

$$z = 10.002208x + 19.998022y + 29.951579$$

To find the noise variance we first found out the predicted z values : $Z_{predicted}$, then subtracted the predicted values from the actual values to get the noise and then found its variance. **the noise variance we got was 23.068503**

4 Question 4

(a)

```
n = 1000;
data = normrnd(0,4,1,n);
indices = randperm(n, 750);
T = data(indices);
remaining_indices = setdiff(1:n, indices);
V = data(remaining_indices);
```

These lines of codes draw $n = 1000$ independent samples from $\mathcal{N}(0,16)$ and then distribute them such that 750 are in subset T and the rest 250 are in V and thus making sure they are disjoint.

(b)

The estimate for the PDF built from T with bandwidth parameter σ is:

$$\hat{p}_n(x; \sigma) = \sum_{j=1}^{750} \frac{e^{-(x-T_j)^2/2\sigma^2}}{750 \cdot \sigma \sqrt{2\pi}}$$

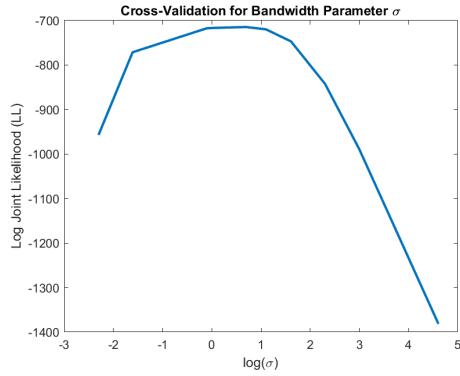
The expression for the joint likelihood of the samples in V, based on the estimate of the PDF built from T with bandwidth parameter σ is:

$$f(x_1, x_2, x_3 \dots x_{250}; \sigma) = \prod_{i=1}^{250} \sum_{j=1}^{750} \frac{e^{-(V_i-T_j)^2/2\sigma^2}}{750 \cdot \sigma \sqrt{2\pi}}$$

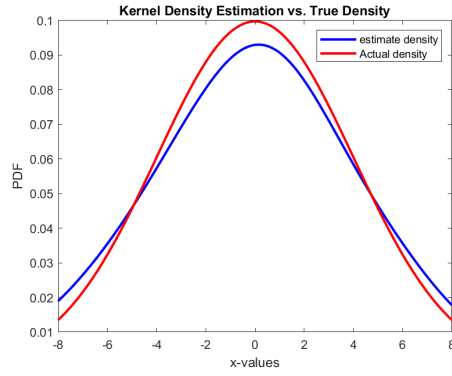
This follows as the samples in V are independent and identically distributed.

(c)

The plots for LL vs $\log(\sigma)$ and the density estimation function vs the true pdf.



(a) Cross-Validation for σ

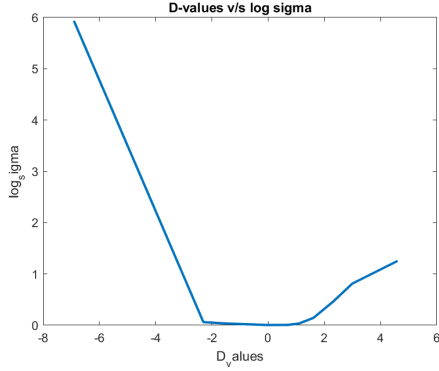


(b) Kernel Density Estimation vs. True Density

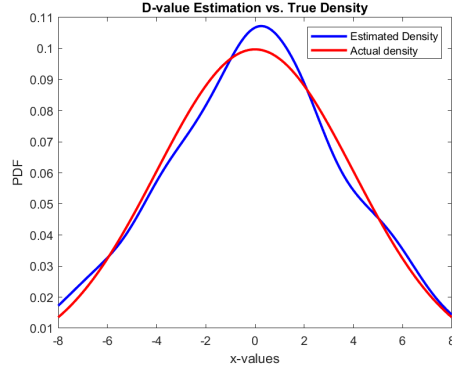
The best value of σ we got was **2.0** and this gave LL as **-714.574944**;

(d)

The plots for LL vs $\log(\sigma)$ and the density estimation function vs the true pdf.



(a) D-values for σ

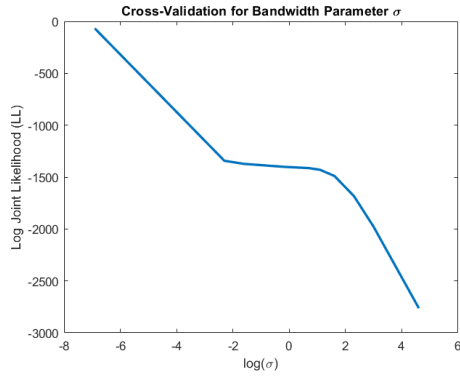


(b) Kernel Density Estimation vs. True Density

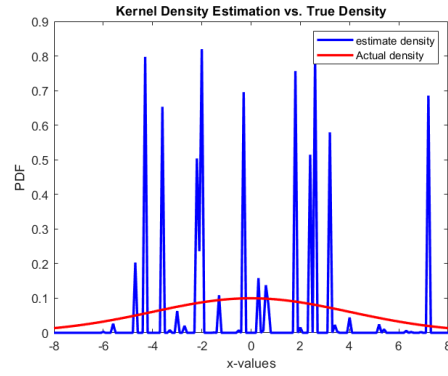
The best value of σ we got was **1.0** and this has the D-value as **0.005436**. The D-value we got for $\sigma = 2.0$ (the one for the best LL) is **0.007035**.

(e)

When we have T and V both equal (with both having 500 elements), we get the following graph for LL vs $\log \sigma$.



(a) LL vs σ



(b) Kernel Density Estimation vs. True Density

This happens because in the term for the joint-likelihood of the samples in V, based on the estimate of the PDF built from T with bandwidth parameter σ , we will have one term as $\frac{e^0}{n\sigma\sqrt{2\pi}} = \frac{1}{\sigma\sqrt{2\pi}}$ (as T and V are equal, $T_j - V_i = 0$ for one term in the summation). Now when $\sigma \rightarrow 0$, $\frac{1}{\sigma\sqrt{2\pi}} \rightarrow \infty$ and $\frac{\exp(-(T_i - V_j)^2 / 2\sigma^2)}{\sigma\sqrt{2\pi}} \rightarrow 0$. Thus the LL function gets huge for small σ and in the cross-validation step, we will always get the best σ as the smallest possible value we are taking, which might not give the best kernel estimation for the distribution, as shown in the (b) subfigure.

5 Question 5

$$P(S_n - E[S_n] > t) = P(e^{s(S_n - E[S_n])} > e^{st}) \text{ for some } s > 0$$

Using Markov's Inequality we know that

$$P(e^{s(S_n - E[S_n])} > e^{st}) \leq \frac{E[e^{s(S_n - E[S_n])}]}{e^{st}}$$

Now consider the random variables $Y_i = e^{s(X_i - E[X_i])}$.

As the X_i are independent, we know that the Y_i s are independent as well.

$$\begin{aligned}
e^{s(S_n - E[S_n])} &= e^{s(\sum_{i=1}^n X_i - \sum_{i=1}^n E[X_i])} \\
&= e^{\sum_{i=1}^n s(X_i - E[X_i])} \\
&= \prod_{i=1}^n e^{s(X_i - E[X_i])} \\
&= \prod_{i=1}^n Y_i
\end{aligned} \tag{19}$$

$$\begin{aligned}
E[e^{s(S_n - E[S_n])}] &= E[\prod_{i=1}^n Y_i] = \prod_{i=1}^n E[Y_i] \\
&= \prod_{i=1}^n E[e^{s(X_i - E[X_i])}] \\
&\leq \prod_{i=1}^n e^{\frac{s^2(b_i^2 - a_i^2)}{8}} \\
&= e^{\frac{s^2}{8} \sum_{i=1}^n (b_i - a_i)^2} \\
&= e^{\frac{s^2 \lambda}{8}}
\end{aligned}$$

where $\lambda = \sum_{i=1}^n (b_i - a_i)^2$

$$\begin{aligned}
P(e^{s(S_n - E[S_n])} > e^t) &\leq \frac{E[e^{s(S_n - E[S_n])}]}{e^{st}} \\
&= \frac{e^{\frac{s^2 \lambda}{8}}}{e^{st}} \\
&= e^{\frac{s^2 \lambda}{8} - st}
\end{aligned}$$

As we vary s , $\frac{s^2 \lambda}{8} - st$ achieves a minima which can be regarded as the upper bound on $P(S_n - E[S_n] > t)$

Minima is achieved at

$$\frac{\partial}{\partial s} \left(\frac{s^2 \lambda}{8} - st \right) = 0 \implies \frac{s \lambda}{4} = t \implies s = \frac{4t}{\lambda}$$

Therefore

$$P(S_n - E[S_n] > t) = P(e^{s(S_n - E[S_n])} > e^{st}) \leq e^{\left(\frac{4t}{\lambda}\right)^2 \frac{\lambda}{8} - \frac{4t}{\lambda} t} = e^{\frac{-2t^2}{\lambda}}$$

$$P(S_n - E[S_n] > t) \leq e^{\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Now, for the proof of the intermediate result

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

$$\begin{aligned}
e^{sx} &\leq e^{sa} + \frac{e^{sb} - e^{sa}}{b - a}(x - a) \\
e^{sx} &\leq \frac{(b - x)e^{sa}}{b - a} + \frac{(x - a)e^{sb}}{b - a}
\end{aligned}$$

2. Taking Expectation on both sides, and using $\mathbb{E}[X] = 0$

$$\begin{aligned}
\mathbb{E}[e^{sx}] &\leq \frac{e^{sa}}{b-a} \mathbb{E}[b-x] + \frac{e^{sb}}{b-a} \mathbb{E}[x-a] \\
\mathbb{E}[e^{sx}] &\leq \frac{be^{sa} - ae^{sb}}{b-a} = \frac{be^{sa} - ae^{sa} + ae^{sa} - ae^{sb}}{b-a} \\
&= e^{sa} \left(1 + \frac{a - ae^{s(b-a)}}{b-a} \right) \\
&= e^{\frac{s(b-a)a}{b-a}} \left(1 + \frac{a - ae^{s(b-a)}}{b-a} \right) \\
&= e^{L(s(b-a))}
\end{aligned}$$

where $L(h) = \frac{ha}{b-a} + \log \left(1 + \frac{a - ae^h}{b-a} \right)$

3.

$$L(h) = \frac{ha}{b-a} + \log \left(1 + \frac{a - ae^h}{b-a} \right)$$

$$\begin{aligned}
L'(h) &= \frac{a}{b-a} + \frac{1}{1 + \frac{a - ae^h}{b-a}} \frac{-ae^h}{b-a} \\
&= \frac{a}{b-a} - \frac{ae^h}{b - ae^h}
\end{aligned}$$

$$\begin{aligned}
L''(h) &= \frac{d}{dh} \left(\frac{-ae^h}{b - ae^h} \right) \\
&= \frac{-abe^h}{(b - ae^h)^2}
\end{aligned}$$

Now $\frac{(b - ae^h)^2}{4} \geq -abe^h$ if a and b are of different signs by AM-GM. If they are of the same sign, then its obviously true. Thus we have that

$$L''(h) \leq \frac{1}{4}$$

4. By Taylor's theorem we can write that

$$L(h) = L(0) + hL'(0) + \frac{h^2}{2} L''(c)$$

for all $h \geq 0$ and $0 \leq c \leq h$. Further as $L(0) = L'(0) = 0$

$$\implies L(h) \leq 0 + 0 + \frac{h^2}{2} \frac{1}{4} \quad \forall h \geq 0$$

Thus we have that

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