Introduction to Machine Learning, Fall 2023
Homework 2
(Due Tuesday Nov. 14 at 11:59pm (CST))

November 9, 2023

1. [10 points] [Convex Optimization Basics]

   (a) Proof any norm $f : \mathbb{R}^n \to \mathbb{R}$ is convex. [2 points]

   (b) Determine the convexity (i.e., convex, concave or neither) of $f(x_1, x_2) = x_1^2/x_2$ on $\mathbb{R} \times \mathbb{R}_{>0}$. [2 points]

   (c) Determine the convexity of $f(x_1, x_2) = x_1/x_2$ on $\mathbb{R}^2_{>0}$. [2 points]

   (d) Recall Jensen's inequality $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$ if f is convex for any random variable X. Proof the log sum inequality:
$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$
   where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are positive numbers. Hints: $f(x) = x \log x$ is strictly convex. [4 points]

Solution:

(a) Since f is a norm function, we have the properties of norm functions that,
1. $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$.
2. $\forall \mathbf{x} \in \mathbb{R}^n, \forall a \in \mathbb{R}, f(a\mathbf{x}) = |a| f(\mathbf{x})$.
So we have, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall \lambda \in [0, 1]$,
From property 1., we can get that
$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq f(\lambda \mathbf{x}) + f((1 - \lambda)\mathbf{y})$$
From property 2., we can get that
$$f(\lambda \mathbf{x}) = |\lambda| f(\mathbf{x}) \; and \; f((1 - \lambda)\mathbf{y}) = |1 - \lambda| f(\mathbf{y})$$
Since $\lambda \in [0, 1]$, so we have $|\lambda| = \lambda$ and $|1 - \lambda| = 1 - \lambda$,
So we can get that $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall \lambda \in [0, 1]$,
$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$
So above all, from the defination, we can get that f is a convex function.

(b) Since $f(x_1, x_2) = \dfrac{x_1^2}{x_2}$, so we have the Hessain matrix of f is
$$H = \nabla^2 f(x_1, x_2) = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix}$$
So $\forall \mathbf{y} \in \mathbb{R}^2$, let $\mathbf{y} = (y_1, y_2)^T$, we have
$$\mathbf{y}^T H \mathbf{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{2y_1^2}{x_2} - \frac{4x_1 y_1 y_2}{x_2^2} + \frac{2x_1^2 y_2^2}{x_2^3} = \frac{2}{x_2}\left(y_1 - \frac{x_1 y_2}{x_2}\right)^2 \geq 0$$
Since $x_2 > 0$, so $\forall \mathbf{y} \in \mathbb{R}^2$, we have $\mathbf{y}^T H \mathbf{y} \geq 0$.
So we can get that $H = \nabla^2 f(x_1, x_2) \succeq 0$, so f is a convex function.

So above all, f is a convex function.

(c) Since $f(x_1,x_2)=\dfrac{x_1}{x_2}$, so we have the Hessain matrix of f is

$$H=\nabla^2 f(x_1,x_2)=\begin{bmatrix}0 & -\frac{1}{x_2^2}\\ -\frac{1}{x_2^2} & \frac{2x_1}{x_2^3}\end{bmatrix}$$

Let $\mathbf{y}\in\mathbb{R}^2=(y_1,y_2)^T$, we have

$$\mathbf{y}^T H\mathbf{y}=\begin{bmatrix}y_1 & y_2\end{bmatrix}\begin{bmatrix}0 & -\frac{1}{x_2^2}\\ -\frac{1}{x_2^2} & \frac{2x_1}{x_2^3}\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}=\frac{2y_2}{x_2^3}(x_1y_2-x_2y_1)$$

Since $x_1>0,x_2>0$, so:
If $\mathbf{y}=(0,0)^T$, then $\mathbf{y}^T H\mathbf{y}=0$.
If $\mathbf{y}=(\dfrac{2x_1}{x_2},1)^T$, then $\mathbf{y}^T H\mathbf{y}=-\dfrac{2x_1}{x_2^3}<0$.
If $\mathbf{y}=(\dfrac{x_1}{2x_2},1)^T$, then $\mathbf{y}^T H\mathbf{y}=\dfrac{x_1}{x_2^3}>0$.

So $\mathbf{y}^T H\mathbf{y}$ can be positive, 0, or negative, when $\mathbf{y}$ takes different values.
so $H=\nabla^2 f(x_1,x_2)$ is neither positive semidefinite nor negative semidefinite.

So above all, f is neither a convex nor a concave function.

(d) We can construct a distribution X s.t.
the domain of X is $\dfrac{a_i}{b_i}, i\in\{1,2,\cdots,n\}$, and the PMF of X is $P(X=\dfrac{a_i}{b_i})=\dfrac{b_i}{\sum\limits_{i=1}^n b_i}$.

Since $\forall i\in\{1,2,\cdots,n\}, a_i>0, b_i>0$,
So $P(X=\dfrac{a_i}{b_i})>0$, and $\sum\limits_{i=1}^n P(X=\dfrac{a_i}{b_i})=1$.
So it's a valid distribution.
So

$$\mathbb{E}(X)=\sum_{i=1}^n(\frac{a_i}{b_i})\cdot P(X=\frac{a_i}{b_i})=\sum_{i=1}^n\frac{a_i}{b_i}\cdot\frac{b_i}{\sum\limits_{k=1}^n b_k}=\frac{\sum\limits_{i=1}^n a_i}{\sum\limits_{i=1}^n b_i}$$

And since $f(x)=x\log x$ is strictly convex, so from the Jensen's inequality, we can get that

$$f(\mathbb{E}(X))\le\mathbb{E}(f(X))$$

i.e.

$$\left(\frac{\sum\limits_{i=1}^n a_i}{\sum\limits_{i=1}^n b_i}\right)\log\left(\frac{\sum\limits_{i=1}^n a_i}{\sum\limits_{i=1}^n b_i}\right)\le\sum_{i=1}^n P(X=\frac{a_i}{b_i})\cdot f(\frac{a_i}{b_i})$$

$$\left(\frac{\sum\limits_{i=1}^n a_i}{\sum\limits_{i=1}^n b_i}\right)\log\left(\frac{\sum\limits_{i=1}^n a_i}{\sum\limits_{i=1}^n b_i}\right)\le\sum_{i=1}^n\frac{b_i}{\sum\limits_{k=1}^n b_k}\cdot(\frac{a_i}{b_i})\log(\frac{a_i}{b_i})$$

Since $b_i>0$, so $\sum\limits_{i=1}^n b_i>0$, so appointment $\sum\limits_{i=1}^n b_i$ on both sides simultaneously, we can get that

$$\left(\sum_{i=1}^n a_i\right)\log\left(\frac{\sum\limits_{i=1}^n a_i}{\sum\limits_{i=1}^n b_i}\right)\le\sum_{i=1}^n a_i\log\left(\frac{a_i}{b_i}\right)$$

i.e.

$$\sum_{i=1}^n a_i\log\frac{a_i}{b_i}\ge\left(\sum_{i=1}^n a_i\right)\log\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

So above all, with such construction, we have proved the inequality

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

2. [10 points] [Linear Methods for Classification] Consider the "Multi-class Logistic Regression" algorithm. Given training set $\mathcal{D} = \{(x^i, y^i) \mid i = 1, \dots, n\}$ where $x^i \in \mathbb{R}^{p+1}$ is the feature vector and $y^i \in \mathbb{R}^k$ is a one-hot binary vector indicating k classes. We want to find the parameter $\hat{\beta} = [\hat{\beta}_1, \dots, \hat{\beta}_k] \in \mathbb{R}^{(p+1)\times k}$ that maximize the likelihood for the training set. Introducing the softmax function, we assume our model has the form

$$p(y_c^i = 1 \mid x^i; \beta) = \frac{\exp(\beta_c^\top x^i)}{\sum_{c'} \exp(\beta_{c'}^\top x^i)},$$

where y_c^i is the c-th element of y^i.

(a) Complete the derivation of the conditional log likelihood for our model, which is

$$\ell(\beta) = \ln \prod_{i=1}^n p(y_t^i \mid x^i; \beta) = \sum_{i=1}^n \sum_{c=1}^k \left[y_c^i(\beta_c^\top x^i) - y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) \right].$$

For simplicity, we abbreviate $p(y_t^i = 1 \mid x^i; \beta)$ as $p(y_t^i \mid x^i; \beta)$, where t is the true class for x^i. [4 points]

(b) Derive the gradient of $\ell(\beta)$ w.r.t. β_1, i.e.,

$$\nabla_{\beta_1} \ell(\beta) = \nabla_{\beta_1} \sum_{i=1}^n \sum_{c=1}^k \left[y_c^i(\beta_c^\top x^i) - y_c^i \ln \left(\sum_{c'} \exp(\beta_{c'}^\top x^i) \right) \right].$$

Remark: Log likelihood is always concave; thus, we can optimize our model using gradient ascent. (The gradient of $\ell(\beta)$ w.r.t. $\beta_2, \dots, \beta_k$ is similar, you don't need to write them) [6 points]

Solution:

(a) Since the model with softmax function is that $p(y_c^i = 1|x^i; \beta) = \dfrac{exp(\beta_c^T x^i)}{\sum\limits_{c'} exp(\beta_{c'}^T x^i)}$.

And since y^i is a one-hot binary vector, so $y_t^i = 1$, and $\forall c \neq t, y_c^i = 0$.
So $\forall i \in \{1, 2, \cdots, n\}, \forall c \in \{1, 2, \cdots, k\}$, we have

$$p(y^i|x^i; \beta) = p(y_t^i|x^i; \beta) = \prod_{c=1}^k p(y_c^i|x^i; \beta)^{y_c^i}$$

So the likelihood is that

$$L(\beta) = \prod_{i=1}^n p(y^i|x^i; \beta) = \prod_{i=1}^n \prod_{c=1}^k p(y_c^i|x^i; \beta)^{y_c^i}$$

So the log-likelihood is that

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^n \sum_{c=1}^k y_c^i \log p(y_c^i|x^i; \beta) = \sum_{i=1}^n \sum_{c=1}^k y_c^i \log \frac{exp(\beta_c^T x^i)}{\sum\limits_{c'} exp(\beta_{c'}^T x^i)}$$

$$= \sum_{i=1}^n \sum_{c=1}^k \left[y_c^i(\beta_c^T x^i) - y_c^i \log(\sum_{c'} exp(\beta_{c'}^T x^i)) \right]$$

So above all, the log-likelihood is that

$$\ell(\beta) = \sum_{i=1}^n \sum_{c=1}^k \left[y_c^i(\beta_c^T x^i) - y_c^i \log(\sum_{c'} exp(\beta_{c'}^T x^i)) \right]$$

(b) The gradient of $\ell(\beta)$ w.r.t. β_1 is that

$$\nabla_{\beta_1} \ell(\beta) = \nabla_{\beta_1} \sum_{i=1}^n \sum_{c=1}^k \left[y_c^i(\beta_c^T x^i) - y_c^i \log(\sum_{c'} exp(\beta_{c'}^T x^i)) \right]$$

If $c \neq 1$, then $\nabla_{\beta_1} y_c^i \beta_c^T x^i = 0$, and if $c = 1$, then $\nabla_{\beta_1} y_c^i \beta_c^T x^i = y_c^i x^i = y_1^i x^i$.
And $\nabla_{\beta_1} \log(\sum\limits_{c'} \exp(\beta_{c'}^T x^i)) = \dfrac{\exp(\beta_1^T x^i) x^i}{\sum\limits_{c'} \exp(\beta_{c'}^T x^i)}$.

So

$$
\begin{aligned}
\nabla_{\beta_1}\ell(\beta) &= \sum_{i=1}^{n}\left(\sum_{c=1}^{k}\nabla_{\beta_1}y_c^i\beta_c^T x^i - \sum_{c=1}^{k}\nabla_{\beta_1}y_c^i\log(\sum_{c'} exp(\beta_{c'}^T x^i))\right) \\
&= \sum_{i=1}^{n}\left(y_1^i x^i - (\sum_{c=1}^{k}y_c^i)\cdot\frac{\exp(\beta_1^T x^i)x^i}{\sum_{c'}\exp(\beta_{c'}^T x^i)}\right)
\end{aligned}
$$

Also, since y^i is a one-hot binary vector, so $y_t^i = 1$, and $\forall c \neq t, y_c^i = 0$.
So $\sum_{c=1}^{k} y_c^i = 1$.
So

$$
\nabla_{\beta_1}\ell(\beta) = \sum_{i=1}^{n}\left(y_1^i x^i - \frac{\exp(\beta_1^T x^i)x^i}{\sum_{c'}\exp(\beta_{c'}^T x^i)}\right)
$$

So above all

$$
\nabla_{\beta_1}\ell(\beta) = \sum_{i=1}^{n}\left(y_1^i x^i - \frac{\exp(\beta_1^T x^i)x^i}{\sum_{c'}\exp(\beta_{c'}^T x^i)}\right)
$$

3. [10 points] [Probability and Estimation] Suppose $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ are i.i.d. samples from exponential distribution with parameter $\lambda > 0$, i.e., $X \sim \text{Expo}(\lambda)$. Recall the PDF of exponential distribution is

$$p(x \mid \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}.$$

(a) To derive the posterior distribution of λ, we assume its prior distribution follows gamma distribution with parameters $\alpha, \beta > 0$, i.e., $\lambda \sim \text{Gamma}(\alpha, \beta)$ (since the range of gamma distribution is also $(0, +\infty)$, thus it's a plausible assumption). The PDF of λ is given by

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta},$$

where $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt,\ \alpha > 0$. Show that the posterior distribution $p(\lambda \mid \mathcal{D})$ is also a gamma distribution and identify its parameters. Hints: Feel free to drop constants. [4 points]

(b) Derive the maximum a posterior (MAP) estimation for λ under $\text{Gamma}(\alpha, \beta)$ prior. [3 points]

(c) For exponential distribution $\text{Expo}(\lambda)$, $\sum_{i=1}^n x_i \sim \text{Gamma}(n, \lambda)$ and the inverse sample mean $\frac{n}{\sum_{i=1}^n x_i}$ is the MLE for λ. Argue that whether $\frac{n-1}{n} \hat{\lambda}_{MLE}$ is unbiased ($\mathbb{E}(\frac{n-1}{n} \hat{\lambda}_{MLE}) = \lambda$). Hints: $\Gamma(z+1) = z\Gamma(z)$, $z > 0$. [3 points]

Solution:
(a) From Bayes' Rule, we can get that

$$p(\lambda|\mathcal{D}) = \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})}$$

Since $\mathcal{D}$ do not contain any λ, so we can get that

$$p(\lambda|\mathcal{D}) \propto p(\mathcal{D}|\lambda)p(\lambda)$$

And since $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$ are i.i.d. samples from exponential distribution with parameter $\lambda > 0$, so we can get that

$$p(\mathcal{D}|\lambda) = p(x_1, x_2, \cdots, x_n|\lambda) = \prod_{i=1}^n p(x_i|\lambda)$$

Since $p(x|\lambda) = \lambda e^{-\lambda x}, x > 0$, so WLOG, we can assume that all the sampling points are positive, i.e. $\forall i, x_i > 0$, then we can get that

$$p(\mathcal{D}|\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum\limits_{i=1}^n x_i}$$

And since we know that the prior distribution of λ is that $\lambda \sim Gamma(\alpha, \beta)$, so we can get that

$$p(\lambda) = p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$$

So

$$p(\lambda|\mathcal{D}) \propto \lambda^n e^{-\lambda \sum\limits_{i=1}^n x_i} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \propto \lambda^{n+\alpha-1} e^{-\lambda(\sum\limits_{i=1}^n x_i + \beta)}$$

Since $p(\lambda|\mathcal{D})$ is in terms of conditional probability, so its distribution must be a valid distribution. And from

$$p(\lambda|\mathcal{D}) \propto \lambda^{n+\alpha-1} e^{-\lambda(\sum\limits_{i=1}^n x_i + \beta)}$$

we can get that the distribution is

$$p(\lambda|\mathcal{D}) \sim Gamma(\alpha + n, \beta + \sum_{i=1}^n x_i)$$

So above all, we have proved that the posterior distribution $p(\lambda|\mathcal{D})$ is also a Gamma distribution, and the parameters is that $p(\lambda|\mathcal{D}) \sim Gamma(\alpha + n, \beta + \sum\limits_{i=1}^n x_i)$

(b) From (a), we get that $p(\lambda|\mathcal{D}) \sim Gamma(\alpha + n, \beta + \sum\limits_{i=1}^{n} x_i)$.

and $p(\lambda|\mathcal{D}) \propto \lambda^{n+\alpha-1} e^{-\lambda(\sum\limits_{i=1}^{n} x_i + \beta)}$.

So the MAP for λ under $Gamma(\alpha, \beta)$ prior is that

$$\hat{\lambda}_{MAP} = \underset{\lambda}{\operatorname{argmax}}\, \lambda^{\alpha+n-1} e^{-\lambda(\beta + \sum\limits_{i=1}^{n} x_i)}$$

Take it into the log-likelyhood function, the result of MAP is the same.

So

$$\hat{\lambda}_{MAP} = \underset{\lambda}{\operatorname{argmax}}(\alpha + n - 1)\log\lambda - (\beta + \sum_{i=1}^{n} x_i)\lambda$$

Let

$$f(\lambda) = (\alpha + n - 1)\log\lambda - (\beta + \sum_{i=1}^{n} x_i)\lambda$$

then

$$f'(\lambda) = \frac{\alpha + n - 1}{\lambda} - (\beta + \sum_{i=1}^{n} x_i)$$

And

$$f''(\lambda) = -(\alpha + n - 1)\frac{1}{\lambda^2} < 0$$

So we could find that the function $f(\lambda)$ is a concave function.

So to get the MAP, we need to find the point where the first derivative of $f(\lambda)$ is equal to 0.

i.e.

$$\frac{\alpha + n - 1}{\lambda} - (\beta + \sum_{i=1}^{n} x_i) = 0$$

So

$$\hat{\lambda}_{MAP} = \frac{\alpha + n - 1}{\beta + \sum\limits_{i=1}^{n} x_i}$$

So above all, the MAP estimation for λ under $Gamma(\alpha, \beta)$ prior is that

$$\hat{\lambda}_{MAP} = \frac{\alpha + n - 1}{\beta + \sum\limits_{i=1}^{n} x_i}$$

(c) Since the MLE is that $\hat{\lambda}_{MLE} = \frac{n}{\sum\limits_{i=1}^{n} x_i}$, so we can get that

$$\mathbb{E}(\frac{n-1}{n}\hat{\lambda}_{MLE}) = \mathbb{E}(\frac{n-1}{n} \cdot \frac{n}{\sum\limits_{i=1}^{n} x_i}) = \mathbb{E}(\frac{n-1}{\sum\limits_{i=1}^{n} x_i})$$

Let $Y = \sum\limits_{i=1}^{n} x_i$, then $Y \sim Gamma(n, \lambda)$, so we can get that

$$\mathbb{E}(\frac{n-1}{n}\hat{\lambda}_{MLE}) = (n-1)\mathbb{E}(\frac{1}{Y})$$

Since $Y \sim Gamma(n, \lambda)$, so with LOTUS, we can get that

$$\mathbb{E}(\frac{1}{Y}) = \int_0^{+\infty} \frac{1}{y} \cdot \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} dy$$

Since $\Gamma(n) = (n-1)\Gamma(n-1)$, so we can get that

$$\mathbb{E}(\frac{1}{Y}) = \int_0^{+\infty} \frac{\lambda \cdot \lambda^{n-1}}{(n-1)\Gamma(n-1)} y^{n-2} e^{-\lambda y} dy$$

$$= \frac{\lambda}{n-1} \int_0^{+\infty} \frac{\lambda^{n-1}}{\Gamma(n-1)} y^{n-2} e^{-\lambda y} dy$$

Since $\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{n-2} e^{-\lambda y}$ is the PDF of $Gamma(n-1, \lambda)$, so

$$\int_0^{+\infty} \frac{\lambda^{n-1}}{\Gamma(n-1)} y^{n-2} e^{-\lambda y} dy = 1$$

so

$$\mathbb{E}(\frac{1}{Y}) = \frac{\lambda}{n-1}$$

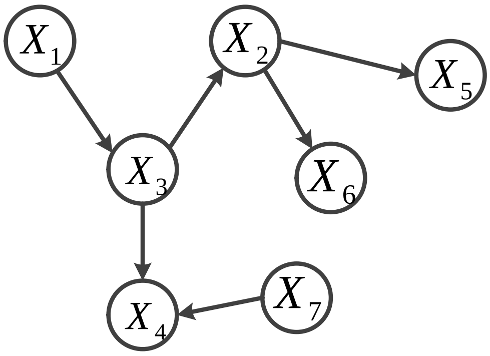
so

$$\mathbb{E}(\frac{n-1}{n} \hat{\lambda}_{MLE}) = (n-1)\mathbb{E}(\frac{1}{Y}) = (n-1) \cdot \frac{\lambda}{n-1} = \lambda$$

But there has a problem that, when $n = 1$, then in the upper steps, many steps include $\frac{1}{n-1}$ become valid, and $\Gamma(n-1) = \Gamma(0)$ is not valid, $Gamma(n-1, \lambda) \sim \Gamma(0, \lambda)$ is also not a valid distribution.
Actually, since $\lambda > 0$, and when $n = 1$, $\mathbb{E}(\frac{n-1}{n} \hat{\lambda}_{MLE}) = 0 \neq \lambda$.

So above all, when $n = 1$, $\frac{n-1}{n} \hat{\lambda}_{MLE}$ is not unbiased.
When $n > 1$, $\frac{n-1}{n} \hat{\lambda}_{MLE}$ is unbiased.

4. [10 points] [Graphical Models] Given the following Bayesian Network,



answer the following questions.

(a) Factorize the joint distribution of $X_1, \cdots, X_7$ according to the given Bayesian Network. [2 points]

(b) Justify whether $X_1 \perp X_5 \mid X_2$? [2 points]

(c) Justify whether $X_5 \perp X_7 \mid X_3, X_4$? [2 points]

(d) Justify whether $X_5 \perp X_7 \mid X_4$? [2 points]

(e) Write down the variables that are in the Markov blanket of X_3. [2 points]

Solution:
(a) $P(X_1, \cdots, X_7) = P(X_1)P(X_2|X_3)P(X_3|X_1)P(X_4|X_3, X_7)P(X_5|X_2)P(X_6|X_2)P(X_7)$

(b) Yes.
During the path of X_1 and X_5 the variables along the path is $\{X_1, X_3, X_2, X_5\}$.
And since X_2 is given, and its the "head to tail" of the path, so $X_1 \perp X_5|X_2$.

(c) Yes.
During the path of X_5 and X_7 the variables along the path is $\{X_5, X_2, X_3, X_4, X_7\}$.
Although X_4 is given, making the path not active, but X_3 is also given, so it blocks the path.
So $X_5 \perp X_7|X_3, X_4$.

(d) No.
During the path of X_5 and X_7 the variables along the path is $\{X_5, X_2, X_3, X_4, X_7\}$.
X_4 is given, making the path not active, so the path is not blocked.
So $X_5 \not\perp X_7|X_4$.

(e) The parents of X_3 is $\{X_1\}$.
The children of X_3 is $\{X_2, X_4\}$.
The other parents of the children of X_3 is $\{X_7\}$.
And the Markov blanket of X_3 the union of the parents, children and the other parents of the children of X_3.

So above all, the variables that are in the Markov blanket of X_3 is $\{X_1, X_2, X_4, X_7\}$.