CS 6190: Probabilistic Modelling Spring 2019

Homework 0

Aishwarya Gupta(01266423)

Handed out: 26 Aug, 2019
Due: 11:59pm, 5 Sep, 2019

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by **midnight of the due date**. Please submit the homework on Canvas.

Warm up[100 points + 10 bonus]

1. [10 points] Given two events A and B, prove that

$$
\begin{aligned}
p(A \cup B) &\leq p(A) + p(B)\\
p(A \cap B) &\leq p(A)\\
p(A \cap B) &\leq p(B)
\end{aligned}
$$

When will the equality conditions hold?

For disjoint events $A_1, A_2, ..., A_\infty$, the probability axiom states :

$$p(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i) \tag{1}$$

Using the set theory, we can write the following equalities :

$$(A \cup B) = (A - B) \cup (B - A) \cup (A \cap B) \tag{2}$$
$$A = (A - B) \cup (A \cap B) \tag{3}$$
$$B = (B - A) \cup (A \cap B) \tag{4}$$

As $(A - B)$, $(B - A)$ and $(A \cap B)$ are disjoint sets, we can use the probability axiom. Thus, using eq(1), (2), (3) and (4), we get :

$$\begin{aligned} p(A) &= p(A-B) + p(A \cap B) \\ p(A-B) &= p(A) - p(A \cap B) \end{aligned} \tag{5}$$

$$\begin{aligned} p(B) &= p(B-A) + p(A \cap B) \\ p(B-A) &= p(B) - p(A \cap B) \end{aligned} \tag{6}$$

$$\begin{aligned} p(A \cup B) &= p(A-B) + p(B-A) + p(A \cap B) \\ &= p(A) - p(A \cap B) + p(B) - p(A \cap B) + p(A \cap B) \\ &= p(A) + p(B) - p(A \cap B) \end{aligned} \tag{7}$$

Since for any event A_i, $p(A_i) \geq 0$, from equation (7), we get

$$p(A \cup B) \leq p(A) + p(B)$$

The equality will hold when A and B are independent events.
From equation (5) and (6), we get :

$$\begin{aligned} p(A \cap B) &\leq p(A) \\ p(A \cap B) &\leq p(B) \end{aligned}$$

The equality will hold when A and B are the same events

2. [5 points] Let $\{A_1, \ldots, A_n\}$ be a collection of events. Show that

$$p(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n p(A_i).$$

When does the equality hold? (Hint: induction)
For $i = 1$, $p(A_1) = p(A_1)$
For $i = 2$, $p(\cup_{i=1}^2 A_i) = p(A_1) + p(A_2) - p(A_1 \cap A_2) \leq p(A_1) + p(A_2)$
Let us assume that the given inequality exist for $i = k$, i.e.

$$p(\cup_{i=1}^k A_i) \leq \sum_{i=1}^k p(A_i) \tag{1}$$

Then, for $i = k + 1$,

$$\begin{aligned} p(\cup_{i=1}^{k+1} A_i) &= p((\cup_{i=1}^k A_i) \cup A_{k+1}) \\ &= p(\cup_{i=1}^k A_i) + p(A_{k+1}) + p((\cup_{i=1}^k A_i) \cap A_{k+1}) \end{aligned} \tag{2}$$

Since for any event A_i, $p(A_i) \geq 0$, from equation (2), we get

$$\begin{aligned} p(\cup_{i=1}^{k+1} A_i) &\leq p(\cup_{i=1}^k A_i) + p(A_{k+1}) \\ &\leq p(\cup_{i=1}^{k+1} A_i) \qquad \text{from equation(1)} \end{aligned}$$

If A_1, A_2, ..., A_n are pairwise disjoint, then the equality will exist i.e. $p(\cup_{i=1}^n A_i) = p(\cup_{i=1}^n A_i)$

3. [20 points] We use $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ to denote a random variable's mean (or expectation) and variance, respectively. Given two discrete random variables X and Y, where $X \in \{0, 1\}$ and $Y \in \{0, 1\}$. The joint probability $p(X, Y)$ is given in as follows:

   (a) [10 points] Calculate the following distributions and statistics.

	$Y=0$	$Y=1$
$X=0$	$3/10$	$1/10$
$X=1$	$2/10$	$4/10$

i. the the marginal distributions $p(X)$ and $p(Y)$

$$\begin{aligned}
p(X=0) &= p(X=0\cap Y=0)+p(X=0\cap Y=1)\\
&= 3/10+1/10\\
&= 0.4\\
p(X=1) &= p(X=1\cap Y=0)+p(X=1\cap Y=1)\\
&= 2/10+4/10\\
&= 0.6\\
p(Y=0) &= p(Y=0\cap X=0)+p(Y=0\cap X=1)\\
&= 3/10+2/10\\
&= 0.5\\
p(Y=1) &= p(Y=1\cap X=0)+p(Y=1\cap X=1)\\
&= 1/10+4/10\\
&= 0.5
\end{aligned}$$

ii. the conditional distributions $p(X|Y)$ and $p(Y|X)$

$$\begin{aligned}
p(X=0|Y=0) &= \frac{p(X=0,Y=0)}{p(Y=0)} = \frac{0.3}{0.5} = 0.6\\
p(X=1|Y=0) &= \frac{p(X=1,Y=0)}{p(Y=0)} = \frac{0.2}{0.5} = 0.4\\
p(X=0|Y=1) &= \frac{p(X=0,Y=1)}{p(Y=1)} = \frac{0.1}{0.5} = 0.2\\
p(X=1|Y=1) &= \frac{p(X=1,Y=1)}{p(Y=1)} = \frac{0.4}{0.5} = 0.8
\end{aligned}$$

	$X=0$	$X=1$
$p(Y=0\|X)$	0.75	$1/3$
$p(Y=1\|X)$	0.25	$2/3$

iii. $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{V}(X)$, $\mathbb{V}(Y)$

$$\begin{aligned}
&E[X] = \sum xp(x) = 0*0.4+1*0.6 = 0.6\\
&E[Y] = 0.5\\
&V(X) = \sum (x-\mathbf{E}(x))^2 p(x) = (0-0.6)^2*0.4+(1-0.6)^2*0.6 = 0.24\\
&V(Y) = 0.25
\end{aligned}$$

iv. $\mathbb{E}(Y|X=0)$, $\mathbb{E}(Y|X=1)$, $\mathbb{V}(Y|X=0)$, $\mathbb{V}(Y|X=1)$

$E[Y|X=0] = \sum yp(y|x=0) = 0*0.75 + 1*0.25 = 0.25$
$E[Y|X=1] = 2/3$
$E[X|Y=0] = 0.4$
$E[X|Y=1] = 0.8$
$V(Y|X=0) = \sum (y - \mathbf{E}(y|x=0))^2 p(y|x=0) = (0-0.25)^2 * 0.75 + (1-0.25)^2 * 0.25 = 0.1875$
$V(Y|X=1) = 2/9$

v. the covariance between X and Y

$$cov(X,Y) = E(XY) - E(X)E(Y)$$
$$E(XY) = 0*3/10 + 0*1/10 + 0*2/10 + 1*4/10 = 0.4$$
$$cov(X,Y) = 0.4 - (0.6)*(0.5) = 0.1$$

(b) [5 points] Are X and Y independent? Why?
X and Y are not independent as the conditional probability is not same as the marginal probability of X and Y i.e. $p(x|y) \neq p(x)$ and $p(y|x) \neq p(y)$

(c) [5 points] When X is not assigned a specific value, are $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ still constant? Why?
When X is not assigned any value, $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ will not be constant as the probability mass/density function $p(Y|X)$ is a function of X and Y. $p(Y|X)$ will be a value only when both X and Y are assigned any value.

4. [10 points] Assume a random variable X follows a standard normal distribution, i.e., $X \sim \mathcal{N}(X|0,1)$. Let $Y = e^{-X^2}$. Calculate the mean and variance of Y.

(a) $\mathbb{E}(Y)$

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} e^{-x^2} \frac{e^{\frac{-(x-0)^2}{2\times(1)}}}{\sqrt{2\pi\times(1)}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-3x^2}{2}} dx$$

Substituting $3x^2/2 = t^2$, we get, $dx = \sqrt{2/3}dt$. Thus, the above equation reduces to :

$$\mathbb{E}(Y) = \frac{1}{\sqrt{3\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt = \frac{1}{\sqrt{3\pi}} \times \sqrt{\pi} = \frac{1}{\sqrt{3}} \tag{8}$$

(b) $\mathbb{V}(Y)$

$$\begin{aligned}
\mathbb{V}(Y) &= \int_{-\infty}^{\infty} \left(e^{-x^2} - \frac{1}{\sqrt{3}}\right)^2 \frac{e^{\frac{-(x-0)^2}{2\times(1)}}}{\sqrt{2\pi\times(1)}} dx \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} e^{\frac{-5x^2}{2}} dx + \int_{-\infty}^{\infty} \frac{1}{3} e^{\frac{-x^2}{2}} dx - \int_{-\infty}^{\infty} \frac{2e^{\frac{-3x^2}{2}}}{\sqrt{3}} dx\right] \\
&= \left[\frac{\sqrt{\frac{1}{5}}}{\sqrt{2\pi\frac{1}{5}}} \int_{-\infty}^{\infty} e^{\frac{-x^2}{\frac{2}{5}}} dx\right] + \left[\frac{1}{3\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx\right] - \left[\frac{2}{3\sqrt{\frac{2\pi}{3}}} \int_{-\infty}^{\infty} e^{\frac{-x^2}{\frac{2}{3}}} dx\right] \\
&= \frac{1}{\sqrt{5}} + \frac{1}{3} - \frac{2}{3} = \frac{1}{\sqrt{5}} - \frac{1}{3} \quad \text{(since pdf of gaussian is 1)}
\end{aligned}$$

5. [10 points] Derive the probability density functions of the following transformed random variables.

(a) $X \sim \mathcal{N}(X|0,1)$ and $Y = X^3$.

$$y = x^3 \tag{1}$$
$$dy = 3x^2 dx$$
$$dx = \frac{1}{3y^{\frac{2}{3}}} dy \quad \text{from equation (1)} \tag{2}$$
$$p_X(x)dx = \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}} dx = \left(\frac{1}{3y^{\frac{2}{3}}}\right) \frac{e^{\frac{-y^{\frac{2}{3}}}{2}}}{\sqrt{2\pi}} dy = \quad p_Y(y)dy \quad \text{from equation(1) and equation(2)}$$
$$p_Y(y) = \left(\frac{1}{3\sqrt{2\pi}y^{\frac{2}{3}}}\right) e^{\frac{-y^{\frac{2}{3}}}{2}}$$

(b) $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix})$ and $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$.

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \frac{6}{7} \begin{bmatrix} 1 & -1/2 \\ 1/3 & 1 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \mathbf{K} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \mathbf{KY} \tag{1}$$
$$\mathbf{J} = \begin{vmatrix} \frac{\partial X_1}{\partial Y_1} & \frac{\partial X_1}{\partial Y_2} \\ \frac{\partial X_2}{\partial Y_1} & \frac{\partial X_2}{\partial Y_2} \end{vmatrix} = \begin{vmatrix} 6/7 & -3/7 \\ 2/7 & 6/7 \end{vmatrix} = \frac{6}{7}$$
$$\partial X = \frac{6}{7} \partial Y \tag{2}$$
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix})$$

From equation (1) and (2), the pdf of Y can be written as :

$$p(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}) = \frac{1}{2\pi|\mathbf{\Sigma}|^{1/2}} e^{\frac{-1}{2}(\mathbf{X}^\top \mathbf{\Sigma}^{-1} \mathbf{X})} (\partial X)$$
$$p(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}) = \frac{6}{14\pi|\mathbf{\Sigma}|^{1/2}} e^{\frac{-1}{2}((\mathbf{KY})^\top \mathbf{\Sigma}^{-1} (\mathbf{KY}))} (\partial Y)$$
$$= \frac{6}{14\pi|\mathbf{\Sigma}|^{1/2}} e^{\frac{-1}{2}(\mathbf{Y}^\top \mathbf{K}^\top \mathbf{\Sigma}^{-1} \mathbf{KY})} (\partial Y)$$
$$= \frac{6}{14\pi|\mathbf{\Sigma}|^{1/2}} e^{\frac{-1}{2}(\mathbf{Y}^\top \mathbf{\Sigma}_{\mathbf{Y}}^{-1} \mathbf{Y})} (\partial Y) \quad \text{where } \mathbf{\Sigma}_{\mathbf{Y}}^{-1} = \mathbf{K}^\top \mathbf{\Sigma}^{-1} \mathbf{K}$$
$$\mathbf{\Sigma}_{\mathbf{Y}}^{-1} = \frac{48}{49} \begin{bmatrix} 13/9 & 1/4 \\ 1/4 & 3/4 \end{bmatrix} \quad \text{and } |\mathbf{\Sigma}_{\mathbf{Y}}^{-1}| = 48/49$$

6. [10 points] Given two random variables X and Y, show that

(a) $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y|X)) &= \int \mathbb{E}(Y|X=x)p_X(x)dx = \int\int y p_{Y|X}(y|X=x)dy p_X(x)dx \\
&= \int\int y p_{Y|X}(y|X=x)p_X(x)dxdy \quad [\text{as } p_{Y|X}(y|X=x)p_X(x) = p_{X,Y}(x,y)] \\
&= \int\int y p_{X,Y}(x,y)dxdy \\
&= \int y \int p_{X,Y}(x,y)dxdy \quad \left[\text{as } \int p_{X,Y}(x,y)dx = p_Y(y)\right] \\
&= \int y p_Y(y)dy = \mathbb{E}(Y)
\end{aligned}$$

(b) $\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$

$$\begin{aligned}
\mathbb{V}(Y|X) &= \mathbb{E}(Y^2|X) - \mathbb{E}(Y|X)^2 \quad \text{using variance definition} \\
\mathbb{E}(\mathbb{V}(Y|X)) &= \mathbb{E}(\mathbb{E}(Y^2|X)) - \mathbb{E}(\mathbb{E}(Y|X)^2) \quad \text{taking expectation on both the sides} \\
&= \mathbb{E}(Y^2) - \mathbb{E}(\mathbb{E}(Y|X)^2) \quad \text{since } \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y|X) \\
&= (\mathbb{E}(Y^2) - (\mathbb{E}(Y))^2) + ((\mathbb{E}(Y))^2 - \mathbb{E}(\mathbb{E}(Y|X)^2)) \\
&= \mathbb{V}(Y) - (\mathbb{E}(\mathbb{E}(Y|X)^2) - \mathbb{E}(\mathbb{E}(Y|X))^2) \\
&= \mathbb{V}(Y) - \mathbb{V}(\mathbb{E}(Y|X)) \quad \text{(By variance definition.)} \qquad (1)
\end{aligned}$$

Thus, from equation(1), we get, $\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$. [Help is taken from this link]

(Hints: using definition.)

7. [15 points] Given a logistic function, $f(\mathbf{x}) = 1/(1+\exp(-\mathbf{a}^\top\mathbf{x}))$ ($\mathbf{x}$ is a vector),

(a) derive $\nabla f(\mathbf{x})$

$$\nabla f(\mathbf{x}) = -\left(\frac{1}{1+\exp(-\mathbf{a}^\top\mathbf{x}}\right)^2 \exp(-\mathbf{a}^\top\mathbf{x})(-\mathbf{a}) = \frac{\mathbf{a}\exp(-\mathbf{a}^\top\mathbf{x})}{(1+\exp(-\mathbf{a}^\top\mathbf{x}))^2}$$

(b) derive $\nabla^2 f(\mathbf{x})$

$$\begin{aligned}
\nabla f(\mathbf{x}) &= \mathbf{a}\frac{((1+\exp(-\mathbf{a}^\top\mathbf{x}))^2\exp(-\mathbf{a}^\top\mathbf{x}) - 2\exp(-\mathbf{a}^\top\mathbf{x})(1+\exp(-\mathbf{a}^\top\mathbf{x}))\exp(-\mathbf{a}^\top\mathbf{x}))}{(1+\exp(-\mathbf{a}^\top\mathbf{x}))^4}(-\mathbf{a}^\top) \\
&= \mathbf{a}\frac{\exp(-\mathbf{a}^\top\mathbf{x}) + \exp(-\mathbf{a}^\top\mathbf{x})^2 - 2\exp(-\mathbf{a}^\top\mathbf{x})^2}{(1+\exp(-\mathbf{a}^\top\mathbf{x}))^3}(-\mathbf{a}^\top) \\
&= \mathbf{a}\frac{(\exp(-\mathbf{a}^\top\mathbf{x}) - 1)\exp(-\mathbf{a}^\top\mathbf{x})}{(1+\exp(-\mathbf{a}^\top\mathbf{x}))^3}\mathbf{a}^\top
\end{aligned}$$

(c) show that $-log(f(\mathbf{x}))$ is convex

$$\begin{aligned}
\nabla(-log(f(\mathbf{x}))) &= -\frac{\nabla f(\mathbf{x})}{f(\mathbf{x})} \\
\nabla^2(-log(f(\mathbf{x}))) &= \nabla(-\frac{\nabla f(\mathbf{x})}{f(\mathbf{x})}) = -\frac{f(\mathbf{x})\nabla^2 f(\mathbf{x}) - (\nabla f(\mathbf{x}))^2}{f(\mathbf{x})^2}
\end{aligned}$$

Then from part (a) and (b), we get :

$$\nabla^2(-log(f(\mathbf{x}))) = -\mathbf{a}\frac{\frac{1}{1+exp(-\mathbf{a}^\top\mathbf{x})}\frac{exp(-\mathbf{a}^\top\mathbf{x})(exp(-\mathbf{a}^\top\mathbf{x})-1)}{(1+exp(-\mathbf{a}^\top\mathbf{x}))^3} - \frac{exp(-\mathbf{a}^\top\mathbf{x})^2}{(1+exp(-\mathbf{a}^\top\mathbf{x}))^4}}{\frac{1}{(1+exp(-\mathbf{a}^\top\mathbf{x}))^2}}\mathbf{a}^\top$$

$$= \mathbf{a}\frac{exp(-\mathbf{a}^\top\mathbf{x})}{(1+exp(-\mathbf{a}^\top\mathbf{x}))^2}\mathbf{a}\top \geq 0$$

Thus, $-log(f(\mathbf{x}))$ is a convex function

Note that $0 \leq f(\mathbf{x}) \leq 1$.

8. [10 points] Derive the convex conjugate for the following functions

(a) $f(x) = -\log(x)$

$$g(\lambda) = \max_x(\lambda x + \log(x))$$
$$\nabla_x g(\lambda) = \lambda + \frac{1}{x} = 0$$
$$x = -\frac{1}{\lambda}$$
$$g(\lambda) = -1 + \log(-\frac{1}{\lambda})$$

(b) $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x}$ where $\mathbf{A} \succ 0$

$$g(\lambda) = \max_x(\lambda^\top\mathbf{x} - \mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x})$$
$$\nabla_\mathbf{x} g(\lambda) = \lambda - 2\mathbf{A}^{-1}\mathbf{x} = 0$$
$$\mathbf{x} = \frac{\mathbf{A}\lambda}{2}$$
$$g(\lambda) = \frac{\lambda^\top\mathbf{A}\lambda - \lambda^\top\mathbf{A}^\top\mathbf{A}^{-1}\mathbf{A}\lambda}{2}$$
$$g(\lambda) = \frac{2\lambda^\top\mathbf{A}\lambda - \lambda^\top\mathbf{A}^\top\lambda}{4} \quad (9)$$

9. [10 points] Derive the (partial) gradient of the following functions

(a) $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log\left(\mathcal{N}(\mathbf{a}|\mathbf{A}\boldsymbol{\mu}, \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top)\right)$, derive $\frac{\partial f}{\partial \boldsymbol{\mu}}$ and $\frac{\partial f}{\partial \boldsymbol{\Sigma}}$,

$$f(\mathbf{u}, \mathbf{V}) = \log\left(\mathcal{N}(\mathbf{a}|\mathbf{u}, \mathbf{V})\right)$$
$$= \log\left(\frac{1}{\sqrt{(2\pi)^{n/2}}}\right) - \frac{\log(|\mathbf{V}|)}{2} + \log(e^{\frac{-1}{2}(\mathbf{x}-\mathbf{u})^\top\mathbf{V}^{-1}(\mathbf{x}-\mathbf{u})})$$
$$= \log\left(\frac{1}{\sqrt{(2\pi)^{n/2}}}\right) - \frac{\log(|\mathbf{V}|)}{2} - \frac{1}{2}(\mathbf{x}-\mathbf{u})^\top\mathbf{V}^{-1}(\mathbf{x}-\mathbf{u})$$
$$\frac{\partial f}{\partial \mathbf{u}} = \frac{-1}{2}\frac{\partial((\mathbf{x}-\mathbf{u})^\top\mathbf{V}^{-1}(\mathbf{x}-\mathbf{u}))}{\partial \mathbf{u}} = \frac{-1}{2}\frac{\partial(\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y})}{\partial \mathbf{u}} \quad \text{where } \mathbf{y} = (\mathbf{x}-\mathbf{u})$$
$$\frac{\partial(\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y})}{\partial \mathbf{u}} = \frac{\partial(\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y})}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{u}}$$
$$= \mathbf{y}^\top(\mathbf{V}^{-1} + (\mathbf{V}^{-1})^\top)(-1)$$
$$\frac{\partial f}{\partial \mathbf{u}} = \frac{1}{2}\mathbf{y}^\top(\mathbf{V}^{-1} + (\mathbf{V}^{-1})^\top) = \frac{1}{2}\mathbf{y}^\top(2\mathbf{V}^{-1}) = \mathbf{y}^\top\mathbf{V}^{-1} \quad \text{Since } \mathbf{V} \text{ is a symmetric matrix}$$
$$= (\mathbf{x}-\mathbf{u})^\top\mathbf{V}^{-1} \quad (1)$$

$$
\begin{aligned}
f(\mathbf{u},\mathbf{V}) &= \log\left(\frac{1}{\sqrt{(2\pi)^{n/2}}}\right) - \frac{\log(|\mathbf{V}|)}{2} - \frac{1}{2}(\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}(\mathbf{x}-\mathbf{u}) \\
\frac{\partial f}{\partial \mathbf{V}} &= -\frac{1}{2}\frac{\partial((\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}(\mathbf{x}-\mathbf{u}))}{\partial \mathbf{V}} - \frac{\partial(\log|\mathbf{V}|)}{2\partial \mathbf{V}} \\
&= -\frac{1}{2}\frac{\partial(tr((\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}(\mathbf{x}-\mathbf{u})))}{\partial \mathbf{V}} - \frac{1}{2|\mathbf{V}|}\frac{\partial|\mathbf{V}|}{\partial \mathbf{V}} \\
&= -\frac{1}{2}\frac{\partial(tr((\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}))}{\partial \mathbf{V}} - \frac{1}{2|\mathbf{V}|}(|\mathbf{V}|tr(\mathbf{V}^{-1})) \\
&= -\frac{1}{2}tr(\frac{\partial((\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1})}{\partial \mathbf{V}}) - \frac{tr(\mathbf{V}^{-1})}{2} \\
&= -\frac{1}{2}tr(-(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}\mathbf{V}^{-1}) - \frac{tr(\mathbf{V}^{-1})}{2} \\
&= \frac{1}{2}tr((\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}\mathbf{V}^{-1}) - \frac{tr(\mathbf{V}^{-1})}{2} \qquad (2)
\end{aligned}
$$

Using the equation (1) and (2), and substituting $\mathbf{A}\boldsymbol{\mu}$ as $\mathbf{u}$ and $\mathbf{S\Sigma S}$ as $\mathbf{V}$ we can write it as

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}} f(\mathbf{A}\boldsymbol{\mu}, \mathbf{S\Sigma S}^\top) &= \frac{\partial f(\mathbf{u},\mathbf{V})}{\partial(\mathbf{u})}\frac{\partial \mathbf{u}}{\partial \boldsymbol{\mu}} \\
&= (\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}\mathbf{A}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}} f(\mathbf{A}\boldsymbol{\mu}, \mathbf{S\Sigma S}^\top) &= \frac{\partial f(\mathbf{u},\mathbf{V})}{\partial(\mathbf{V})}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\Sigma}} \\
&= \left[\frac{1}{2}tr((\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}\mathbf{V}^{-1}) - \frac{tr(\mathbf{V}^{-1})}{2}\right](\mathbf{SS}^\top)
\end{aligned}
$$

(b) $f(\boldsymbol{\Sigma}) = \log\left(\mathcal{N}(\mathbf{a}|\mathbf{b}, \mathbf{K}\otimes\boldsymbol{\Sigma})\right)$ where $\otimes$ is the Kronecker product (Hint: check Minka's notes).

$$
\begin{aligned}
f(\boldsymbol{\Sigma}) &= \log\left(\mathcal{N}(\mathbf{a}|\mathbf{b}, \mathbf{K}\otimes\boldsymbol{\Sigma})\right) \\
&= \log\left(\frac{1}{\sqrt{(2\pi)^{n/2}}}\right) - \frac{\log(|\mathbf{K}\otimes\boldsymbol{\Sigma}|)}{2} - \frac{1}{2}(\mathbf{a}-\mathbf{b})^\top(\mathbf{K}\otimes\boldsymbol{\Sigma})^{-1}(\mathbf{a}-\mathbf{b})
\end{aligned}
$$

The differentiation of Kronecker product is :

$$
\begin{aligned}
\frac{\partial(\mathbf{K}\otimes\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}} &= \frac{\partial \mathbf{K}}{\partial \boldsymbol{\Sigma}}\otimes\boldsymbol{\Sigma} + \mathbf{K}\otimes\frac{\partial}{\partial\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \\
&= K\otimes I \qquad (3)
\end{aligned}
$$

Using the equation (2) and (3), and substituting $\mathbf{b}$ as $\mathbf{u}$ and $\mathbf{K}\otimes\boldsymbol{\Sigma}$ as $\mathbf{V}$ we can write it as :

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\Sigma}} f(\boldsymbol{\Sigma}) &= \frac{\partial f(\mathbf{u},\mathbf{V})}{\partial(\mathbf{V})}\frac{\partial \mathbf{V}}{\partial\boldsymbol{\Sigma}} \\
&= \left(\frac{1}{2}tr((\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^\top \mathbf{V}^{-1}\mathbf{V}^{-1}) - \frac{tr(\mathbf{V}^{-1})}{2}\right)(\mathbf{K}\otimes\mathbf{I})
\end{aligned}
$$

10. [**Bonus**][10 points] Show that for any square matrix $\mathbf{X} \succ 0$, $\log|\mathbf{X}|$ is concave to $\mathbf{X}$.

$f(t) = \log|\mathbf{X} + t\mathbf{Y}| \quad \text{where } (\mathbf{X} + t\mathbf{Y}) \succ 0$

As $\mathbf{X}$ is positive definite matrix, there exists $\mathbf{X}^{1/2}$ such that $\mathbf{X} = \mathbf{X}^{1/2}\mathbf{X}^{1/2}$

$f(t) = \log|\mathbf{X}^{1/2}\mathbf{X}^{1/2} + t\mathbf{X}^{1/2}\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2}\mathbf{X}^{1/2}|$

$= \log|\mathbf{X}^{1/2}(\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}| \quad \text{since} : |\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$

$= \log|\mathbf{X}| + log|\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2}|$

Let $\lambda_1, \lambda_2 ... \lambda_d$ be the eigenvalues of $\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2}$, then

$$log|\mathbf{I} + t\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2}| = \log\prod_i^d (1 + t\lambda_i)$$

$$= \sum_i^d \log(1 + t\lambda_i)$$

$$g(t) = \log|\mathbf{X}| + \sum_i^d \log(1 + t\lambda_i)$$

$$\nabla^2 g(t) = -\sum_i^d \frac{\lambda_i^2}{(1 + t\lambda_i)^2)}$$

$$\nabla^2(-g(t)) = \sum_i^d \frac{\lambda_i^2}{(1 + t\lambda_i)^2)} \geq 0$$

Thus $-g(t)$ is convex i.e. $-\log|\mathbf{X}|$ is convex or $\log|\mathbf{X}|$ is concave [Help is taken from from this link]