CS 6190: Probabilistic Modelling Spring 2019

Homework 0

Abhinav Kumar (u1209853)

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

$$p(A \cup B) \le p(A) + p(B)$$
$$p(A \cap B) \le p(A)$$
$$p(A \cap B) \le p(B)$$

When will the equality conditions hold?

We know that

$$A \cup B = (A - B) \cup (A \cap B) \cup (B - A) \tag{1}$$

The third axiom of probability says that probability of union of mutually exclusive events is equal to the sum of the probabilities and so

$$p(A \cup B) = p(A - B) + p(A \cap B) + p(B - A) \tag{2}$$

However, $A = (A \cap B) \cup (A - B)$ and therefore $p(A) = p(A \cap B) + p(A - B)$ and so

$$p(A - B) = p(A) - p(A \cap B) \tag{3}$$

Similarly,

$$p(B - A) = p(B) - p(A \cap B) \tag{4}$$

.

Substituting equations (3) and (4) in (2), we get

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

(a) Using equation (5), since all probabilities are non-negative (1st axiom of probability), we can say that

$$p(A \cup B) \leq p(A) + p(B) \tag{6}$$

Strict equality holds when $p(A \cap B) = 0$ or when A and B are mutually disjoint.

(b) Using equation (3),

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

Strict equality holds when $p(A - B) = 0$ or when A is inside B.

(c) Using equation (4),

$$\begin{aligned} p(A \cap B) &= p(B) - p(B - A) \\ p(A \cap B) &\leq p(B) \end{aligned} \tag{8}$$

Strict equality holds when $p(B - A) = 0$ or when B is inside A.

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

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

When does the equality hold? (Hint: induction)

The base case definitely holds. $p(A \cup B) \leq p(A) + p(B)$. This has been shown in question 1(a).

The inductive step is as follows. Assume it is true for some k. So, we have Now, we have

$$p\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} p(A_i) \tag{9}$$

Now, we need to show it for $k+1$. We have,

$$\begin{aligned} p\left(\bigcup_{i=1}^{k+1} A_i\right) &= p\left(\bigcup_{i=1}^{k} A_i \cup A_{k+1}\right) \\ &\leq p\left(\bigcup_{i=1}^{k} A_i\right) + p(A_{k+1}) \text{ Using base step} \\ &\leq \sum_{i=1}^{k} p(A_i) + p(A_{k+1}) \text{ Using induction assumption} \\ &\leq \sum_{i=1}^{k+1} p(A_i) \end{aligned} \tag{10}$$

Thus, we satisfy the induction step as well.

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:

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

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

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

$$p(X) = \begin{cases} 0.4 & \text{for } X=0 \\ 0.6 & \text{for } X=1 \end{cases}$$

$$p(Y) = \begin{cases} 0.5 & \text{for } Y=0 \\ 0.5 & \text{for } Y=1 \end{cases}$$

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

$$p(X/Y=0) = \begin{cases} 0.6 & \text{for } X=0 \\ 0.4 & \text{for } X=1 \end{cases}$$

$$p(X/Y=1) = \begin{cases} 0.2 & \text{for } X=0 \\ 0.8 & \text{for } X=1 \end{cases}$$

$$p(Y/X=0) = \begin{cases} 0.75 & \text{for } Y=0 \\ 0.25 & \text{for } Y=1 \end{cases}$$

$$p(Y/X=1) = \begin{cases} 0.33 & \text{for } Y=0 \\ 0.67 & \text{for } Y=1 \end{cases}$$

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

$\mathbb{E}(X) = 0.6$
$\mathbb{E}(Y) = 0.5$
$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 0.24$
$\mathbb{V}(X) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = 0.25$

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

$\mathbb{E}(Y|X=0) = 0.25$
$\mathbb{E}(Y|X=1) = 0.67$
$\mathbb{V}(Y|X=0) = 0.1875$
$\mathbb{V}(Y|X=1) = 0.2222$

v. the covariance between X and Y

$Cov(XY) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = 0.1$

(b) [5 points] Are X and Y independent? Why?

$\mathbb{E}(XY) = 0.4 \neq \mathbb{E}(X)\mathbb{E}(Y)$. Since these are not equal, so they are not independent.

(c) [5 points] When X is not assigned a specific value, are $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ still constant? Why?

They are not constant since X and Y are not independent. Had they been independent $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ would have been same for all value of X

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) $\mathbb{E}(Y)$

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

As, $X \sim \mathcal{N}(X|0,1)$, $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Clearly, Y can only take positive values. Since, $Y = e^{-X^2}$, therefore $X = g(Y) = \pm\sqrt{\log \dfrac{1}{y}}$. Now, the pdf of the transformed random variable Y is given by

$$
\begin{aligned}
f_Y(y) &= \sum \left| \frac{dg(Y)}{dy} \right| f_X\left(x = g(Y)\right) \\
&= \begin{cases} \frac{1}{y\sqrt{\log\left(\frac{1}{y}\right)}} \frac{\sqrt{y}}{\sqrt{2\pi}} & \text{for } y \in [0,1] \\ 0 & \text{elsewhere} \end{cases} \qquad (11) \\
&= \begin{cases} \frac{1}{\sqrt{y\log\left(\frac{1}{y}\right)}} \frac{1}{\sqrt{2\pi}} & \text{for } y \in [0,1] \\ 0 & \text{elsewhere} \end{cases} \qquad (12)
\end{aligned}
$$

(a)

$$
\begin{aligned}
\mathbb{E}(Y) &= \int_{-\infty}^{\infty} y f_Y(y) dy \\
&= \int_0^1 \frac{1}{\sqrt{\log\left(\frac{1}{y}\right)}} \frac{\sqrt{y}}{\sqrt{2\pi}} dy
\end{aligned}
\qquad (13)
$$

Substitute $t = \sqrt{\log \frac{1}{y}}$ and so $\frac{dy}{\sqrt{\log \frac{1}{y}}} = -2ydt = -2e^{-t^2}dt$

$$
\begin{aligned}
\mathbb{E}(Y) &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} 2e^{-t^2} dt e^{-t^2/2} \\
&= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{3t^2}{2}} dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{3t^2}{2}} dt \\
&= \frac{\sqrt{\frac{1}{3}}}{\sqrt{2\pi\frac{1}{3}}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2\frac{1}{3}}} dt \\
&= \sqrt{\frac{1}{3}} \quad \text{(Integration of PDF of Gaussian RV is 1)} \qquad (14)
\end{aligned}
$$

(b)

$$\begin{aligned}\mathbb{E}(Y^2) &= \int_{-\infty}^{\infty} y^2 f_Y(y)dy \\ &= \int_0^1 \frac{1}{\sqrt{\log\left(\frac{1}{y}\right)}}\frac{y^{1.5}}{\sqrt{2\pi}}dy\end{aligned} \tag{15}$$

Substitute $t = \sqrt{\log \frac{1}{y}}$ and so $\frac{dy}{\sqrt{\log \frac{1}{y}}} = -2ydt = -2e^{-t^2}dt$

$$\begin{aligned}\mathbb{E}(Y^2) &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} 2e^{-t^2}dte^{-3t^2/2} \\ &= \frac{2}{\sqrt{2\pi}}\int_0^{\infty} e^{-\frac{5t^2}{2}}dt \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{5t^2}{2}}dt \\ &= \frac{\sqrt{\frac{1}{5}}}{\sqrt{2\pi\frac{1}{5}}}\int_{-\infty}^{\infty} e^{-\frac{t^2}{2\frac{1}{5}}}dt \\ &= \sqrt{\frac{1}{5}} \quad \text{(Integration of PDF of Gaussian RV is 1)}\end{aligned} \tag{16}$$

$$\tag{17}$$

So, $\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \frac{1}{\sqrt{5}} - \frac{1}{3}$

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

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

 (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 $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$.

 (a) As, $X \sim \mathcal{N}(X|0,1)$, $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$.
 Since, $Y = X^3$ which is a monotonic function and therefore $X = g(Y) = Y^{1/3}$. Now, the pdf of the transformed random variable Y is given by

$$\begin{aligned}f_Y(y) &= \left|\frac{dg(Y)}{dy}\right| f_X\left(x = g(Y)\right) \\ &= \frac{1}{3y^{2/3}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^{2/3}}{2}} \quad \text{for } y \neq 0\end{aligned} \tag{18}$$

(b) • For 2D multivariate random variable,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{e^{-\dfrac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}}}{2\pi|\boldsymbol{\Sigma}|^{0.5}} \tag{19}$$

Substituting the values of $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}$. Therefore, $|\boldsymbol{\Sigma}| = \frac{3}{4}$ and $\boldsymbol{\Sigma}^{-1} = \frac{4}{3}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$ Substituting, we get

$$f_{\mathbf{X}} = \frac{e^{-\frac{2}{3}(x_1^2+x_1x_2+x_2^2)}}{2\pi\left(\frac{3}{4}\right)^{1/2}} \tag{20}$$

• Now, we know that if we use a transformation $\mathbf{Y} = \mathbf{A}\mathbf{X}+\mathbf{b}$, then the pdf of the transformed random variable is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\mathbf{A}|} f_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{Y}-\mathbf{b})) \tag{21}$$

. Reference-ECE Notes http://ece-research.unm.edu/bsanthan/ece340/note1.pdf
$|\mathbf{A}| = 7/6$ and $\mathbf{b} = 0$ Also, $\mathbf{A}^{-1} = \frac{6}{7}\begin{bmatrix} 1 & -1/2 \\ 1/3 & 1 \end{bmatrix}$ and so $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} = \frac{6}{7}\begin{bmatrix} y_1 - \frac{y_2}{2} \\ \frac{y_1}{3} + y_2 \end{bmatrix}$. Substituting,

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{e^{-\frac{2}{3}\left(\frac{6}{7}\right)^2\left(\frac{13}{9}y_1^2+\frac{1}{2}y_1y_2+\frac{3}{4}y_2^2\right)}}{2\pi\left(\frac{3}{4}\right)^{1/2}\frac{7}{6}} \tag{22}$$

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

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

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

(Hints: using definition.)

(a)

$$
\begin{aligned}
\mathbb{E}(\mathbb{E}(Y|X)) &= \int_{-\infty}^{\infty} \mathbb{E}(Y|X=x) f_X(x) dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{Y|X}(y|X=x) dy f_X(x) dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{Y|X}(y|X=x) f_X(x) dy dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{Y,X}(y,x) dy dx \\
&= \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{Y,X}(y,x) dx dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) dy \\
&= \mathbb{E}(Y)
\end{aligned} \tag{23}
$$

(b)

$$
\begin{aligned}
\mathbb{V}(Y) &= \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 \\
&= \mathbb{E}(\mathbb{E}(Y^2|X)) - [\mathbb{E}(\mathbb{E}(Y|X))]^2 \\
&= \mathbb{E}(\mathbb{V}(Y|X) + [\mathbb{E}(Y|X)]^2) - [\mathbb{E}(\mathbb{E}(Y|X))]^2 \\
&= \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{E}([\mathbb{E}(Y|X)]^2) - [\mathbb{E}(\mathbb{E}(Y|X))]^2 \\
&= \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))
\end{aligned} \tag{24}
$$

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 $\nabla f(\mathbf{x})$
   (b) derive $\nabla^2 f(\mathbf{x})$
   (c) show that $-log(f(\mathbf{x}))$ is convex

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

(a)

$$
\begin{aligned}
\nabla f(\mathbf{x}) &= \frac{-1}{(1+\exp(-\mathbf{a}^\top\mathbf{x}))^2}\exp(-\mathbf{a}^\top\mathbf{x})(-\mathbf{a}) \\
&= f(\mathbf{x})(1-f(\mathbf{x}))\mathbf{a}
\end{aligned} \tag{25}
$$

(b)

$$\begin{aligned}
\nabla^2 f(\mathbf{x}) &= \nabla.\nabla^T f(\mathbf{x}) \\
&= \nabla.[f(\mathbf{x})(1-f(\mathbf{x}))\mathbf{a}^T] \\
&= [\nabla f(\mathbf{x})](1-f(\mathbf{x})).\mathbf{a}^T - f(\mathbf{x})[\nabla f(\mathbf{x})].\mathbf{a}^T \\
&= [\nabla f(\mathbf{x})](1-2f(\mathbf{x})).\mathbf{a}^T \\
&= f(\mathbf{x})(1-f(\mathbf{x}))(1-2f(\mathbf{x}))\mathbf{a}\mathbf{a}^T
\end{aligned} \tag{26}$$

(c) To show that $-\log(f(\mathbf{x}))$ is convex, it is sufficient to show that $\nabla^2(-\log(f(\mathbf{x}))) \succeq 0$ or $\mathbf{z}^T\nabla^2(-\log(f(\mathbf{x})))\mathbf{z} \geq 0$ for arbitrary $\mathbf{z}$.

Now,

$$\nabla(-\log(f(\mathbf{x}))) = \frac{-1}{f(\mathbf{x})} f(\mathbf{x})(1-f(\mathbf{x}))\mathbf{a} = (f(\mathbf{x})-1)\mathbf{a} \quad \text{Using (25)}$$

We can now easily calculate $\nabla^2(-\log(f(\mathbf{x})))$ which is given by

$$\begin{aligned}
\nabla^2(-\log(f(\mathbf{x}))) &= \nabla.\nabla^T(-\log(f(\mathbf{x}))) \\
&= \nabla.[(f(\mathbf{x})-1)\mathbf{a}^T] \\
&= f(\mathbf{x})(1-f(\mathbf{x}))\mathbf{a}\mathbf{a}^T \quad \text{Using (25)}
\end{aligned} \tag{27}$$

Now, let $\mathbf{z}$ be an arbitrary vector. Then $\mathbf{z}^T\nabla^2(-\log(f(\mathbf{x})))\mathbf{z} = f(\mathbf{x})(1-f(\mathbf{x}))\mathbf{z}^T\mathbf{a}\mathbf{a}^T\mathbf{z} = f(\mathbf{x})(1-f(\mathbf{x}))(\mathbf{z}^T\mathbf{a})^2$. Since, $0 \leq f(\mathbf{x}) \leq 1$ so, we have $f(\mathbf{x}) \geq 0$ and $1-f(\mathbf{x}) \geq 0$. Also, $(\mathbf{z}^T\mathbf{a})^2 \geq 0$. Hence, we have $\mathbf{z}^T\nabla^2(-\log(f(\mathbf{x})))\mathbf{z} \geq 0$ for arbitrary $\mathbf{z}$.

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

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

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

The convex conjugate $f^*(y)$ is given by

$$\max_x (y^T x - f(x))$$

(a) When $f(x) = -log(x), x \in \mathbf{R}$, we have

$$f^*(y) = \max_x (yx + \log(x))$$

Differentiating wrt x and equating to 0, we get $x = -1/y$. Substituting, we get

$$f^*(y) = -\log(-y) - 1, y \in (-\infty, 0) \tag{28}$$

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

$$f^*(\mathbf{y}) = \max_{\mathbf{x}} (\mathbf{y}^T\mathbf{x} - \mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x})$$

Differentiating wrt $\mathbf{x}$ and equating to 0, we get

$$\begin{aligned}
\mathbf{y} - 2\mathbf{A}^{-1}\mathbf{x} &= 0 \\
or, \mathbf{x} &= \frac{1}{2}\mathbf{A}\mathbf{y}
\end{aligned}$$

Substituting, we get

$$
\begin{aligned}
f^*(\mathbf{y}) &= \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} - \frac{1}{4}\mathbf{y}^T\mathbf{A}^T\mathbf{A}^{-1}\mathbf{A}\mathbf{y} \\
&= \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} - \frac{1}{4}\mathbf{y}^T\mathbf{A}^T\mathbf{y} \qquad (29)
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
g_{\mathbf{X}}(\mathbf{x};\mathbf{m},\mathbf{C}) &= \frac{e^{-\frac{(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}{2}}}{(2\pi)^{k/2}|\mathbf{C}|^{0.5}} \\
f_{\mathbf{X}}(\mathbf{x};\mathbf{m},\mathbf{C}) = \log g_{\mathbf{X}}(\mathbf{x};\mathbf{m},\mathbf{C}) &= -\frac{(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}{2} - \frac{1}{2}\log|\mathbf{C}| + \mathit{constant} \qquad (30)
\end{aligned}
$$

Clearly,

$$
\frac{\partial f}{\partial \mathbf{m}} = \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) \qquad (31)
$$

.

The derivative w.r.t. $\mathbf{C}$ is a bit tricky to compute. First, we use the Matrix Cookbook to get the following formulae:

$$
\begin{aligned}
\frac{\partial|\mathbf{C}|}{\partial \mathbf{C}} &= |\mathbf{C}|(\mathbf{C}^{-T}) \\
\frac{\partial \mathbf{a}^T\mathbf{C}^{-1}\mathbf{a}}{\partial \mathbf{C}} &= -\mathbf{C}^{-T}\mathbf{a}\mathbf{a}^T\mathbf{C}^{-T}
\end{aligned}
$$

Clearly,

$$
\begin{aligned}
\frac{\partial f}{\partial \mathbf{C}} &= \frac{1}{2}\mathbf{C}^{-T}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-T} - \frac{1}{2|\mathbf{C}|}|\mathbf{C}|\mathbf{C}^{-T} \\
&= \frac{1}{2}\mathbf{C}^{-T}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\mathbf{C}^{-T} - \frac{1}{2}\mathbf{C}^{-T} \qquad (32)
\end{aligned}
$$

   (a) Clearly, $\mathbf{m} = \mathbf{A}\boldsymbol{\mu}$. We use the chain rule

$$
\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\mu}} &= \frac{\partial f}{\partial \mathbf{m}}\frac{\partial \mathbf{m}}{\partial \boldsymbol{\mu}} \\
&= \mathbf{C}^{-1}(\mathbf{x}-\mathbf{A}\boldsymbol{\mu})\mathbf{A} \qquad (33)
\end{aligned}
$$

   Next, we have $\mathbf{C} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top$ and so we again use the chain rule

$$
\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\Sigma}} &= \frac{\partial f}{\partial \mathbf{C}}\frac{\partial \mathbf{C}}{\partial \boldsymbol{\Sigma}} \\
&= \left[\frac{1}{2}\mathbf{C}^{-T}(\mathbf{x}-\mathbf{A}\boldsymbol{\mu})(\mathbf{x}-\mathbf{A}\boldsymbol{\mu})^T\mathbf{C}^{-T} - \frac{1}{2}\mathbf{C}^{-T}\right]\mathbf{S}\mathbf{S}^\top \qquad (34)
\end{aligned}
$$

   where $\mathbf{C} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top$

(b) We have $\mathbf{m} = \mathbf{b}$ and $\mathbf{C} = \mathbf{K} \otimes \mathbf{\Sigma}$ and so we again use the chain rule

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{\Sigma}} &= \frac{\partial f}{\partial \mathbf{C}} \frac{\partial \mathbf{C}}{\partial \mathbf{\Sigma}} \\ &= \left[\frac{1}{2}\mathbf{C}^{-T}(\mathbf{x} - \mathbf{b})(\mathbf{x} - \mathbf{b})^T\mathbf{C}^{-T} - \frac{1}{2}\mathbf{C}^{-T}\right] \mathbf{K} \otimes \mathbf{I} \end{aligned} \tag{35}$$

where $\mathbf{C} = \mathbf{K} \otimes \mathbf{\Sigma}$

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

We apply the fact that a function is convex if and only if its restriction to any line is convex to prove that log determinant function is a concave function. Define $g(t) = \log |\mathbf{X} + t\mathbf{V}|$ where $\mathbf{X} + t\mathbf{V} \succ 0$. Since, $\mathbf{X}$ is positive definite, we can split $\mathbf{X} = \mathbf{X}^{\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$ and substitute as follows

$$\begin{aligned} g(t) &= \log \left|\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} + t\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{\frac{-1}{2}}\mathbf{V}\mathbf{X}^{\frac{-1}{2}}\mathbf{X}^{\frac{1}{2}}\right| \\ &= \log \left|\mathbf{X}^{\frac{1}{2}}(\mathbf{I} + t\mathbf{X}^{\frac{-1}{2}}\mathbf{V}\mathbf{X}^{\frac{-1}{2}})\mathbf{X}^{\frac{1}{2}}\right| \\ &= \log |\mathbf{X}| + \log \left|\mathbf{I} + t\mathbf{X}^{\frac{-1}{2}}\mathbf{V}\mathbf{X}^{\frac{-1}{2}}\right| \quad \text{using } |\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \end{aligned} \tag{36}$$

Since, $\mathbf{X} \succ 0$ and $\mathbf{X} + t\mathbf{V} \succ 0$, hence we also have $\mathbf{X}^{\frac{-1}{2}}\mathbf{V}\mathbf{X}^{\frac{-1}{2}} \succ 0$. Let $\lambda_i > 0$ be the eigen values of the matrix $\mathbf{X}^{\frac{-1}{2}}\mathbf{V}\mathbf{X}^{\frac{-1}{2}}$ and so we have

$$\begin{aligned} g(t) &= \log |\mathbf{X}| + \log \left[\prod_i (1 + t\lambda_i)\right] \\ &= \log |\mathbf{X}| + \sum_i \log (1 + t\lambda_i) \end{aligned} \tag{37}$$

The second order derivative is then given by

$$g''(t) = -\sum_i \frac{\lambda_i}{(1 + t\lambda_i)^2} < 0 \tag{38}$$

Hence, $g(t)$ is concave and therefore, $\log |\mathbf{X}|$ is concave to $\mathbf{X}$.

Reference: Piazza https://piazza-resources.s3.amazonaws.com/is58gs5cfya7ft/itawqt5undn2bv/lecture8.pdf