CS 6190: Probabilistic Machine Learning Spring 2022

Solutions to Homework 0
by: Meysam Alishahi
(UNID: u01323606)

Handed out: 10 Jan, 2022
Due: 11:59pm, 21 Jan, 2022

- 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 + 5 bonus]

1. [2 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 does the equality hold?

In view of the third probability Axiom, known as the assumption of σ-additivity, we have

$$p(A \cup B) = P(A \setminus B) + P(B) \qquad \text{and} \qquad p(A) = p(A \setminus B) + p(A \cap B) \tag{1}$$

since $(A \setminus B) \cap B = \varnothing$ and $(A \setminus B) \cap B = \varnothing$. Since $p(B)$ is non-negative, these equalities imply

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

To have the equality, we need $p(A) - p(A \cap B) + p(B) = p(A) + p(B)$ which concludes $p(A \cap B) = 0$. In particular, when $A \cap B = \varnothing$ we have the equlity.

In view of the second equality in Equality (1),

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

Again, to have the equality, we should have $p(A) - p(A \setminus B) = p(A)$ or equivalently, $p(A \setminus B) = 0$. In particular, when $A \subseteq B$, the equality holds. With a similar approach, we can prove $p(A \cap B) \leq p(B)$ and for which the equality holds if and only if $p(B \setminus A) = 0$. In particular, when $B \subseteq A$, we have the equality.

2. [2 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)
We are going to prove the following statement:

> *If $\{A_1, \ldots, A_n\}$ are a collection of events, then $p(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} p(A_i)$ and equality holds if and only if $p(A_i \cap A_j) = 0$ for each $i \neq j \in \{1, \ldots, n\}$. In particular, if $A_1, \ldots, A_n$ are pairwise disjoint, i.e., $A_i \cap A_j = \varnothing$ for each $i \neq j \in \{1, \ldots, n\}$, then we have the equality.*

Proof. The induction base ($n = 2$) is holding due to Exercise 1. Assume that the statement is true for $n = k \geq 2$ and we want to prove it for $n = k + 1$. Set $A = \cup_{i=1}^{k} A_i$ and $B = A_{k+1}$. By induction base and hypothesis, we have

$$\begin{aligned} p(\cup_{i=1}^{n} A_i) &= p(A \cup B) \\ &\leq p(A) + p(B) \\ &= p(\cup_{i=1}^{k} A_i) + p(A_{k+1}) \\ &\leq \sum_{i=1}^{k} p(A_i) + p(A_{k+1}). \end{aligned}$$

To have the equality, we should have

$$p(A \cup B) = p(A) + p(B) \qquad \text{and} \qquad p(\cup_{i=1}^{k} A_i) = \sum_{i=1}^{k} p(A_i).$$

Again using induction base, we conclude $p((\cup_{i=1}^{k} A_i) \cap A_{k+1}) = 0$, which implies $p(A_i \cap A_{k+1}) = 0$ for each $i \in \{1, \cdots, k\}$. Also, from the induction hypothesis, since $p(\cup_{i=1}^{k} A_i) = \sum_{i=1}^{k} p(A_i)$, we got

$$p(A_i \cap A_j) = 0 \quad \text{for each } i \neq j \in \{1, \ldots, k\}.$$

Overall, we proved that if $p(\cup_{i=1}^{k+1} A_i) = \sum_{i=1}^{k+1} p(A_i)$, then

$$p(A_i \cap A_j) = 0 \quad \text{for each } i \neq j \in \{1, \ldots, k+1\}.$$

3. [14 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 $p(Y)$
Marginal distributions $p(X)$:

$$p(X=0) = p(X=0, Y=0) + p(X=0, Y=1) = 3/10 + 1/10 = 4/10$$

$$p(X=1) = p(X=1, Y=0) + p(X=1, Y=1) = 2/10 + 4/10 = 6/10$$

Marginal distributions $p(Y)$:

$$p(Y=0) = p(X=0, Y=0) + p(X=1, Y=0) = 3/10 + 2/10 = 5/10$$

$$p(Y=1) = p(X=0, Y=1) + p(X=0, Y=1) = 1/10 + 4/10 = 5/10$$

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

$$p(X=0|Y=0) = 3/5 \qquad \text{and} \qquad p(X=1|Y=0) = 2/5$$

$$p(X=0|Y=1) = 1/5 \qquad \text{and} \qquad p(X=1|Y=1) = 4/5$$

$$p(Y=0|X=0) = 3/4 \qquad \text{and} \qquad p(Y=1|X=0) = 1/4$$

$$p(Y=0|X=1) = 1/3 \qquad \text{and} \qquad p(Y=1|X=1) = 2/3$$

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

$$\mathbb{E}(X) = 0.6 \qquad \text{and} \qquad \mathbb{E}(Y) = 0.5$$

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 0.6 - 0.36 = 0.24$$

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

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

$$\mathbb{E}(Y|X=0) = 1/4 \qquad \text{and} \qquad \mathbb{E}(Y|X=1) = 2/3$$

$$\mathbb{V}(Y|X=0) = \mathbb{E}(Y^2|X=0) - \mathbb{E}(Y|X=0)^2 = 1/4 - 1/16 = 3/16$$

$$\mathbb{V}(Y|X=1) = \mathbb{E}(Y^2|X=1) - \mathbb{E}(Y|X=1)^2 = 2/3 - 4/9 = 2/9$$

v. the covariance between X and Y

$$cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0.4 - 0.6 \times 0.5 = 0.37$$

(b) [2 points] Are X and Y independent? Why? No, since

$$0.3 = p(X=0, Y=0) \neq p(X=0)p(Y=0) = 0.4 \times 0.5 = 0.2$$

(c) [2 points] When X is not assigned a specific value, are $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ still constant? Why? No, since X and Y are not independent, $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ are both functions of X and therefore, they are two random variables depending on X.

4. [9 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) $\mathbb{E}(Y)$

Note that if $Z \sim \mathcal{N}(Z|\mu, \sigma)$, then the probability density function of Z is $f_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$.

$$\begin{aligned}\mathbb{E}(Y) &= \int_{-\infty}^{+\infty} e^{-X^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dx \\ &= \frac{1}{\sqrt{3}} \int_{-\infty}^{+\infty} \frac{1}{\frac{1}{\sqrt{3}}\sqrt{2\pi}} e^{-\frac{X^2}{2/3}} dx \\ &= \frac{1}{\sqrt{3}} \times 1 = \frac{1}{\sqrt{3}}.\end{aligned}$$

Note that $g_X(x) = \frac{1}{\frac{1}{\sqrt{3}}\sqrt{2\pi}} e^{-\frac{x^2}{2/3}}$ is the probability density function of $\mathcal{N}(X|0, 1/\sqrt{3})$ and thus $\int_{-\infty}^{+\infty} \frac{1}{\frac{1}{\sqrt{3}}\sqrt{2\pi}} e^{-\frac{X^2}{2/3}} dx = 1$.

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

To compute $\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$, we need to know $\mathbb{E}(Y^2)$. Lets compute it,

$$\begin{aligned}\mathbb{E}(Y^2) &= \int_{-\infty}^{+\infty} e^{-2X^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dx \\ &= \frac{1}{\sqrt{5}} \int_{-\infty}^{+\infty} \frac{1}{\frac{1}{\sqrt{5}}\sqrt{2\pi}} e^{-\frac{X^2}{2/5}} dx \\ &= \frac{1}{\sqrt{5}} \times 1 = \frac{1}{\sqrt{5}}.\end{aligned}$$

Therefore,

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

(c) $\text{cov}(X, Y)$

We know that $\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. As we know $\mathbb{E}(X) = 0$ and $\mathbb{E}(Y) = \frac{1}{\sqrt{3}}$, to compute cov(X, Y), it suffices to know the value of $\mathbb{E}(XY)$.

$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{+\infty} Xe^{-X^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dx \\ &= \frac{1}{\sqrt{3}} \int_{-\infty}^{+\infty} X \frac{1}{\frac{1}{\sqrt{3}}\sqrt{2\pi}} e^{-\frac{X^2}{2/3}} dx \\ &= \frac{1}{\sqrt{3}} \mathbb{E}(X) \qquad \text{where } X \sim \mathcal{N}(X|0, 1/\sqrt{3}) \\ &= 0.\end{aligned}$$

Therefore, $\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0 - 0 \times \frac{1}{\sqrt{3}} = 0$.

5. [8 points] Derive the probability density functions of the following transformed random variables.

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

We remind that if $X \sim f(X)$, then by changing the variable $X = g(Y)$ where g is monotonic function, $Y \sim g'(Y)f(g(Y))$. Therefore, since here $g(Y) = Y^{1/3}$ and $X \sim \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}}$, we have

$$Y \sim \frac{1}{3} Y^{-2/3} \frac{1}{\sqrt{2\pi}} e^{-\frac{Y^{2/3}}{2}} = \frac{1}{3\sqrt{2\pi}\sqrt[3]{Y^2}} e^{-\frac{Y^{2/3}}{2}}.$$

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

We remind that if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, then $\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$. Therefore, $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ has a Gaussian distribution with mean

$$\begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and covariance

$$\begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix}^\top = \begin{bmatrix} 3/4 & -1/4 \\ -1/4 & 13/9 \end{bmatrix}.$$

In other words,
$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3/4 & -1/4 \\ -1/4 & 13/9 \end{bmatrix} \right).$$

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

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

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y|X)) &= \int \mathbb{E}(Y|X = x) f_X(x) dx \\
&= \int \left(\int y f_{Y|X}(y|x) dy \right) f_X(x) dx \\
&= \int \int y f_X(x) f_{Y|X}(y|x) dy dx \\
&= \int \int y f_{X,Y}(x, y) dy dx \\
&= \int y \left(\int f_{X,Y}(x, y) dx \right) dy \\
&= \int y f_Y(y) dy = \mathbb{E}(Y) \qquad \text{(we used the fact that } f_Y(y) = \textstyle\int F_{X,Y}(x, y) dx)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{V}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\
&= \mathbb{E}(\mathbb{E}(Y^2|X)) - \mathbb{E}(\mathbb{E}(Y|X))^2 \\
&= \mathbb{E}\Big(\mathbb{E}(Y^2|X) - \mathbb{E}(Y|X)^2 + \mathbb{E}(Y|X)^2\Big) - \mathbb{E}(\mathbb{E}(Y|X))^2 \\
&= \mathbb{E}(\mathbb{E}(Y^2|X)) - \mathbb{E}(Y|X)^2) + \mathbb{E}(\mathbb{E}(Y|X)^2) - \mathbb{E}(\mathbb{E}(Y|X))^2 \qquad \text{linearity of expectation} \\
&= \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)).
\end{aligned}$$

7. [9 points] Given a logistic function, $f(\mathbf{x}) = 1/(1 + \exp(-\mathbf{a}^\top \mathbf{x}))$ ($\mathbf{x}$ is a vector),

(a) derive $\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}}$

We know $\frac{\mathrm{d}\sigma}{\mathrm{d}z} = \sigma(z)(1 - \sigma(z))$ if $\sigma(z) = 1/(1 + \exp(-z))$. Set $z = \mathbf{a}^\top \mathbf{x}$ and note that $f(\mathbf{x}) = \sigma(z)$.

$$\begin{aligned}
\mathrm{d}f(\mathbf{x}) &= \frac{\mathrm{d}\sigma}{\mathrm{d}z} \mathrm{d}z \\
&= \sigma(z)(1 - \sigma(z)) \mathrm{d}(\mathbf{a}^\top \mathbf{x}) \\
&= \sigma(z)(1 - \sigma(z)) \mathbf{a}^\top \mathrm{d}\mathbf{x} \quad \implies \frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}} = \sigma(z)(1 - \sigma(z)) \mathbf{a}^\top
\end{aligned}$$

(b) derive $\frac{\mathrm{d}^2 f(\mathbf{x})}{\mathrm{d}\mathbf{x}^2}$, i.e., the Hessian matrix
Set $t(\mathbf{x}) = f(\mathbf{x})(1 - f(\mathbf{x}))$. The vector $\mathbf{a}$ is $n \times 1$ and here, we consider t as a 1×1 matrix. Clearly, $\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}} = t(\mathbf{x})\mathbf{a}^\top$. Therefore,

$$\begin{aligned}
\mathrm{d}\left(\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}}\right) &= \mathrm{d}\left(\mathbf{a}t(\mathbf{x})\right) \\
&= \mathbf{a}(\mathrm{d}t(\mathbf{x})) \\
&= \mathbf{a}f(\mathbf{x})(1 - f(\mathbf{x}))(1 - 2f(\mathbf{x}))\mathbf{a}^\top \mathrm{d}\mathbf{x} \quad \Longrightarrow \frac{\mathrm{d}^2 f(\mathbf{x})}{\mathrm{d}\mathbf{x}^2} = f(\mathbf{x})(1 - f(\mathbf{x}))(1 - 2f(\mathbf{x}))\mathbf{a}\mathbf{a}^\top
\end{aligned}$$

(c) show that $-\log f(\mathbf{x})$ is convex Note that $0 \le f(\mathbf{x}) \le 1$.
To prove that $-\log f(\mathbf{x})$ is convex, we use a result asserting that a function $g(\mathbf{x})$ is convex if its Hessian matrix $H(\mathbf{x})$ is semi-positive for each x, i.e, $\frac{\mathrm{d}^2 g(\mathbf{x})}{\mathrm{d}\mathbf{x}^2} \succeq 0$.

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}(-\log f(\mathbf{x})) = -\frac{1}{f(\mathbf{x})}\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}} = -(1 - f(\mathbf{x}))\mathbf{a} \qquad \text{Using Part (a)}$$

Therefore,

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{x}^2}(-\log f(\mathbf{x})) &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\Big(\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}(-\log f(\mathbf{x}))\Big) \\
&= \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\Big(-(1 - f(\mathbf{x}))\mathbf{a}\Big) \\
&= f(\mathbf{x})(1 - f(\mathbf{x}))\mathbf{a}\mathbf{a}^t
\end{aligned}$$

Notice $A_{n \times n}$ is semi-positive if and only if for each vector $\mathbf{z} \in \mathbb{R}^n$, we have $z^\top A\mathbf{z} \ge 0$.

$$\begin{aligned}
\mathbf{z}^\top\Big(f(\mathbf{x})(1 - f(\mathbf{x}))\mathbf{a}\mathbf{a}^\top\Big)\mathbf{z} &= f(\mathbf{x})(1 - f(\mathbf{x}))\mathbf{z}^\top\mathbf{a}\mathbf{a}^\top\mathbf{z} \\
&= f(\mathbf{x})(1 - f(\mathbf{x}))\mathbf{z}^\top\mathbf{a}(\mathbf{z}^\top\mathbf{a})^\top \\
&= f(\mathbf{x})(1 - f(\mathbf{x}))(\mathbf{z}^\top\mathbf{a})^2 \ge 0 \qquad \text{We used } f(\mathbf{x})(1 - f(\mathbf{x})) \ge 0 \text{ as well.}
\end{aligned}$$

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

(a) $f(x) = -\log(x)$
By the definition, the convex conjugate of $f(x)$ is $f^*(\lambda) = \max\limits_{x \in (0,\infty)} (\lambda x + \log x)$. Note that since, for each arbitrary fixed $\lambda \in \mathbb{R}$, $\lambda x + \log x$ is a strictly concave function, either it takes its maximum in a unique point (can be found by setting its derivative to zero) or its maximum is $+\infty$. When $\lambda \ge 0$, clearly $\max\limits_x (\lambda x + \log x) = \infty$. When $\lambda < 0$, if we set the derivative of $\max\limits_{x \in (0,\infty)} (\lambda x + \log x)$ to zero, then

$$(\lambda x + \log x)' = \lambda + 1/x = 0 \qquad \Longrightarrow \qquad x = -1/\lambda.$$

Substituting $x = -1/\lambda$, we get $\max\limits_{x \in (0,\infty)} (\lambda x + \log x) = -(1 + \log(\lambda))$. Therefore,

$$f^*(\lambda) = \begin{cases} +\infty & \lambda \ge 0 \\ -(1 + \log(\lambda)) & \lambda < 0 \end{cases}$$

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

By the definition,

$$f^*(\mathbf{y}) = \max_{\mathbf{x}\in\mathbb{R}^n} \left(\mathbf{y}^\top \mathbf{x} - \mathbf{x}^\top \mathbf{A}^{-1}\mathbf{x}\right).$$

Since $\left(\mathbf{y}^\top \mathbf{x} - \mathbf{x}^\top \mathbf{A}^{-1}\mathbf{x}\right)$ is strictly concave, it takes its maximum in a point making its derivative zero.

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(\mathbf{y}^\top \mathbf{x} - \mathbf{x}^\top \mathbf{A}^{-1}\mathbf{x}\right) = \mathbf{y} - 2\mathbf{A}^{-1}\mathbf{x} = 0 \quad \Longrightarrow \mathbf{x} = \frac{1}{2}\mathbf{A}\mathbf{y}.$$

Substituting $\mathbf{x} = \frac{1}{2}\mathbf{A}\mathbf{y}$, we get $f^*(\mathbf{y}) = \frac{1}{4}\mathbf{y}^\top \mathbf{A}\mathbf{y}$.

9. [20 points] Derive the (partial) gradient of the following functions. Note that bold small letters represent vectors, bold captial letters matrices, and non-bold letters just scalars.

(a) $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$, derive $\frac{\partial f}{\partial \mathbf{x}}$

$$\begin{aligned}
\mathrm{d}f(\mathbf{x}) &= \mathrm{d}\left(\mathbf{x}^\top \mathbf{A}\mathbf{x}\right) \\
&= (\mathrm{d}\mathbf{x})^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathrm{d}(\mathbf{A}\mathbf{x}) \\
&= (\mathrm{d}\mathbf{x})^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}\mathrm{d}\mathbf{x} \\
&= \mathbf{x}^\top \mathbf{A}^\top \mathrm{d}\mathbf{x} + \mathbf{x}^\top \mathbf{A}\mathrm{d}\mathbf{x} \qquad \text{Since } (\mathrm{d}\mathbf{x})^\top \mathbf{A}\mathbf{x} \text{ is scaler, it is equal to its transpose.} \\
&= (\mathbf{x}^\top \mathbf{A}^\top + \mathbf{x}^\top \mathbf{A})\mathrm{d}\mathbf{x} \quad \Longrightarrow \quad \frac{\partial f}{\partial \mathbf{x}} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)
\end{aligned}$$

(b) $f(\mathbf{x}) = \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}$, derive $\frac{\partial f}{\partial \mathbf{x}}$

$$\begin{aligned}
\mathrm{d}f(\mathbf{x}) &= \mathrm{d}\left(\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}\right) \\
&= \mathrm{d}\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x} + \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathrm{d}\mathbf{x} \\
&= -\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathrm{d}\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x} + \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathrm{d}\mathbf{x} \\
&= -\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\left((\mathrm{d}\mathbf{x})\mathbf{x}^\top + \mathbf{x}(\mathrm{d}\mathbf{x})^\top\right)\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x} + \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathrm{d}\mathbf{x} \\
&= -\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}(\mathrm{d}\mathbf{x})\underbrace{\mathbf{x}^\top\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}}_{\text{Scaler}} - \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}\underbrace{(\mathrm{d}\mathbf{x})^\top\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}}_{\text{Scaler}} + \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathrm{d}\mathbf{x} \\
&= -\underbrace{\left(\mathbf{x}^\top\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}\right)}_{\text{Scaler}}\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}(\mathrm{d}\mathbf{x}) - \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}\mathbf{x}^\top\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathrm{d}\mathbf{x} + \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathrm{d}\mathbf{x} \\
&= \left[-\left(\mathbf{x}^\top\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}\right)\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1} - \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}\mathbf{x}^\top\left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1} + \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\right]\mathrm{d}\mathbf{x}
\end{aligned}$$

(c) $f(\alpha) = \log|\mathbf{K} + \alpha\mathbf{I}|$, where $|\cdot|$ means the determinant. Derive $\frac{\partial f}{\partial \alpha}$

We know that $\partial(\ln|X|) = \mathrm{tr}(X^{-1}\partial\mathbf{X})$

$$\begin{aligned}
\mathrm{d}\log|\mathbf{K} + \alpha\mathbf{I}| &= \mathrm{tr}\left((\mathbf{K} + \alpha\mathbf{I})^{-1}\mathrm{d}(\mathbf{K} + \alpha\mathbf{I})\right) \\
&= \mathrm{tr}\left((\mathbf{K} + \alpha\mathbf{I})^{-1}(\mathrm{d}\alpha)\mathbf{I}\right) \\
&= \mathrm{tr}\left((\mathbf{K} + \alpha\mathbf{I})^{-1}\right)\mathrm{d}\alpha \qquad \Longrightarrow \qquad \frac{\mathrm{d}\log|\mathbf{K} + \alpha\mathbf{I}|}{\mathrm{d}\alpha} = \mathrm{tr}\left((K + \alpha\mathbf{I})^{-1}\right)
\end{aligned}$$

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

We remind that

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{-1}{2}(\mathbf{a} - \mathbf{m})^t\boldsymbol{\Omega}^{-1}(\mathbf{a} - \mathbf{m}) - \frac{1}{2}\log|\boldsymbol{\Omega}| + C,$$

where $\mathbf{m} = \mathbf{A}\mu$ and $\mathbf{\Omega} = \mathbf{S\Sigma S}^\top$ and C is a constant.

$$\begin{aligned}
\mathrm{d}f &= \frac{-1}{2}\mathrm{d}(\mathbf{a}-\mathbf{m})^\top \mathbf{\Omega}^{-1}(\mathbf{a}-\mathbf{m}) + \frac{-1}{2}(\mathbf{a}-\mathbf{m})^\top \mathbf{\Omega}^{-1}\mathrm{d}(\mathbf{a}-\mathbf{m}) \\
&= \frac{-1}{2}\mathrm{d}(\mathbf{m})^\top \mathbf{\Omega}^{-1}(\mathbf{a}-\mathbf{m}) + \frac{1}{2}(\mathbf{a}-\mathbf{m})^\top \mathbf{\Omega}^{-1}\mathrm{d}(\mathbf{m}) \\
&= \frac{1}{2}(\mathbf{a}-\mathbf{m})^\top (\mathbf{\Omega}^{-1})^\top \mathrm{d}(\mathbf{m}) + \frac{1}{2}(\mathbf{a}-\mathbf{m})^\top \mathbf{\Omega}^{-1}\mathrm{d}(\mathbf{m}) \quad \text{Since } \mathrm{d}(\mathbf{m})^\top \mathbf{\Omega}^{-1}(\mathbf{a}-\mathbf{m}) \text{ is scaler.} \\
&= (\mathbf{a}-\mathbf{m})^\top \mathbf{\Omega}^{-1}\mathrm{d}(\mathbf{m}) \\
&= (\mathbf{a}-\mathbf{m})^\top \mathbf{\Omega}^{-1}\mathbf{A}\mathrm{d}\boldsymbol{\mu} \quad \Longrightarrow \quad \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\mu}} = (\mathbf{a}-\mathbf{m})^\top \mathbf{\Omega}^{-1}\mathbf{A}
\end{aligned}$$

$$\begin{aligned}
\mathrm{d}f &= -\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^\top (\mathrm{d}\mathbf{\Omega}^{-1})(\mathbf{a}-\boldsymbol{\mu}) - \frac{1}{2}\mathrm{d}(\log|\mathbf{\Omega}|) \\
&= -\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^\top \mathbf{\Omega}^{-1}(\mathrm{d}\mathbf{\Omega})\mathbf{\Omega}^{-1}(\mathbf{a}-\boldsymbol{\mu}) - \frac{1}{2}\mathrm{tr}(\mathbf{\Omega}^{-1}(\mathrm{d}\mathbf{\Omega})) \\
&= -\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^\top \mathbf{\Omega}^{-1}\mathbf{S}(\mathrm{d}\mathbf{\Sigma})\mathbf{S}^\top \mathbf{\Omega}^{-1}(\mathbf{a}-\boldsymbol{\mu}) - \frac{1}{2}\mathrm{tr}(\mathbf{\Omega}^{-1}\mathbf{S}(\mathrm{d}\mathbf{\Sigma})\mathbf{S}^\top) \\
&= -\mathrm{tr}\left(\frac{1}{2}\mathbf{S}^\top \mathbf{\Omega}^{-1}(\mathbf{a}-\boldsymbol{\mu})(\mathbf{a}-\boldsymbol{\mu})^\top \mathbf{\Omega}^{-1}\mathbf{S}(\mathrm{d}\mathbf{\Sigma})\right) - \frac{1}{2}\mathrm{tr}(\mathbf{S}^\top \mathbf{\Omega}^{-1}\mathbf{S}(\mathrm{d}\mathbf{\Sigma}))
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}\mathbf{\Sigma}} &= -\frac{1}{2}\mathbf{S}^\top \mathbf{\Omega}^{-1}(\mathbf{a}-\boldsymbol{\mu})(\mathbf{a}-\boldsymbol{\mu})^\top \mathbf{\Omega}^{-1}\mathbf{S} - \frac{1}{2}\mathbf{S}^\top \mathbf{\Omega}^{-1}\mathbf{S} \\
&= -\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{S}^{-1}(\mathbf{a}-\boldsymbol{\mu})(\mathbf{a}-\boldsymbol{\mu})^\top (\mathbf{S}^{-1})^\top \mathbf{\Sigma}^{-1} - \frac{1}{2}\mathbf{\Sigma}^{-1}
\end{aligned}$$

(e) $f(\mathbf{\Sigma}) = \log\left(\mathcal{N}(\mathbf{a}|\mathbf{b}, \mathbf{K}\otimes\mathbf{\Sigma})\right)$ where $\otimes$ is the Kronecker product (Hint: check Minka's notes).
Note that $f(\mathbf{\Sigma}) = -\frac{1}{2}(\mathbf{a}-\mathbf{b})^t\mathbf{\Omega}^{-1}(\mathbf{a}-\mathbf{b}) - \frac{1}{2}\log|\mathbf{\Omega}| + C$, where $\mathbf{\Omega} = \mathbf{K}\otimes\mathbf{\Sigma}$ and C is a constant.

$$\begin{aligned}
\mathrm{d}f &= -\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^\top (\mathrm{d}\mathbf{\Omega}^{-1})(\mathbf{a}-\boldsymbol{\mu}) - \frac{1}{2}\mathrm{d}\log|\mathbf{\Omega}| \\
&= -\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^\top \mathbf{\Omega}^{-1}(\mathrm{d}\mathbf{\Omega})\mathbf{\Omega}^{-1}(\mathbf{a}-\boldsymbol{\mu}) - \frac{1}{2}\mathrm{tr}(\mathbf{\Omega}^{-1}(\mathrm{d}\mathbf{\Omega})) \\
&= -\frac{1}{2}\mathrm{tr}\left((\mathbf{a}-\boldsymbol{\mu})^\top \mathbf{\Omega}^{-1}(\mathbf{K}\otimes\mathrm{d}\mathbf{\Sigma})\mathbf{\Omega}^{-1}(\mathbf{a}-\boldsymbol{\mu})\right) - \frac{1}{2}\mathrm{tr}(\mathbf{\Omega}^{-1}(\mathbf{K}\otimes\mathrm{d}\mathbf{\Sigma})) \quad \text{Using } \mathrm{d}\mathbf{\Omega} = \mathbf{K}\otimes\mathrm{d}\mathbf{\Sigma} \\
&= -\frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega}^{-1}(\mathbf{a}-\boldsymbol{\mu})(\mathbf{a}-\boldsymbol{\mu})^\top \mathbf{\Omega}^{-1}(\mathbf{K}\otimes\mathrm{d}\mathbf{\Sigma})\right) - \frac{1}{2}\mathrm{tr}(\mathbf{\Omega}^{-1}(\mathbf{K}\otimes\mathrm{d}\mathbf{\Sigma}))
\end{aligned}$$

10. [2 points] Given the multivariate Gaussian probability density,

$$p(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Sigma}) = |2\pi\mathbf{\Sigma}|^{-\frac{1}{2}}\exp\left(-(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Show that the density function achieves the maximum when $\mathbf{x} = \boldsymbol{\mu}$.
Since $\log(\cdot)$ is an increasing function, to make computation easy, we prove that $\log(p(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Sigma}))$ takes it maximum in $\boldsymbol{\mu}$.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(\log(p(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Sigma}))\right) &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(-(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + C\right) \qquad C \text{ here is a constant} \\
&= -2\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = 0 \qquad \Longrightarrow x = \boldsymbol{\mu}
\end{aligned}$$

Moreover, since $\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{x}^2}\left(\log(p(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Sigma}))\right) = -2\Sigma^{-1} \prec 0$, the function $\log(p(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Sigma}))$ is a strictly concave function taking its maximum in $x = \boldsymbol{\mu}$, as desired.

11. [5 points] Show that

$$\int \exp(-\frac{1}{2\sigma^2}x^2)\mathrm{d}x = \sqrt{2\pi\sigma^2}.$$

Note that this is about how the normalization constant of the Gaussian density is obtained. Hint: consider its square and use double integral.

Set $I = \int \exp(-\frac{1}{2\sigma^2}x^2)\mathrm{d}x$. In the following we prove that $I^2 = 2\pi\sigma^2$.

$$\begin{aligned}
I^2 &= (\int_{-\infty}^{\infty} \exp(-\frac{1}{2\sigma^2}x^2)\mathrm{d}x)^2 \\
&= \left(\int_{-\infty}^{\infty} \exp(-\frac{1}{2\sigma^2}x^2)\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} \exp(-\frac{1}{2\sigma^2}y^2)\mathrm{d}y\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2+y^2)\right)\mathrm{d}x\mathrm{d}y \\
&= \int_{0}^{2\pi}\int_{0}^{\infty} r\exp\left(-\frac{1}{2\sigma^2}r^2\right)\mathrm{d}r\mathrm{d}\theta \\
&= 2\pi\left[-\sigma^2\exp\left(-\frac{1}{2\sigma^2}r^2\right)\right]_0^{\infty} = 2\pi\sigma^2 \quad\Longrightarrow\quad I = \sqrt{2\pi\sigma^2}
\end{aligned}$$

12. [5 points] The gamma function is defined as

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-u}\mathrm{d}u.$$

Show that $\Gamma(1) = 1$ and $\Gamma(x+1) = x\Gamma(x)$. Hint: using integral by parts.

$$\begin{aligned}
\Gamma(1) &= \int_0^{\infty} e^{-u}\mathrm{d}u \\
&= \Big[-e^{-u}\Big]_0^{\infty} = 0-(-1) = 1.
\end{aligned}$$

$$\begin{aligned}
\Gamma(x+1) &= \int_0^{\infty} u^x e^{-u}\mathrm{d}u \qquad (U = u^x \text{ and } V' = e^{-u}) \\
&= \Big[-u^x e^{-u}\Big]_0^{\infty} + \int_0^{\infty} xu^{x-1}e^{-u}\mathrm{d}u \\
&= 0 + x\int_0^{\infty} u^{x-1}e^{-u}\mathrm{d}u = x\Gamma(x).
\end{aligned}$$

13. [2 points] By using Jensen's inequality with $f(x) = \log(x)$, show that for any collection of positive numbers $\{x_1,\ldots,x_N\}$,

$$\frac{1}{N}\sum_{n=1}^{N} x_n \geq \left(\prod_{n=1}^{N} x_n\right)^{\frac{1}{N}}.$$

Since $f(x) = -\log(x)$ is convex, using Jensen's inequality, we conclude

$$\begin{aligned}
-\log\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right) &\leq \frac{1}{N}\sum_{n=1}^{N} -\log(x_n) \\
&= -\log\left(\left(\prod_{n=1}^{N} x_n\right)^{1/N}\right)
\end{aligned}$$

Since $\log(\cdot)$ is an increasing monotonic function, it implies

$$\frac{1}{N}\sum_{n=1}^{N} x_n \geq \left(\prod_{n=1}^{N} x_n\right)^{\frac{1}{N}}.$$

14. [2 points] Given two probability density functions $p(\mathbf{x})$ and $q(\mathbf{x})$, show that

$$\int p(\mathbf{x})\log\frac{p(\mathbf{x})}{q(\mathbf{x})}\mathrm{d}\mathbf{x} \geq 0.$$

Since $f(x) = -\log(x)$ is convex, by Jensen's inequality, we have

$$\begin{aligned}
\int p(\mathbf{x})\log\frac{p(\mathbf{x})}{q(\mathbf{x})}\mathrm{d}\mathbf{x} &= \mathbb{E}_{x\sim p}\left(-\log\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) \\
\text{(using Jensen's inequality for } -\log(x)\text{)} \quad &\geq -\log\left(\mathbb{E}_{x\sim p}\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right)\right) \\
&= -\log\left(\int p(x)\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right)\mathrm{d}x\right) \\
&= -\log\left(\int q(\mathbf{x})\mathrm{d}x\right) = -\log 1 = 0.
\end{aligned}$$

15. [**Bonus**][5 points] Show that for any square matrix $\mathbf{X} \succ 0$, $\log|\mathbf{X}|$ is concave to $\mathbf{X}$.

Consider $\mathbf{X}$ as a vector in $\mathbb{R}^{2n}$ whose coordinates are indexed by $ij \in [n]\times[n]$. We know that $\frac{\mathrm{d}\log|\mathbf{X}|}{\mathrm{d}\mathbf{X}} = X^{-1}$ and consequently,

$$\frac{\mathrm{d}^2\log|\mathbf{X}|}{\mathrm{d}\mathbf{X}^2} = \frac{\mathrm{d}\mathbf{X}^{-1}}{\mathrm{d}\mathbf{X}}.$$

By the Matrix Cookboo,

$$\frac{\mathrm{d}(\mathbf{X}^{-1})_{kl}}{\mathrm{d}(\mathbf{X})_{ij}} = -(X^{-1})_{ki}(X^{-1})_{jl}.$$

If we consider $\log(|\mathbf{X}|)$ as a function from $\mathbb{R}^{n^2} \longrightarrow \mathbb{R}$, then $L = \frac{\mathrm{d}^2\log|\mathbf{X}|}{\mathrm{d}\mathbf{X}^2}$ is a $n^2\times n^2$ matrix whose rows and columns are indexed by $ij \in [n]\times[n]$. In what follows, we prove that this matrix is semi-negative definite. Consider $\mathbf{Z}\in\mathbb{R}^{n^2}$, as a vector, such that its entries are indexed by indices $ij \in [n]\times[n]$. To fulfill the proof,

$$\begin{aligned}
\mathbf{Z}^\top L\mathbf{Z} &= \sum_{kl}\sum_{ij}\mathbf{Z}_{kl}L_{kl,ij}\mathbf{Z}_{ij} \\
&= -\sum_{kl}\sum_{ij}\mathbf{Z}_{kl}(\mathbf{X}^{-1})_{ki}(\mathbf{X}^{-1})_{jl}\mathbf{Z}_{ij} \\
&= -\sum_{i}\sum_{l}\left(\sum_{j}(\mathbf{X}^{-1})_{jl}\mathbf{Z}_{ij}\right)\left(\sum_{k}\mathbf{Z}_{kl}(\mathbf{X}^{-1})_{ki}\right) \\
&= -\sum_{i}\sum_{l}(\mathbf{Z}\mathbf{X}^{-1})_{il}\left((\mathbf{X}^{-1})^\top\mathbf{Z}\right)_{il} \quad \text{Hereafter, we see } \mathbf{X} \text{ and } \mathbf{Z} \text{ as Matrices!} \\
&= -\sum_{i}\sum_{l}(\mathbf{X}^{-1}\mathbf{Z})_{il}^2 \leq 0 \qquad \text{since } \mathbf{X} \text{ is symmetric.}
\end{aligned}$$