# Homework Set 1 Solution

Introduction to Artificial Intelligence with Mathematics (MAS473)

1. (10pts) The ABC test is a new test for diagnosing depression. An extensive clinical evaluation was performed of this instrument, whereby participants were interviewed by psychiatrists and a definitive clinical diagnosis of depression was made. The table below shows the number of participants with or without depression based on test scores. If one gets the test score higher than $x$, we are going to diagnose depression for him/her.

| | With depression (positive) | Without depression (negative) |
|---|---|---|
| $0 \sim 5$ | 0 | 6 |
| $6 \sim 10$ | 1 | 20 |
| $11 \sim 15$ | 1 | 9 |
| $16 \sim 20$ | 3 | 4 |
| $21 \sim 25$ | 5 | 1 |

(a) If $x = 15$, find the sensitivity and the specificity.

(b) Calculate F1 score for each $x = 5, 10, 15, 20$ and find the best value $x$ among $5, 10, 15, 20$ with respect to the F1 score.

*Solution.*

(a)

$$\text{(Sensitivity)} = \frac{8}{8+2} = \frac{4}{5}$$
$$\text{(Specificity)} = \frac{35}{35+5} = \frac{7}{8}$$

(b) F1 score for each $x = 5, 10, 15, 20$ is on the table below.
Since 16/23 is the highest among F1 scores, $x = 15$ is the best value.

| $x$ | Recall(=Sensitivity) | Precision | F1 score |
|---|---|---|---|
| 5 | 10/(10+0) | 10/(10+34) | 10/27 |
| 10 | 9/(9+1) | 9/(9+14) | 6/11 |
| 15 | 8/(8+2) | 8/(8+5) | 16/23 |
| 20 | 5/(5+5) | 5/(5+1) | 5/8 |

2. (10pts) A distribution $\text{Dir}(\alpha)$ (called a Dirichlet distribution) is a distribution whose probability density function is given by

$$\text{Dir}(\mathbf{x}|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mathbf{x}_k^{\alpha_k - 1}$$

on $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_K)^\top \in \triangle^{K-1} = \{(y_1, \cdots, y_K) \in \mathbb{R}^K : y_1 + \cdots + y_K = 1 \text{ and } y_i \geq 0 \text{ for any } i = 1, \cdots, K\}$ where $\alpha = (\alpha_1, \cdots, \alpha_K)$ ($\alpha_i > 0$ for any $i$), $\alpha_0 = \alpha_1 + \cdots + \alpha_K$, and the gamma function $\Gamma$

is defined by

$$\Gamma(y) = \int_0^\infty t^{y-1}e^{-t}\, dt.$$

(a) Prove that $\mathrm{Dir}(\alpha)$ is normalized, i.e.

$$\int_{\triangle^{K-1}} \mathrm{Dir}(\mathbf{x}|\alpha)\, d\mathbf{x} = 1.$$

You can use the fact that the Beta distribution is normalized.

(b) Find $\mathbf{E}[\mathbf{x}_i]$, $\mathrm{Var}(\mathbf{x}_j)$ and $\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)$ $(i \neq j)$ where $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_K) \sim \mathrm{Dir}(\alpha)$.
(Hint : You may use a property of gamma function $\Gamma(x+1) = x\Gamma(x)$ for $x > 0$)

*Solution.*

(a) Let $a, m, n$ be positive constants. Then,

$$\begin{aligned}\int_0^a x^{m-1}(a-x)^{n-1}dx &= a^{m+n-2}\int_0^a \left(\frac{x}{a}\right)^{m-1}\left(1-\frac{x}{a}\right)^{n-1} dx \\ &= a^{m+n-1}\int_0^1 x^{m-1}(1-x)^{n-1}dx \\ &= a^{m+n-1}\frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}\end{aligned}$$

We used the fact that Beta distribution is normalized for the last equality. By using this equality, we get

$$\begin{aligned}&\int_{\triangle^{K-1}} \left(\prod_{k=1}^{K} \mathbf{x}_k^{\alpha_k-1}\right) d\mathbf{x} \\ &= \int_0^1\int_0^{1-\mathbf{x}_1}\cdots\int_0^{1-\mathbf{x}_1-\cdots-\mathbf{x}_{K-2}} \left(\prod_{k=1}^{K-2} \mathbf{x}_k^{\alpha_k-1}\right) \mathbf{x}_{K-1}^{\alpha_{K-1}-1}(1-\mathbf{x}_1-\cdots-\mathbf{x}_{K-1})^{\alpha_K-1}\, d\mathbf{x}_{K-1}\cdots d\mathbf{x}_2\, d\mathbf{x}_1 \\ &= \frac{\Gamma(\alpha_{K-1})\Gamma(\alpha_K)}{\Gamma(\alpha_{K-1}+\alpha_K)}\int_0^1\int_0^{1-\mathbf{x}_1}\cdots\int_0^{1-\mathbf{x}_1-\cdots-\mathbf{x}_{K-3}} \left(\prod_{k=1}^{K-3} \mathbf{x}_k^{\alpha_k-1}\right) \mathbf{x}_{K-2}^{\alpha_{K-2}-1}(1-\mathbf{x}_1-\cdots-\mathbf{x}_{K-2})^{\alpha_{K-1}+\alpha_K-1}\, d\mathbf{x}_{K-2}\cdots d\mathbf{x}_2\, d\mathbf{x}_1 \\ &= \cdots \\ &= \frac{\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0-\alpha_1)}\int_0^1 \mathbf{x}_1^{\alpha_1-1}(1-\mathbf{x}_1)^{\alpha_2+\cdots+\alpha_K-1}\, d\mathbf{x}_1 \\ &= \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)}\end{aligned}$$

Therefore, $\mathrm{Dir}(\alpha)$ is normalized.

(b) First,

$$\begin{aligned}\mathbf{E}[\mathbf{x}_i] &= \int_{\triangle^{K-1}} \mathbf{x}_i \mathrm{Dir}(\mathbf{x}|\alpha)\, d\mathbf{x} \\ &= \int_{\triangle^{K-1}} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\left(\prod_{\substack{k=1,\cdots,K\\ k\neq i}} \mathbf{x}_k^{\alpha_k-1}\right)\mathbf{x}_i^{\alpha_i}\, d\mathbf{x} \\ &= \frac{\Gamma(\alpha_0)\Gamma(\alpha_i+1)}{\Gamma(\alpha_0+1)\Gamma(\alpha_i)}\int_{\triangle^{K-1}} \frac{\Gamma(\alpha_0+1)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_{i-1})\Gamma(\alpha_i+1)\Gamma(\alpha_{i+1})\cdots\Gamma(\alpha_K)}\left(\prod_{\substack{k=1,\cdots,K\\ k\neq i}} \mathbf{x}_k^{\alpha_k-1}\right)\mathbf{x}_i^{\alpha_i}\, d\mathbf{x} \\ &= \frac{\Gamma(\alpha_0)\Gamma(\alpha_i+1)}{\Gamma(\alpha_0+1)\Gamma(\alpha_i)} = \frac{\alpha_i}{\alpha_0}\end{aligned}$$

Likewise, if $i \neq j$,

$$\mathbf{E}[\mathbf{x}_i^2] = \int_{\triangle^{K-1}} \mathbf{x}_i^2 \mathrm{Dir}(\mathbf{x}|\alpha)\, d\mathbf{x} = \frac{\Gamma(\alpha_0)\Gamma(\alpha_i+2)}{\Gamma(\alpha_0+2)\Gamma(\alpha_i)} = \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)},$$
$$\mathbf{E}[\mathbf{x}_i\mathbf{x}_j] = \int_{\triangle^{K-1}} \mathbf{x}_i\mathbf{x}_j \mathrm{Dir}(\mathbf{x}|\alpha)\, d\mathbf{x} = \frac{\Gamma(\alpha_0)\Gamma(\alpha_i+1)\Gamma(\alpha_j+1)}{\Gamma(\alpha_0+2)\Gamma(\alpha_i)\Gamma(\alpha_j)} = \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)}$$

Thus,

$$\mathrm{Var}(\mathbf{x}_j) = \mathbf{E}[\mathbf{x}_i^2] - \mathbf{E}[\mathbf{x}_i]^2 = \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)} - \frac{\alpha_i^2}{\alpha_0^2} = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)},$$
$$\mathrm{Cov}(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{E}[\mathbf{x}_i\mathbf{x}_j] - \mathbf{E}[\mathbf{x}_i]\mathbf{E}[\mathbf{x}_j] = \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)} - \frac{\alpha_i\alpha_j}{\alpha_0^2} = -\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$$

3. (10pts) A $K$-dimensional multivariate normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ is a distribution whose probability density function is given by

$$\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{K/2}} \cdot \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

on $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_K) \in \mathbb{R}^K$ where $\mu = (\mu_1, \cdots, \mu_K)^\top \in \mathbb{R}^K$ and $\Sigma \in \mathbb{R}^{K\times K}$ is a positive definite matrix. Let $X_1, \cdots, X_n$ be i.i.d. random variables with the probability density function $\mathcal{N}(\mu, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is known.

(a) Find the maximum likelihood estimator (MLE) of $\mu$.

(b) Suppose $\mu$ has a prior distribution $p(\mu) = \mathcal{N}(\mu|\mu_0, \mathbf{\Sigma}_0)$. Find the posterior distribution of $\mu$ and the maximum a posterior (MAP) estimator of $\mu$.

*Solution.*

(a) Since $X_1, \cdots, X_n$ are i.i.d., the log-likelihood function becomes

$$\begin{aligned}\ell(\mu|\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}) &= \log \prod_{i=1}^n \mathcal{N}_{X_i}(\mathbf{x}^{(i)}|\mu,\mathbf{\Sigma}) \\ &= \log \prod_{i=1}^n \frac{1}{(2\pi)^{K/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(i)}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)\right) \\ &= -\frac{nK}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\sum_{i=1}^n (\mathbf{x}^{(i)}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)\end{aligned}$$

Then,

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \sum_{i=1}^n \mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu), \\ \frac{\partial^2 \ell}{\partial \mu^2} &= -n\mathbf{\Sigma}^{-1}\end{aligned}$$

Note that since $\mathbf{\Sigma}$ is positive definite, we can diagonalize $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = UDU^\top$$

where $U$ is an orthogonal matrix and $D$ is a diagonal matrix with positive diagonal elements. Then,

$$\mathbf{\Sigma}^{-1} = UD^{-1}U^{\top}$$

and thus $\mathbf{\Sigma}^{-1}$ is also positive definite. Finally, setting the first order derivative of $\ell$ w.r.t. $\mu$ equal to zero, we get

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

since the second order derivative of $\ell$ w.r.t. $\mu$ $(= -n\mathbf{\Sigma}^{-1})$ is negative definite.

(b) Note that $p(\mu|\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}) \propto p(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}|\mu)p(\mu)$.

$$\log\left(p(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}|\mu)p(\mu)\right) = \log\left(\mathcal{N}(\mu|\mu_0, \mathbf{\Sigma}_0)\prod_{i=1}^{n}\mathcal{N}(\mathbf{x}^{(i)}|\mu, \mathbf{\Sigma})\right)$$

$$= -\frac{(n+1)K}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\log|\mathbf{\Sigma}_0| - \frac{1}{2}(\mu-\mu_0)^{\top}\mathbf{\Sigma}_0^{-1}(\mu-\mu_0) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}^{(i)}-\mu)^{\top}\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu)$$

$$= -\frac{1}{2}\mu^{\top}\mathbf{\Sigma}_0^{-1}\mu + \mu^{\top}\mathbf{\Sigma}_0^{-1}\mu_0 - \frac{n}{2}\mu^{\top}\mathbf{\Sigma}^{-1}\mu + \sum_{i=1}^{n}\mu^{\top}\mathbf{\Sigma}^{-1}\mathbf{x}^{(i)} + (constant\ w.r.t.\ \mu)$$

$$= -\frac{1}{2}\mu^{\top}\left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)\mu + \mu^{\top}\left(\mathbf{\Sigma}_0^{-1}\mu_0 + \mathbf{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}^{(i)}\right) + (constant\ w.r.t.\ \mu)$$

$$= -\frac{1}{2}\left(\mu - \left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}\left(\mathbf{\Sigma}_0^{-1}\mu_0 + \mathbf{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}^{(i)}\right)\right)^{\top}\left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)$$

$$\left(\mu - \left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}\left(\mathbf{\Sigma}_0^{-1}\mu_0 + \mathbf{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}^{(i)}\right)\right) + (constant\ w.r.t.\ \mu)$$

Thus, the posterior distribution of $\mu$ is $K$-dimensional multivariate normal distribution

$$\mathcal{N}\left(\left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}\left(\mathbf{\Sigma}_0^{-1}\mu_0 + \mathbf{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}^{(i)}\right), \left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}\right)$$

We can easily check that $\left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}$ is positive definite using the fact that $\mathbf{\Sigma}$ is positive definite if and only if $\mathbf{\Sigma}^{-1}$ is positive definite. (We proved this fact in (a).) Then, it is clear that the density of a multivariate normal distribution is maximized at the mean, the MAP estimator of $\mu$ is

$$\left(n\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}_0^{-1}\right)^{-1}\left(\mathbf{\Sigma}_0^{-1}\mu_0 + \mathbf{\Sigma}^{-1}\sum_{i=1}^{n}X_i\right)$$

4. (10pts) Answer the followings.

(a) Let $X$ be a discrete random variable that has a probability mass

$$\mathbf{P}(X = x_i) = p_i \quad (i = 1, \cdots, m)$$

for $p_i > 0$ and $\sum_{i=1}^{m} p_i = 1$. Prove that

$$H(X) \le \log m$$

where $H(X)$ denotes the entropy of $X$.

(b) Calculate the KL divergence between two univariate Gaussian distributions $D_{KL}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(m, s^2))$.

*Solution.*

(a) By Jensen's inequality,

$$H(X) = -\sum_{i=1}^{m} p_i \log p_i = \sum_{i=1}^{m} p_i \log \frac{1}{p_i} \leq \log \left( \sum_{i=1}^{m} p_i \frac{1}{p_i} \right) = \log m$$

Note that the function $\log x$ is concave (i.e., $-\log x$ is convex).

(b) Let $\sigma, s > 0$.

$$\begin{aligned} D_{KL}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(m, s^2)) &= \int \mathcal{N}(x|\mu, \sigma^2) \log \frac{\mathcal{N}(x|\mu, \sigma^2)}{\mathcal{N}(x|m, s^2)} \, dx \\ &= \int \left( \log s - \log \sigma - \frac{(x-\mu)^2}{2\sigma^2} + \frac{(x-m)^2}{2s^2} \right) \mathcal{N}(x|\mu, \sigma^2) \, dx \\ &= \log s - \log \sigma - \frac{1}{2\sigma^2} \left( (\mu^2 + \sigma^2) - 2\mu^2 + \mu^2 \right) + \frac{1}{2s^2} \left( (\mu^2 + \sigma^2) - 2m\mu + m^2 \right) \\ &= \log \frac{s}{\sigma} - \frac{1}{2} + \frac{\mu^2 + \sigma^2 - 2m\mu + m^2}{2s^2} \end{aligned}$$

5. (10pts) Note that there are two different layout conventions : the numerator-layout notation and the denominator-layout notation. In our lecture, we use the numerator-layout notation. With the numerator-layout notation,

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{x}_1} & \frac{\partial y}{\partial \mathbf{x}_2} & \cdots & \frac{\partial y}{\partial \mathbf{x}_n} \end{bmatrix}, \qquad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial x} & \frac{\partial \mathbf{y}_2}{\partial x} & \cdots & \frac{\partial \mathbf{y}_m}{\partial x} \end{bmatrix}^\top$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_n} \\ \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n} \end{bmatrix}, \qquad \frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{X}_{11}} & \frac{\partial y}{\partial \mathbf{X}_{21}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{p1}} \\ \frac{\partial y}{\partial \mathbf{X}_{12}} & \frac{\partial y}{\partial \mathbf{X}_{22}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{p2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial \mathbf{X}_{1q}} & \frac{\partial y}{\partial \mathbf{X}_{2q}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{pq}} \end{bmatrix},$$

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial \mathbf{Y}_{11}}{\partial x} & \frac{\partial \mathbf{Y}_{12}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{1n}}{\partial x} \\ \frac{\partial \mathbf{Y}_{21}}{\partial x} & \frac{\partial \mathbf{Y}_{22}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{Y}_{m1}}{\partial x} & \frac{\partial \mathbf{Y}_{m2}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{mn}}{\partial x} \end{bmatrix}$$

where

$$x, y \in \mathbb{R}, \; \mathbf{x} \in \mathbb{R}^n, \; \mathbf{y} \in \mathbb{R}^m, \; \mathbf{X} \in \mathbb{R}^{p \times q}, \; \mathbf{Y} \in \mathbb{R}^{m \times n}.$$

On the other hand, with the denominator-layout notation,

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{x}_1} & \frac{\partial y}{\partial \mathbf{x}_2} & \cdots & \frac{\partial y}{\partial \mathbf{x}_n} \end{bmatrix}^\top, \qquad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial x} & \frac{\partial \mathbf{y}_2}{\partial x} & \cdots & \frac{\partial \mathbf{y}_m}{\partial x} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_1} \\ \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_2} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_n} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_n} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n} \end{bmatrix}, \qquad \frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{X}_{11}} & \frac{\partial y}{\partial \mathbf{X}_{12}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{1q}} \\ \frac{\partial y}{\partial \mathbf{X}_{21}} & \frac{\partial y}{\partial \mathbf{X}_{22}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial \mathbf{X}_{p1}} & \frac{\partial y}{\partial \mathbf{X}_{p2}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{pq}} \end{bmatrix}$$

where

$$x, y \in \mathbb{R},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^m,\ \mathbf{X} \in \mathbb{R}^{p\times q},\ \mathbf{Y} \in \mathbb{R}^{m\times n}.$$

(Please read `https://en.wikipedia.org/wiki/Matrix_calculus#Layout_conventions` for details.) In this problem, we use the numerator-layout notation as in the lecture. Answer the followings.

(a) Prove that

$$\frac{\partial}{\partial A}\mathrm{tr}(AB) = B$$

for $n \times n$ matrices $A$ and $B$. In particular,

$$\frac{\partial}{\partial A}\mathrm{tr}(A) = I.$$

(b) Prove that

$$\frac{\partial}{\partial x}\log|A| = \mathrm{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)$$

for a $n \times n$ invertible matrix $A(x)$ where $x \in \mathbb{R}$ and $|A| = \det(A) > 0$.
(Hint: Prove when $A$ is symmetric, and then prove the general case. Note that if $A$ is symmetric, we can represent

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$$

where $\lambda_i$ is an eigenvalue of $A$, $u_i$ is a corresponding eigenvector, and $u_i$'s are orthonormal.)

(c) Prove that

$$\frac{\partial}{\partial A}\log|A| = A^{-1}$$

for a $n \times n$ invertible matrix $A$ where $|A| = \det(A) > 0$.

*Solution.*

(a) Let $a_{ij}, b_{ij}$ be $(i,j)$ entry of $A$ and $B$, respectively. Note that

$$\mathrm{tr}(AB) = \sum_{1\le i,j\le n} a_{ij}b_{ji}$$

So,

$$\frac{\partial}{\partial a_{ij}}\mathrm{tr}(AB) = b_{ji}$$

and we get the desired result.

(b) Assume that $A$ is symmetric. Then, we can diagonalize $A$ as

$$A = U\Lambda U^\top = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$$

and also

$$A^{-1} = U\Lambda^{-1}U^\top = \sum_{i=1}^{n} \frac{1}{\lambda_i} u_i u_i^\top$$

Note that we use the same notation as in the Hint. Then,

$$\begin{aligned}
\operatorname{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right) &= \operatorname{tr}\left(\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}u_iu_i^\top\right)\left(\sum_{i=1}^{n}\left[\lambda_i\left(\frac{\partial u_i}{\partial x}u_i^\top + u_i\frac{\partial u_i^\top}{\partial x}\right) + \frac{\partial \lambda_i}{\partial x}u_iu_i^\top\right]\right)\right)\\
&= \operatorname{tr}\left(\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}u_iu_i^\top\right)\left(\sum_{i=1}^{n}\lambda_i\frac{\partial u_i}{\partial x}u_i^\top\right)\right) + \operatorname{tr}\left(\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}u_iu_i^\top\right)\left(\sum_{i=1}^{n}\lambda_i u_i\frac{\partial u_i^\top}{\partial x}\right)\right)\\
&+ \operatorname{tr}\left(\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}u_iu_i^\top\right)\left(\sum_{i=1}^{n}\frac{\partial \lambda_i}{\partial x}u_iu_i^\top\right)\right)\\
&= \operatorname{tr}\left(\left(\sum_{i=1}^{n}\lambda_i\frac{\partial u_i}{\partial x}u_i^\top\right)\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}u_iu_i^\top\right)\right) + \operatorname{tr}\left(\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}u_iu_i^\top\right)\left(\sum_{i=1}^{n}\lambda_i u_i\frac{\partial u_i^\top}{\partial x}\right)\right)\\
&+ \operatorname{tr}\left(\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}u_iu_i^\top\right)\left(\sum_{i=1}^{n}\frac{\partial \lambda_i}{\partial x}u_iu_i^\top\right)\right)\\
&= \operatorname{tr}\left(\sum_{i=1}^{n}\frac{\partial u_i}{\partial x}u_i^\top\right) + \operatorname{tr}\left(\sum_{i=1}^{n}u_i\frac{\partial u_i^\top}{\partial x}\right) + \operatorname{tr}\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x}u_iu_i^\top\right) \quad (\because u_i\text{'s are orthonormal})\\
&= 2\sum_{i=1}^{n}\operatorname{tr}\left(u_i\frac{\partial u_i^\top}{\partial x}\right) + \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x}\operatorname{tr}\left(u_iu_i^\top\right)\\
&= 2\sum_{i=1}^{n}\operatorname{tr}\left(\frac{\partial u_i^\top}{\partial x}u_i\right) + \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x}\operatorname{tr}\left(u_i^\top u_i\right)\\
&= \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x}
\end{aligned}$$

The last equality holds since $u_i^\top u_i = 1$. Since $|A| = \prod_{i=1}^{n}\lambda_i$,

$$\frac{\partial}{\partial x}\log|A| = \sum_{i=1}^{n}\frac{\partial}{\partial x}\log|\lambda_i| = \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x}$$

Thus, we proved when $A$ is symmetric. Let $A$ be an invertible matrix (not necessarily symmetric). However, since $AA^\top$ is symmetric,

$$\frac{\partial}{\partial x}\log|AA^\top| = \operatorname{tr}\left((AA^\top)^{-1}\frac{\partial(AA^\top)}{\partial x}\right)$$

holds as we proved. Since

$$\frac{\partial}{\partial x}\log|AA^\top| = \frac{\partial}{\partial x}\log\left(|A||A|\right) = 2\frac{\partial}{\partial x}\log|A|$$

and

$$\begin{aligned}
\operatorname{tr}\left((AA^\top)^{-1}\frac{\partial(AA^\top)}{\partial x}\right) &= \operatorname{tr}\left((A^\top)^{-1}A^{-1}\left(\frac{\partial A}{\partial x}A^\top + A\frac{\partial A^\top}{\partial x}\right)\right)\\
&= \operatorname{tr}\left((A^\top)^{-1}A^{-1}\frac{\partial A}{\partial x}A^\top\right) + \operatorname{tr}\left((A^\top)^{-1}A^{-1}A\frac{\partial A^\top}{\partial x}\right)\\
&= \operatorname{tr}\left(\frac{\partial A}{\partial x}A^{-1}\right) + \operatorname{tr}\left((A^{-1})^\top\frac{\partial A^\top}{\partial x}\right)\\
&= 2\operatorname{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)
\end{aligned}$$

Therefore, we get the desired result.

(c) Let $a_{ij}, a'_{ij}$ be $(i, j)$ entry of $A$ and $A^{-1}$, respectively. Then,

$$\frac{\partial}{\partial a_{ij}} \log |A| = \operatorname{tr}\left( A^{-1} \frac{\partial A}{\partial a_{ij}} \right) = \operatorname{tr}\left( A^{-1} E_{ij} \right) = a'_{ji}$$

where $E_{ij}$ denotes a $n \times n$ matrix with all entries are 0 except the $(i, j)$ entry which is 1. Thus,

$$\frac{\partial}{\partial A} \log |A| = A^{-1}.$$