DDA3020 Assignment 1 Student ID: 121090429 Ou Ziyi 欧子裔

Writing Problems:

1.1 Matrix Derivation

For the matrix derivation, denominator-layout is adopted

(2) Derivation by definition

① (5 pts)

$$f(w) = w^T A w$$

$$= [w_1, w_2, \cdots, w_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \ddots & & \vdots \\ \vdots & & & \\ a_{n1} & \cdots & & a_{nn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$= [a_{11}w_1 + a_{21}w_2 + \cdots + a_{n1}w_n, \cdots, a_{1n}w_1 + a_{2n}w_2 + \cdots + a_{nn}w_n] \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$= a_{11}w_1^2 + a_{21}w_2w_1 + \cdots + a_{n1}w_nw_1 + \cdots + a_{1n}w_1w_n + a_{2n}w_2w_n + \cdots + a_{nn}w_n^2$$

$$\frac{df}{dw} = \begin{bmatrix} a_{11}w_1 + a_{21}w_2 + \cdots + a_{n1}w_n + a_{11}w_1 + a_{12}w_2 + \cdots + a_{1n}w_n \\ a_{12}w_1 + a_{22}w_2 + \cdots + a_{n2}w_n + a_{21}w_1 + a_{22}w_2 + \cdots + a_{2n}w_n \\ \vdots \\ a_{n1}w_1 + a_{n2}w_2 + \cdots + a_{nn}w_n + a_{1n}w_1 + a_{2n}w_2 + \cdots + a_{nn}w_n \end{bmatrix}_{n\times 1}$$

$$= \begin{bmatrix} a_{11}w_1 + a_{21}w_2 + \cdots + a_{n1}w_n \\ a_{12}w_1 + a_{22}w_2 + \cdots + a_{n2}w_n \\ \vdots \\ a_{n1}w_1 + a_{n2}w_2 + \cdots + a_{nn}w_n \end{bmatrix} + \begin{bmatrix} a_{11}w_1 + a_{12}w_2 + \cdots + a_{1n}w_n \\ a_{21}w_1 + a_{22}w_2 + \cdots + a_{2n}w_n \\ \vdots \\ a_{1n}w_1 + a_{2n}w_2 + \cdots + a_{nn}w_n \end{bmatrix}$$

$$= A^T w + A w$$

② (5 pts)

$$f(w) = Aw$$

$$= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \ddots & & \vdots \\ \vdots & & & \\ a_{n1} & \cdots & & a_{nn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}w_1 + a_{12}w_2 + \cdots + a_{1n}w_n \\ a_{21}w_1 + a_{22}w_2 + \cdots + a_{2n}w_n \\ \cdots \\ a_{n1}w_1 + a_{n2}w_2 + \cdots + a_{nn}w_n \end{bmatrix}_{n\times 1}$$

$$\frac{df}{dw} = \begin{bmatrix} a_{11} & a_{21} & a_{31} & \cdots & a_{n1} \\ a_{12} & a_{22} & a_{32} & \cdots & a_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & a_{3n} & \cdots & a_{nn} \end{bmatrix}$$

$$= A^T$$

(3) (Bonus track: 5 pts) Derivation by differentiation

① (2 pts) $f(w) = w^T A w$ is a scalar

Let $g(w) = w^T$, $h(w) = Aw$

Since $d(XY) = (dX)Y + X(dY)$, $tr(A+B) = tr(A) + tr(B)$

$$
\begin{aligned}
d(w^T A w) &= (dw^T) Aw + w^T (dAw) \\
&= (dw^T) Aw + w^T \cdot dA \cdot w + w^T A (dw) \\
&= (dw^T) Aw + w^T A (dw) \\
&= tr((dw^T) Aw + w^T A (dw)) \\
&= tr((dw^T) Aw) + tr(w^T A (dw)) \\
&= tr(w^T A^T (dw)) + tr(w^T A (dw)) \\
&= tr[w^T A^T (dw) + w^T A (dw)] \\
&= tr([(A^T + A) w]^T (dw)) \\
&= [(A^T + A) w]^T (dw)
\end{aligned}
$$

Therefore $\frac{df}{dw} = \frac{d}{dw}(w^T A w) = (A^T + A) w$

② (3 pts) $f(W) = tr(W^T A W)$

Since $d(tr(X)) = tr(dX)$,

$df = d(tr(W^T A W)) = tr(d(W^T A W))$

$$
\begin{aligned}
tr(d(W^T A W)) &= tr(d(W^T) \cdot AW + W^T d(A) W + W^T A (dW)) \\
&= tr(d(W^T) \cdot AW + W^T A (dW)) \\
&= tr(d(W^T) \cdot AW) + tr(W^T A (dW)) \\
&= tr(W^T A^T (dW)) + tr(W^T A (dW)) \\
&= tr([(A^T + A) W]^T (dW))
\end{aligned}
$$

For the formula $df = tr((\frac{df}{dW})^T dW)$,

we get: $\frac{df}{dW} = (A^T + A) W$.

(4) (a) (10 pts)

Given that $l = \| Xw - y \|_2^2$, $z = Xw - y$.

Since $l = z^T z$, $\frac{dl}{dz} = 2z$.

$$
\begin{aligned}
\frac{\partial z}{\partial w^T} &= \frac{\partial}{\partial w^T}(Xw - y) \\
&= X
\end{aligned}
$$

$$\Rightarrow \frac{dl}{dw} = \left(\frac{\partial z}{\partial w^T}\right)^T \cdot \frac{\partial l}{\partial z}$$

$$= X^T \cdot 2z$$

$$= 2X^T(Xw - y)$$

(b) (Bonus Task, 5 pts)

Suppose $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times p}$, $\Rightarrow B \in \mathbb{R}^{m\times p}$

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}, \quad X = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}. \quad B = \begin{bmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mp} \end{bmatrix}$$

For each entry of Y,

$$Y_{kl} = \Sigma_s A_{ks} X_{sl}$$

$$\Rightarrow \frac{\partial Y_{kl}}{\partial X_{ij}} = \frac{\partial \Sigma_s A_{ks} X_{sl}}{\partial X_{ij}} = \frac{\partial A_{ki} X_{il}}{\partial X_{ij}} = A_{ki}\delta_{lj}$$

where $\delta_{lj} = 1$ when $l = j$, otherwise $\delta_{lj} = 0$.

Therefore $\frac{\partial l}{\partial X_{ij}} = \sum_{kl} \frac{\partial l}{\partial Y_{kl}} (A_{ki}\delta_{lj}) = \sum_k \frac{\partial l}{\partial Y_{kj}} A_{ki}$

$$\Rightarrow \frac{\partial l}{\partial X} = \begin{bmatrix} \Sigma_k \left(\frac{\partial l}{\partial Y_{k1}}\right) \cdot A_{k1} & \Sigma_k \left(\frac{\partial l}{\partial Y_{k1}}\right) \cdot A_{k2} & \cdots & \Sigma_k \left(\frac{\partial l}{\partial Y_{k1}}\right) \cdot A_{kn} \\ \Sigma_k \left(\frac{\partial l}{\partial Y_{k2}}\right) \cdot A_{k1} & \Sigma_k \left(\frac{\partial l}{\partial Y_{k2}}\right) \cdot A_{k2} & \cdots & \Sigma_k \left(\frac{\partial l}{\partial Y_{k2}}\right) \cdot A_{kn} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_k \left(\frac{\partial l}{\partial Y_{kp}}\right) \cdot A_{k1} & \Sigma_k \left(\frac{\partial l}{\partial Y_{kp}}\right) \cdot A_{k2} & \cdots & \Sigma_k \left(\frac{\partial l}{\partial Y_{kp}}\right) \cdot A_{kn} \end{bmatrix}$$

1.2 Convexity of functions, 10 pts

Proof: (1) Proved by definition.

Without loss of generally, for $x, y \in \mathbb{R}$, assume $x \leq y$.

We have the following 3 situations:

① $x \leq y \leq 0$

$f(x) = f(y) = 0$.

$\forall \theta \in [0,1]$, $\theta x + (1-\theta) y \leq 0$.

Therefore $f(\theta x + (1-\theta)y) = 0 = \theta f(x) + (1-\theta) f(y)$ $\forall \theta \in [0,1]$, $x \leq y \leq 0$.

② $x \leq 0 < y$

$f(x) = 0$. $f(y) = y > 0$.

(i) If $\theta x + (1-\theta) y \leq 0$,

Then $f(\theta x + (1-\theta) y) = 0 < \theta f(x) + (1-\theta) f(y) = (1-\theta) y$

(ii) If $\theta x + (1-\theta) y > 0$,

$f(\theta x + (1-\theta) y) = \theta x + (1-\theta) y < (1-\theta) y = \theta f(x) + (1-\theta) f(y)$

Hence we conclude that: for $x \leq 0 < y$, $\forall \theta \in [0,1]$,

$$f(\theta x+(1-\theta)y) \leq \theta f(x)+(1-\theta)f(y)$$

③ $0 < x \leq y$

$f(x)=x$. $f(y)=y$.

$\theta x+(1-\theta)y>0$,

$$f(\theta x+(1-\theta)y) = \theta x+(1-\theta)y = \theta \cdot f(x)+(1-\theta)f(y)$$

In conclusion, $\forall x,y \in \mathbb{R}$, $\theta \in [0,1]$,

$$f(\theta x+(1-\theta)y) \leq \theta f(x)+(1-\theta)f(y).$$

So $f(x)=\max(0,x)$ is convex for $x \in \mathbb{R}$.

(2) Proved by second-order conditions.

$$f(x)=|x|. \Rightarrow f(x)=\begin{cases} x & \text{for } x \geq 0 \\ -x & \text{for } x<0. \end{cases}$$

$$\nabla f(x)=\begin{cases} 1 & \text{for } x \geq 0 \\ -1 & \text{for } x<0. \end{cases}$$

$\nabla^2 f(x)=0$ for all $x \in \mathbb{R}$.

By the second-order conditions, $f(x)=|x|$ is convex.

(3) Proved by second-order conditions.

$f(x)=\|Ax-b\|_2^2$

By 1.1 (4)-(a), $\nabla f(x)=2A^T(Ax-b)$

$\nabla^2 f(x)=2A^TA \geq 0$ for $\forall x \in \mathbb{R}^n$

By the second-order conditions, $f(x)=\|Ax-b\|_2^2$ is convex.

1.3 Gradient Descent, 10pts

(1) Define $f(W)=tr[(Y-XW)^TA(Y-XW)]$

where $Y=(y_1,y_2,\cdots,y_N)^T \in \mathbb{R}^{N\times k}$, $X=[(x_1^T,1)^T,(x_2^T,1)^T,\cdots,(x_N^T,1)^T] \in \mathbb{R}^{N\times(d+1)}$

$W=(w,b)^T \in \mathbb{R}^{(d+1)\times k}$, $A=diag(a_1,a_2,\cdots,a_N)$.

To get the closed-form solution, let $\frac{df}{dW}=0$

$$df = d(tr[(Y-XW)^TA(Y-XW)])$$
$$= tr[d((Y-XW)^TA(Y-XW))]$$

$$d((Y-XW)^TA(Y-XW))$$
$$= d((Y-XW)^T)A(Y-XW)+(Y-XW)^Td(A(Y-XW))$$
$$= (d(Y-XW))^TA(Y-XW)+(Y-XW)^Td(A(Y-XW))$$

$$= (d(-XW))^T A(Y-XW) + (Y-XW)^T A(d(Y-XW))$$

$$= [(Y-XW)^T A^T (d(Y-XW))]^T + [(Y-XW)^T A(d(Y-XW))]$$

$$\Rightarrow df = tr[d((Y-XW)^T A(Y-XW))]$$

$$= tr\left([(Y-XW)^T A^T (d(Y-XW))]^T + [(Y-XW)' A(d(Y-XW))]\right)$$

$$= tr([(Y-XW)^T A^T + (Y-XW)^T A](d(Y-XW)))$$

$\Rightarrow$ Use differentiation method.

we get: $$\frac{df}{d(Y-XW)} = [(Y-XW)^T A^T + (Y-XW)^T A]^T$$

$$= A(Y-XW) + A^T(Y-XW)$$

$$= 2A(Y-XW) \quad \text{since } A^T = A.$$

Since $\frac{d(-XW)}{dW^T} = -X$, we get:

$$\frac{df}{dW} = \left(\frac{d(-XW)}{dW^T}\right)^T \cdot \frac{df}{d(-XW)} = 2(-X^T)A(Y-XW) = 0$$

$$\Rightarrow X^T A Y = X^T A X W$$

$$\Rightarrow W = (X^T A X)^{-1} X^T A Y$$

(2) The goal is to minimize $f(W) = tr[(Y-XW)^T A(Y-XW)]$

where $Y = (y_1, y_2, \cdots, y_N)^T \in \mathbb{R}^{N\times k}$, $X = [(x_1^T, 1)^T, (x_2^T, 1)^T, \cdots, (x_N^T, 1)^T] \in \mathbb{R}^{N\times(d+1)}$

$W = (W, b)^T \in \mathbb{R}^{(d+1)\times k}$ $A = diag(a_1, a_2, \cdots, a_N)$.

The step of gradient descent is as follows:

① Set an initial $W \in \mathbb{R}^{(d+1)\times k}$

② Find the gradient matrix of $f(W)$: $\frac{df}{dW}$

③ Choose α s.t. $\alpha = \underset{\alpha}{argmin}\, f(W + \alpha \cdot \frac{df}{dW})$

④ Set $W = W + \alpha \cdot \frac{df}{dW}$, repeat the above steps until $\nabla f(W)$ is small enough.

1.4 Maximum Likelihood Estimation, 10 pts

Proof: The likelihood function is:

$$L(\mu, \sigma^2) = \prod_{n=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \cdot \exp\left(-\frac{(X_n-\mu)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \cdot \exp\left(-\frac{1}{2\sigma^2} \cdot \sum_{n=1}^{N} (X_n-\mu)^2\right)$$

$$\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2)$$

$$= -\frac{N}{2} \cdot \log(2\pi\sigma^2) + \left(-\frac{1}{2\sigma^2}\right) \cdot \sum_{n=1}^{N} (x_n - \mu)^2.$$

$$\hat{\mu}_{MLE} := \arg\max_{\mu} \ell(\mu, \sigma^2)$$

$$\hat{\sigma}^2_{MLE} := \arg\max_{\sigma^2} \ell(\mu, \sigma^2)$$

$$\Rightarrow \left.\frac{\partial f}{\partial \mu}\right|_{\mu = \hat{\mu}_{MLE}} = -\frac{1}{2\sigma^2} \cdot \sum_{n=1}^{N} (x_n - \mu) \cdot (-2) \Big|_{\mu = \hat{\mu}_{MLE}} = 0$$

Therefore $\sum_{n=1}^{N} (x_n - \hat{\mu}_{MLE}) = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{1}{N} \sum_{n=1}^{N} x_n$

$$\left.\frac{\partial f}{\partial \sigma^2}\right|_{\sigma^2 = \hat{\sigma}^2_{MLE},\, \mu = \hat{\mu}_{MLE}} = \left(-\frac{N}{2}\right) \cdot \frac{2\pi}{2\pi\sigma^2} + \left(-\frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2\right) \cdot \frac{1}{\sigma^4} (-1)$$

$$\Rightarrow N = \frac{\sum_{n=1}^{N} (x_n - \hat{\mu}_{MLE})^2}{\hat{\sigma}^2_{MLE}}$$

Therefore $\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu}_{MLE})^2.$