1. (2)

① $f(w) = w^T A w$ given $w \in R^n$, $A \in R^{n \times n}$

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \quad A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \quad w^T = [w_1, \ldots, w_n]$$

$$f(w) = [w_1, \ldots, w_n] \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$= \left[\sum_{i=1}^{n} w_i a_{i1}, \ldots, \sum_{i=1}^{n} w_i a_{in} \right] \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$= w_1 \sum_{i=1}^{n} w_i a_{i1} + w_2 \sum_{i=1}^{n} w_i a_{i2} + \cdots + w_n \sum_{i=1}^{n} w_i a_{in}$$

$$= \sum_{j=1}^{n} w_j \sum_{i=1}^{n} w_i a_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j a_{ji} \qquad i=1$$

since f is a scalar function, w is a vector

the size of $\frac{\partial f}{\partial w}$ will be the same as w, i.e. R^n

since. $$\frac{\partial f}{\partial w_i} = \left(\sum_{j=1}^{n} w_j a_{ji} + w_i a_{ii} \right) + \sum_{\substack{j=1 \\ j \neq i}}^{n} w_j a_{ij} = \sum_{j=1}^{n} w_j a_{ji} + \sum_{j=1}^{n} w_j a_{ij}$$

Therefore, $\frac{\partial f}{\partial w} = A^T w + A w$.

we can examine the size of $\frac{\partial f}{\partial w}$, which is R^n

② $f(w) = Aw$ given $w \in R^n$, $A \in R^{m\times n}$.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

let $f(w) = y$. we can tell that y is a vector. of size R^m.

so $\frac{\partial f}{\partial w}$ will be a matrix of size $R^{m\times n}$

$$f(w) = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{n} a_{1j} w_j \\ \sum_{j=1}^{n} a_{2j} w_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj} w_j \end{bmatrix} = y$$

since $\frac{\partial y_i}{\partial w_j} = a_{ij}$

therefore. $\frac{\partial f}{\partial w} = A$

we can examine the correctness of the answer by the size.

1. (3) $f(w) = w^T A w$ given $A \in R^{n \times n}$ $w \in R^n$

① $A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$ $w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$

$$df = d(w^T A w) = d(w^T) A w + w^T (d(Aw))$$
$$= d(w^T) A w + w^T (d(A) w + A d(w))$$
$$= d(w^T) A w + w^T A d(w)$$

since f is a scalar function

$$df = tr(dy)$$

$$df = tr(d(w^T) A w + w^T A d(w))$$
$$= tr(d(w^T) A w) + tr(w^T A d(w))$$

Since $tr(X^T) = tr(X)$

$tr(XY) = tr(YX)$

therefore, $tr(d(w^T) A w) = tr((dw)^T A w)$
$$= tr((Aw)^T dw)$$

$$tr(w^T A dw) = tr((A^T w)^T dw)$$

$$df = tr((Aw + A^T w)^T dw)$$

$$\frac{df}{dw} = Aw + A^T w$$

① $f=(W)=tr(W^TAW)$ given. $W\in R^{m\times n}$, $A\in R^{m\times m}$

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & & a_{mm} \end{bmatrix} \quad W=\begin{bmatrix} W_1 & \cdots & W_{1n} \\ \vdots & & \\ W_{m1}, & \cdots, & W_{mn} \end{bmatrix}$$

$$df = d(tr(W^TAW))$$

$$= tr(d(W^TAW))$$

$$= tr(d(W^T)AW + W^Td(A)dW + W^TAdW)$$

$$= tr(d(W^T)AW + W^TAdw)$$

$$= tr((dw)^TAW + W^TAdw)$$

$$= tr((dw)^TAW) + tr(W^TAdw)$$

$$= tr((AW)^Tdw) + tr(W^TAdw)$$

$$= tr((AW + A^TW)^Tdw)$$

therefore, $\frac{df}{dW} = AW + A^TW$.

(4)

a. Since $\frac{dl}{dW} = \left(\frac{dz}{dW^T}\right)^T \frac{dl}{dz}$

$$z = XW - y.$$

$$\frac{dz}{dW} = X^T \qquad \frac{dz}{dW^T} = X$$

$$l = z^T z$$

$$\frac{dl}{dz} = 2z$$

Therefore: $\frac{dl}{dW} = X^T \cdot 2(XW - y)$

$$= 2X^T \cdot XW - 2X^T y.$$

(4). b. since $\frac{\partial Y_{kl}}{\partial X_{ij}} = \frac{\partial \sum_s A_{ks} X_{sl}}{\partial X_{ij}} = \frac{\partial A_{ki} X_{il}}{\partial X_{ij}} = A_{ki}\delta_{lj}$

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

$$\frac{\partial l}{\partial X_{ij}} = \sum_{kl} \frac{\partial l}{\partial Y_{kl}} A_{ki}\delta_{lj} = \sum_{k} \frac{\partial l}{\partial Y_{kj}} A_{ki} = A^{T}_{:,i} \left(\frac{\partial l}{\partial Y}\right)_{:,j}$$

from $\frac{\partial l}{\partial X_{ij}}$, we can derive $\frac{\partial l}{\partial X}$

$$\frac{\partial l}{\partial X} = A \frac{\partial l}{\partial Y}$$

1.2.

(1). $f(x) = Relu(x) = \begin{cases} x & x \geqslant 0 \\ 0 & x < 0 \end{cases}$ $x \in R$. $\nabla f(x) = \begin{cases} 1 & x > 0 \\ 0 & x < 0 \end{cases}$

f is not differentiable.

By the definition of convexity,

$$f(\theta x + (1-\theta)y) = \begin{cases} \theta x + (1-\theta)y & \text{if } \theta x + (1-\theta)y \geq 0 \\ 0 & \theta x + (1-\theta)y < 0 \end{cases}$$

$\theta \in [0,1]$

since $f(x) \geqslant x$ $f(y) \geqslant y$

Therefore, if $\theta x + (1-\theta)y \geqslant 0$.

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

$$\Rightarrow f(\theta x + (1-\theta)y) \leqslant \theta f(x) + (1-\theta) f(y)$$

if $\theta x + (1-\theta)y < 0$

since $f(x) \geqslant 0$, $f(y) \geqslant 0$

$$\theta f(x) + (1-\theta) f(y) \geqslant 0 = f(\theta x + (1-\theta)y),$$

Therefore, Relu(x) is convex.

(2) $f(x) = |x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$

f is not differentiable at $x=0$.

By the definition of convexity,

$$f(\theta x + (1-\theta)y) = \begin{cases} \theta x + (1-\theta)y & \text{if } \theta x + (1-\theta)y \geq 0 \\ (\theta-1)y - \theta x. & \theta x + (1-\theta)y < 0 \end{cases}$$

$\theta \in [0,1]$

since $f(x) \geq x \quad f(y) \geq y$

Therefore, $\underline{\text{if } \theta x + (1-\theta)y \geq 0.}$

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

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

$\underline{\text{if } \theta x + (1-\theta)y < 0}$

① if $x<0 \quad y>0$.

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

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

② if $x<0,\ y<0,$

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

③ if $x>0, y<0$,

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

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

Therefore, $f(x)=|x|$ is convex for $x\in R$.

(3). $f(x) = \|Ax - b\|_2^2$ $\quad$ dom $f \in R$. given $A \in R^{m \times n}$ $x \in R^n$

f is twice differentiable

By 2nd-order conditions, f is convex iff $\nabla^2 f(x) \geq 0$.

$\nabla f(x) = 2A^T(Ax - b)$

$\nabla^2 f(x) = 2A^T A$ $\quad$ (from the note)

notice that $A^T A$ is non-negative

therefore, $\nabla^2 f(x) \geq 0$

$\Rightarrow f(x) = \|Ax - b\|_2^2$ is convex.

1.3,

$$\sum_{i=1}^{N} \alpha_i \| y_i - Wx_i - b \|_2^2$$

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

$$Y = (y_1, y_2, \ldots, y_N)^T \in R^{N \times k}$$

$$X = [(x_1^T, 1), (x_2^T, 1)^T, \ldots (x_N^T, 1)^T]^T \in R^{N \times (d+1)}$$

$$W = (W, b)^T \in R^{(d+1) \times k}$$

$$A = diag(\alpha_1, \alpha_2, \ldots, \alpha_N).$$

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

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

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

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

$$W^T (X^T A X) W$$

let $\frac{d}{dW} J(W) = 0$

$$-(X^T A Y)^T - (Y^T A X)^T + (X^T A X) W + (X^T A^T X) W = 0.$$

$$(X^T A X + X^T A^T X) W = Y^T A^T X + X^T A^T Y$$

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

(2) given that

$$\frac{dJ(W)}{dW} = -(X^TAY)^T - (Y^TAX)^T + (X^TAX)W + (X^TA^TX)W$$

where

$$Y = (y_1, y_2, \dots, y_N)^T \in R^{N\times k}$$

$$X = [(x_1^T, 1), (x_2^T, 1)^T, \dots (x_N^T, 1)^T]^T \in R^{N\times(d+1)}$$

$$W = (w, b)^T \in R^{(d+1)\times k}$$

$$A = diag(\alpha_1, \alpha_2, \dots, \alpha_N).$$

step 1: initialize w and b. , α (learning rate)

step 2: iterate:

$$J(W) = WX.$$

$$W \leftarrow W - \alpha \frac{dJ(W)}{dW}$$

if breaking criteria. satisfy

(e.g. $\left|\alpha \frac{dJ(W)}{dW}\right| < \varepsilon_0$)

then

save w, b.

break.

1.4.

since $X_i \sim N(\mu, \sigma^2)$

set $J(w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$

when $\frac{dJ(w)}{d\sigma^2} = 0$, we can get the optimal σ^2_{MLE}

when $\frac{dJ(w)}{d\mu} = 0$, we can get the optimal μ_{MLE}.

for simlicity. we make $z = \sigma^2$

$$J(w) = \prod_{i=1}^{N} (2\pi z)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(X_i-\mu)^2 z^{-1}\right)$$

$$= (2\pi z)^{-\frac{N}{2}} \prod_{i=1}^{N} \exp\left(-\frac{1}{2}(X_i-\mu)^2 z^{-1}\right)$$

$$\ln J(w) = -\frac{N}{2} \ln 2\pi z - z^{-1} \sum_{i=1}^{N} \frac{1}{2}(X_i-\mu)^2$$

$$\frac{d \ln J(w)}{d\mu} = -2z^{-1} \sum_{i=1}^{N} \frac{1}{2}(X_i-\mu) = 0$$

$$\Rightarrow \mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$\frac{d \ln J(w)}{dz} = -\frac{N}{2} \cdot z^{-1} + z^{-2} \sum_{i=1}^{N} \frac{1}{2} (x_i - \mu)^2 = 0.$$

$$\frac{N}{2} \cdot z^{-1} = z^{-2} \sum_{i=1}^{N} \frac{1}{2} (x_i - \mu)^2$$

$$z = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

When $\mu = \mu_{MLE}$

$$\sigma^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{MLE})^2$$