(1.1) We can denote w and A as:

(1)
$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

Therefore, $w^T A w$ can be written as $\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j a_{ij}$, which is a scalar

$$\frac{d(w^T A w)}{dw} = \frac{d\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j a_{ij}\right)}{dw}$$

Then we can get that

$$\frac{d(w^T A w)}{dw} = \begin{bmatrix} \sum_{i=1}^{n} w_i a_{i1} + \sum_{j=1}^{n} w_j a_{1j} \\ \sum_{i=1}^{n} w_i a_{i2} + \sum_{j=1}^{n} w_j a_{2j} \\ \vdots \\ \sum_{i=1}^{n} w_i a_{in} + \sum_{j=1}^{n} w_j a_{nj} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} x_{i1} + \sum_{j=1}^{n} x_{1j} \\ \sum_{i=1}^{n} x_{i2} + \sum_{j=1}^{n} x_{2j} \\ \vdots \\ \sum_{i=1}^{n} x_{in} + \sum_{j=1}^{n} x_{nj} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = (A + A^T) w$$

(1.1) Derivative of a matrix-vector product w.r.t the vector

(2) Let's write $f(w)$ explicitly as a summation:

$$f(w) = Aw = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} w_j$$

Compute the derivative $\frac{df}{dw}$ by taking partial derivative of $f(w)$ w.r.t w:

$$\frac{df}{dw} = \frac{d}{dw}\left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} w_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{d}{dw}(A_{ij} w_j)$$

Consider A_{ij} is a constant w.r.t w, treat w_j as variable, derivative of a constant times a variable is the constant itself:

$$\frac{d}{dw}(A_{ij} w_j) = A_{ij}$$

$$\therefore \frac{df}{dw} = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} = A$$

With each element of the matrix is equal to the corr element of A

(1.1) $\frac{df}{dw} = d(w^T \times Aw)$

(3a) $= d(w^T) \times Aw + w^T \times d(Aw)$

$= (dw)^T \times Aw + w^T \times (dA \times w + A \times dw)$

$= (dw)^T \times Aw + w^T \times dA \times w + w^T \times A \times dw$

$tr \frac{df}{dw} = tr((dw)^T \times Aw + w^T \times dA \times w + w^T \times A \times dw)$

$= tr((dw)^T \times Aw) + tr(w^T \times dA \times w) + tr(w^T \times A \times dw)$

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

$tr \frac{df}{dw} = tr(dw^T \times Aw) + tr(w^T \times dA \times w) + tr(w^T \times A \times dw)$

since $tr(XY) = tr(YX)$,

$tr \frac{df}{dw} = tr(w^T \times dw \times A) + tr(w^T \times dA \times w) + tr(w^T \times A \times dw)$

$\therefore \frac{df}{dw} = w^T \times dw \times A + w^T \times dA \times w + w^T \times A \times dw$

(1.1)

$$\frac{df}{dw} = d(tr(W^T \times AW))$$

(3b)

$$= tr(d(W^T \times AW))$$

$$= tr((dW)^T \times AW + W^T \times d(AW))$$

$$= tr((dW)^T \times AW) + tr(W^T \times d(AW))$$

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

$$\frac{df}{dw} = tr(dW^T \times AW) + tr(W^T \times d(AW))$$

Since $tr(XY) = tr(YX)$,

$$\frac{df}{dw} = tr(W^T \times dW \times A) + tr(W^T \times d(AW))$$

$$\therefore \frac{df}{dw} = W^T \times dW \times A + W^T \times d(AW)$$

(1.1) Given that $\ell = \|z\|_2^2 = z^T z$,

(4a) then $\frac{\partial \ell}{\partial z} = 2z$

We now find $\frac{\partial z}{\partial w}$ and $\left(\frac{\partial z}{\partial w^T}\right)^T$, given that $z = Xw - y$,

$$\frac{dz}{dw} = X \quad \text{(Linear)}$$

$$\therefore \left(\frac{\partial z}{\partial w^T}\right)^T = X^T$$

Now using chain rule,

$$\frac{\partial \ell}{\partial w} = \left(\frac{\partial z}{\partial w^T}\right)^T \frac{\partial \ell}{\partial z}$$

$$= X^T(2z) = 2X^T w - 2X^T y$$

(1.2) To show that ReLU is convex, we need to prove that

(1) for any x and $y \in \text{dom } f$, $0 \le \theta \le 1$

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

Consider 4 cases:

Case 1: $x_1 \ge 0$ and $x_2 \ge 0$, $\text{ReLU}(x_1) = x_1$ and $\text{ReLU}(x_2) = x_2$

$$f(\theta x_1 + (1-\theta)x_2) = \theta x_1 + (1-\theta)x_2$$

$$\theta f(x_1) + (1-\theta) f(x_2) = \theta x_1 + (1-\theta)x_2$$

Since θ and $(1-\theta)$ are both non-negative, $\therefore$ satisfied

Case 2: $x_1 \ge 0$ and $x_2 \le 0$: $\text{ReLU}(x_1) = x_1$ and $\text{ReLU}(x_2) = 0$

$$f(\theta x_1 + (1-\theta)x_2) = \theta x_1$$

$$\theta f(x_1) + (1-\theta) f(x_2) = \theta x_1$$

$\therefore$ Satisfied

Case 3: $x_1 < 0$ and $x_2 \ge 0$, $\text{ReLU}(x_1) = 0$ and $\text{ReLU}(x_2) = x_2$

$$f(\theta x_1 + (1-\theta) x_2) = (1-\theta) x_2$$

$$\theta f(x_1) + (1-\theta) f(x_2) = (1-\theta) x_2$$

$\therefore$ Satisfied

Case 4: $x_1 < 0$ and $x_2 < 0$, $\text{ReLU}(x_1) = 0$ and $\text{ReLU}(x_2)$

$$f(\theta x_1 + (1-\theta) x_2 = 0$$

$$\theta f(x_1) + (1-\theta) x_2 = 0$$

$\therefore$ Satisfied

$\therefore$ $f(x)$ is convex for $x \in \mathbb{R}$

(1.2) f is strictly convex on $\mathbb{R}$ if and only if:

(2) $f(\alpha x+\beta y) \leq \alpha f(x)+\beta f(y)$

$\forall x, y \in \mathbb{R} : \forall \alpha, \beta \in \mathbb{R}_{\geq 0}, \alpha+\beta=1$ Let α be θ, β be $(1-\theta)$

$f(\alpha x+\beta y) = |\alpha x+\beta y|$

$\leq |\alpha x| + |\beta y|$ Triangle inequality

$= |\alpha||x| + |\beta||y|$ Absolute value is multiplicative

$= \alpha f(x) + \beta f(y)$ By Definition

$\therefore$ Convex

(3) Least square objective

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

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

$\nabla^2 f(x) = 2A^TA$

Since the second order derivative is positive semi-definite matrix, f is a convex function

(1.3) To solve $\min_{W,b} \sum_{i=1}^{N} a_i \|y_i - Wx_i - b\|_2^2$

(1) $$\frac{d}{dW} tr[(Y-XW)^T A(Y-XW)] = 0$$

Apply the trace of scalar property, $tr(AB) = tr(BA)$,

$$\frac{d}{dW} tr[(Y-XW)^T AY - (Y-XW)^T AXW] = 0$$

$$\frac{d}{dW} tr(Y^T AY - X^T W^T AY - Y^T AXW + X^T W^T AXW) = 0$$

Since trace operation is linear,

$$-2X^T AY + 2X^T AXW = 0$$

$$X^T AY = X^T AXW$$

$$W = (X^T AX)^{-1} X^T AY \quad //$$

$$\frac{d}{db} tr[(Y-XW)^T A(Y-XW)] = 0$$

$$-2\sum_{i=1}^{N} a_i (y_i - (Wx_i + b)) = 0$$

$$b = \frac{\sum_{i=1}^{N} a_i y_i - \sum_{i=1}^{N} a_i Wx_i}{\sum_{i=1}^{N} a_i} \quad //$$

(1.3) By gradient descent algorithm:

(2) 1. Choose learning rate α and initialise W_0 and b_0

2. Calculate the gradient of objective function w.r.t W and b using derivatives from (1)

$$J(W,b) = tr[(Y-XW)^T A(Y-XW)]$$

$$\frac{dJ(W,b)}{dW} = -2X^T A(Y-W)$$

$$\frac{dJ(W,b)}{db} = -2\sum_{i=1}^{N} a_i(y_i - (Wx_i + b))$$

3. Update W and b using gradients and learning rate

$$W_{i+1} = W_i - \alpha \frac{dJ(W,b)}{dW}$$

$$b_{i+1} = b_i - \alpha \frac{dJ(W,b)}{db}$$

4. Repeat steps 2 & 3 until convergence criteria are met

(1.4) Start with likelihood function, which is the PDF of the normal distribution:

$$L(\mu,\sigma^2|x_1,x_2,\ldots,x_N)=\prod_{n=1}^{N}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\times\exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right]$$

To find MLE, maximise likelihood function wrt σ^2,

Simplify maximisation with logarithm likelihood:

$$\ln(L)=\ln\left(\frac{1}{(2\pi)^{\frac{N}{2}}}\right)+\ln\left(\frac{1}{\sigma^N}\right)+\left[\sum_{n=1}^{N}\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)\right]$$

$$=\frac{N}{2}\ln(2\pi)+N\ln(\sigma)-\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}$$

Differentiate $\ln(L)$ wrt σ^2,

$$\frac{d}{d(\sigma^2)}\ln(L)=-\frac{N}{2\sigma^2}+\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^4}=0$$

$$\frac{N}{2\sigma^2}=\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^4}$$

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

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