1.1.(2) ①

$$w^TAw = A_{11}w_1w_1 + A_{12}w_1w_2 + \cdots + A_{1n}w_1w_n$$
$$+ \cdots$$
$$+ A_{n1}w_nw_1 + A_{n2}w_nw_2 + \cdots + A_{nn}w_nw_n.$$

$$\frac{df}{dw} = \begin{bmatrix} \frac{w^TAw}{\partial w_1} \\ \frac{w^TAw}{\partial w_2} \\ \vdots \\ \frac{w^TAw}{\partial w_n} \end{bmatrix} = \begin{bmatrix} (A_{11}w_1 + A_{12}w_2 + \cdots + A_{1n}w_n) + (A_{11}w_1 + A_{21}w_2 + \cdots + A_{n1}w_n) \\ \vdots \\ (A_{n1}w_n + A_{n2}w_2 + \cdots + A_{nn}w_n) + (A_{1n}w_1 + A_{2n}w_2 + \cdots + A_{nn}w_n) \end{bmatrix}$$

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

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

② . $$Aw = \begin{bmatrix} A_{11}w_1 & A_{12}w_2 & \cdots & A_{1n}w_n \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1}w_1 & A_{m2}w_2 & \cdots & A_{mn}w_n \end{bmatrix}$$

$$\frac{df}{dw} = \left[\frac{\partial A_{ij}w_j}{\partial w_j}\right] = [A_{ij}] = A.$$

1.1.(3)

① $$\frac{df}{dw} = \frac{d(w^TA)}{dw} w + w^T \frac{d(Aw)}{dw}$$
$$= Aw + w^TA$$
$$= (A + A^T)w$$

② $$\frac{df}{dw} = tr(d(w^TAw))$$
$$= tr[(A + A^T)w]$$

1.1.(4) (a) $$\frac{\partial l}{\partial w} = \left(\frac{\partial z}{\partial w^T}\right)^T \frac{\partial l}{\partial z}$$
$$= (x)^T \cdot 2z$$
$$= 2x^Tz$$

1.2 (1) $f(\lambda x_1 + (1-\lambda)x_2)$: $(\lambda \in (0,1))$

1) x_1, x_2 share same Positivity / Negativity

As Relu is linear at $x \le 0$ or $x \ge 0$, it is easy to find that :

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

2) Take $x_1 < 0$, $x_2 < 0$ for example.

$$f(\lambda x_1 + (1-\lambda)x_2) \le f((1-\lambda)x_2)$$
$$= (1-\lambda)f(x_2) + \lambda f(x_1)$$ which is 0.

As $\lambda x_1 + (1-\lambda)x_2 < (1-\lambda)x_2$

=> As a conclusion, $f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)x_2$.

it is convex

2)

1) For x_1, x_2 with same Positivity/Negativity

Apparently $f(\lambda x_1 + (1-\lambda)x_2) = \lambda f(x_1) + (1-\lambda)x_2$

2) take $x_1 < 0$, $x_2 \geq 0$

$$f(\lambda x_1 + (1-\lambda)x_2) = |\lambda x_1 + (1-\lambda)x_2| \leq |\lambda x_1| + |(1-\lambda)x_2|$$
$$= \lambda|x_1| + (1-\lambda)|x_2|$$
$$= \lambda f(x_1) + (1-\lambda)f(x_2)$$

As a conclusion, it is convex

3) $f(x) = (Ax-b)^T(Ax-b)$

$$= (x^T A^T - b^T)(Ax-b)$$
$$= x^T A^T Ax - x^T A^T b - b^T Ax + b^T b$$

$$\frac{\partial f(x)}{\partial x} = 2A^T Ax - 2A^T b$$

$$\frac{\partial^2 f(x)}{\partial^2 x} = 2A^T A$$, $A^T A$ is a square matrix => symmetric.

Let's prove $A^T A$ is semidefinite <=> Prove $\forall$ vector V, $V^T(A^T A)V \geq 0$

$$V^T(A^T A)V = 2(AV)^T(AV)$$
$$\Rightarrow \|AV\|_2^2 \geq 0 \Rightarrow \text{Proved.}$$

=> $\frac{\partial^2 f(x)}{\partial^2 x} = 2A^T A$ is semidefinite => $f(x)$ is convex

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

$$= tr[Y^T AY - Y^T AXW - W^T X^T AY + W^T X^T AXW]$$
$$= tr[Y^T AY] - 2tr[W^T X^T AY] + tr[W^T X^T AXW]$$

Taking derivative W.R.T. W:

$$\ldots = \frac{\partial}{\partial W}[-2tr[W^T X^T AY] + tr[W^T X^T AXW]]$$
$$= -2X^T AY + 2X^T AXW$$

let $\frac{\partial}{\partial W} = 0$ => $X^T AXW = X^T AY$ => $W = (X^T AX)^{-1} X^T AY$.

(2) $\frac{\partial}{\partial W} = -2X^T AY + 2X^T AXW$

=> $W^{(t+1)} = W^{(t)} - \alpha(-2X^T AY + 2X^T AXW^{(t)})$, α is the learn rate.

1.4

Likelihood func for given data:

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

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

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

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