1 Written Problems

1.1. ① Proof: Assume X is a $d \times h$ matrix, then

$$\frac{d(x_j^T w)}{dw_i} = \frac{d(x_{j1}w_1 + x_{j2}w_2 + \cdots + x_{jd}w_d)}{dw_i} = x_{ji}$$

$$\frac{d(X^T w)}{dw} = \begin{bmatrix} \frac{dx_1^T w}{dw_1} & \cdots & \frac{dx_h^T w}{dw_1} \\ \vdots & & \vdots \\ \frac{dx_1^T w}{dw_d} & & \frac{dx_h^T w}{dw_d} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{h1} \\ \vdots & & \vdots \\ x_{1d} & \cdots & x_{hd} \end{bmatrix} = X$$

② Proof: We let $\alpha^T = y^T X$, then $\alpha = X^T y$, since α is also not a function of w, we can obtain:

$$\frac{d(y^T X w)}{dw} = \frac{d(\alpha^T w)}{dw} \xrightarrow[\text{①}]{\text{Apply}} \alpha = X^T y$$

Thus, $\frac{d(y^T X w)}{dw} = X^T y$

③ Proof: w is $d \times 1$, X is $d \times d$, then by definition,

$$w^T X w = \sum_{j=1}^{d} \sum_{i=1}^{d} x_{ij} w_i w_j$$

Then, if we differentiate to the k^{th} element of w,

$$\frac{d(w^T X w)}{dw_k} = \sum_{j=1}^{d} x_{kj} w_j + \sum_{i=1}^{d} x_{ik} w_i \quad \forall k.$$

$$= w^T X_k^T + w^T X_k$$

Consequently, $\frac{d(w^T X w)}{dw} = w^T X^T + w^T X = w^T(X^T + X) = (X + X^T) w$

1.2 (1) We can pack $f_{W,b}(x) = Wx + b$ as

$f_{W,b}(X) = X\bar{w}$, where

$$X = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix} \in R^{N \times (d+1)}, \quad \bar{w} = [b \ w_1 \ w_2 \cdots w_d]^T \in R^{(d+1) \times k}$$

and also $Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} \in R^{N \times k}$, $A = \begin{bmatrix} a_1 & & 0 \\ & a_2 & \\ 0 & & \ddots \ a_N \end{bmatrix} \in R^{N \times N}$

We can reformulate $\sum_{i=1}^{N} a_i \|y_i - Wx_i - b\|^2$ as:

$$= (X\bar{w} - Y)^T A (X\bar{w} - Y)$$
$$= \bar{w}^T X^T A X \bar{w} - \bar{w}^T X^T A Y - Y^T A X \bar{w} + Y^T A Y$$
$$= f(\bar{w})$$

$$\frac{df(\bar{w})}{d\bar{w}} = 2X^T A X \bar{w} - 2X^T A Y$$

Let $\frac{df(\bar{w})}{d\bar{w}} = 0 \Rightarrow 2X^T A X \bar{w} = 2X^T A Y$

$$\hat{\bar{w}} = (X^T A X)^{-1} X^T A Y$$

Thus, the closed form solution is $\hat{\bar{w}} = (X^T A X)^{-1} X^T A Y$,

where $\bar{w} = [b \ w_1 \ w_2 \cdots w_d]^T$

(2) From (1), $J(\bar{w}) = (X\bar{w} - Y)^T A (X\bar{w} - Y)$,

$$\frac{dJ(\bar{w})}{d\bar{w}} = 2X^T A X \bar{w} - 2X^T A Y$$

We can update $\bar{w}$: $\bar{w}^* = \bar{w} - 2r \cdot X^T A (X\bar{w} - Y)$

where r is the appropriate step size, which can be determined by backtracking algorithm.

1.3 (1) $f'(x) = 4x^3 \quad f''(x) = 12x^2 \geq 0$

Since the second-order derivative is non-negative,

then $f(x) = x^4$ is convex.

(2) $f(x) = |x|$ is not second-order differentiable on R, we prove by definition: $\forall x_1, x_2 \in R$, $\lambda \in [0,1]$:

$$\begin{aligned} f(\lambda x_1 + (1-\lambda)x_2) &= |\lambda x_1 + (1-\lambda)x_2| \\ &\leq |\lambda x_1| + |(1-\lambda)x_2| \\ &= \lambda|x_1| + (1-\lambda)|x_2| \\ &= \lambda f(x_1) + (1-\lambda) f(x_2) \end{aligned}$$

We conclude that $f(x) = |x|$ is convex.

(3) $f(x) = ||Ax-b||^2 = (Ax-b)^T(Ax-b) = x^TA^TAx - x^TA^Tb - b^TAx + b^Tb$

$$\nabla f(x) = 2A^TAx - 2A^Tb$$

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

Since $||Ax||^2 = (Ax)^TAx = x^T(A^TA)x \geq 0$,

A^TA is always a PSD matrix.

Thus, $f(x) = ||Ax-b||^2$ is convex.

1.4. The pdf is given by $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

The Likelihood function is

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

We take natural log on both sides,

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

Take partial derivative w.r.t μ, σ^2.

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

$$\frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 = 0$$

Set derivatives equal to 0 and solve them, we get:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n = \bar{x} \ ,$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{ML})^2$$