### HW1- Answer

### Mathematic Background (0.8%)

- 1.  $\left(AA^{ op}\right)^{ op}=AA^{ op}\Rightarrow AA^{ op}$  is a symmetric matrix Moreover,  $orall x \in \mathbb{R}, \quad x^ op AA^ op x = \left(A^ op x
  ight)^ op \left(A^ op x
  ight) = \|A^ op x\|\geqslant 0$ Hence,  $AA^{\top}$  is positive semi-definite
- 2.  $f(x_1, x_2) = x_1 \sin(x_2) \exp(-x_1 x_2)$  $rac{\partial f(x_1,x_2)}{\partial x_1} = \sin(x_2) \exp(-x_1 x_2) - x_1 x_2 \exp(-x_1 x_2) \sin(x_2)$  $rac{\partial f(x_1, x_2)}{\partial x_2} = x_1 \cos(x_2) \exp(-x_1 x_2) + x_1 \sin(x_2) \exp(-x_1 x_2) \left(-x_1
  ight)$  $\Rightarrow 
  abla f(x) = \left[ egin{array}{l} \sin(x_2) \exp(-x_1 x_2) - x_1 x_2 \sin(x_2) \exp(-x_1 x_2) \ x_1 \cos(x_2) \exp(-x_1 x_2) - x_1^2 \sin(x_2) \exp(-x_1 x_2) \ \end{array} 
  ight]$
- 3. Consider the likelihood function

$$egin{aligned} &\prod_{i=1}^n f\left(x_i;p
ight) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \ \hat{p}_{mle} \in \operatorname{argmax}_p \prod_{i=1}^n f\left(x_i;p
ight) = \operatorname{argmax}_p \log \left(\prod_{i=1}^n f\left(x_i;p
ight)
ight) \ \log \left(\prod_{i=1}^n f\left(x_i;p
ight)
ight) = \sum_{i=1}^n x_i \log p + \left(n-\sum_{i=1}^n x_i^i
ight) \log (1-p) \end{aligned}$$
 Consider the first order condition:

Consider the first order condition:

$$egin{aligned} rac{\partial}{\partial p} \log \left( \prod_{i=1}^n f\left(x_i; p
ight) 
ight) &= rac{1}{p} \sum_{i=1}^n x_i - rac{1}{1-p} \left(n - \sum_{i=1}^n x_i 
ight) = 0 \ \Rightarrow pn - p \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i - p \sum_{i=1}^n x_i &\Rightarrow \hat{p}_{mle} = rac{1}{n} \sum_{i=1}^n x_i \ rac{\partial^2}{\partial p^2} \log \left( \prod_{i=1}^n f\left(x_i; p
ight) 
ight) &= rac{-1}{p^2} \sum_{i=1}^n x_i - rac{1}{(1-p)^2} \left(n - \sum_{i=1}^n x_i 
ight) \leq 0. \end{aligned}$$

# Closed-Form Linear Regression Solution (0.8%)

1 of 7 12/22/2022, 4:05 PM 1. Consider

$$egin{aligned} L(oldsymbol{ heta}) := \sum_{i=1}^n w_i (y_i - \mathbf{X}_i oldsymbol{ heta})^2 &= (\mathbf{y} - \mathbf{X} oldsymbol{ heta})^T oldsymbol{\Omega} \, (\mathbf{y} - \mathbf{X} oldsymbol{ heta}) \ &= \mathbf{y}^T oldsymbol{\Omega} \mathbf{y} - oldsymbol{ heta}^T \mathbf{X}^T oldsymbol{\Omega} \mathbf{y} - \mathbf{y}^T oldsymbol{\Omega} \mathbf{X} oldsymbol{ heta} + oldsymbol{ heta}^T \mathbf{X}^T oldsymbol{\Omega} \mathbf{X} oldsymbol{ heta} \end{aligned}$$

**Implies** 

$$abla L(oldsymbol{ heta}) = -2 \mathbf{X}^T \mathbf{\Omega} \mathbf{y} + 2 (\mathbf{X}^T \mathbf{\Omega} \mathbf{X}) oldsymbol{ heta} := 0$$

Then  $m{ heta}^* = \left(\mathbf{X}^T m{\Omega} \mathbf{X} \right)^{-1} \mathbf{X}^T m{\Omega} \mathbf{y}$ . Now, check  $m{ heta}^*$  is the optimal solution.

$$\begin{split} L(\boldsymbol{\theta}) &= \mathbf{y}^T \boldsymbol{\Omega} \mathbf{y} - \boldsymbol{\theta}^T \mathbf{X}^T \boldsymbol{\Omega} \mathbf{y} - \mathbf{y}^T \boldsymbol{\Omega} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{X}^T \boldsymbol{\Omega} \mathbf{X} \boldsymbol{\theta} \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{X}^T \boldsymbol{\Omega} \mathbf{X} \left( \boldsymbol{\theta} - \boldsymbol{\theta}^* \right) + \mathbf{y}^T \boldsymbol{\Omega} \mathbf{y} - \mathbf{y}^T \boldsymbol{\Omega} \mathbf{X} \left( \mathbf{X}^T \boldsymbol{\Omega} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Omega} \mathbf{y} \end{split}$$

Since  $\mathbf{X}^T \mathbf{\Omega} \mathbf{X}$  is positive semi-definte, the optimal solution appears at  $oldsymbol{ heta}^*$ 

2. Consider

$$egin{aligned} L(oldsymbol{ heta}) := \sum_i (y_i - \mathbf{X}_i oldsymbol{ heta})^2 + \lambda \sum_j w_j^2 &= (\mathbf{y} - \mathbf{X} oldsymbol{ heta})^T (\mathbf{y} - \mathbf{X} oldsymbol{ heta}) + \lambda oldsymbol{ heta}^T oldsymbol{ heta} \ &= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} oldsymbol{ heta} - oldsymbol{ heta}^T \mathbf{X}^T \mathbf{y} + oldsymbol{ heta}^T \mathbf{X}^T \mathbf{X} oldsymbol{ heta} + \lambda oldsymbol{ heta}^T oldsymbol{ heta} \end{aligned}$$

Since  $\lambda \boldsymbol{\theta}^T \boldsymbol{\theta} = \lambda \boldsymbol{\theta}^T I_m \boldsymbol{\theta}$ , where  $I_m$  is  $m \times m$  identity matrix. Moreover,  $\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^T I_m \boldsymbol{\theta} = \boldsymbol{\theta}^T (\mathbf{X}^T \mathbf{X} + \lambda I_m) \boldsymbol{\theta}$ 

$$abla L(oldsymbol{ heta}) = -2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X} + \lambda I_m)oldsymbol{ heta} := 0$$

Then  $m{ heta}^* = (\mathbf{X}^T\mathbf{X} + \lambda I_m)^{-1} m{X}^T\mathbf{y}$ . Now, check  $m{ heta}^*$  is the optimal solution.

$$\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \boldsymbol{\theta} - \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^T \boldsymbol{\theta}$$
  
=  $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T (\mathbf{X}^T \mathbf{X} + \lambda I_m) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \mathbf{y}^T \mathbf{y} - (\boldsymbol{\theta}^*)^T (\mathbf{X}^T \mathbf{X} + \lambda I_m) (\boldsymbol{\theta}^*)$ 

Clearly,  $\mathbf{X}^T\mathbf{X} + \lambda I_m$  is positive semi-definite matrix, since  $\lambda$  is a positive scaler. Hence,  $\boldsymbol{\theta}^*$  is the optimal solution.

(Bonus) Let  $X' = [X, \mathbf{1}] \in \mathbb{R}^{n imes (m+1)}$ , where  $\mathbf{1} = [1, 1, \cdots, 1]^T$ . Then

2 of 7

$$egin{aligned} L(oldsymbol{ heta}) := \sum_i (y_i - \mathbf{X}_i oldsymbol{ heta})^2 + \lambda \sum_j w_j^2 &= (\mathbf{y} - \mathbf{X}' oldsymbol{ heta})^T (\mathbf{y} - \mathbf{X}' oldsymbol{ heta}) + \lambda oldsymbol{ heta}^T oldsymbol{ heta} - \lambda b^2 \ &= (\mathbf{y} - \mathbf{X}' oldsymbol{ heta})^T (\mathbf{y} - \mathbf{X}' oldsymbol{ heta}) + oldsymbol{ heta}^T \lambda D oldsymbol{ heta} \end{aligned}$$

where 
$$D = egin{bmatrix} 1 & 0 & \cdots & 0 & 0 \ 0 & 1 & \cdots & 0 & 0 \ dots & dots & \ddots & dots & dots \ 0 & 0 & \cdots & 1 & 0 \ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^{(m+1) imes (m+1)}$$
 i.e.  $\mathrm{diag}(1,1,\cdots,1,0)$ 

Thus, we follow the same argument in (b), then we have

$$oldsymbol{ heta}^* = \left[egin{array}{c} \mathbf{w}^* \ b^* \end{array}
ight] = (\mathbf{X}^T\mathbf{X} + \lambda D)^{-1} oldsymbol{X}^T\mathbf{y}$$

# Logistic Sigmoid Function and Hyperbolic Tangent Function (0.8%)

1. By assumption, we have

$$2\sigma(2a) - 1 = rac{2}{1 + e^{-2a}} - 1$$

$$= rac{2}{1 + e^{-2a}} - rac{1 + e^{-2a}}{1 + e^{-2a}}$$

$$= rac{1 - e^{-2a}}{1 + e^{-2a}}$$

$$= rac{e^a - e^{-a}}{e^a + e^{-a}}$$

$$= anh(a)$$

3 of 7

2. If we now take  $a_j=rac{(x-\mu_j)}{2s}$  , we can rewrite as

$$egin{aligned} y(\mathbf{x},\mathbf{w}) &= w_0 + \sum_{j=1}^M w_j \sigma\left(2a_j
ight) \ &= w_0 + \sum_{j=1}^M rac{w_j}{2} \left(2\sigma\left(2a_j
ight) - 1 + 1
ight) \ &= u_0 + \sum_{j=1}^M u_j anhig(a_jig) \end{aligned}$$

where  $u_j=rac{1}{2}w_j$ , for  $j=1,\ldots,M$ , and  $u_0=w_0+rac{1}{2}\sum_{j=1}^M w_j$ .

## Noise and Regulation (0.8%)

By definition,

$$\begin{split} \tilde{L}_{ss}(\mathbf{w},b) &= \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_{i}+\eta_{i})-y_{i})^{2}\right] \\ &= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\{(\mathbf{w}^{T}(\mathbf{x}_{i}+\eta_{i})-y_{i})^{2}\} \\ &= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[\{(\mathbf{w}^{T}\mathbf{x}_{i}-y_{i})+\mathbf{w}^{T}\eta_{i}\}^{2}\right] \\ &= \frac{1}{2N}\sum_{i=1}^{N}\mathbb{E}\left[(\mathbf{w}^{T}\mathbf{x}_{i}-y_{i})^{2}\right]-2\mathbb{E}\{\mathbf{w}^{T}\eta_{i}(\mathbf{w}^{T}\mathbf{x}_{i}-y_{i})\}+\mathbb{E}\left[(\mathbf{w}^{T}\eta_{i})^{2}\right] \\ &= \frac{1}{2N}\sum_{i=1}^{N}(\mathbf{w}^{T}\mathbf{x}_{i}-y_{i})^{2}-2\mathbf{w}^{T}(\mathbf{w}^{T}\mathbf{x}_{i}-y_{i})\mathbb{E}(\eta_{i})+\mathbb{E}\left[(\mathbf{w}^{T}\eta_{i})^{2}\right] \\ &= \frac{1}{2N}\sum_{i=1}^{N}(\mathbf{w}^{T}\mathbf{x}_{i}-y_{i})^{2}+\mathbb{E}\left[(\mathbf{w}^{T}\eta_{i})^{2}\right] \end{split}$$

Note that  $\mathbb{E}(\eta_i) = 0$ Now, calculate  $\mathbb{E}\left[(\mathbf{w}^T\eta_i)^2
ight]$ 

4 of 7 12/22/2022, 4:05 PM

$$egin{aligned} \sum_{i=1}^{N} \mathbb{E}(\mathbf{w}^{T}\eta_{i})^{2} &= \sum_{i=1}^{N} \mathbb{E}(\sum_{j=1}^{k} w_{j}\eta_{i,j}) \ &= \sum_{i=1}^{N} \mathbb{E}(\sum_{j=1}^{k} \sum_{l=1}^{k} w_{j}w_{l}\eta_{i,j}\eta_{i,l}) \ &= \sum_{j=1}^{k} \sum_{l=1}^{k} w_{j}w_{l} \sum_{i=1}^{N} \mathbb{E}(\eta_{i,j}\eta_{i,l}) \ &= N\sigma^{2} \sum_{j=1}^{k} \sum_{l=1}^{k} w_{j}w_{l} = N\sigma^{2} \|w\|^{2} \end{aligned}$$

Hence,

$$egin{aligned} ilde{L}_{ss}(\mathbf{w},b) &= rac{1}{2N} \sum_{i=1}^{N} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 + rac{1}{2N} N \sigma^2 \|w\|^2 \ &= rac{1}{2N} \sum_{i=1}^{N} (f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + rac{\sigma^2}{2} \|\mathbf{w}\|^2 \end{aligned}$$

## Logistic Regression (0.8%)

1.

$$egin{align} \mathbf{w}^{ op}\mathbf{x} + b &= egin{bmatrix} -1 & 2 & -1 & 5 \end{bmatrix} egin{bmatrix} 7 \ 0 \ 3 \ 10 \end{bmatrix}^{ op} + 3 &= -7 + 0 - 3 + 50 + 3 = 43 \ P\left(c_1 \mid x\right) &= \sigma(43) = rac{1}{1 + e^{-43}} = 1 \ \Rightarrow P\left(c_2 \mid x\right) &= 1 - P\left(c_1 \mid x\right) = 1 - rac{1}{1 + e^{-43}} = rac{e^{-43}}{1 + e^{-43}} = 0 \ \end{cases}$$

2.

$$egin{aligned} P\left(y_i \mid \mathbf{x}_i
ight) &= f_{\mathbf{w},b}(\mathbf{x}_i)^{y_i} \cdot \left(1 - f_{\mathbf{w},b}(\mathbf{x}_i)^{1-y_i}
ight), y_i \in \{0,1\}. \ P(\mathbf{y} \mid \mathbf{x}) &= \prod_i P\left(y_i \mid \mathbf{x}_i
ight) = \prod_i f_{\mathbf{w},b}(\mathbf{x}_i)^{y_i} (1 - f_{\mathbf{w},b}\left(\mathbf{x}_i
ight))^{1-y_i} \end{aligned}$$

5 of 7 12/22/2022, 4:05 PM

Loss function 
$$L(\mathbf{w},b) = -\log p(\mathbf{y}|\mathbf{x}) = -\sum_{i} \left(y_{i} \log f_{\mathbf{w},b}\left(\mathbf{x}_{i}\right) + (1-y_{i}) \log(1-f_{\mathbf{w},b}\left(\mathbf{x}_{i}\right)\right)$$

3. Note that

$$egin{aligned} rac{d}{dx}\sigma(x) &= rac{d}{dx}rac{1}{1+\exp(-x)} \ &= (-1)(1+\exp(-x))^{-2}(-\exp(-x)) \ &= rac{\exp(-x)}{(1+\exp(-x))^2} \ &= rac{1}{(1+\exp(-x))}(1-rac{1}{(1+\exp(-x))}) = \sigma(x)(1-\sigma(x)) \end{aligned}$$

Consider  $\mathbf{z} = \mathbf{w}^T \mathbf{x} + b$ . Then

$$\bullet \frac{\partial \log \sigma(\mathbf{z})}{\partial \mathbf{w}} = \frac{\partial \log \sigma(\mathbf{z})}{\partial \sigma(\mathbf{z})} \frac{\partial \sigma(\mathbf{z})}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{w}} = (1 - \sigma(\mathbf{z}))\mathbf{x}$$

$$\bullet \frac{\partial \log(1 - \sigma(\mathbf{z}))}{\partial \mathbf{w}} = \frac{\partial \log(1 - \sigma(\mathbf{z}))}{\partial (1 - \sigma(\mathbf{z}))} \frac{\partial (1 - \sigma(\mathbf{z}))}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{w}} = -\sigma(\mathbf{z})\mathbf{x}$$

$$\bullet \frac{\partial \log \sigma(\mathbf{z})}{\partial b} = \frac{\partial \log \sigma(\mathbf{z})}{\partial \sigma(\mathbf{z})} \frac{\partial \sigma(\mathbf{z})}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial b} = (1 - \sigma(\mathbf{z}))$$

$$\bullet \frac{\partial \log(1 - \sigma(\mathbf{z}))}{\partial b} = \frac{\partial \log(1 - \sigma(\mathbf{z}))}{\partial (1 - \sigma(\mathbf{z}))} \frac{\partial (1 - \sigma(\mathbf{z}))}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial b} = -\sigma(\mathbf{z})$$

By above, we get

$$egin{aligned} rac{\partial L(\mathbf{w},b)}{\partial \mathbf{w}} &= -\sum_{i=1}^n y_i \cdot \left(1 - f_{\mathbf{w},b}\left(\mathbf{x}_i
ight)
ight) \mathbf{x}_i - \left(1 - y_i
ight) \cdot \left(-f_{\mathbf{w},b}\left(\mathbf{x}_i
ight)
ight) \mathbf{x}_i \\ &= \sum_{i=1}^n \mathbf{x}_i \left(f_{\mathbf{w},b}\left(\mathbf{x}_i
ight) - y_i
ight) \\ rac{\partial L(\mathbf{w},b)}{\partial b} &= \sum_{i=1}^n f_{\mathbf{w},b}\left(\mathbf{x}_i
ight) - y_i \end{aligned}$$

Hence,

$$egin{aligned} \mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta \sum_{i=1}^n \mathbf{x}_i \left( f_{\mathbf{w},b} \left( \mathbf{x}_i 
ight) - y_i 
ight) \ b^{t+1} \leftarrow b^t - \eta \sum_{i=1}^n f_{\mathbf{w},b} \left( \mathbf{x}_i 
ight) - y_i \end{aligned}$$

6 of 7 12/22/2022, 4:05 PM

7 of 7