HW1- Answer

Mathematic Background (0.8%)

1. $\left(AA^{\top}\right)^{\top} = AA^{\top} \Rightarrow AA^{\top}$ is a symmetric matrix
   Moreover, $\forall x \in \mathbb{R}, \quad x^{\top}AA^{\top}x = \left(A^{\top}x\right)^{\top}\left(A^{\top}x\right) = \|A^{\top}x\| \geqslant 0$
   Hence, $AA^{\top}$ is positive semi-definite
2. $f(x_1, x_2) = x_1 \sin(x_2)\exp(-x_1x_2)$
   $\frac{\partial f(x_1,x_2)}{\partial x_1} = \sin(x_2)\exp(-x_1x_2) - x_1x_2\exp(-x_1x_2)\sin(x_2)$
   $\frac{\partial f(x_1,x_2)}{\partial x_2} = x_1\cos(x_2)\exp(-x_1x_2) + x_1\sin(x_2)\exp(-x_1x_2)(-x_1)$
$$\Rightarrow \nabla f(x) = \begin{bmatrix} \sin(x_2)\exp(-x_1x_2) - x_1x_2\sin(x_2)\exp(-x_1x_2) \\ x_1\cos(x_2)\exp(-x_1x_2) - x_1^2\sin(x_2)\exp(-x_1x_2) \end{bmatrix}$$
3. Consider the likelihood function
   $\prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$
   $\hat{p}_{mle} \in \operatorname{argmax}_p \prod_{i=1}^{n} f(x_i; p) = \operatorname{argmax}_p \log\left(\prod_{i=1}^{n} f(x_i; p)\right)$
   $\log\left(\prod_{i=1}^{n} f(x_i; p)\right) = \sum_{i=1}^{n} x_i \log p + \left(n - \sum_{i=1}^{n} x_i^i\right)\log(1-p)$
   Consider the first order condition:
$$\frac{\partial}{\partial p}\log\left(\prod_{i=1}^{n} f(x_i; p)\right) = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right) = 0$$
$$\Rightarrow pn - p\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i \quad \Rightarrow \hat{p}_{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\frac{\partial^2}{\partial p^2}\log\left(\prod_{i=1}^{n} f(x_i; p)\right) = \frac{-1}{p^2}\sum_{i=1}^{n} x_i - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n} x_i\right) \leq 0.$$

Closed-Form Linear Regression Solution (0.8%)

1. Consider

$$
\begin{aligned}
L(\boldsymbol{\theta}) := \sum_{i=1}^{n} w_i(y_i - \mathbf{X}_i\boldsymbol{\theta})^2 &= (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T \boldsymbol{\Omega} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\right) \\
&= \mathbf{y}^T\boldsymbol{\Omega}\mathbf{y} - \boldsymbol{\theta}^T\mathbf{X}^T\boldsymbol{\Omega}\mathbf{y} - \mathbf{y}^T\boldsymbol{\Omega}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{X}^T\boldsymbol{\Omega}\mathbf{X}\boldsymbol{\theta}
\end{aligned}
$$

Implies

$$
\nabla L(\boldsymbol{\theta}) = -2\mathbf{X}^T\boldsymbol{\Omega}\mathbf{y} + 2(\mathbf{X}^T\boldsymbol{\Omega}\mathbf{X})\boldsymbol{\theta} := 0
$$

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

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

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

2. Consider

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

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

$$
\nabla L(\boldsymbol{\theta}) = -2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X} + \lambda I_m)\boldsymbol{\theta} := 0
$$

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

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

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

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

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

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

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

$$\boldsymbol{\theta}^* = \begin{bmatrix} \mathbf{w}^* \\ b^* \end{bmatrix} = (\mathbf{X}^T\mathbf{X} + \lambda D)^{-1}\boldsymbol{X}^T\mathbf{y}$$

Logistic Sigmoid Function and Hyperbolic Tangent Function (0.8%)

1. By assumption, we have

$$\begin{aligned} 2\sigma(2a) - 1 &= \frac{2}{1+e^{-2a}} - 1 \\ &= \frac{2}{1+e^{-2a}} - \frac{1+e^{-2a}}{1+e^{-2a}} \\ &= \frac{1-e^{-2a}}{1+e^{-2a}} \\ &= \frac{e^a - e^{-a}}{e^a + e^{-a}} \\ &= \tanh(a) \end{aligned}$$

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

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma\left(2a_j\right) \\ &= w_0 + \sum_{j=1}^{M} \frac{w_j}{2}\left(2\sigma\left(2a_j\right) - 1 + 1\right) \\ &= u_0 + \sum_{j=1}^{M} u_j \tanh\left(a_j\right) \end{aligned}$$

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

Noise and Regulation (0.8%)

By definition,

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

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

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

Hence,

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

Logistic Regression (0.8%)

1.

$$\begin{aligned}&\mathbf{w}^\top\mathbf{x} + b = \begin{bmatrix}-1 & 2 & -1 & 5\end{bmatrix}\begin{bmatrix}7\\0\\3\\10\end{bmatrix}^\top + 3 = -7 + 0 - 3 + 50 + 3 = 43 \\ &P\left(c_1 \mid x\right) = \sigma(43) = \frac{1}{1+e^{-43}} = 1 \\ \Rightarrow &P\left(c_2 \mid x\right) = 1 - P\left(c_1 \mid x\right) = 1 - \frac{1}{1+e^{-43}} = \frac{e^{-43}}{1+e^{-43}} = 0\end{aligned}$$

2.

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

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

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

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

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

By above, we get

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

Hence,

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