1 Problem 1

1.1

$$
\begin{aligned}
L(w,b) &:= -\sum_{n=1}^{N}\Big[y_n \log\big(\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)\big) + (1-y_n)\log\big(1-\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)\big)\Big] \\
&= -\sum_{n=1}^{N}\left[y_n \log\left(\frac{1}{1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}}\right) + (1-y_n)\log\left(\frac{e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}}{1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}}\right)\right] \\
&= -\Big[-\sum_{n=1}^{N}\big(y_n\log(1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)})\big) - \sum_{n=1}^{N}\big((1-y_n)\log(1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)})\big) + \sum_{n=1}^{N}\big((1-y_n)\log(e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)})\big)\Big] \\
&= \sum_{n=1}^{N}\log(1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}) - \sum_{n=1}^{N}\log\big((1-y_n)e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}\big) \\
\implies \nabla_\mathbf{w} L(w,b) &= \sum_{n=1}^{N} -\mathbf{x}_n\frac{e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}}{1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}} - \sum_{n=1}^{N}(1-y_n)(-\mathbf{x}_n)\cancel{\frac{e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}}{e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}}} \\
&= \sum_{n=1}^{N} -\mathbf{x}_n\frac{e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}}{1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}} + \sum_{n=1}^{N}\mathbf{x}_n - \sum_{n=1}^{N}\mathbf{y}_n\mathbf{x}_n \\
&= \sum_{n=1}^{N}\left(\frac{1}{1+e^{-(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b)}} - y_n\right)\mathbf{x}_n \\
&= \sum_{n=1}^{N}\Big(\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b) - y_n\Big)\mathbf{x}_n
\end{aligned}
$$

Gradient Descent:

$$
\begin{aligned}
\mathbf{w}^{(t+1)} &\leftarrow \mathbf{w}^{(t)} - \lambda\nabla_\mathbf{w} L(w,b) \\
&= \mathbf{w}^{(t)} - \lambda\sum_{n=1}^{N}\Big(\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_n+b) - y_n\Big)\mathbf{x}_n
\end{aligned}
$$

$(\mathbf{w}^{(0)} = [0,0]^\mathsf{T}$, $b=0$, $\lambda = 0.1) \implies$

$$
\begin{aligned}
\mathbf{w}^{(1)} &= -0.1\left((\sigma(0)-1)\begin{bmatrix}0\\1\end{bmatrix} + (\sigma(0)-0)\begin{bmatrix}2\\2\end{bmatrix} + (\sigma(0)-1)\begin{bmatrix}1\\0\end{bmatrix}\right) \\
&= -0.1\left(-\frac{1}{2}\begin{bmatrix}0\\1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}2\\2\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1\\0\end{bmatrix}\right) \\
&= -0.1\begin{bmatrix}1/2\\1/2\end{bmatrix} \\
&= \begin{bmatrix}-1/20\\-1/20\end{bmatrix}
\end{aligned}
$$

Let the classification rule be the following:

$$\mathbb{P}(y_n|\mathbf{x},\mathbf{w}) = \sigma({\mathbf{w}^{(1)}}^\mathsf{T}\mathbf{x}_n) \geq 0.5$$

$$\begin{aligned}
\sigma({\mathbf{w}^{(1)}}^\mathsf{T}\mathbf{x}_1) &= \frac{1}{1+e^{1/20}} \approx 0.488 < 0.5 \implies \hat{y}_1 = 0 \neq y_1 \\
\sigma({\mathbf{w}^{(1)}}^\mathsf{T}\mathbf{x}_2) &= \frac{1}{1+e^{1/5}} \approx 0.450 < 0.5 \implies \hat{y}_2 = 0 = y_2 \\
\sigma({\mathbf{w}^{(1)}}^\mathsf{T}\mathbf{x}_3) &= \frac{1}{1+e^{1/20}} \approx 0.488 < 0.5 \implies \hat{y}_3 = 0 \neq y_3
\end{aligned}$$

Training Accuracy $= 1/3$

1.2

$$\begin{aligned}
\sigma({\mathbf{w}^{(1)}}^\mathsf{T}\mathbf{x}_1) &= \frac{1}{1+e^{3/20}} \approx 0.463 < 0.5 \implies \hat{y}_1 = 0 = y_1 \\
\sigma({\mathbf{w}^{(1)}}^\mathsf{T}\mathbf{x}_2) &= \frac{1}{1+e^{-1/10}} \approx 0.525 > 0.5 \implies \hat{y}_2 = 1 = y_2 \\
\sigma({\mathbf{w}^{(1)}}^\mathsf{T}\mathbf{x}_3) &= \frac{1}{1+e^{1/10}} \approx 0.475 < 0.5 \implies \hat{y}_3 = 0 \neq y_3
\end{aligned}$$

Testing Accuracy $= 2/3$

2 Problem 2

2.1

$$\mathbb{P}(y_i|\mathbf{x}_i)|_{\widetilde{\mathbf{w}}=\mathbf{w}^*} = \mathbb{P}(\mathbf{w}^{*\mathsf{T}}\mathbf{x}_i + \epsilon_i|\mathbf{x}_i)$$

Given $\mathbf{x}_i$, $\mathbf{w}^{*\mathsf{T}}\mathbf{x}_i$ becomes a constant, and
$\epsilon_i \sim$ i.i.d. Laplace$(0, b) \implies \mathbf{w}^{*\mathsf{T}}\mathbf{x}_i + \epsilon_i \sim$ Laplace$(\mathbf{w}^{*\mathsf{T}}\mathbf{x}_i, b)$

,

$$\begin{aligned}
\implies \mathbb{P}(\mathbf{w}^{*\mathsf{T}}\mathbf{x}_i + \epsilon_i|\mathbf{x}_i) &= \frac{1}{2b}\exp\left(\frac{-|\mathbf{w}^{*\mathsf{T}}\mathbf{x}_i + \epsilon_i - \mathbf{w}^{*\mathsf{T}}\mathbf{x}_i|}{b}\right) \\
&= \frac{1}{2b}\exp\left(\frac{-|\epsilon_i|}{b}\right) \\
&= \boxed{\frac{1}{2b}\exp\left(\frac{-|y_i - \mathbf{w}^{*\mathsf{T}}\mathbf{x}_i|}{b}\right)}
\end{aligned}$$

2.2

$$\begin{aligned}
L(\mathbf{w}) &= \mathbb{P}(\mathbf{y}|\boldsymbol{X})|_{\widetilde{\mathbf{w}}=\mathbf{w}} \\
&= \mathbb{P}(\boldsymbol{X}\mathbf{w} + \boldsymbol{\epsilon}|\boldsymbol{X}) \\
&= \mathbb{P}(\boldsymbol{X}\mathbf{w} + \boldsymbol{\epsilon}) && \text{(Given } \boldsymbol{X}, \boldsymbol{X}\mathbf{w} \text{ is constant)} \\
&= \mathbb{P}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_1 + \epsilon_1|\epsilon_2, ..., \epsilon_n)\mathbb{P}(\mathbf{w}^{\mathsf{T}}(\mathbf{x}_2, ..., \mathbf{x}_n) + (\epsilon_2, ..., \epsilon_n)) \\
&= \mathbb{P}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_1 + \epsilon_1|\epsilon_2, ..., \epsilon_n)\mathbb{P}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_2 + \epsilon_2|\epsilon_3, ..., \epsilon_n)\ldots\mathbb{P}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + \epsilon_n) \\
&= \mathbb{P}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_1 + \epsilon_1)\mathbb{P}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_2 + \epsilon_2)\ldots\mathbb{P}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + \epsilon_n) && (\epsilon_i \perp \epsilon_j \ \forall\, i, j \in \{1, ..., n\}) \\
&= \prod_{i=1}^{n}\frac{1}{2b}\exp\left(\frac{-|y_i - \mathbf{w}^{\mathsf{T}}\mathbf{x}_i|}{b}\right)
\end{aligned}$$

2.3

Let's denote $\mathbf{w}^{\max}$ as the vector that maximizes the likelihood function.

$$\begin{aligned}
\mathbf{w}^{\max} &= \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{2b} \exp\left(\frac{-|y_i - \mathbf{w}^\intercal \mathbf{x}_i|}{b}\right) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} \log\left(\frac{1}{2b} \exp\left(\frac{-|y_i - \mathbf{w}^\intercal \mathbf{x}_i|}{b}\right)\right) && \text{(log monot. } \uparrow\text{)} \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} \left[\log(1) - \log(2b) + \log\left(\exp\left(\frac{-|y_i - \mathbf{w}^\intercal \mathbf{x}_i|}{b}\right)\right)\right] \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \left[- \cancelto{0}{N \log(1)} + N \log(2b) + \frac{1}{b} \sum_{i=1}^{n} |y_i - \mathbf{w}^\intercal \mathbf{x}_i| \right] \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - \mathbf{w}^\intercal \mathbf{x}_i| && \text{(Since b and N are constant)} \\
&= \mathbf{w}^*
\end{aligned}$$

3 Problem 3

3.1

$$h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\begin{aligned}
\implies \frac{\delta h(x)}{\delta x} &= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} && \text{(Using Quotient Rule)} \\
&= \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2} \\
&= 1 - \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)^2 \\
&= \boxed{1 - \big[h(x)\big]^2}
\end{aligned}$$

3.2

The change in L with respect to weights can be expressed as a chain of derivatives. I define δ_k to be the chain of derivatives starting from z_k. I.e.,

$$\frac{\partial L}{\partial v_{jk}} = \frac{\partial L}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial v_{jk}}$$

$$\frac{\partial L}{\partial w_{ki}} = \underbrace{\frac{\partial L}{\partial z_k}}_{:=\, \delta_k} \cdot \frac{\partial z_k}{\partial w_{ki}}$$

$$
\begin{aligned}
\delta_k := \frac{\partial L}{\partial z_k} &= \sum_{j=1}^{2} \frac{\partial L}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial z_k} \\
&= \boxed{\sum_{j=1}^{2} -(y_j - \hat{y}_j) v_{jk}}
\end{aligned}
$$

3.3

$$
\begin{aligned}
\frac{\partial L}{\partial v_{jk}} &= \frac{\partial L}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial v_{jk}} \\
&= \boxed{-(y_j - \hat{y}_j) z_k}
\end{aligned}
$$

3.4

$$
\begin{aligned}
\frac{\partial L}{\partial w_{ki}} &= \frac{\partial L}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_{ki}} \\
&= \delta_k \cdot \frac{\partial h(z_k)}{\partial w_{ki}} \\
&= \boxed{\sum_{j=1}^{2} -(y_j - \hat{y}_j) v_{jk} (1 - z_k^2) x_i}
\end{aligned}
$$

4 Problem 4

subpart i)

Without bias: $2 \cdot 2 \cdot 2 \cdot 3 + 4 \cdot 3 \cdot 3 \cdot 2 = 24 + 72 =$ $\boxed{96 \text{ parameters}}$

With bias: $2 \cdot 2 \cdot 2 \cdot 3 + 1 \cdot 2 + 4 \cdot 3 \cdot 3 \cdot 2 + 1 \cdot 4 =$ $\boxed{102 \text{ parameters}}$

subpart ii)

First CONV Layer: $(8 - 2)/1 + 1 = 7 \implies 7 \times 7 \times 2$
Second CONV Layer: $(7 - 3)/2 + 1 = 3 \implies 3 \times 3 \times 4$
Pooling Layer: $(3 - 2)/1 + 1 = 2 \implies 2 \times 2 \times 4$ (Depth preserved)

Thus, the final dimension is $\boxed{2 \times 2 \times 4}$

5 Problem 5

5.1

Consider the 2×2 upper left determinant of $\mathbf{K}$.

$$\begin{vmatrix} [2f(\mathbf{x}_1)]^2 & [f(\mathbf{x}_1)+f(\mathbf{x}_2)]^2 \\ [f(\mathbf{x}_1)+f(\mathbf{x}_2)]^2 & [2f(\mathbf{x}_2)]^2 \end{vmatrix} = \underbrace{16f(\mathbf{x}_1)^2 f(\mathbf{x}_2)^2}_{\geq 0} - \underbrace{[f(\mathbf{x}_1)+f(\mathbf{x}_2)]^4}_{\geq 0}$$

The sign of the determinant depends on $f(\mathbf{x})$. For example if we have $\mathbf{x} \in \mathbb{R}$, $f(\mathbf{x}) = \mathbf{x}$, $\mathbf{x_1} = 1$, $\mathbf{x_2} = 2$, then $16f(\mathbf{x}_1)^2 f(\mathbf{x}_2)^2 - [f(\mathbf{x}_1)+f(\mathbf{x}_2)]^4 = -17 < 0$. The determinant is less than zero, so the kernel matrix is not positive semi-definite and $k(\mathbf{x}_1, \mathbf{x}_2)$ is not a valid kernel.

5.2

$k(\mathbf{x}_1, \mathbf{x}_2)$ is a valid kernel $\implies$ $k(\mathbf{x}_1, \mathbf{x}_2)k(\mathbf{x}_1, \mathbf{x}_2) = k^2(\mathbf{x}_1, \mathbf{x}_2)$ is a valid kernel.

Suppose $k^n(\mathbf{x}_1, \mathbf{x}_2)$ is a valid kernel, then

$k^{n+1}(\mathbf{x}_1, \mathbf{x}_2) = k^n(\mathbf{x}_1, \mathbf{x}_2)k(\mathbf{x}_1, \mathbf{x}_2)$ is a product of two kernels, so it is also a kernel.

Since this is true for $n = 1$ and $n = 2$, $k^n(\mathbf{x}_1, \mathbf{x}_2)$ is a valid kernel $\forall n \in \mathbb{Z}^+$.

$$\begin{aligned} f(k(\mathbf{x}_1, \mathbf{x}_2)) &= \sum_{i=0}^{p} c_i k^i(\mathbf{x}_1, \mathbf{x}_2) \\ &= c_0 + c_1 k(\mathbf{x}_1, \mathbf{x}_2) + c_2 k^2(\mathbf{x}_1, \mathbf{x}_2) + \ldots + c_p k^p(\mathbf{x}_1, \mathbf{x}_2) \end{aligned}$$

$c_0 \geq 0$. Then let $\phi(\mathbf{x}) = \sqrt{\mathbf{c_0}} \;\; \forall \mathbf{x}$
$\implies \phi(\mathbf{x_1})\phi(\mathbf{x_2}) = \sqrt{\mathbf{c_0}} \cdot \sqrt{\mathbf{c_0}}$.
$\implies c_0$ can be represented as a kernel function.

Thus, the whole expression is a linear combination of valid kernels with positive coefficients, so $f(k(\mathbf{x}_1, \mathbf{x}_2))$ is a kernel.