2.

1. We have

$$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{n}\mu_i\xi_i,$$

where

$$\begin{aligned}
&\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}, \quad \boldsymbol{\xi} = \{\xi_1, \ldots, \xi_n\}, \quad \boldsymbol{\mu} = \{\mu_1, \ldots, \mu_n\};\\
&\alpha_i \geq 0, \quad \xi_i \geq 0, \quad \mu_i \geq 0, \quad y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, n;\\
&\alpha_i\left(y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right) = 0, \qquad \mu_i\xi_i = 0.
\end{aligned}$$

Differentiate L with respect to $b, \mathbf{w}, \xi_i$ and let them be 0, we get

$$\begin{aligned}
&\frac{\partial L}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i = 0 && \Longrightarrow \quad \sum_{i=1}^{n}\alpha_i y_i = 0;\\
&\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i = 0 && \Longrightarrow \quad \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i;\\
&\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 && \Longrightarrow \quad \alpha_i + \mu_i = C.
\end{aligned}$$

2. From part (1), we know that $\frac{\partial L}{\partial b} = 0$ gives $\sum_{i=1}^{n}\alpha_i y_i = 0$, $\frac{\partial L}{\partial \mathbf{w}} = 0$ gives $\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i$, $\frac{\partial L}{\partial \xi_i}$ gives $\alpha_i + \mu_i = C$, then we get

$$\max_{\boldsymbol{\xi}}\min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j.$$

It follows that

$$\begin{aligned}
\max_{\boldsymbol{\alpha},\boldsymbol{\mu}}\left(\max_{\boldsymbol{\xi}}\min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})\right) &= \max_{\boldsymbol{\alpha}}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j\\
&= \min_{\boldsymbol{\alpha}}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{n}\alpha_i.
\end{aligned}$$

According to the KKT condition, we have

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad \mu_i \geq 0, \quad \alpha_i + \mu_i = C, \quad i = 1, \ldots, n,$$

this means that this dual optimization problem is subject to

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n.$$

Therefore, the dual optimization problem is

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{n}\alpha_i,\\
&s.t. \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n.
\end{aligned}$$

3.

. a) *Proof.* Let $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$ be the corresponding feature maps for x_i and x_j respectively. Then we get

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) = \Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_i) = K(\mathbf{x}_j, \mathbf{x}_i).$$

This just means that the $K(\mathbf{x}_i, \mathbf{x}_j)$ is symmetric. □

b) *Proof.* Let $\Phi(\mathbf{x}_i)$ be the feature map for the i-th example and define the matrix $B = [\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)]$. Then we would get $A = B^T B$ and

$$\mathbf{x}^T A \mathbf{x} = \mathbf{x}^T B^T B \mathbf{x} = (B\mathbf{x})^T (B\mathbf{x}) = \|B\mathbf{x}\|_2^2 \geq 0, \qquad \forall \mathbf{x} \in \mathbb{R}^n.$$

This just means that the kernel matrix A is semi-positive definite. □

4.

1. It does not matter in Model 1, because the $w_1X_1 + w_2X_2 = 0$ always holds for $\mathbf{x}^{(3)} = (0,0)^T$. It follows that $P(Y = 1|\mathbf{x}^{(3)}, w_1, w_2) = \frac{1}{1+e^0} = \frac{1}{2}$, thus, the corresponding part in the MLE equation is

$$P(Y = 1|\mathbf{x}^{(3)}, w_1, w_2)^{[y^{(3)}=1]} \left(1 - P(Y = 1|\mathbf{x}^{(3)}, w_1, w_2)\right)^{[y^{(3)}=-1]} = \left(\frac{1}{2}\right)^{[y^{(3)}=1]} \left(\frac{1}{2}\right)^{[y^{(3)}=-1]},$$

   which is the same for both $y^{(3)} = \pm 1$. Therefore, the learned value of $\mathbf{w} = (w_1, w_2)$ would be the same when we change the label of the third example to -1.
   It does matter in Model 2, since $w_0 + w_1X_1 + w_2X_2 = w_0$ holds.

2. We would get the penalized log-likelihood of the labels be

$$\begin{aligned} l(\mathbf{w}) &= \sum_i \log P(y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}) - \frac{\lambda}{2}\|\mathbf{w}\|^2 \\ &= \sum_i \log g(y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)}) - \frac{\lambda}{2}\|\mathbf{w}\|^2 \\ &\approx \frac{1}{2}\sum_i y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)} - \frac{\lambda}{2}\|\mathbf{w}\|^2. \end{aligned}$$

   Differentiate l with respect to $\mathbf{w}$ and let it be 0, we get

$$\frac{\partial l}{\partial \mathbf{w}} \approx \frac{1}{2}\sum_i y^{(i)}\mathbf{x}^{(i)} - \lambda\mathbf{w} = 0.$$

   It gives that

$$\hat{\mathbf{w}} \approx \frac{1}{2\lambda}\sum_i y^{(i)}\mathbf{x}^{(i)}.$$

   Hence, the magnitude of $\mathbf{w}$ decreases as λ increases.

5.

1. From

$$z^{(l+1)} = W^{(l)}a^{(l)} + b^{(l)},$$

we have

$$z_i^{(l+1)} = \left(\sum_j w_{ij}^{(l)} a_j^{(l)}\right) + b_i^{(l)}.$$

It follows that

$$\frac{\partial z_i^{(l+1)}}{\partial w_{ij}^{(l)}} = a_j^{(l)} \qquad \text{and} \qquad \frac{\partial z_i^{(l+1)}}{\partial b_i^{(l)}} = 1.$$

Hence, we get

$$\frac{\partial}{\partial w_{ij}^{(l)}} J(W, b; \mathbf{x}, y) = \frac{\partial J(W, b; \mathbf{x}, y)}{\partial z_i^{(l+1)}} \frac{\partial z_i^{(l+1)}}{\partial w_{ij}^{(l)}} = \delta_i^{(l+1)} a_j^{(l)} = a_j^{(l)} \delta_i^{(l+1)}$$

and

$$\frac{\partial}{\partial b_i^{(l)}} J(W, b; \mathbf{x}, y) = \frac{\partial J(W, b; \mathbf{x}, y)}{\partial z_i^{(l+1)}} \frac{\partial z_i^{(l+1)}}{\partial b_i^{(l)}} = \delta_i^{(l+1)}.$$

Therefore, we have

$$\begin{aligned}\frac{\partial}{\partial w_{ij}^{(l)}} J(W, b) &= \frac{1}{n} \sum_{\text{sample id}=1}^{n} \frac{\partial}{\partial w_{ij}^{(l)}} J(W, b; \mathbf{x}, y) + \frac{\lambda}{2} \frac{\partial}{\partial w_{ij}^{(l)}} \sum_{l=1}^{L} \sum_{\bar{i}=1}^{s_l} \sum_{\bar{j}=1}^{s_{l+1}} (w_{\bar{j}\bar{i}}^{(l)})^2 \\ &= \frac{1}{n} \sum_{\text{sample id}=1}^{n} a_j^{(l)} \delta_i^{(l+1)} + \lambda w_{ij}^{(l)}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial b_i^{(l)}} J(W, b) &= \frac{1}{n} \sum_{\text{sample id}=1}^{n} \frac{\partial}{\partial b_i^{(l)}} J(W, b; \mathbf{x}, y) + \frac{\lambda}{2} \frac{\partial}{\partial b_i^{(l)}} \sum_{l=1}^{L} \sum_{\bar{i}=1}^{s_l} \sum_{\bar{j}=1}^{s_{l+1}} (w_{\bar{j}\bar{i}}^{(l)})^2 \\ &= \frac{1}{n} \sum_{\text{sample id}=1}^{n} \delta_i^{(l+1)}.\end{aligned}$$

2. Notice that

$$J(W, b; \mathbf{x}, y) = \frac{1}{2}\|h_{W,b}(\mathbf{x}) - y\|^2, \quad h_{W,b}(\mathbf{x}) = a^{(L)} \quad \text{and} \quad a^{(l+1)} = f(z^{(l+1)}).$$

We would get

$$\begin{aligned}\delta_i^{(L)} &= \frac{\partial J(W, b; \mathbf{x}, y)}{\partial z_i^{(L)}} = \frac{\partial \frac{1}{2}\|a^{(L)} - y\|^2}{\partial z_i^{(L)}} \\ &= (a_i^{(L)} - y_i) \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} = -(y_i - a_i^{(L)}) \frac{\partial f(z_i^{(L)})}{\partial z_i^{(L)}} = -(y_i - a_i^{(L)}) f'(z_i^{(L)}),\end{aligned}$$

and

$$\begin{aligned}\delta_i^{(l)} &= \frac{\partial J(W, b; \mathbf{x}, y)}{\partial z_i^{(l)}} = \sum_{j=1}^{s_l+1} \frac{\partial J(W, b; \mathbf{x}, y)}{\partial z_j^{(l+1)}} \frac{\partial z_j^{(l+1)}}{\partial z_i^{(l)}} \\ &= \sum_{j=1}^{s_l+1} \delta_j^{(l+1)} \frac{\partial \left(\sum_k w_{jk}^{(l)} a_k^{(l)}\right) + b_j^{(l)}}{\partial z_i^{(l)}} = \sum_{j=1}^{s_l+1} \delta_j^{(l+1)} \frac{\partial \left(\sum_k w_{jk}^{(l)} f(z_k^{(l)})\right) + b_j^{(l)}}{\partial z_i^{(l)}} \\ &= \sum_{j=1}^{s_l+1} \delta_j^{(l+1)} w_{ji}^{(l)} f'(z_i^{(l)}) = \left(\sum_{j=1}^{s_l+1} w_{ji}^{(l)} \delta_j^{(l+1)}\right) f'(z_i^{(l)}), \quad \text{for } l = L-1, \ldots, 2.\end{aligned}$$