Q2. (1)

$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \overbrace{\underbrace{\frac{1}{2n}\|Xw+b\mathbf{1}-y\|_2^2}_{\text{error}}+\lambda\|w\|_2^2}^{\text{loss}}, \tag{1}$$

where $X\in\mathbb{R}^{n\times d}$ and $y\in\mathbb{R}^n$ are the given dataset and $\lambda\geq 0$ is the regularization hyperparameter. If $\lambda=0$, then this is the standard linear regression problem. Observe the distinction between the *error* (which does not include the regularization term) and the *loss* (which does).

Prove (1) is equivalent to

$$\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2n}\left\|\begin{bmatrix} X & \mathbf{1}_n \\ \sqrt{2\lambda n}\, I_d & 0_d\end{bmatrix}\begin{bmatrix} w \\ b\end{bmatrix}-\begin{bmatrix} y \\ 0_d\end{bmatrix}\right\|_2^2$$

Simplifying the expression inside the norm,

$$\begin{aligned}
&\begin{bmatrix} X & \mathbf{1}_n \\ \sqrt{2\lambda n}\, I_d & 0_d\end{bmatrix}\begin{bmatrix} w \\ b\end{bmatrix}-\begin{bmatrix} y \\ 0_d\end{bmatrix}\\
&=\begin{bmatrix} Xw+b\mathbf{1}_n \\ \sqrt{2\lambda n}\, I_d\, w+0_d\cdot b\end{bmatrix}-\begin{bmatrix} y \\ 0_d\end{bmatrix}\\
&=\begin{bmatrix} Xw+b\mathbf{1}_n-y \\ \sqrt{2\lambda n}\, w-0_d\end{bmatrix}\\
&=\begin{bmatrix} Xw+b\mathbf{1}_n-y \\ \sqrt{2\lambda n}\, w\end{bmatrix}
\end{aligned}$$

So the full expression becomes

$$\begin{aligned}
&\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2n}\left\|\begin{bmatrix} X & \mathbf{1}_n \\ \sqrt{2\lambda n}\, I_d & 0_d\end{bmatrix}\begin{bmatrix} w \\ b\end{bmatrix}-\begin{bmatrix} y \\ 0_d\end{bmatrix}\right\|_2^2\\
&=\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2n}\left\|\begin{bmatrix} Xw+b\mathbf{1}_n-y \\ \sqrt{2\lambda n}\, w\end{bmatrix}\right\|_2^2\\
&=\min_{w\in\mathbb{R}^d,\, b\in\mathbb{R}} \frac{1}{2n}\left\|\begin{bmatrix} Xw+b\mathbf{1}_n-y \\ \sqrt{2\lambda n}\, w\end{bmatrix}\right\|_2^2
\end{aligned}$$

(2) The loss function is

$$L(w,b)=\frac{1}{2n}\|Xw+b\mathbf{1}-y\|_2^2+\lambda\|w\|_2^2.$$

Derivative with respect to w:

First, expand the squared norm:

$$\|Xw+b\mathbf{1}-y\|_2^2=(Xw+b\mathbf{1}-y)^T(Xw+b\mathbf{1}-y).$$

Then, taking the derivative:

$$\frac{\partial L}{\partial w} = \frac{1}{2n}\frac{\partial}{\partial w}\left[(Xw + b\mathbf{1} - y)^T(Xw + b\mathbf{1} - y)\right] + \lambda\frac{\partial}{\partial w}(w^T w).$$

Using the chain rule:

$$\frac{\partial}{\partial w}\left[(Xw + b\mathbf{1} - y)^T(Xw + b\mathbf{1} - y)\right] = 2X^T(Xw + b\mathbf{1} - y),$$

$$\frac{\partial}{\partial w}(w^T w) = 2w.$$

So we get:

$$\frac{\partial L}{\partial w} = \frac{1}{n}X^T(Xw + b\mathbf{1} - y) + 2\lambda w.$$

Derivative with respect to b:
Similarly, for b:

$$\frac{\partial L}{\partial b} = \frac{1}{2n}\frac{\partial}{\partial b}\left[(Xw + b\mathbf{1} - y)^T(Xw + b\mathbf{1} - y)\right] + \lambda\frac{\partial}{\partial b}(w^T w).$$

The second term is zero because $w^T w$ does not depend on b. For the first term:

$$\frac{\partial}{\partial b}\left[(Xw + b\mathbf{1} - y)^T(Xw + b\mathbf{1} - y)\right] = 2\mathbf{1}^T(Xw + b\mathbf{1} - y).$$

Thus:

$$\frac{\partial L}{\partial b} = \frac{1}{n}\mathbf{1}^T(Xw + b\mathbf{1} - y).$$

(3) In separate PDF

(4) In separate PDF

(5)
- Standardization improves gradient descent convergence; without it, errors can be large.
- For $\lambda = 0$, closed-form: training error 9.69, test 128.40; gradient descent: training 10.02, test 119.76. For $\lambda = 2$, closed-form: training 9.71, test 126.39; gradient descent: training 62.96, test 51.14.
- Closed-form is much faster (under 1 ms) than gradient descent (few ms).
- Gradient descent shows gradual decrease in training loss; standardization improves stability.

Conclusion: Closed-form is faster, stable, and more accurate for this dataset. Gradient descent is slower but can generalize better with proper standardization and is suitable for large datasets.