Q1

We first represent J in matrix form

$$\begin{aligned}J &= ||\mathbf{X}\boldsymbol{w} - \boldsymbol{t}||_2^2 + ||\boldsymbol{w}||^2 \\ &= (\mathbf{X}\boldsymbol{w} - t)^\top(\mathbf{X}\boldsymbol{w} - t) + w^\top w \\ &= \boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - \boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{t} - \boldsymbol{t}^\top\mathbf{X}\boldsymbol{w} + \boldsymbol{t}^\top\boldsymbol{t} + \boldsymbol{w}^\top\boldsymbol{w}\end{aligned}$$

Take the derivative of J with respect to $\boldsymbol{w}$

$$\begin{aligned}\nabla_{\boldsymbol{w}} J &= \nabla_{\boldsymbol{w}}(\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - \boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{t} - \boldsymbol{t}^\top\mathbf{X}\boldsymbol{w} + \boldsymbol{t}^\top\boldsymbol{t} + \boldsymbol{w}^\top\boldsymbol{w}) \\ &= 2(\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - \mathbf{X}^\top\boldsymbol{t} + \boldsymbol{w})\end{aligned}$$

We know J is a convex function, the closed-form solution can be found by letting $\nabla_{\boldsymbol{w}} J = 0$. Thus,

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{w} - \mathbf{X}^\top\boldsymbol{t} + \boldsymbol{w} = 0$$

Therefore, we have

$$\boldsymbol{w} = (\mathbf{X}^\top\mathbf{X} + \mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{t}$$

Q2

Suppose we have a function $f(x) = wx$, and we want to optimize it using L_1 loss: $J(w) = |f(x) - t)|$. If we have only one data point in our dataset $\mathcal{D} = \{(1, 0)\}$, and we want to find the value w that fits this dataset. In this case, it is easy to see that the global optimal is $w^* = 0$.

With the gradient descent algorithm, each time step t we update w with $w^{(t)} = w^{(t-1)} - \nabla_w J(w) = w^{(t-1)} - \alpha^{(t-1)}\nabla_w |w^{(t-1)}|$, where $w \neq 0$.

Suppose our gradient descent starts with $w^{(0)} = 1$, and it has a small initial learning rate $\alpha^{(0)} = 0.1$. Therefore, the gradient descent is converging to $w^* = 0$ from the $w > 0$ side. Thus, we have $\nabla_w |w^{(t-1)}| = 1$.

Now, let us use an annealed learning rate $\alpha^{(t)} = \frac{1}{2^t}\alpha^{(0)}$. The annealed gradient descent computes $w^{(t)}$ as following

$$w^{(t)} = w^{(0)} - \alpha^{(0)}(\frac{1}{2^0} + \frac{1}{2^1} + \cdots + \frac{1}{2^{t-1}}) \tag{1}$$

We know that

$$\lim_{t\to\infty}(\frac{1}{2^0} + \frac{1}{2^1} + \cdots + \frac{1}{2^{t-1}}) = 2 \tag{2}$$

Therefore,

$$\lim_{t\to\infty} w^{(t)} = w^{(0)} - 2\alpha^{(0)} = 0.8 > 0 \tag{3}$$

This example shows that, a decayed learning rate may prevents the gradient descent algorithm from having enough energy for finding better local/global optimums.