Q1

The KL divergence is:

$$KL(p \parallel q_\theta) = \sum_x p(X) \log \frac{p(X)}{q_\theta(X)} \tag{1}$$

The cross entropy is:

$$H(p, q_\theta) = \sum_x p(X) \log \frac{1}{q_\theta(X)} \tag{2}$$

The relationship is showed as formula below:

$$\begin{aligned} KL(p \parallel q_\theta) &= \sum_x p(X) \log \frac{p(X)}{q_\theta(X)} \\ &= -\sum_x p(x) \log \frac{1}{p(x)} + \sum_x p(x) \log \frac{1}{q_\theta(x)} \\ &= H(p, q_\theta) - H(p) \end{aligned} \tag{3}$$

$l(\theta \mid \mathcal{D}) = \log\ L(\theta \mid \mathcal{D})\ =\ \log\ q_\theta(\mathcal{D})$

In the machine learning process, we wish to minimize the KL divergence (1). From (3), we know that minimize the $KL(p \parallel q_\theta)$ is same as minimize the cross entropy $H(p, q_\theta)$ since $H(p)$ is unchanged. Approximating the cross entropy using data, we have:

$$\begin{aligned} H(p, q_\theta) &= -\int p(\mathbf{x}) \log\ q_\theta(\mathbf{x}) d\mathbf{x} \\ &\approx -\frac{1}{N} \sum_{i=1}^{N} \log\ q_\theta(\mathbf{x}_i) \\ &= -\frac{1}{N} \log\ q_\theta(\mathcal{D}) \end{aligned}$$

Which is same as maximizing the log likelihood.

Q2

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \quad y = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

$$\begin{aligned} \hat{\mathbf{w}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \begin{bmatrix} 2.0 \\ 1.5 \\ -1.5 \end{bmatrix} \end{aligned}$$

Therefore, $\hat{y} = 1.5 * x_1 - 1.5 * x_2 + 2.0$

Q3

1.c. There is no penalty on w_0, thus $w_0 = 1$

2.b. The strong penalty on the weight will force the function have small slope.

3.a. The penalty is given to w_0 either, resulting $w_0 < 1$.

4. d. The regularization scheme put strong penalty on both the intercept and weight, results the intercept in figure d) relatively low and the model underfit.

Q4

The prediction of before iteration is:

X_1	0	0	1	1
X_2	0	1	0	1
Y	1	0	0	1
$\sigma(\mathbf{w}^\top \mathbf{x}_i)$	0.1192	0.2689	0.2689	0.5

According to the batch gradient descent formula:

$$w_j \leftarrow w_j + \alpha \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)] x_{i,j}$$

W_0 = -2 + 0.1 * 0.25 * [(1 – 0.1192)*1 + (0 – 0.2689)*1 + (0 – 0.2689)*1 + (1 – 0.5)*1]
= -2 + 0.025 * 0.843
= -1.978925

W_1 = 1 + 0.1 * 0.25 * [(1 – 0.1192)*0 + (0 – 0.2689)*0 + (0 – 0.2689)*1 + (1 – 0.5)*1]
= 1 + 0.025 * 0.2311
= 1.0058

W_2 = 1 + 0.1 * 0.25 * [(1 – 0.1192)*0 + (0 – 0.2689)*1 + (0 – 0.2689)*0 + (1 – 0.5)*1]
= 1 + 0.025 * 0.2311
= 1.0058

The prediction of After iteration is:

X_1	0	0	1	1
X_2	0	1	0	1
Y	0	0	0	1
$Y_{predict}$	0	0	0	1

The training error is 0.

Q5

1.

The minimum achievable training error is 0.25, since a straight line cannot separate the dataset and classify the data correctly.

A possible weight is w_0 = -1.1, $w_1 = w_2 = 1$

2.

The minimum achievable training error is 0

A possible weight is $w_0 = 1$, $w_1 = w_2 = -1$, $w_3 = 2$

Q6

1.

For the first updating rule, since $x_1 = 0$ in all examples with $y = 0$ and large fraction of examples with $y = 1$, the update term $\sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}$ will be 0.

For $x_1 = 1$ and $y = 1$, we can infer that $[y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)] = [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1} \geq 0$.

Generally, examples with $x_1 = 1$ and $y = 1$ count a small fraction of training set, which means $\alpha \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}$ will be very small or even 0.

Therefore, w_1 will increase very slowly or even keep constant with iterations.

2.

From $w_1 \leftarrow w_1 + \alpha \left[-\lambda w_1 + \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}\right]$, we can infer that $|(1 - \alpha\lambda)w_j| < |w_j|$, which means the regularization will forces the weights to be smaller with $1 > \alpha\lambda > 0$.

Combine with conclusion drawn in q1, the w_1 will increase slower than the speed in q1. When the iteration is terminated, the final value of w_1 will be smaller than the final value of w_1 in q1.