Q1:

(a) $KL(p\|q_\theta) = \sum_{x=1}^{N} p(x) \log \frac{p(x)}{q_\theta(x)}$

(b) $H(p, q_\theta) = \sum_{x=1}^{N} p(x) \log \frac{1}{q_\theta(x)}$

(c) $KL(p\|q_\theta) = \sum_{x=1}^{N} p(x) \log \frac{p(x)}{q_\theta(x)} = E_p[\log p(x)] - E_p[\log q_\theta(x)] = H(P, Q) - H(P)$

where $H(p) = \sum_{x=1}^{N} p(x) \log \frac{1}{p(x)}$

(d) $l(\theta \mid \mathcal{D}) = \log L(\theta \mid \mathcal{D}) = \log \theta^{m_h}(1-\theta)^{m_t} = m_h \log \theta + m_t \log(1-\theta)$

The pair (m_h, m_t) here is a sufficient statistic.

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

According to this formula,due to the negative $-\frac{1}{N}$ parameter, in order to maximize $l(\theta \mid \mathcal{D})$, we need to minimize $H(p, q_\theta)$ and minimize the $KL(p\|q_\theta)$. It means that for the parameter θ and distribution D, with bigger $l(\theta \mid \mathcal{D})$ has a smaller $KL(p\|q_\theta)$ and $H(p, q_\theta)$.

Q2:

(a) Assume that $y = \omega x + b$, and the formula for the vector form is $y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ where **weights** $\mathbf{w} = (w_0, w_1, \ldots, w_D)^\top$and ω_0 is the bias term.

and $x_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$ therefore $X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$ $\mathbf{X}^\top = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$

According to the lecture, to get compute the ordinary least squares solution,

$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

$A = (\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 2 & -\frac{3}{2} & -\frac{3}{2} \\ \frac{3}{2} & \frac{3}{2} & 1 \\ \frac{3}{2} & 1 & \frac{3}{2} \end{bmatrix}$ $A \times \mathbf{X}^\top = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -1 & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}$ $\hat{\mathbf{w}} = \begin{bmatrix} 2 \\ \frac{3}{2} \\ -\frac{3}{2} \end{bmatrix}$

So, the result is $y = f(x) = \frac{3}{2}x_1 - \frac{3}{2}x_2 + 2$

Q3: In my opinion, 1 matches (c), 2 matches (b), 3 matches (a), 4 matches (d)

The right penalized least squares is $J(w_1, w_0) = \frac{1}{3}\sum_{i=1}^{3}(y_i - w_0 - w_1 x_i) + \lambda w_1^2$

and $\frac{\partial J}{\partial \omega_1} = \frac{2}{3}\omega_1 \sum_{i=1}^{3}(x_i^2) + 2\lambda\omega_1 - \frac{2}{3}\sum_{i=1}^{3}(y_i - \omega_0)x_i$

when $\frac{\partial J}{\partial \omega_1} = 0 \quad \omega_1 = \frac{\sum_{i=1}^{3}(y_i - \omega_0)x_i}{3\lambda + x_1^2 + x_2^2 + x_3^2}$

From the above equation we can conclude that as λ is larger, ω is smaller and the slope is smaller. so we can initially determine that (a) and (c) can correspond to 1, 3, while (b) and (d) can correspond to 2, 4. The variance of the 2 and 4 formulas is greater due to the presence of ω_0^2. This means that the variance of the distance between the point on the axis and the line is greater, so 2 matches (b),4 matches (d), and similarly 1 matches (c), 3 matches (a).

Q4:
According to the lecture, in batch gradient descent algorithm $w_j = w_j + \alpha\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \sigma\left(\mathbf{w}^\top \mathbf{x}_i\right)\right]x_{i,j}$
In this case, let;s turn this formula into vector form
Assume that $z = \omega_0 + \omega_1 x_1 + w_2 x_2$, therefore $\frac{\partial J(\omega)}{\partial \omega_0} = -\frac{1}{4}\sum_{i=1}^{4}[y_i - \sigma(z)]$ $\frac{\partial J(\omega)}{\partial \omega_1} = -\frac{1}{4}\sum_{i=1}^{4}[y_i - \sigma(z)]x_1 i$

$\frac{\partial J(\omega)}{\partial \omega_2} = -\frac{1}{4}\sum_{i=1}^{4}[y_i - \sigma(z)]x_2 i$

So $\omega_0^1 = -2 + 0.1 \times \frac{1}{4}[[0 - \sigma(-2)] \times 1 + [0 - \sigma(-1)] \times 1 + [0 - \sigma(-1)] \times 1 + [1 - \sigma(0)] \times 1] = -2.0039$
Similarly, $\omega_1^1 = 1 + 0.1 \times \frac{1}{4}[[0 - \sigma(-2)] \times 0 + [0 - \sigma(-1)] \times 0 + [0 - \sigma(-1)] \times 1 + [1 - \sigma(0)] \times 1] = 1.006$

$\omega_2^1 = 1 + 0.1 \times \frac{1}{4}[[0 - \sigma(-2)] \times 0 + [0 - \sigma(-1)] \times 1 + [0 - \sigma(-1)] \times 0 + [1 - \sigma(0)] \times 1] = 1.006$

$\omega^{\mathbf{1}} = [-2.0039, 1.006, 1.006]^T$

The first point: (x_1, x_2) = (0,0) $p(y = 1|x, w) = \sigma(-2.0039) = 0.119 < 0.5$ y=0, correct
The second point: (x_1, x_2) = (0,1) $p(y = 1|x, w) = \sigma(-0.973) = 0.274 < 0.5$ y=0, correct
The third point: (x_1, x_2) = (1,0) $p(y = 1|x, w) = \sigma(-0.973) = 0.274 < 0.5$ y=0, correct
The last point: (x_1, x_2) = (1,1) $p(y = 1|x, w) = \sigma(0.0081) = 0.502 > 0.5$ y=1, correct
The training error is 0.

Q5:
1.
The first point: y=1, (x_1, x_2) = (0,0) $p(y = 1|x, w) = \sigma(\omega_0)$ if this training point is correct, then ω_0>0.
The second point: y=0, (x_1, x_2) = (0,1) $p(y = 1|x, w) = \sigma(\omega_0 + \omega_2)$ if this training point is correct, then $\omega_0 + \omega_2$<0.
The third point: y=0, (x_1, x_2) = (1,0) $p(y = 1|x, w) = \sigma(\omega_0 + \omega_1)$ if this training point is correct, then $\omega_0 + \omega_1$<0.
The last point: y=1, (x_1, x_2) = (1,1) $p(y = 1|x, w) = \sigma(\omega_0 + \omega_1 + \omega_2)$ if this training point is correct, then $\omega_0 + \omega_1 + \omega_2$>0.

There is no way to satisfy all four conditions at the same time, at most three conditions at the same time. So the minimum achievable training error in this case is 25%. An example that satisfies the minimum training error rate is $\omega_0 = 1, \omega_1 = -2, \omega_2 = -2$.

2.

The first point: y=1, (x_1, x_2) = (0,0) $p(y = 1|x, w) = \sigma(\omega_0)$ if this training point is correct, then ω_0>0.

The second point: y=0, (x_1, x_2) = (0,1) $p(y = 1|x, w) = \sigma(\omega_0 + \omega_2)$ if this training point is correct, then $\omega_0 + \omega_2$<0.

The third point: y=0, (x_1, x_2) = (1,0) $p(y = 1|x, w) = \sigma(\omega_0 + \omega_1)$ if this training point is correct, then $\omega_0 + \omega_1$<0.

The last point: y=1, (x_1, x_2) = (1,1) $p(y = 1|x, w) = \sigma(\omega_0 + \omega_1 + \omega_2 + \omega_3)$ if this training point is correct, then $\omega_0 + \omega_1 + \omega_2 + \omega_3 > 0$.

It is easy to satisfy the above four conditions simultaneously, and the minimum training error is 0。An example that satisfies the minimum training error rate is $\omega_0 = 1$, $\omega_1 = -2$, $\omega_2 = -2$, $\omega_3 = 4$

Q6

1. According to the question, when x_1 equals is 0, y equals 0, and when x_1 equals 1, y equals 1, And the data set with x_1 equal to 0 accounts for most of the data, which means that for most of the $[y_i - \sigma(\omega^T x_i)]x_1 = 0$. And for the data that $x_1 = 1, y = 1$, 1> $\sigma(\omega^T x_i) = \sigma(\omega_1 x_1 + z)$ > 0.5, where $z = \sum_{i=2}^{N} \omega_i x_i + \omega_0$ and invariant.

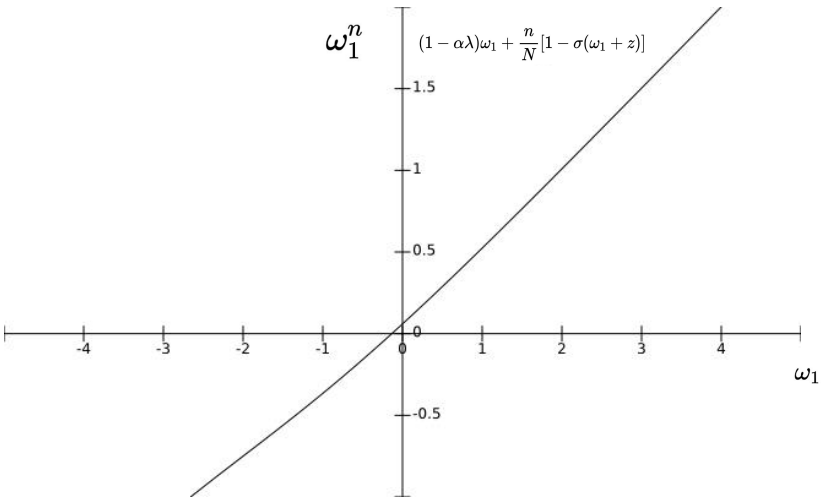
Therefore, $\alpha \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}$ >0(i.e. $\alpha \frac{\partial J(\omega)}{\partial \omega_1} > 0$) and equals to a small value a little larger than 0, ω_1 will increase little by little.

Conclusion：ω_1 will not converge to a point, but increase all the time.

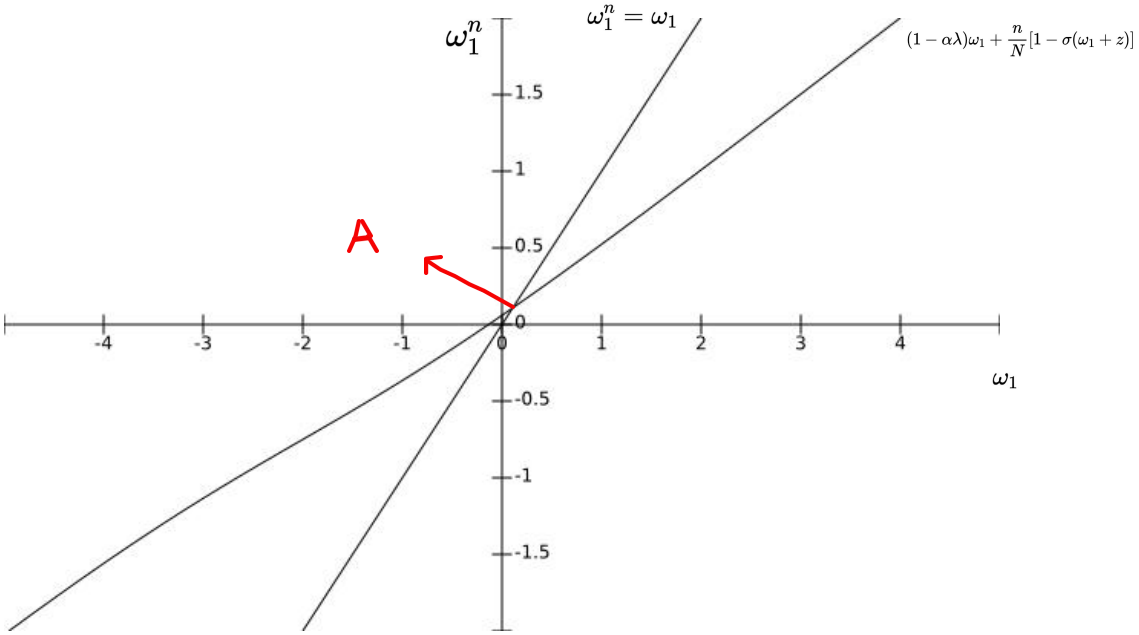
2. when we use the new update rule $w_1 \leftarrow w_1 + \alpha[-\lambda w_1 + \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}]$, we assume that there are a total count of n that $x_1 = 1, y = 1$, so $\frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}$ = $\frac{n}{N}[1 - \sigma(\omega_1 + z)]$ where $z = \sum_{i=2}^{N} \omega_i x_i + \omega_0$, so $-\lambda w_1 + \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}]$ = $\frac{n}{N}[1 - \sigma(\omega_1 + z)] - \lambda \omega_1$

$f(\omega_1)$ = $(1 - \alpha\lambda)\omega_1 + \frac{n}{N}[1 - \sigma(\omega_1 + z)]$

And we can get a figure for ω_1



And than we compare it with $\omega_1^n = \omega_1$



The two lines intersect at the point A(a,a), and we can find that f(ω_1)>ω_1 when ω_1 is initially less than a, which means that this f(ω_1) will keep increasing until f(ω_1) =ω_1, that is, to reach the location of point A.

Similarly, when ω_1 is initially bigger than a, which means that this f(ω_1) will keep decreasing until f(ω_1) =ω_1, that is, to reach the location of point A. Actually, point A（a,a）is value that $\frac{n}{N}[1-\sigma(\omega_1+z)] - \lambda\omega_1 = 0$.

Conclusion: ω_1 will converge to a point where $\omega_1 = \dfrac{\frac{n}{N}[1-\sigma(\omega_1+z)]}{\lambda}$