Q1. (a) Let's get the negative logarithm of the likelihood function

$$E(w, \Sigma) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)^T \Sigma^{-1} (y(x_n, w) - t_n) + \frac{N}{2} \ln|\Sigma| + C \quad (C \text{ is a const})$$

Since Σ is fixed and known, we can get

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)^T \Sigma^{-1} (y(x_n, w) - t_n)$$

Thus, the error function is the MSE

(b) Since Σ is determined from the data,

we get the derivative of Σ

$$\frac{\partial}{\partial \Sigma} E(w, \Sigma) = \frac{N}{2} \Sigma^{-1} + \frac{1}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)' \Sigma^{-2} (y(x_n, w) - t_n)$$

Let $\frac{\partial E}{\partial \Sigma} = 0$, we get

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (y(x_n, w) - t_n)^T (y(x_n - w) - t_n)$$

Q2. Since $\sigma(a) \in [0, 1]$, and the network output $\in [-1, 1]$,

then we can design $h(a) = 2\sigma(a) - 1 \in [-1, 1]$

Since $-1 \leq y(x, w) \leq 1$, $C \in (-1, 1)$

then we have

$$E(w) = -\sum_{n=1}^{N} \left(\frac{1+t_n}{2} \ln \frac{1+y_n}{2} + \frac{1-t_n}{2} \ln \frac{1-y_n}{2} \right)$$

$$= -\frac{1}{2} \sum_{n=1}^{N} ((1+t_n) \ln(1+y_n) + (1-t_n) \cdot \ln(1-y_n)) + N \cdot \ln 2$$

Q3. (a) $E(t) = \int t \cdot N(t \mid \mu, \sigma^2 I)\, dt = \mu$

$E(\|t\|^2) = \int \|t\|^2 N(t \mid \mu, \sigma^2 I)\, dt = L\sigma^2 + \|\mu\|^2$

Here L is the dimension

Thus, $E(t \mid x) = \int t\, p(t \mid x)\, dt$

$$= \int t \cdot \sum_{k=1}^{K} \pi_k N(t \mid \mu_k, \sigma^2)\, dt$$

$$= \sum_{k=1}^{K} \pi_k \int t\, N(t \mid \mu_k, \sigma^2)\, dt$$

$$= \sum_{k=1}^{K} \pi_k \cdot \mu_k$$

$$= \sum_{k=1}^{K} \pi_k(x)\, \mu_k(x)$$

(b). $s^2(x) = E[\|t - E(t \mid x)\|^2 \mid x]$

$$= E[(t^2 - 2t \cdot E(t \mid x) + E[t \mid x]^2) \mid x]$$

$$= E[t^2 \mid x] - E[2t E(t \mid x) \mid x] + E[t \mid x]^2$$

$$= E[t^2 \mid x] - E[t \mid x]^2$$

$$= \int \|t\|^2 \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)\, dt - \|\sum_{l=1}^{K} \pi_l \cdot \mu_l\|$$

$$= \sum_{k=1}^{K} \pi_k (L\sigma_k^2 + \|\mu_k\|^2) - \|\sum_{l=1}^{K} \pi_l \cdot \mu_l\|$$

$$= L \cdot \sum_{k=1}^{K} \pi_k \cdot \sigma_k^2 + \sum_{k=1}^{K} \pi_k \|\mu_k\|^2$$

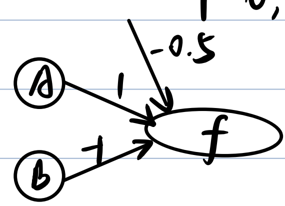
$$- 2\left(\sum_{l=1}^{K} \pi_l \mu_l\right) \cdot \left(\sum_{k=1}^{K} \pi_k \mu_k\right) + \sum_{k=1}^{K} \pi_k \|\sum_{l=1}^{K} \pi_l \mu_l\|^2$$

$$= L \cdot \sum_{k=1}^{K} \pi_k \cdot \sigma_k^2 + \sum_{k=1}^{K} \pi_k \|\mu_k - \sum_{l=1}^{K} \pi_l \cdot \mu_l\|^2$$

$$= \sum_{k=1}^{K} \pi_k \left(L\sigma_k^2 + \|\mu_k - \sum_{l=1}^{K} \pi_l \cdot \mu_l\|^2\right)$$

Q4. Yes, since it is linearly separable.

$$f(A, B) = \begin{cases} 1, & A - B - 0.5 \geq 0 \\ 0, & A - B - 0.5 \leq 0. \end{cases}$$



Q5. (a) $3 \times 4 \times 4 \times 1 = 48$

(b) $3 \times 5 \times 5 = 75$

(c) $N_{conv} = 3 \times 4 \times 4 \times 1 = 48$

$N_{pool} = 0$

$N_{ReLU} = 0$

$N_{FCL} = 3 \times 5 \times 5 \times 4 = 300$

$N = 48 + 0 + 0 + 300 = 348$

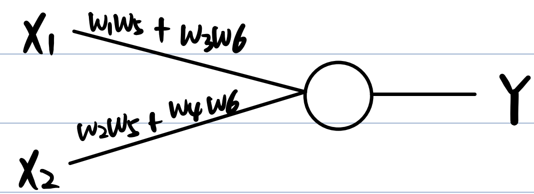
(d) True

(e) There are too many weights to learn.

Q6. (a) Let's define $W_1 = \begin{bmatrix} w_1 & w_3 \\ w_2 & w_4 \end{bmatrix}$ $W_2 = \begin{bmatrix} w_5 \\ w_6 \end{bmatrix}$ $A = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

then $Y = C \cdot W_2^T W_1^T \cdot A = C[w_1w_5 + w_3w_6,\ w_2w_5 + w_4w_6] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Thus, we get $W = \begin{bmatrix} w_1w_5 + w_3w_6 \\ w_2w_5 + w_4w_6 \end{bmatrix}$ $A = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$



(b) Yes, the entire network can be seen to perform a chain of matrix multiplication.

(c) Let $w_1 = w_3 = -5$, $w_2 = w_4 = -10$, $w_5 = 5$, $w_6 = -6$ then we have the whole network working like XOR