Machine Learning Techniques - Homework 2

資工四 B05902023 李澤諺

Descent Methods for Probabilistic SVM

1. 令

$$s_n = -y_n(Az_n + B)$$

因此

$$\begin{aligned} z_n &= \mathbf{w}_{SVM}^T \boldsymbol{\phi}(\mathbf{x}_n) + b_{SVM} \\ s_n &= -y_n(Az_n + B) \\ p_n &= \theta(s_n) \end{aligned}$$

以及

$$F(A, B) = \frac{1}{N} \sum_{n=1}^{N} ln\left(1 + exp\left(s_n\right)\right)$$

故

$$\begin{aligned} \frac{\partial F}{\partial A} &= \frac{\partial}{\partial A} \left(\frac{1}{N} \sum_{n=1}^{N} ln(1 + exp(s_n)) \right) \\ &= \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial}{\partial A} ln(1 + exp(s_n)) \right) \\ &= \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{1 + exp(s_n)} \cdot \frac{\partial}{\partial A} (1 + exp(s_n)) \right) \\ &= \frac{1}{N} \sum_{n=1}^{N} \left(\frac{exp(s_n)}{1 + exp(s_n)} \cdot \frac{\partial}{\partial A} s_n \right) \\ &= \frac{1}{N} \sum_{n=1}^{N} \left(\theta(s_n) \cdot \frac{\partial}{\partial A} (-y_n(Az_n + B)) \right) \\ &= \frac{1}{N} \sum_{n=1}^{N} (p_n \cdot (-y_n z_n)) = -\frac{1}{N} \sum_{n=1}^{N} y_n z_n p_n \end{aligned}$$

$$
\begin{aligned}
\frac{\partial F}{\partial B} &= \frac{\partial}{\partial B}\left(\frac{1}{N}\sum_{n=1}^{N} ln(1+exp(s_n))\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial}{\partial B} ln(1+exp(s_n))\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{1+exp(s_n)}\cdot\frac{\partial}{\partial B}(1+exp(s_n))\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{exp(s_n)}{1+exp(s_n)}\cdot\frac{\partial}{\partial B}s_n\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\theta(s_n)\cdot\frac{\partial}{\partial B}(-y_n(Az_n+B))\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}(p_n\cdot(-y_n)) = -\frac{1}{N}\sum_{n=1}^{N} y_n p_n
\end{aligned}
$$

因此可得

$$
\nabla F = \begin{pmatrix} \frac{\partial F}{\partial A} \\ \frac{\partial F}{\partial B} \end{pmatrix} = \begin{pmatrix} -\frac{1}{N}\sum_{n=1}^{N} y_n z_n p_n \\ -\frac{1}{N}\sum_{n=1}^{N} y_n p_n \end{pmatrix}
$$

2. 因爲

$$
\begin{aligned}
\theta'(s) = \left(\frac{exp(s)}{1+exp(s)}\right)' &= \frac{(exp(s))'\cdot(1+exp(s)) - exp(s)\cdot(1+exp(s))'}{(1+exp(s))^2} \\
&= \frac{exp(s)\cdot(1+exp(s)) - exp(s)\cdot exp(s)}{(1+exp(s))^2} \\
&= \frac{exp(s)}{(1+exp(s))^2} = \frac{exp(s)}{1+exp(s)}\left(1-\frac{exp(s)}{1+exp(s)}\right) \\
&= \theta(s)(1-\theta(s))
\end{aligned}
$$

所以

$$
\begin{aligned}
\frac{\partial^2 F}{\partial A^2} &= \frac{\partial}{\partial A}\left(-\frac{1}{N}\sum_{n=1}^{N} y_n z_n p_n\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n\cdot\frac{\partial}{\partial A}p_n\right) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n\cdot\frac{\partial}{\partial A}\theta(s_n)\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n \theta'(s_n)\cdot\frac{\partial}{\partial A}s_n\right) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n \theta(s_n)(1-\theta(s_n))\cdot\frac{\partial}{\partial A}(-y_n(Az_n+B))\right) \\
&= -\frac{1}{N}\sum_{n=1}^{N}(y_n z_n p_n(1-p_n)\cdot(-y_n z_n)) = \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n^2 p_n(1-p_n)
\end{aligned}
$$

$$\begin{aligned}\frac{\partial^2 F}{\partial A \partial B} &= \frac{\partial}{\partial A}\left(-\frac{1}{N}\sum_{n=1}^{N} y_n p_n\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \cdot \frac{\partial}{\partial A} p_n\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \cdot \frac{\partial}{\partial A}\theta(s_n)\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \theta'(s_n) \cdot \frac{\partial}{\partial A} s_n\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \theta(s_n)(1-\theta(s_n)) \cdot \frac{\partial}{\partial A}(-y_n(Az_n + B))\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}(y_n p_n(1-p_n) \cdot (-y_n z_n)) = \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n p_n(1-p_n)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 F}{\partial B \partial A} &= \frac{\partial}{\partial B}\left(-\frac{1}{N}\sum_{n=1}^{N} y_n z_n p_n\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n \cdot \frac{\partial}{\partial B} p_n\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n \cdot \frac{\partial}{\partial B}\theta(s_n)\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n \theta'(s_n) \cdot \frac{\partial}{\partial B} s_n\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n z_n \theta(s_n)(1-\theta(s_n)) \cdot \frac{\partial}{\partial B}(-y_n(Az_n + B))\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}(y_n z_n p_n(1-p_n) \cdot (-y_n)) = \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n p_n(1-p_n)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 F}{\partial B^2} &= \frac{\partial}{\partial B}\left(-\frac{1}{N}\sum_{n=1}^{N} y_n p_n\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \cdot \frac{\partial}{\partial B} p_n\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \cdot \frac{\partial}{\partial B}\theta(s_n)\right) = -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \theta'(s_n) \cdot \frac{\partial}{\partial B} s_n\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\left(y_n \theta(s_n)(1-\theta(s_n)) \cdot \frac{\partial}{\partial B}(-y_n(Az_n + B))\right) \\ &= -\frac{1}{N}\sum_{n=1}^{N}(y_n p_n(1-p_n) \cdot (-y_n)) = \frac{1}{N}\sum_{n=1}^{N} y_n^2 p_n(1-p_n)\end{aligned}$$

因此可得

$$\begin{aligned}H(F) &= \begin{pmatrix} \frac{\partial^2 F}{\partial A^2} & \frac{\partial^2 F}{\partial A \partial B} \\ \frac{\partial^2 F}{\partial B \partial A} & \frac{\partial^2 F}{\partial B^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n^2 p_n(1-p_n) & \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n p_n(1-p_n) \\ \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n p_n(1-p_n) & \frac{1}{N}\sum_{n=1}^{N} y_n^2 p_n(1-p_n) \end{pmatrix} \\ &= \frac{1}{N}\sum_{n=1}^{N} y_n^2 p_n(1-p_n) \begin{pmatrix} z_n^2 & z_n \\ z_n & 1 \end{pmatrix}\end{aligned}$$

3. 因爲 $\forall\ \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$ 且 $\mathbf{x} \neq \mathbf{0}$，皆有

$$\begin{aligned}
\mathbf{x}^T H(F)\ \mathbf{x} &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \left(\frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n (1-p_n) \begin{pmatrix} z_n^2 & z_n \\ z_n & 1 \end{pmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n (1-p_n) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} z_n^2 & z_n \\ z_n & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n (1-p_n) \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} z_n^2 & z_n \\ z_n & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n (1-p_n) \begin{pmatrix} z_n^2 x_1 + z_n x_2 & z_n x_1 + x_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n (1-p_n)(z_n^2 x_1^2 + 2 z_n x_1 x_2 + x_2^2) \\
&= \frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n (1-p_n)(z_n x_1 + x_2)^2
\end{aligned}$$

其中 $\frac{1}{N} \geq 0$、$y_n^2 \geq 0$、$(z_n x_1 + x_2)^2 \geq 0$，並且

$$p_n = \theta(s_n) = \frac{exp(s_n)}{1 + exp(s_n)} \geq 0$$

$$1 - p_n = 1 - \frac{exp(s_n)}{1 + exp(s_n)} = \frac{1}{1 + exp(s_n)} \geq 0$$

因此可得

$$\mathbf{x}^T H(F)\ \mathbf{x} = \frac{1}{N} \sum_{n=1}^{N} y_n^2 p_n (1-p_n)(z_n x_1 + x_2)^2 \geq 0$$

故 $H(F)$ 爲 positive semi-definite。

Neural Network

4. 取 $w_1 = w_2 = \cdots = w_d = 1$，$w_0 = d - 1$，即

$$g_A(\mathbf{x}) = sign(x_1 + x_2 + \cdots + x_d + d - 1)$$

則當 x_1、x_2、$\cdots$、x_d 皆爲 -1 時，可得

$$x_1 + x_2 + \cdots + x_d + d - 1 = d \times (-1) + d - 1 = -1$$

因此 $g_A(\mathbf{x}) = sign(x_1 + x_2 + \cdots + x_d + d - 1) = -1$，而當 x_1、x_2、$\cdots$、x_d 之中至少有一者爲 1 時，可得

$$x_1 + x_2 + \cdots + x_d + d - 1 \geq (d-1) \times (-1) + 1 \times 1 + d - 1 = 1$$

因此 $g_A(\mathbf{x}) = sign(x_1 + x_2 + \cdots + x_d + d - 1) = 1$，由此可知
$g_A(\mathbf{x}) = sign(x_1 + x_2 + \cdots + x_d + d - 1) = OR(x_1, x_2, \cdots, x_d)$。

5. 因爲

$$\frac{\partial e_n}{\partial w_{ij}^{(l)}} = x_i^{(l-1)} \delta_j^{(l)}$$

其中，當 $2 \leq l \leq L$ 時，因爲

$$s_i^{(l-1)} = \sum_{k=0}^{d^{(l-2)}} w_{ki}^{(l-1)} x_k^{(l-2)} = \sum_{k=0}^{d^{(l-2)}} 0 \cdot x_k^{(l-2)} = 0$$

所以

$$x_i^{(l-1)} = tanh(s_i^{(l-1)}) = tanh 0 = 0$$

此外，當 $1 \leq l \leq L-1$ 時

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)} w_{jk}^{(l+1)} tanh'(s_j^{(l)}) = \sum_k \delta_k^{(l+1)} \cdot 0 \cdot tanh'(s_j^{(l)}) = 0$$

因此可得 $\forall\ l \in \{1, 2, \cdots, L\}$，皆有

$$\frac{\partial e_n}{\partial w_{ij}^{(l)}} = 0$$

意即，loss function 對整個 Neural Network 的 gradient 爲 **0**。

6. 首先，說明當 $L \geq 4$，意即 hidden layer 至少有 3 層時，可以找到一個 layer 更少的 neural network，其 weight 的數量會比原本的 neural network 更多。考慮 **Figure. 1** 中的 neural network，設其中 $L \geq 4$

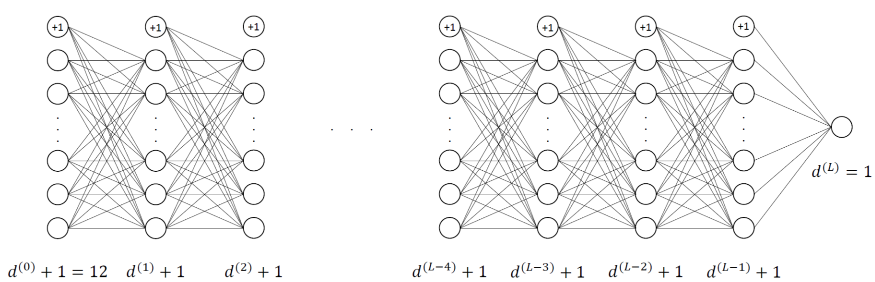


Figure. 1

若將該 neural network 中倒數第 2 層的 neuron 全部移除，意即，移除 **Figure. 2** 中紅色的 neuron

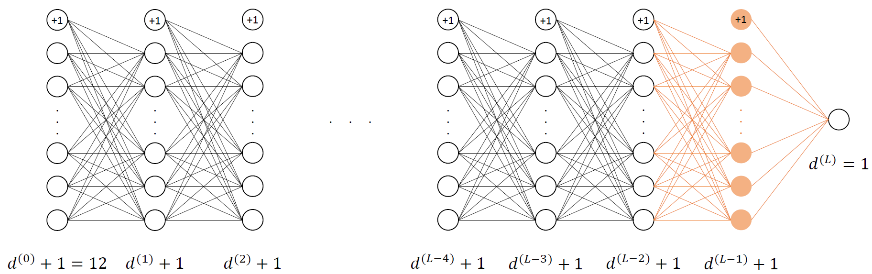


Figure. 2

則減少的 weight 的數量，即 **Figure. 2** 中紅色線段的數量爲

$$\begin{aligned} n^- &= (d^{(L-2)}+1)d^{(L-1)} + (d^{(L-1)}+1) \\ &= d^{(L-1)}d^{(L-2)} + 2d^{(L-1)} + 1 \end{aligned}$$

若將這些被移除的 $d^{(L-1)}+1$ 個 neuron，其中 $d^{(L-1)}$ 個 neuron 加入倒數第 4 層中，剩下的 1 個 neuron 加入倒數第 3 層中，如 **Figure. 3** 中藍色的 neuron 所示

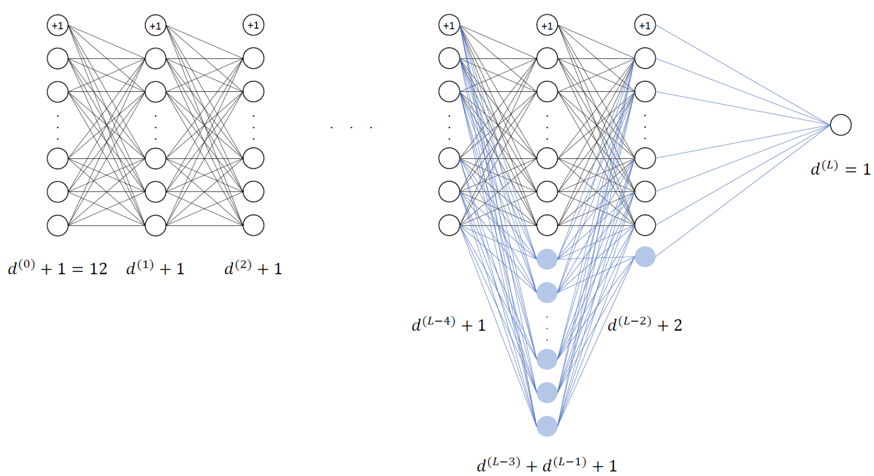


Figure. 3

則增加的 weight 的數量，即 **Figure. 3** 中藍色線段的數量爲

$$\begin{aligned} n^+ &= (d^{(L-4)}+1)d^{(L-1)} + d^{(L-1)}(d^{(L-2)}+1) + (d^{(L-2)}+2) \\ &= d^{(L-1)}d^{(L-2)} + 2d^{(L-1)} + d^{(L-1)}d^{(L-4)} + d^{(L-2)} + 2 \end{aligned}$$

因爲

$$\begin{aligned} n^+ &= d^{(L-1)}d^{(L-2)} + 2d^{(L-1)} + d^{(L-1)}d^{(L-4)} + d^{(L-2)} + 2 \\ &= n^- + d^{(L-1)}d^{(L-4)} + d^{(L-2)} + 1 > n^- \end{aligned}$$

因此，依照以上方式將倒數第 2 層的 neuron 全部移到倒數第 3 層和倒數第 4 層之後，可以得到一個 layer 更少但 weight 數量更多的 neural network，由此可知，當 $L \geq 4$，意即 hidden layer 至少有 3 層時，weight 的數量不可能有最大值，因此，只需考慮 hidden layer 爲 1 層或 2 層的 neural network 即可。當 $L = 2$，意即 hidden layer 爲 1 層時，weight 的數量爲

$$(d^{(0)} + 1)d^{(1)} + (d^{(1)} + 1) = 12 \times 47 + 48 = 612$$

而當 $L = 3$，意即 hidden layer 爲 2 層時，因爲

$$\begin{aligned} (d^{(1)} + 1) + (d^{(2)} + 1) &= 48 \\ d^{(2)} &= 46 - d^{(1)} \end{aligned}$$

所以 weight 的數量爲

$$\begin{aligned} &(d^{(0)} + 1)d^{(1)} + (d^{(1)} + 1)d^{(2)} + (d^{(2)} + 1) \\ =& 12d^{(1)} + (d^{(1)} + 1)(46 - d^{(1)}) + (47 - d^{(1)}) \\ =& -(d^{(1)})^2 + 56d^{(1)} + 93 \\ =& -(d^{(1)} - 28)^2 + 877 \end{aligned}$$

因此，當 $d^{(1)} = 28$，$d^{(2)} = 18$ 時，weight 的數量有最大值 877。綜合以上所述，可得當 $L = 3$，意即 hidden layer 爲 2 層，且當 $d^{(1)} = 28$，$d^{(2)} = 18$ 時，weight 的數量有最大值 877。

Autoencoder

7. 因爲

$$\begin{aligned} err_n(\mathbf{w}) &= \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 \\ &= (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) \\ &= (\mathbf{x}_n^T - \mathbf{x}_n^T\mathbf{w}\mathbf{w}^T)(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) \\ &= \mathbf{x}_n^T\mathbf{x}_n - 2\mathbf{x}_n^T\mathbf{w}\mathbf{w}^T\mathbf{x}_n + \mathbf{x}_n^T\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T\mathbf{x}_n \\ &= \mathbf{x}_n^T\mathbf{x}_n - 2(\mathbf{x}_n^T\mathbf{w})(\mathbf{w}^T\mathbf{x}_n) + (\mathbf{x}_n^T\mathbf{w})(\mathbf{w}^T\mathbf{w})(\mathbf{w}^T\mathbf{x}_n) \\ &= \mathbf{x}_n^T\mathbf{x}_n - 2(\mathbf{w}^T\mathbf{x}_n)^2 + (\mathbf{w}^T\mathbf{w})(\mathbf{w}^T\mathbf{x}_n)^2 \end{aligned}$$

所以

$$\begin{aligned} \nabla_{\mathbf{w}} err_n(\mathbf{w}) &= \frac{\partial err_n(\mathbf{w})}{\partial \mathbf{w}} \\ &= \frac{\partial}{\partial \mathbf{w}}\left(\mathbf{x}_n^T\mathbf{x}_n - 2(\mathbf{w}^T\mathbf{x}_n)^2 + (\mathbf{w}^T\mathbf{w})(\mathbf{w}^T\mathbf{x}_n)^2\right) \\ &= \frac{\partial}{\partial \mathbf{w}}(\mathbf{x}_n^T\mathbf{x}_n) - 2 \cdot \frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{x}_n)^2 + \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{w}^T\mathbf{w})(\mathbf{w}^T\mathbf{x}_n)^2\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{0} - 4(\mathbf{w}^T\mathbf{x}_n)\cdot\frac{\partial}{\partial\mathbf{w}}(\mathbf{w}^T\mathbf{x}_n) + (\mathbf{w}^T\mathbf{x}_n)^2\cdot\frac{\partial}{\partial\mathbf{w}}(\mathbf{w}^T\mathbf{w}) + \\
&\quad (\mathbf{w}^T\mathbf{w})\cdot\frac{\partial}{\partial\mathbf{w}}(\mathbf{w}^T\mathbf{x}_n)^2 \\
&= -4(\mathbf{w}^T\mathbf{x}_n)\cdot\mathbf{x}_n + (\mathbf{w}^T\mathbf{x}_n)^2\cdot 2\mathbf{w} + (\mathbf{w}^T\mathbf{w})\cdot 2(\mathbf{w}^T\mathbf{x}_n)\cdot \\
&\quad \frac{\partial}{\partial\mathbf{w}}(\mathbf{w}^T\mathbf{x}_n) \\
&= -4(\mathbf{w}^T\mathbf{x}_n)\cdot\mathbf{x}_n + (\mathbf{w}^T\mathbf{x}_n)^2\cdot 2\mathbf{w} + (\mathbf{w}^T\mathbf{w})\cdot 2(\mathbf{w}^T\mathbf{x}_n)\cdot\mathbf{x}_n \\
&= -4(\mathbf{w}^T\mathbf{x}_n)\mathbf{x}_n + 2(\mathbf{w}^T\mathbf{x}_n)^2\mathbf{w} + 2(\mathbf{w}^T\mathbf{w})(\mathbf{w}^T\mathbf{x}_n)\mathbf{x}_n
\end{aligned}
$$

8. 首先，證明在第 8 題的過程之中會用到的一個 property。

Property $\forall\ \mathbf{u}$、$\mathbf{v} \in \mathbb{R}^n$，$\mathbf{u}^T\mathbf{v} = trace(\mathbf{u}\mathbf{v}^T)$。

Proof 因爲

$$
\mathbf{u}\mathbf{v}^T = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}^T = \begin{pmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_n \\ u_2v_1 & u_2v_2 & \cdots & u_2v_n \\ \vdots & \vdots & & \vdots \\ u_nv_1 & u_nv_2 & \cdots & u_nv_n \end{pmatrix}
$$

所以

$$
trace(\mathbf{u}\mathbf{v}^T) = u_1v_1 + u_2v_2 + \cdots + u_nv_n = \mathbf{u}^T\mathbf{v}
$$

以下開始說明第 8 題，因爲

$$
\begin{aligned}
E_{in}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n + \boldsymbol{\epsilon}_n)\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\|(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) - \mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}((\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) - \mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n)^T((\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) - \mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n) \\
&= \frac{1}{N}\sum_{n=1}^{N}((\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T - (\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n)^T)((\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) - \mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n) \\
&= \frac{1}{N}\sum_{n=1}^{N}((\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) - (\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n) - \\
&\qquad (\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n)^T(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n) + (\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n)^T(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n))
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{n=1}^{N}(\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 - 2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n) + \\
&\qquad trace((\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n)(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n)^T)) \\
&= \frac{1}{N}\sum_{n=1}^{N}(\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 - 2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T)\boldsymbol{\epsilon}_n + \\
&\qquad trace(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T\mathbf{w}\mathbf{w}^T))
\end{aligned}
$$

所以

$$
\begin{aligned}
\mathbb{E}\left[E_{in}(\mathbf{w})\right] &= \mathbb{E}\Bigg[\frac{1}{N}\sum_{n=1}^{N}(\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 - 2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T)\boldsymbol{\epsilon}_n + \\
&\qquad trace(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T\mathbf{w}\mathbf{w}^T))\Bigg] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 - 2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T)\boldsymbol{\epsilon}_n + \right. \\
&\qquad \left. trace(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T\mathbf{w}\mathbf{w}^T)\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}\left[\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2\right] - \mathbb{E}\left[2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T)\boldsymbol{\epsilon}_n\right] + \right. \\
&\qquad \left. \mathbb{E}\left[trace(\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T\mathbf{w}\mathbf{w}^T)\right]\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 - 2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T)\mathbb{E}\left[\boldsymbol{\epsilon}_n\right] + \right. \\
&\qquad \left. trace\left(\mathbb{E}\left[\mathbf{w}\mathbf{w}^T\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T\mathbf{w}\mathbf{w}^T\right]\right)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 - 2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T)\mathbb{E}\left[\boldsymbol{\epsilon}_n\right] + \right. \\
&\qquad \left. trace\left(\mathbf{w}\mathbf{w}^T\mathbb{E}\left[\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n^T\right]\mathbf{w}\mathbf{w}^T\right)\right)
\end{aligned}
$$

其中，因爲 $\boldsymbol{\epsilon}_n$ 是 i.i.d 從 zero mean 且 unit variance 的 Gaussian distribution 所產生，因此 $\mathbb{E}\left[\boldsymbol{\epsilon}_n\right] = \mathbf{0}$、$\mathbb{E}\left[\boldsymbol{\epsilon}_n^T\boldsymbol{\epsilon}_n\right] = I_d$，故

$$
\begin{aligned}
\mathbb{E}\left[E_{in}(\mathbf{w})\right] &= \frac{1}{N}\sum_{n=1}^{N}\left(\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 - 2(\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n)^T(\mathbf{w}\mathbf{w}^T)\ \mathbf{0}\ + \right. \\
&\qquad \left. trace\left(\mathbf{w}\mathbf{w}^T I_d\mathbf{w}\mathbf{w}^T\right)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 + trace\left(\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T\right)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 + trace(\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T)
\end{aligned}
$$

因此可得

$$
\begin{aligned}
\Omega(\mathbf{w}) &= trace(\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T) = trace((\mathbf{w}\mathbf{w}^T\mathbf{w})\mathbf{w}^T) \\
&= (\mathbf{w}\mathbf{w}^T\mathbf{w})^T\mathbf{w} = \mathbf{w}^T\mathbf{w}\mathbf{w}^T\mathbf{w} = (\mathbf{w}^T\mathbf{w})^2
\end{aligned}
$$

9. $\mathbf{x}_n$ 經過 encode 之後會變爲

$$
\mathbf{x}_n^{(1)} = \begin{pmatrix} tanh\left(\sum_{p=1}^{d} w_{p1}^{(1)} x_{np}\right) \\ tanh\left(\sum_{p=1}^{d} w_{p2}^{(1)} x_{np}\right) \\ \vdots \\ tanh\left(\sum_{p=1}^{d} w_{p\tilde{d}}^{(1)} x_{np}\right) \end{pmatrix}
$$

$\mathbf{x}_n^{(1)}$ 經過 decode 之後會變爲

$$
\mathbf{x}_n^{(2)} = \begin{pmatrix} \sum_{q=1}^{\tilde{d}} w_{q1}^{(2)} x_{nq}^{(1)} \\ \sum_{q=1}^{\tilde{d}} w_{q2}^{(2)} x_{nq}^{(1)} \\ \vdots \\ \sum_{q=1}^{\tilde{d}} w_{qd}^{(2)} x_{nq}^{(1)} \end{pmatrix} = \begin{pmatrix} \sum_{q=1}^{\tilde{d}} w_{q1}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \\ \sum_{q=1}^{\tilde{d}} w_{q2}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \\ \vdots \\ \sum_{q=1}^{\tilde{d}} w_{qd}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \end{pmatrix}
$$

因此 error function 爲

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{x}_n^{(2)}\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} \left\| \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nd} \end{pmatrix} - \begin{pmatrix} \sum_{q=1}^{\tilde{d}} w_{q1}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \\ \sum_{q=1}^{\tilde{d}} w_{q2}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \\ \vdots \\ \sum_{q=1}^{\tilde{d}} w_{qd}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \end{pmatrix} \right\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} \left\| \begin{pmatrix} x_{n1} - \sum_{q=1}^{\tilde{d}} w_{q1}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \\ x_{n2} - \sum_{q=1}^{\tilde{d}} w_{q2}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \\ \vdots \\ x_{nd} - \sum_{q=1}^{\tilde{d}} w_{qd}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right) \end{pmatrix} \right\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{d} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right)\right)^2
\end{aligned}
$$

若 $u_{ij} = w_{ij}^{(1)} = w_{ji}^{(2)}$，則此時 error function 爲

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{d}\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2$$

10. 由第 9 題可知

$$E_9 = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{d}\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2$$

$$E_{10} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{d}\left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh\left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np}\right)\right)^2$$

所以

$$\begin{aligned}
\frac{\partial E_9}{\partial u_{ij}} &= \frac{\partial}{\partial u_{ij}}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{d}\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial}{\partial u_{ij}}\sum_{k=1}^{d}\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial}{\partial u_{ij}}\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2 + \right. \\
&\qquad \left. \frac{\partial}{\partial u_{ij}}\sum_{k\neq i}\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2\right)
\end{aligned}$$

其中

$$\begin{aligned}
&\frac{\partial}{\partial u_{ij}}\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2 \\
&= 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)\cdot \\
&\quad \frac{\partial}{\partial u_{ij}}\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \\
&\quad \frac{\partial}{\partial u_{ij}} \left(x_{ni} - u_{ij} tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right) - \sum_{q \neq j} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \\
&= 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \\
&\quad \left(\frac{\partial}{\partial u_{ij}} x_{ni} - \frac{\partial}{\partial u_{ij}} u_{ij} tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right) - \frac{\partial}{\partial u_{ij}} \sum_{q \neq j} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \\
&= 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \\
&\quad \left(0 - \left(\frac{\partial}{\partial u_{ij}} u_{ij} \cdot tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right) + u_{ij} \cdot \frac{\partial}{\partial u_{ij}} tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right) - 0\right) \\
&= 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \\
&\quad \left(-1 \cdot tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right) - u_{ij} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right) \cdot \frac{\partial}{\partial u_{ij}} \sum_{p=1}^{d} u_{pj} x_{np}\right) \\
&= 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \\
&\quad \left(-1 \cdot tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right) - u_{ij} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right) \cdot x_{ni}\right) \\
&= 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \left(-tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right) + \\
&\quad 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \left(-u_{ij} x_{ni} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right)
\end{aligned}
$$

並且

$$
\frac{\partial}{\partial u_{ij}} \sum_{k \neq i} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)^2
$$

$$
\begin{aligned}
&= \sum_{k \neq i} \frac{\partial}{\partial u_{ij}} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right)^2 \\
&= \sum_{k \neq i} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \frac{\partial}{\partial u_{ij}} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \right) \\
&= \sum_{k \neq i} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(\frac{\partial}{\partial u_{ij}} x_{nk} - \frac{\partial}{\partial u_{ij}} \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \right) \\
&= \sum_{k \neq i} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(0 - \frac{\partial}{\partial u_{ij}} u_{kj} tanh \left(\sum_{p=1}^{d} u_{pj} x_{np} \right) \right) \right) \\
&= \sum_{k \neq i} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(-u_{kj} tanh' \left(\sum_{p=1}^{d} u_{pj} x_{np} \right) \cdot \frac{\partial}{\partial u_{ij}} \sum_{p=1}^{d} u_{pj} x_{np} \right) \right) \\
&= \sum_{k \neq i} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(-u_{kj} tanh' \left(\sum_{p=1}^{d} u_{pj} x_{np} \right) \cdot x_{ni} \right) \right) \\
&= \sum_{k \neq i} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(-u_{kj} x_{ni} tanh' \left(\sum_{p=1}^{d} u_{pj} x_{np} \right) \right) \right)
\end{aligned}
$$

因此

$$
\begin{aligned}
\frac{\partial E_9}{\partial u_{ij}} = \frac{1}{N}\sum_{n=1}^{N} & \left(2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \right. \\
& \left(-tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right) + \\
& 2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \\
& \left(-u_{ij} x_{ni} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right) + \\
& \sum_{k \neq i}\left(2\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \right. \\
& \left.\left.\left(-u_{kj} x_{ni} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right)\right)\right) \\
= \frac{1}{N}\sum_{n=1}^{N} & \left(2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \right. \\
& \left(-tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right) + \\
& \sum_{k=1}^{d}\left(2\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \right. \\
& \left.\left.\left(-u_{kj} x_{ni} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right)\right)\right) \\
= \frac{1}{N}\sum_{n=1}^{N} & \left(2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \right. \\
& \left.\left(-tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right)\right) + \\
\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{d} & \left(2\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right) \cdot \right. \\
& \left.\left(-u_{kj} x_{ni} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right)\right)
\end{aligned}
$$

接著，因爲

$$
\begin{aligned}
\frac{\partial E_{10}}{\partial w_{ij}^{(1)}} &= \frac{\partial}{\partial w_{ij}^{(1)}} \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right)^2 \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \frac{\partial}{\partial w_{ij}^{(1)}} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right)^2 \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad\qquad \left. \frac{\partial}{\partial w_{ij}^{(1)}} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad\qquad \left. \left(\frac{\partial}{\partial w_{ij}^{(1)}} x_{nk} - \frac{\partial}{\partial w_{ij}^{(1)}} \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad\qquad \left. \left(0 - \frac{\partial}{\partial w_{ij}^{(1)}} w_{jk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pj}^{(1)} x_{np} \right) \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad\qquad \left. \left(-w_{jk}^{(2)} tanh' \left(\sum_{p=1}^{d} w_{pj}^{(1)} x_{np} \right) \cdot \frac{\partial}{\partial w_{ij}^{(1)}} \sum_{p=1}^{d} w_{pj}^{(1)} x_{np} \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad\qquad \left. \left(-w_{jk}^{(2)} tanh' \left(\sum_{p=1}^{d} w_{pj}^{(1)} x_{np} \right) \cdot x_{ni} \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(2 \left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad\qquad \left. \left(-u_{kj} x_{ni} tanh' \left(\sum_{p=1}^{d} u_{pj} x_{np} \right) \right) \right)
\end{aligned}
$$

以及

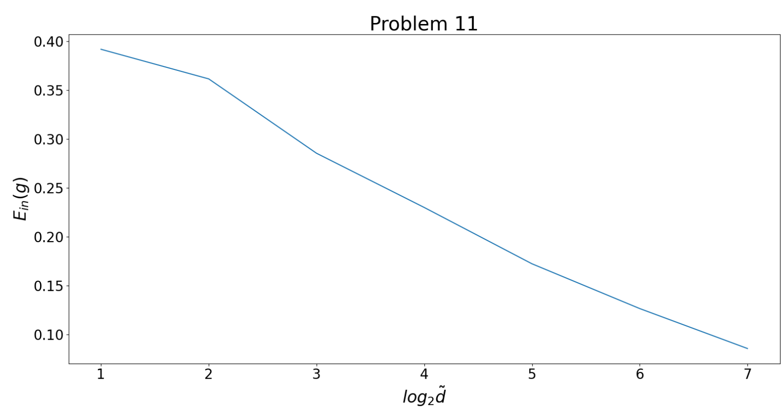
$$
\begin{aligned}
\frac{\partial E_{10}}{\partial w_{ji}^{(2)}} &= \frac{\partial}{\partial w_{ji}^{(2)}} \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right)^2 \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial}{\partial w_{ji}^{(2)}} \sum_{k=1}^{d} \left(x_{nk} - \sum_{q=1}^{\tilde{d}} w_{qk}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right)^2 \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial}{\partial w_{ji}^{(2)}} \left(x_{ni} - \sum_{q=1}^{\tilde{d}} w_{qi}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right)^2 \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(2 \left(x_{ni} - \sum_{q=1}^{\tilde{d}} w_{qi}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \frac{\partial}{\partial w_{ji}^{(2)}} \left(x_{ni} - \sum_{q=1}^{\tilde{d}} w_{qi}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(2 \left(x_{ni} - \sum_{q=1}^{\tilde{d}} w_{qi}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(\frac{\partial}{\partial w_{ji}^{(2)}} x_{ni} - \frac{\partial}{\partial w_{ji}^{(2)}} \sum_{q=1}^{\tilde{d}} w_{qi}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(2 \left(x_{ni} - \sum_{q=1}^{\tilde{d}} w_{qi}^{(2)} tanh \left(\sum_{p=1}^{d} w_{pq}^{(1)} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(0 - tanh \left(\sum_{p=1}^{d} w_{pj}^{(1)} x_{np} \right) \right) \right) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(2 \left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh \left(\sum_{p=1}^{d} u_{pq} x_{np} \right) \right) \cdot \right. \\
&\qquad \left. \left(-tanh \left(\sum_{p=1}^{d} u_{pj} x_{np} \right) \right) \right)
\end{aligned}
$$

因此可得

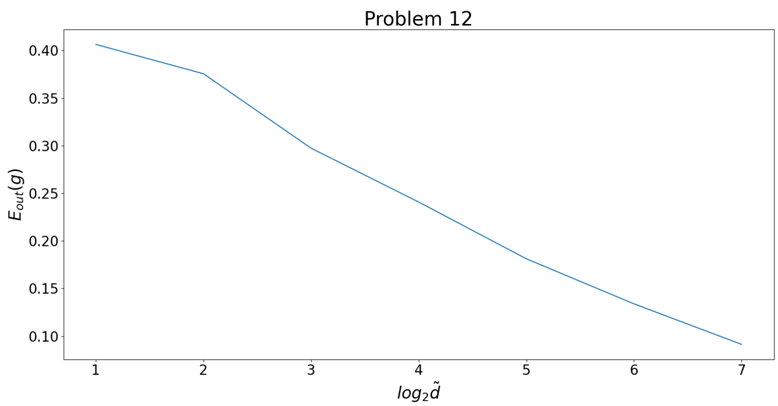
$$\frac{\partial E_9}{\partial u_{ij}} = \frac{1}{N}\sum_{n=1}^{N}\left(2\left(x_{ni} - \sum_{q=1}^{\tilde{d}} u_{iq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)\cdot\right.$$
$$\left.\left(-tanh\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right)\right) +$$
$$\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{d}\left(2\left(x_{nk} - \sum_{q=1}^{\tilde{d}} u_{kq} tanh\left(\sum_{p=1}^{d} u_{pq} x_{np}\right)\right)\cdot\right.$$
$$\left.\left(-u_{kj} x_{ni} tanh'\left(\sum_{p=1}^{d} u_{pj} x_{np}\right)\right)\right)$$
$$= \frac{\partial E_{10}}{\partial w_{ij}^{(1)}} + \frac{\partial E_{10}}{\partial w_{ji}^{(2)}}$$

Experiments with Autoencoder

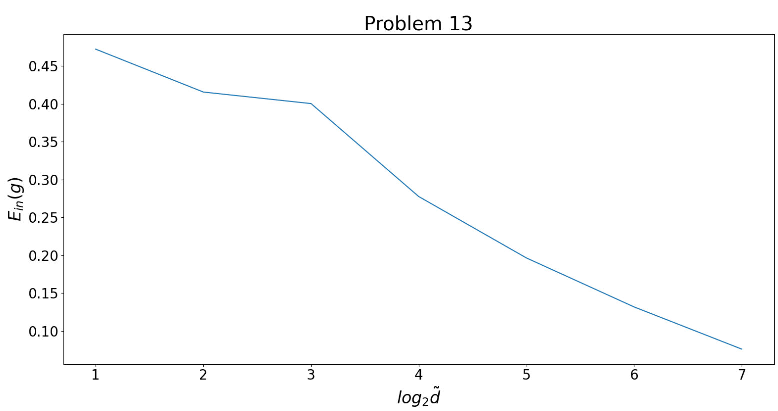
11. 當 $\tilde{d}$ 越大時，$E_{in}(g)$ 越小。



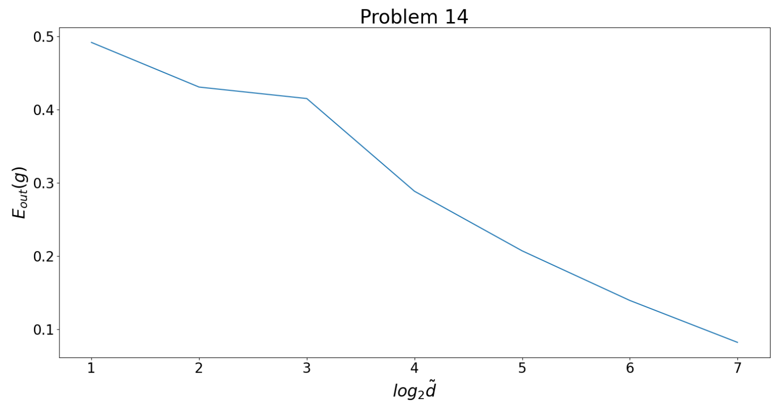
12. 當 $\tilde{d}$ 越大時，$E_{out}(g)$ 越小。



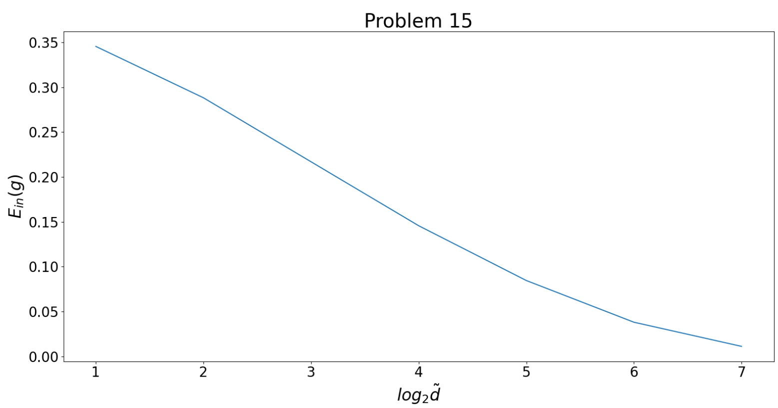
13. 當 $\tilde{d}$ 越大時，$E_{in}(g)$ 越小，此點和第 11 題相同，不過第 13 題的 $E_{in}(g)$ 皆比第 11 題的 $E_{in}(g)$ 還要再高一些。



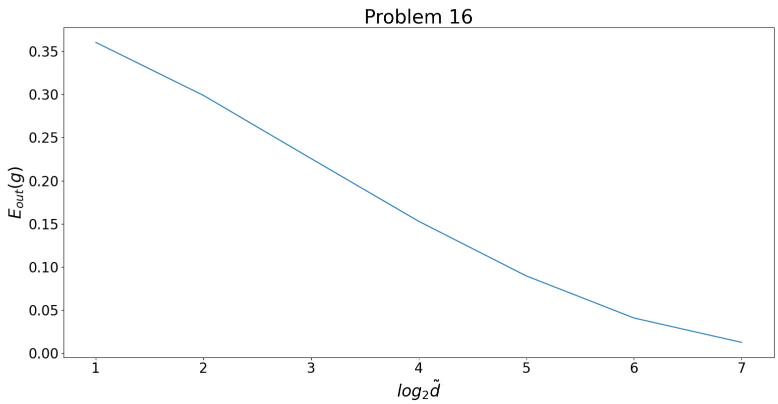
14. 當 $\tilde{d}$ 越大時，$E_{out}(g)$ 越小，此點和第 12 題相同，不過第 14 題的 $E_{out}(g)$ 皆比第 12 題的 $E_{out}(g)$ 還要再高一些。



15. 當 $\tilde{d}$ 越大時，$E_{in}(g)$ 越小，此點和第 13 題相同，不過第 15 題的 $E_{in}(g)$ 皆比第 13 題的 $E_{in}(g)$ 還要再低一些。



16. 當 $\tilde{d}$ 越大時，$E_{out}(g)$ 越小，此點和第 14 題相同，不過第 16 題的 $E_{out}(g)$ 皆比第 14 題的 $E_{out}(g)$ 還要再低一些。



Bonus: VC Dimension of Neural Networks

17. 首先，證明若 $N \geq 3\Delta log_2\Delta$，則有

$$\Delta lnN + \frac{1}{2} < Nln2$$

令

$$f(x) = \Delta lnx - xln2 + \frac{1}{2} \qquad (x > 0)$$

因爲當 $\Delta \geq 2$ 時

$$3(\sqrt{e})^{\frac{1}{\Delta}} log_2\Delta \leq 3(\sqrt{e})^{\frac{1}{2}} log_2\Delta = 3e^{\frac{1}{4}} log_2\Delta < \Delta^2 \qquad (3e^{\frac{1}{4}} \approx 3.852)$$

所以

$$\begin{gathered}3(\sqrt{e})^{\frac{1}{\Delta}}\Delta log_2\Delta < \Delta^3\\(3(\sqrt{e})^{\frac{1}{\Delta}}\Delta log_2\Delta)^{\Delta} < \Delta^{3\Delta}\\\sqrt{e}(3\Delta log_2\Delta)^{\Delta} < 2^{3\Delta log_2\Delta}\\ln(\sqrt{e}(3\Delta log_2\Delta)^{\Delta}) < ln(2^{3\Delta log_2\Delta})\\\frac{1}{2} + \Delta ln(3\Delta log_2\Delta) < (3\Delta log_2\Delta)ln2\\f(3\Delta log_2\Delta) = \Delta ln(3\Delta log_2\Delta) - (3\Delta log_2\Delta)ln2 + \frac{1}{2} < 0\end{gathered}$$

並且，因爲

$$f'(x) = \frac{\Delta}{x} - ln2 \qquad (x > 0)$$

所以當 $0 < x < \frac{\Delta}{ln2}$ 時，$f'(x) > 0$，而當 $x > \frac{\Delta}{ln2}$ 時，$f'(x) < 0$，意即，$f(x)$ 會在 $(0, \frac{\Delta}{ln2})$ 上嚴格遞增，並在 $(\frac{\Delta}{ln2}, \infty)$ 上嚴格遞減，注意當 $\Delta \geq 2$ 時

$$3\Delta log_2\Delta = 3\Delta\frac{ln\Delta}{ln2} \geq 3ln2\frac{\Delta}{ln2} > \frac{\Delta}{ln2} \qquad (3ln2 \approx 2)$$

所以 $f(x)$ 亦會在 $(3\Delta log_2\Delta, \infty) \subset (\frac{\Delta}{ln2}, \infty)$ 上嚴格遞減，因此，若 $N \geq 3\Delta log_2\Delta$，則有

$$f(N) \leq f(3\Delta log_2\Delta) < 0$$

即

$$\Delta lnN - Nln2 + \frac{1}{2} < 0$$

$$\Delta lnN + \frac{1}{2} < Nln2$$

接著，證明若 $N \geq 3\Delta log_2\Delta$，則有

$$ln(N^{\Delta} + 1) < \Delta lnN + \frac{1}{2}$$

令

$$g(x) = ln(x+1) - lnx - \frac{1}{2} \quad (x > 0)$$

因爲當 $x > 0$ 時

$$g'(x) = \frac{1}{x+1} - \frac{1}{x} < 0$$

所以 g 會在 $(0, \infty)$ 上嚴格遞減，並且，因爲 $\Delta \geq 2$，所以若 $N \geq 3\Delta log_2\Delta$，則有

$$N \geq 3\Delta log_2\Delta \geq 3 \times 2 \times log_2 2 = 6$$

$$N^{\Delta} \geq 6^2 = 36$$

因此

$$g(N^{\Delta}) \leq g(36) = ln37 - ln36 - \frac{1}{2} \approx -0.4726 < 0$$

即

$$ln(N^{\Delta} + 1) - lnN^{\Delta} - \frac{1}{2} < 0$$

$$ln(N^{\Delta} + 1) < \Delta lnN + \frac{1}{2}$$

綜合以上所述，可得若 $N \geq 3\Delta log_2\Delta$，則有

$$ln(N^{\Delta} + 1) < \Delta lnN + \frac{1}{2} < Nln2$$

故

$$N^{\Delta} + 1 < 2^N$$

18. 首先，證明以下的 **Lemma**。

Lemma $\forall N \in \mathbb{N}$，以及 $m \in \{0, 1, 2, \cdots, N\}$，皆有 $\sum_{i=0}^{m} \begin{pmatrix} N \\ i \end{pmatrix} \leq N^m$。

Proof 當 $m=0$ 時，因爲 $\sum_{i=0}^{0}\left(\begin{array}{c} N \\ i \end{array}\right)=1=N^0$，所以 $\sum_{i=0}^{0}\left(\begin{array}{c} N \\ i \end{array}\right)\leq N^0$ 成立。接著，設當 $m=k$ 時，$\sum_{i=0}^{k}\left(\begin{array}{c} N \\ i \end{array}\right)\leq N^k$ 成立，則當 $m=k+1$ 時

$$
\begin{aligned}
\sum_{i=0}^{k+1}\left(\begin{array}{c} N \\ i \end{array}\right) &= \sum_{i=0}^{k}\left(\begin{array}{c} N \\ i \end{array}\right)+\left(\begin{array}{c} N \\ k+1 \end{array}\right) \\
&\leq N^k+\left(\begin{array}{c} N \\ k+1 \end{array}\right) \\
&= N^k+\frac{N!}{(k+1)!(N-k-1)!} \\
&= N^k+\frac{N(N-1)(N-2)(N-3)\cdots(N-k)}{(k+1)!} \\
&\leq N^k+N(N-1)(N-2)(N-3)\cdots(N-k) \\
&\leq N^k+N\cdot(N-1)\cdot N\cdot N\cdot\cdots\cdot N \\
&= N^k+N^k(N-1) \\
&= N^{k+1}
\end{aligned}
$$

因此由數學歸納法可知 $\forall\ m\in\{0,1,2,\cdots,N\}$，皆有 $\sum_{i=0}^{m}\left(\begin{array}{c} N \\ i \end{array}\right)\leq N^m$。

以下開始證明第 18 題。考慮 $\mathcal{H}_{3A}$ 在任意的 N 筆資料 $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$ 上可以產生的 dichotomy 數量，其中 $N\geq 3\Delta log_2\Delta$，$\Delta=3(d+1)+1$。首先，考慮 $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$ 經過 hidden layer 的 transformation 之後，$\{\mathbf{x}_1^{(1)},\mathbf{x}_2^{(1)},\cdots,\mathbf{x}_N^{(1)}\}$ 有多少種不同的可能，若僅看 hidden layer 的單一個 neuron，由於其輸出爲 input layer 的 $d+1$ 個 dimension (包含 bias) 的 weighted sum 的正負，因此其可以視爲 dimension 爲 d 的 perceptron，由機器學習基石的課程內容，可知 dimension 爲 d 的 perceptron，其 VC dimension 爲 $d+1$，意即其最小的 break point 爲 $d+2$，因此，所有 dimension 爲 d 的 perceptron 所形成的 hypothesis set，其在 $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$ 上可以產生的 dichotomy 數量，不超過 $B(N,d+2)$，其中 B 爲 bounding function，又由機器學習基石的課程內容以及以上的 **Lemma**，可得

$$B(N,d+2)\leq\sum_{i=0}^{d+1}\left(\begin{array}{c} N \\ i \end{array}\right)\leq N^{d+1}$$

因此，若僅看 hidden layer 的單一個 neuron，將其視爲 dimension 爲 d 的 perceptron，則其在 $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$ 上可以產生的 dichotomy 數量，不超過 N^{d+1}，而 hidden layer 中有三個 neuron，因此 $\{\mathbf{x}_1^{(1)},\mathbf{x}_2^{(1)},\cdots,\mathbf{x}_N^{(1)}\}$ 可能的數量，可以視爲三個 dimension 皆爲 d 的 perceptron，其各自在相同的 $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$ 上可以產生的 dichotomy 作重複組合的數量，其不超過

$$(N^{d+1})^3=N^{3(d+1)}\leq N^{3(d+1)+1}+1=N^{\Delta}+1$$

又 $\mathcal{H}_{3A}$ 中的 neural network，其 hidden layer 和 output layer 之間的 weight 已經固定，因此 $\mathcal{H}_{3A}$ 在 $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$ 上可以產生的 dichotomy 數量亦不超過

$N^{\Delta}+1$，注意 $\Delta \geq 2$，且 $N \geq 3\Delta log_2\Delta$，因此由第 17 題可知 $N^{\Delta}+1 < 2^N$，故對於任意的 N 筆資料 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$，其中 $N \geq 3\Delta log_2\Delta$，$\mathcal{H}_{3A}$ 在其上可以產生的 dichotomy 數量小於 N^{Δ}，意即 $\mathcal{H}_{3A}$ 無法將其 shatter，因此可得 $\mathcal{H}_{3A}$ 的 VC dimension 小於 $3\Delta log_2\Delta = 3(3(d+1)+1)log_2(3(d+1)+1)$。