Machine Learning Techniques - Homework 3

資工四 B05902023 李澤諺

Deep Learning Techniques

1. (以下皆爲在已知 $\mathbf{x}^{(l-1)}$ 的條件下進行推導，爲簡單起見，以下將 $x_i^{(l-1)}$ 皆視爲 constant，而省去 conditional probability、conditional distribution function、conditional density function、conditional expectation 的符號)

因爲

$$\begin{aligned}\mathbb{E}\left[s_j^{(l)}\right] &= \mathbb{E}\left[\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}\right] \\ &= \sum_{i=1}^{d^{(l-1)}} \mathbb{E}\left[w_{ij}^{(l)} x_i^{(l-1)}\right] \\ &= \sum_{i=1}^{d^{(l-1)}} \mathbb{E}\left[w_{ij}^{(l)}\right] \cdot x_i^{(l-1)} \\ &= \sum_{i=1}^{d^{(l-1)}} 0 \cdot x_i^{(l-1)} = 0\end{aligned}$$

因此可得 $s_j^{(l)}$ 爲 zero-mean。接著，$\forall\ n \in \mathbb{N}$ 且 $2 \leq n \leq d^{(l)}$，任取 n 個 random variable $s_{j_1}^{(l)}$、$s_{j_2}^{(l)}$、$\cdots$、$s_{j_n}^{(l)}$，令 $s_{j_k}^{(l)}$ 的 distribution function 和 density function 分別爲 $F_{j_k}\left(t_{j_k}\right)$ 和 $f_{j_k}\left(t_{j_k}\right)$，$s_{j_1}^{(l)}$、$s_{j_2}^{(l)}$、$\cdots$、$s_{j_n}^{(l)}$ 的 joint distribution function 和 joint density function 分別爲 $F_{j_1 j_2 \cdots j_n}\left(t_{j_1}, t_{j_2}, \cdots, t_{j_n}\right)$ 和 $f_{j_1 j_2 \cdots j_n}\left(t_{j_1}, t_{j_2}, \cdots, t_{j_n}\right)$，並令 $\mathbf{w}_{j_k} = \left(w_{1j_k}^{(l)}, w_{2j_k}^{(l)}, \cdots, w_{d^{(l-1)}j_k}^{(l)}\right)$，其 joint density function 爲 $g_{j_k}\left(\mathbf{u}_{j_k}\right) = g_{j_k}\left(u_{j_k 1}, u_{j_k 2}, \cdots, u_{j_k d^{(l-1)}}\right)$，$\mathbf{w}_{j_1}$、$\mathbf{w}_{j_2}$、$\cdots$、$\mathbf{w}_{j_n}$ 的 joint density function 爲 $g_{j_1 j_2 \cdots j_n}\left(\mathbf{u}_{j_1}, \mathbf{u}_{j_2}, \cdots, \mathbf{u}_{j_n}\right)$，並且，$\forall\ t_{j_k} \in \mathbb{R}$，令 $A_{j_k} = \left\{\mathbf{w}_{j_k}\ \middle|\ \sum_{i=1}^{d^{(l-1)}} w_{ij_k}^{(l)} x_i^{(l-1)} \leq t_{j_k}\right\}$。注意因爲 $\mathbf{w}_{j_1}$、$\mathbf{w}_{j_2}$、$\cdots$、$\mathbf{w}_{j_n}$ 爲 independent random vector，所以

$$g_{j_1 j_2 \cdots j_n}\left(\mathbf{u}_{j_1}, \mathbf{u}_{j_2}, \cdots, \mathbf{u}_{j_n}\right) = g_{j_1}\left(\mathbf{u}_{j_1}\right) g_{j_2}\left(\mathbf{u}_{j_2}\right) \cdots g_{j_n}\left(\mathbf{u}_{j_n}\right)$$

因此可得

$$
\begin{aligned}
& F_{j_1 j_2 \cdots j_n}\left(t_{j_1}, t_{j_2}, \cdots, t_{j_n}\right) \\
= & P\left(s_{j_1}^{(l)} \leq t_{j_1}, s_{j_2}^{(l)} \leq t_{j_2}, \cdots, s_{j_n}^{(l)} \leq t_{j_n}\right) \\
= & P\left(\sum_{i=1}^{d^{(l-1)}} w_{i j_1}^{(l)} x_i^{(l-1)} \leq t_{j_1}, \sum_{i=1}^{d^{(l-1)}} w_{i j_2}^{(l)} x_i^{(l-1)} \leq t_{j_2}, \cdots, \sum_{i=1}^{d^{(l-1)}} w_{i j_n}^{(l)} x_i^{(l-1)} \leq t_{j_n}\right) \\
= & P\left(\mathbf{w}_{j_1} \in A_{j_1}, \mathbf{w}_{j_2} \in A_{j_2}, \cdots, \mathbf{w}_{j_n} \in A_{j_n}\right) \\
= & \int_{A_{j_1} \times A_{j_2} \times \cdots \times A_{j_n}} g_{j_1 j_2 \cdots j_n}\left(\mathbf{u}_{j_1}, \mathbf{u}_{j_2}, \cdots, \mathbf{u}_{j_n}\right) d \mathbf{u}_{j_1} d \mathbf{u}_{j_2} \cdots d \mathbf{u}_{j_n} \\
= & \int_{A_{j_1} \times A_{j_2} \times \cdots \times A_{j_n}} g_{j_1}\left(\mathbf{u}_{j_1}\right) g_{j_2}\left(\mathbf{u}_{j_2}\right) \cdots g_{j_n}\left(\mathbf{u}_{j_n}\right) d \mathbf{u}_{j_1} d \mathbf{u}_{j_2} \cdots d \mathbf{u}_{j_n} \\
= & \int_{A_{j_1}} g_{j_1}\left(\mathbf{u}_{j_1}\right) d \mathbf{u}_{j_1} \int_{A_{j_2}} g_{j_2}\left(\mathbf{u}_{j_2}\right) d \mathbf{u}_{j_2} \cdots \int_{A_{j_n}} g_{j_n}\left(\mathbf{u}_{j_n}\right) d \mathbf{u}_{j_n} \\
= & P\left(\mathbf{w}_{j_1} \in A_{j_1}\right) P\left(\mathbf{w}_{j_2} \in A_{j_2}\right) \cdots P\left(\mathbf{w}_{j_n} \in A_{j_n}\right) \\
= & P\left(\sum_{i=1}^{d^{(l-1)}} w_{i j_1}^{(l)} x_i^{(l-1)} \leq t_{j_1}\right) P\left(\sum_{i=1}^{d^{(l-1)}} w_{i j_2}^{(l)} x_i^{(l-1)} \leq t_{j_2}\right) \cdots \\
& P\left(\sum_{i=1}^{d^{(l-1)}} w_{i j_n}^{(l)} x_i^{(l-1)} \leq t_{j_n}\right) \\
= & P\left(s_{j_1}^{(l)} \leq t_{j_1}\right) P\left(s_{j_2}^{(l)} \leq t_{j_2}\right) \cdots P\left(s_{j_n}^{(l)} \leq t_{j_n}\right) \\
= & F_{j_1}\left(t_{j_1}\right) F_{j_2}\left(t_{j_2}\right) \cdots F_{j_n}\left(t_{j_n}\right)
\end{aligned}
$$

故

$$
\begin{aligned}
& f_{j_1 j_2 \cdots j_n}\left(t_{j_1}, t_{j_2}, \cdots, t_{j_n}\right) \\
= & \frac{\partial^n}{\partial t_{j_1} \partial t_{j_2} \cdots \partial t_{j_n}} F_{j_1 j_2 \cdots j_n}\left(t_{j_1}, t_{j_2}, \cdots, t_{j_n}\right) \\
= & \frac{\partial^n}{\partial t_{j_1} \partial t_{j_2} \cdots \partial t_{j_n}} F_{j_1}\left(t_{j_1}\right) F_{j_2}\left(t_{j_2}\right) \cdots F_{j_n}\left(t_{j_n}\right) \\
= & f_{j_1}\left(t_{j_1}\right) f_{j_2}\left(t_{j_2}\right) \cdots f_{j_n}\left(t_{j_n}\right)
\end{aligned}
$$

由此可得 $s_1^{(l)}$、$s_2^{(l)}$、$\cdots$、$s_{d^{(l)}}^{(l)}$ 爲 independent。

2. 因爲

$$
\begin{aligned}
Var\left(x_i^{(l-1)}\right) &= \mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] - \mathbb{E}\left[x_i^{(l-1)}\right]^2 \\
\sigma_x^2 &= \mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] - \bar{x}^2 \\
\mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] &= \sigma_x^2 + \bar{x}^2
\end{aligned}
$$

以及

$$
\begin{aligned}
Var\left(w_{ij}^{(l)}\right) &= \mathbb{E}\left[\left(w_{ij}^{(l)}\right)^2\right] - \mathbb{E}\left[w_{ij}^{(l)}\right]^2 \\
\sigma_w^2 &= \mathbb{E}\left[\left(w_{ij}^{(l)}\right)^2\right] - 0^2 \\
\mathbb{E}\left[\left(w_{ij}^{(l)}\right)^2\right] &= \sigma_w^2
\end{aligned}
$$

所以

$$
\begin{aligned}
Var\left(s_j^{(l)}\right) &= \mathbb{E}\left[\left(s_j^{(l)}\right)^2\right] - \mathbb{E}\left[s_j^{(l)}\right]^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}\right)^2\right] - 0^2 \\
&= \mathbb{E}\left[\sum_{i=1}^{d^{(l-1)}} \sum_{k=1}^{d^{(l-1)}} w_{ij}^{(l)} w_{kj}^{(l)} x_i^{(l-1)} x_k^{(l-1)}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d^{(l-1)}} \left(w_{ij}^{(l)}\right)^2 \left(x_i^{(l-1)}\right)^2 + \sum_{i \neq k} w_{ij}^{(l)} w_{kj}^{(l)} x_i^{(l-1)} x_k^{(l-1)}\right] \\
&= \sum_{i=1}^{d^{(l-1)}} \mathbb{E}\left[\left(w_{ij}^{(l)}\right)^2 \left(x_i^{(l-1)}\right)^2\right] + \sum_{i \neq k} \mathbb{E}\left[w_{ij}^{(l)} w_{kj}^{(l)} x_i^{(l-1)} x_k^{(l-1)}\right]
\end{aligned}
$$

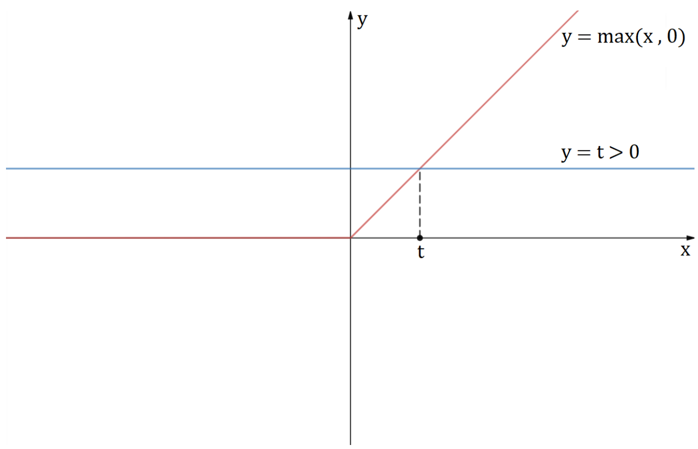
其中，因爲 independence，因此可得

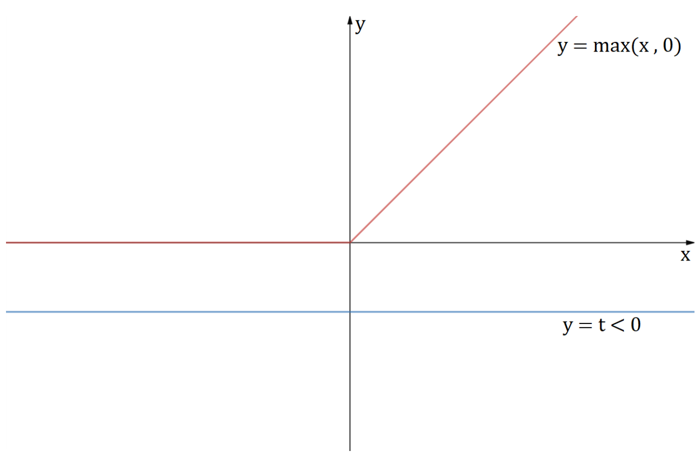
$$
\begin{aligned}
Var\left(s_j^{(l)}\right) &= \sum_{i=1}^{d^{(l-1)}} \mathbb{E}\left[\left(w_{ij}^{(l)}\right)^2 \left(x_i^{(l-1)}\right)^2\right] + \sum_{i \neq k} \mathbb{E}\left[w_{ij}^{(l)} w_{kj}^{(l)} x_i^{(l-1)} x_k^{(l-1)}\right] \\
&= \sum_{i=1}^{d^{(l-1)}} \mathbb{E}\left[\left(w_{ij}^{(l)}\right)^2\right] \mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] + \\
&\quad \sum_{i \neq k} \mathbb{E}\left[w_{ij}^{(l)}\right] \mathbb{E}\left[w_{kj}^{(l)}\right] \mathbb{E}\left[x_i^{(l-1)}\right] \mathbb{E}\left[x_k^{(l-1)}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{d^{(l-1)}} \sigma_w^2 \left(\sigma_x^2 + \bar{x}^2\right) + \sum_{i \neq k} 0 \cdot 0 \cdot \bar{x} \cdot \bar{x} \\
&= d^{(l-1)} \sigma_w^2 \left(\sigma_x^2 + \bar{x}^2\right)
\end{aligned}
$$

3. 令 $s_i^{(l-1)}$ 的 distribution function 和 density function 分別爲 $F_{s_i^{(l-1)}}(t)$ 和 $f_{s_i^{(l-1)}}(t)$，而 $x_i^{(l-1)}$ 的 distribution function 和 density function 分別爲 $F_{x_i^{(l-1)}}(t)$ 和 $f_{x_i^{(l-1)}}(t)$。因爲

$$
\begin{aligned}
F_{x_i^{(l-1)}}(t) &= P\left(x_i^{(l-1)} \leq t\right) = P\left(max\left(s_i^{(l-1)}, 0\right) \leq t\right) \\
&= \begin{cases} P\left(s_i^{(l-1)} \leq t\right) = F_{s_i^{(l-1)}}(t) & \text{若 } t \geq 0 \\ 0 & \text{若 } t < 0 \end{cases}
\end{aligned}
$$





所以

$$f_{x_i^{(l-1)}}(t) = \frac{d}{dt}F_{x_i^{(l-1)}}(t) = \begin{cases} \frac{d}{dt}F_{s_i^{(l-1)}}(t) = f_{s_i^{(l-1)}}(t) & 若\ t > 0 \\ \frac{d}{dt}0 = 0 & 若\ t < 0 \end{cases}$$

因此

$$\begin{aligned}
\mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] &= \int_{-\infty}^{\infty} t^2 f_{x_i^{(l-1)}}(t)dt \\
&= \int_{-\infty}^{0} t^2 f_{x_i^{(l-1)}}(t)dt + \int_{0}^{\infty} t^2 f_{x_i^{(l-1)}}(t)dt \\
&= \int_{-\infty}^{0} t^2 \cdot 0dt + \int_{0}^{\infty} t^2 f_{s_i^{(l-1)}}(t)dt \\
&= \int_{0}^{\infty} t^2 f_{s_i^{(l-1)}}(t)dt
\end{aligned}$$

其中，因爲 $s_i^{(l-1)}$ 爲 symmetric random variable，所以 $f_{s_i^{(l-1)}}(-t) = f_{s_i^{(l-1)}}(t)$ [ref]，因此 $(-t)^2 f_{s_i^{(l-1)}}(-t) = t^2 f_{s_i^{(l-1)}}(t)$，即 $t^2 f_{s_i^{(l-1)}}(t)$ 爲 even function，故

$$\begin{aligned}
\mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] &= \int_{0}^{\infty} t^2 f_{s_i^{(l-1)}}(t)dt \\
&= \frac{1}{2}\int_{-\infty}^{\infty} t^2 f_{s_i^{(l-1)}}(t)dt \\
&= \frac{1}{2}\mathbb{E}\left[\left(s_i^{(l-1)}\right)^2\right]
\end{aligned}$$

4. 由第 2 題和第 3 題可知

$$\mathbb{E}\left[\left(s_i^{(l-1)}\right)^2\right] = 2\mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] = 2\left(\sigma_x^2 + \bar{x}^2\right)$$

所以

$$\begin{aligned}
Var\left(s_i^{(l-1)}\right) &= \mathbb{E}\left[\left(s_i^{(l-1)}\right)^2\right] - \mathbb{E}\left[s_i^{(l-1)}\right]^2 \\
&= 2\left(\sigma_x^2 + \bar{x}^2\right) - 0^2 \\
&= 2\left(\sigma_x^2 + \bar{x}^2\right)
\end{aligned}$$

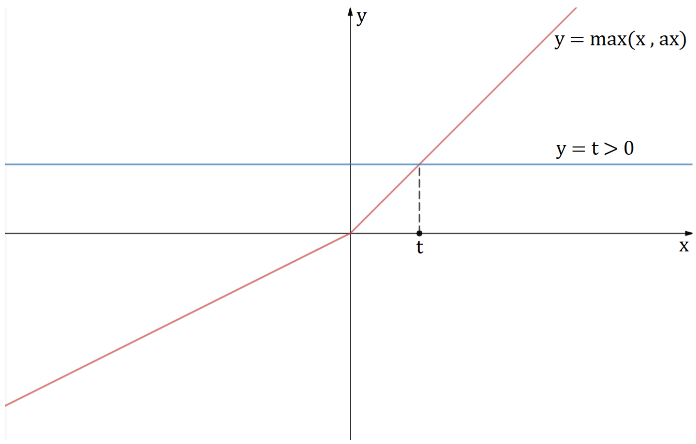
$$\sigma_x^2 + \bar{x}^2 = \frac{1}{2}Var\left(s_i^{(l-1)}\right)$$

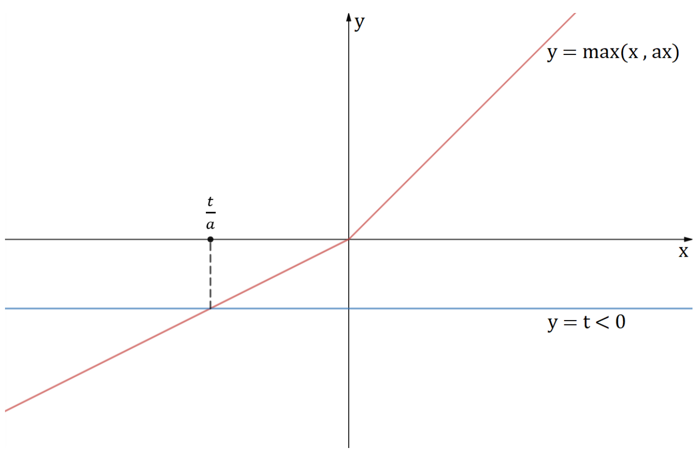
故由第 2 題的結果可得

$$Var\left(s_j^{(l)}\right) = d^{(l-1)}\sigma_w^2\left(\sigma_x^2 + \bar{x}^2\right) = \frac{d^{(l-1)}}{2}\sigma_w^2 Var\left(s_i^{(l-1)}\right)$$

5. 令 $s_j^{(l-1)}$ 的 distribution function 和 density function 分別爲 $F_{s_j^{(l-1)}}(t)$ 和 $f_{s_j^{(l-1)}}(t)$，而 $x_j^{(l-1)}$ 的 distribution function 和 density function 分別爲 $F_{x_j^{(l-1)}}(t)$ 和 $f_{x_j^{(l-1)}}(t)$。因爲

$$
\begin{aligned}
F_{x_j^{(l-1)}}(t) &= P\left(x_j^{(l-1)} \leq t\right) = P\left(max\left(s_j^{(l-1)}, a \cdot s_j^{(l-1)}\right) \leq t\right) \\
&= \begin{cases} P\left(s_j^{(l-1)} \leq t\right) = F_{s_j^{(l-1)}}(t) & \text{若 } t \geq 0 \\ P\left(s_j^{(l-1)} \leq \frac{t}{a}\right) = F_{s_j^{(l-1)}}\left(\frac{t}{a}\right) & \text{若 } t < 0 \end{cases}
\end{aligned}
$$





所以

$$
f_{x_j^{(l-1)}}(t) = \frac{d}{dt} F_{x_j^{(l-1)}}(t) = \begin{cases} \frac{d}{dt} F_{s_j^{(l-1)}}(t) = f_{s_j^{(l-1)}}(t) & \text{若 } t > 0 \\ \frac{d}{dt} F_{s_j^{(l-1)}}\left(\frac{t}{a}\right) = \frac{1}{a} f_{s_j^{(l-1)}}\left(\frac{t}{a}\right) & \text{若 } t < 0 \end{cases}
$$

因此

$$\begin{aligned}
\mathbb{E}\left[\left(x_j^{(l-1)}\right)^2\right] &= \int_{-\infty}^{\infty} t^2 f_{x_j^{(l-1)}}(t)dt \\
&= \int_{-\infty}^{0} t^2 f_{x_j^{(l-1)}}(t)dt + \int_{0}^{\infty} t^2 f_{x_j^{(l-1)}}(t)dt \\
&= \int_{-\infty}^{0} \frac{t^2}{a} f_{s_j^{(l-1)}}\left(\frac{t}{a}\right)dt + \int_{0}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt
\end{aligned}$$

令 $t = au$，則有

$$\begin{aligned}
\mathbb{E}\left[\left(x_j^{(l-1)}\right)^2\right] &= \int_{-\infty}^{0} \frac{t^2}{a} f_{s_j^{(l-1)}}\left(\frac{t}{a}\right)dt + \int_{0}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt \\
&= \int_{-\infty}^{0} a^2u^2 f_{s_j^{(l-1)}}(u)du + \int_{0}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt \\
&= a^2\int_{-\infty}^{0} t^2 f_{s_j^{(l-1)}}(t)dt + \int_{0}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt
\end{aligned}$$

其中，因爲 $s_j^{(l-1)}$ 爲 symmetric random variable，所以 $f_{s_j^{(l-1)}}(-t) = f_{s_j^{(l-1)}}(t)$ [ref]，因此 $(-t)^2 f_{s_j^{(l-1)}}(-t) = t^2 f_{s_j^{(l-1)}}(t)$，即 $t^2 f_{s_j^{(l-1)}}(t)$ 爲 even function，故

$$\begin{aligned}
\mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] &= a^2\int_{-\infty}^{0} t^2 f_{s_j^{(l-1)}}(t)dt + \int_{0}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt \\
&= \frac{a^2}{2}\int_{-\infty}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt + \frac{1}{2}\int_{-\infty}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt \\
&= \frac{a^2+1}{2}\int_{-\infty}^{\infty} t^2 f_{s_j^{(l-1)}}(t)dt \\
&= \frac{a^2+1}{2}\mathbb{E}\left[\left(s_j^{(l-1)}\right)^2\right]
\end{aligned}$$

由上式以及第 2 題，可得

$$\mathbb{E}\left[\left(s_j^{(l-1)}\right)^2\right] = \frac{2}{a^2+1}\mathbb{E}\left[\left(x_i^{(l-1)}\right)^2\right] = \frac{2}{a^2+1}\left(\sigma_x^2 + \bar{x}^2\right)$$

所以

$$\begin{aligned}
Var\left(s_i^{(l-1)}\right) &= \mathbb{E}\left[\left(s_i^{(l-1)}\right)^2\right] - \mathbb{E}\left[s_i^{(l-1)}\right]^2 \\
&= \frac{2}{a^2+1}\left(\sigma_x^2 + \bar{x}^2\right) - 0^2 \\
&= \frac{2}{a^2+1}\left(\sigma_x^2 + \bar{x}^2\right)
\end{aligned}$$

$$\sigma_x^2 + \bar{x}^2 = \frac{a^2+1}{2} Var\left(s_i^{(l-1)}\right)$$

故由第 2 題的結果可得

$$Var\left(s_j^{(l)}\right) = d^{(l-1)} \sigma_w^2 \left(\sigma_x^2 + \bar{x}^2\right) = \frac{d^{(l-1)}(a^2+1)}{2} \sigma_w^2 Var\left(s_i^{(l-1)}\right)$$

因此，只要在滿足 Problem 1 到 Problem 4 的所有條件下，並取 $w_{ij}^{(l)}$ 的 variance 爲 $\sigma_w^2 = \frac{2}{d^{(l-1)}(a^2+1)}$，以此進行 initialization，即可得到 $Var\left(s_j^{(l)}\right) = Var\left(s_i^{(l-1)}\right)$。

6. 以下說明 $\forall\ T \in \mathbb{N}$，$\mathbf{v}_T = \sum_{t=1}^{T} \beta^{T-t}(1-\beta)\mathbf{\Delta}_t$。當 $T = 1$ 時

$$\begin{aligned}
\mathbf{v}_1 &= \beta \mathbf{v}_0 + (1-\beta)\mathbf{\Delta}_1 \\
&= \beta \cdot \mathbf{0} + (1-\beta)\mathbf{\Delta}_1 \\
&= (1-\beta)\mathbf{\Delta}_1 \\
&= \sum_{t=1}^{1} \beta^{1-t}(1-\beta)\mathbf{\Delta}_t
\end{aligned}$$

所以 $\mathbf{v}_1 = \sum_{t=1}^{1} \beta^{1-t}(1-\beta)\mathbf{\Delta}_t$ 成立。設當 $T = k$ 時，$\mathbf{v}_k = \sum_{t=1}^{k} \beta^{k-t}(1-\beta)\mathbf{\Delta}_t$ 成立，則當 $T = k+1$ 時

$$\begin{aligned}
\mathbf{v}_{k+1} &= \beta \mathbf{v}_k + (1-\beta)\mathbf{\Delta}_{k+1} \\
&= \beta \left(\sum_{t=1}^{k} \beta^{k-t}(1-\beta)\mathbf{\Delta}_t\right) + (1-\beta)\mathbf{\Delta}_{k+1} \\
&= \left(\sum_{t=1}^{k} \beta^{k-t+1}(1-\beta)\mathbf{\Delta}_t\right) + (1-\beta)\mathbf{\Delta}_{k+1} \\
&= \sum_{t=1}^{k+1} \beta^{k-t+1}(1-\beta)\mathbf{\Delta}_t
\end{aligned}$$

所以 $\mathbf{v}_{k+1} = \sum_{t=1}^{k+1} \beta^{k-t+1}(1-\beta)\mathbf{\Delta}_t$ 成立，由數學歸納法可知，$\forall\ T \in \mathbb{N}$，皆有 $\mathbf{v}_T = \sum_{t=1}^{T} \beta^{T-t}(1-\beta)\mathbf{\Delta}_t$，因此可得 $\alpha_t = \beta^{T-t}(1-\beta)$。

7. 注意 $0<\beta<1$，所以 $log_2\beta<0$，因此

$$\begin{aligned}
\alpha_1 &< \frac{1}{2}\\
\beta^{T-1}(1-\beta) &< \frac{1}{2}\\
log_2\beta^{T-1}(1-\beta) &< log_2\frac{1}{2}\\
(T-1)log_2\beta + log_2(1-\beta) &< -1\\
(T-1)log_2\beta &< -1-log_2(1-\beta)\\
T-1 &> \frac{-1-log_2(1-\beta)}{log_2\beta}\\
T &> \frac{-1-log_2(1-\beta)}{log_2\beta}+1
\end{aligned}$$

故可取 $T=\left\lceil \frac{-1-log_2(1-\beta)}{log_2\beta}+1\right\rceil$。

8.

$$\alpha'_t=\frac{\alpha_t}{\sum_{t=1}^{T}\alpha_t}=\frac{\beta^{T-t}(1-\beta)}{\sum_{t=1}^{T}\beta^{T-t}(1-\beta)}=\frac{\beta^{T-t}(1-\beta)}{\beta^{T-1}(1-\beta)\frac{1-\beta^{-T}}{1-\beta^{-1}}}=\beta^{1-t}\frac{1-\beta^{-1}}{1-\beta^{-T}}$$

9. 注意 $0<\beta<1$，即 $\beta^{-1}>1$，所以 $1-\beta^{-T}<0$，$ln\beta^{-1}>0$，因此

$$\begin{aligned}
\alpha'_1 &< \frac{1}{2}\\
\frac{1-\beta^{-1}}{1-\beta^{-T}} &< \frac{1}{2}\\
2\left(1-\beta^{-1}\right) &> 1-\beta^{-T}\\
\beta^{-T} &> 2\beta^{-1}-1\\
Tln\beta^{-1} &> ln\left(2\beta^{-1}-1\right)\\
T &> \frac{ln\left(2\beta^{-1}-1\right)}{ln\beta^{-1}}
\end{aligned}$$

故可取 $T=\left\lceil \frac{ln\left(2\beta^{-1}-1\right)}{ln\beta^{-1}}\right\rceil$。

10. 令

$$D_{\mathbf{w}}=\begin{pmatrix} w_0 & 0 & 0 & \cdots & 0\\ 0 & w_1 & 0 & \cdots & 0\\ 0 & 0 & w_2 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & w_d \end{pmatrix}$$

則

$$\mathbf{w}\odot\mathbf{p}=\begin{pmatrix}w_0\\w_1\\w_2\\\vdots\\w_d\end{pmatrix}\odot\begin{pmatrix}p_0\\p_1\\p_2\\\vdots\\p_d\end{pmatrix}=\begin{pmatrix}w_0p_0\\w_1p_1\\w_2p_2\\\vdots\\w_dp_d\end{pmatrix}$$
$$=\begin{pmatrix}w_0&0&0&\cdots&0\\0&w_1&0&\cdots&0\\0&0&w_2&\cdots&0\\\vdots&\vdots&\vdots&&\vdots\\0&0&0&\cdots&w_d\end{pmatrix}\begin{pmatrix}p_0\\p_1\\p_2\\\vdots\\p_d\end{pmatrix}=D_{\mathbf{w}}\mathbf{p}$$

因此

$$\begin{aligned}
&E_{\mathbf{p}}\left[\|\mathbf{y}-X(\mathbf{w}\odot\mathbf{p})\|^2\right]\\
&=E_{\mathbf{p}}\left[\|\mathbf{y}-XD_{\mathbf{w}}\mathbf{p}\|^2\right]\\
&=E_{\mathbf{p}}\left[(\mathbf{y}-XD_{\mathbf{w}}\mathbf{p})^T(\mathbf{y}-XD_{\mathbf{w}}\mathbf{p})\right]\\
&=E_{\mathbf{p}}\left[\left(\mathbf{y}^T-(XD_{\mathbf{w}}\mathbf{p})^T\right)(\mathbf{y}-XD_{\mathbf{w}}\mathbf{p})\right]\\
&=E_{\mathbf{p}}\left[\mathbf{y}^T\mathbf{y}-\mathbf{y}^T(XD_{\mathbf{w}}\mathbf{p})-(XD_{\mathbf{w}}\mathbf{p})^T\mathbf{y}+(XD_{\mathbf{w}}\mathbf{p})^T(XD_{\mathbf{w}}\mathbf{p})\right]\\
&=E_{\mathbf{p}}\left[\mathbf{y}^T\mathbf{y}-2\mathbf{y}^TXD_{\mathbf{w}}\mathbf{p}+\mathbf{p}^TD_{\mathbf{w}}^TX^TXD_{\mathbf{w}}\mathbf{p}\right]\\
&=E_{\mathbf{p}}\left[\mathbf{y}^T\mathbf{y}\right]-E_{\mathbf{p}}\left[2\mathbf{y}^TXD_{\mathbf{w}}\mathbf{p}\right]+E_{\mathbf{p}}\left[\mathbf{p}^TD_{\mathbf{w}}^TX^TXD_{\mathbf{w}}\mathbf{p}\right]\\
&=\mathbf{y}^T\mathbf{y}-2\mathbf{y}^TXD_{\mathbf{w}}E_{\mathbf{p}}[\mathbf{p}]+E_{\mathbf{p}}\left[\mathbf{p}^TD_{\mathbf{w}}^TX^TXD_{\mathbf{w}}\mathbf{p}\right]
\end{aligned}$$

注意 $\forall\ \mathbf{u}=\begin{pmatrix}u_0\\u_1\\u_2\\\vdots\\u_d\end{pmatrix}$、$\mathbf{v}=\begin{pmatrix}v_0\\v_1\\v_2\\\vdots\\v_d\end{pmatrix}\in\mathbb{R}^{d+1}$，皆有

$$\begin{aligned}
\mathbf{u}^T\mathbf{v}&=u_0v_0+u_1v_1+u_2v_2+\cdots+u_dv_d\\
&=tr\begin{pmatrix}u_0v_0&u_0v_1&u_0v_2&\cdots&u_0v_d\\u_1v_0&u_1v_1&u_1v_2&\cdots&u_1v_d\\u_2v_0&u_2v_1&u_2v_2&\cdots&u_2v_d\\\vdots&\vdots&\vdots&&\vdots\\u_dv_0&u_dv_1&u_dv_2&\cdots&u_dv_d\end{pmatrix}=tr\left(\mathbf{u}\mathbf{v}^T\right)
\end{aligned}$$

因此可得

$$
\begin{aligned}
& E_{\mathbf{p}}\left[\|\mathbf{y}-X(\mathbf{w}\odot\mathbf{p})\|^2\right] \\
=\ & \mathbf{y}^T\mathbf{y}-2\mathbf{y}^TXD_{\mathbf{w}}E_{\mathbf{p}}\left[\mathbf{p}\right]+E_{\mathbf{p}}\left[\mathbf{p}^TD_{\mathbf{w}}^TX^TXD_{\mathbf{w}}\mathbf{p}\right] \\
=\ & \mathbf{y}^T\mathbf{y}-2\mathbf{y}^TXD_{\mathbf{w}}E_{\mathbf{p}}\left[\mathbf{p}\right]+E_{\mathbf{p}}\left[\left(D_{\mathbf{w}}^TX^TXD_{\mathbf{w}}\mathbf{p}\right)^T\mathbf{p}\right] \\
=\ & \mathbf{y}^T\mathbf{y}-2\mathbf{y}^TXD_{\mathbf{w}}E_{\mathbf{p}}\left[\mathbf{p}\right]+E_{\mathbf{p}}\left[tr\left(D_{\mathbf{w}}^TX^TXD_{\mathbf{w}}\mathbf{p}\mathbf{p}^T\right)\right] \\
=\ & \mathbf{y}^T\mathbf{y}-2\mathbf{y}^TXD_{\mathbf{w}}E_{\mathbf{p}}\left[\mathbf{p}\right]+tr\left(E_{\mathbf{p}}\left[D_{\mathbf{w}}^TX^TXD_{\mathbf{w}}\mathbf{p}\mathbf{p}^T\right]\right) \\
=\ & \mathbf{y}^T\mathbf{y}-2\mathbf{y}^TXD_{\mathbf{w}}E_{\mathbf{p}}\left[\mathbf{p}\right]+tr\left(D_{\mathbf{w}}^TX^TXD_{\mathbf{w}}E_{\mathbf{p}}\left[\mathbf{p}\mathbf{p}^T\right]\right)
\end{aligned}
$$

其中，因爲

$$E_{\mathbf{p}}\left[p_i\right]=0\cdot P\left(p_i=0\right)+1\cdot P\left(p_i=1\right)=0\cdot\frac{1}{2}+1\cdot\frac{1}{2}=\frac{1}{2}$$

$$E_{\mathbf{p}}\left[p_ip_j\right]=0\cdot P\left(p_ip_j=0\right)+1\cdot P\left(p_ip_j=1\right)=\begin{cases}0\cdot\frac{3}{4}+1\cdot\frac{1}{4}=\frac{1}{4} & \text{若 } i\neq j\\ 0\cdot\frac{1}{2}+1\cdot\frac{1}{2}=\frac{1}{2} & \text{若 } i=j\end{cases}$$

所以

$$E_{\mathbf{p}}\left[\mathbf{p}\right]=E_{\mathbf{p}}\left[\begin{pmatrix}p_0\\p_1\\p_2\\\vdots\\p_d\end{pmatrix}\right]=\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\\\frac{1}{2}\\\vdots\\\frac{1}{2}\end{pmatrix}=\frac{1}{2}\mathbf{1}$$

$$
\begin{aligned}
E_{\mathbf{p}}\left[\mathbf{p}\mathbf{p}^T\right] &= E_{\mathbf{p}}\left[\begin{pmatrix}p_0p_0 & p_0p_1 & p_0p_2 & \cdots & p_0p_d\\ p_1p_0 & p_1p_1 & p_1p_2 & \cdots & p_1p_d\\ p_2p_0 & p_2p_1 & p_2p_2 & \cdots & p_2p_d\\ \vdots & \vdots & \vdots & & \vdots\\ p_dp_0 & p_dp_1 & p_dp_2 & \cdots & p_dp_d\end{pmatrix}\right]\\
&= \begin{pmatrix}\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \cdots & \frac{1}{4}\\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \cdots & \frac{1}{4}\\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \cdots & \frac{1}{4}\\ \vdots & \vdots & \vdots & & \vdots\\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \cdots & \frac{1}{2}\end{pmatrix}=\frac{1}{4}\mathbf{1}\mathbf{1}^T+\frac{1}{4}I
\end{aligned}
$$

此外，注意

$$D_{\mathbf{w}}\mathbf{1}=\begin{pmatrix}w_0 & 0 & 0 & \cdots & 0\\ 0 & w_1 & 0 & \cdots & 0\\ 0 & 0 & w_2 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \cdots & w_d\end{pmatrix}\begin{pmatrix}1\\1\\1\\\vdots\\1\end{pmatrix}=\begin{pmatrix}w_0\\w_1\\w_2\\\vdots\\w_d\end{pmatrix}=\mathbf{w}$$

故

$$
\begin{aligned}
& E_{\mathbf{p}}\left[\|\mathbf{y}-X(\mathbf{w}\odot\mathbf{p})\|^2\right] \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T X D_{\mathbf{w}} E_{\mathbf{p}}[\mathbf{p}] + tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}} E_{\mathbf{p}}\left[\mathbf{p}\mathbf{p}^T\right]\right) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T X D_{\mathbf{w}} \cdot \frac{1}{2}\mathbf{1} + tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\left(\frac{1}{4}\mathbf{1}\mathbf{1}^T + \frac{1}{4}I\right)\right) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X D_{\mathbf{w}}\mathbf{1} + tr\left(\frac{1}{4}D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\mathbf{1}\mathbf{1}^T + \frac{1}{4}D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\mathbf{w} + \frac{1}{4}tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\mathbf{1}\mathbf{1}^T\right) + \frac{1}{4}tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= \mathbf{y}^T\mathbf{y} - \left(X^T\mathbf{y}\right)^T\mathbf{w} + \frac{1}{4}tr\left(\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\mathbf{1}\right)\mathbf{1}^T\right) + \frac{1}{4}tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T\mathbf{y} + \frac{1}{4}\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\mathbf{1}\right)^T\mathbf{1} + \frac{1}{4}tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T\mathbf{y} + \frac{1}{4}\mathbf{1}^T D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\mathbf{1} + \frac{1}{4}tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T\mathbf{y} + \frac{1}{4}\left(D_{\mathbf{w}}\mathbf{1}\right)^T X^T X\left(D_{\mathbf{w}}\mathbf{1}\right) + \frac{1}{4}tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T\mathbf{y} + \frac{1}{4}\mathbf{w}^T X^T X\mathbf{w} + \frac{1}{4}tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right)
\end{aligned}
$$

其中，若令 $\|\cdot\|_F$ 爲 Frobenius norm，則

$$
\begin{aligned}
& tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= tr\Big(\left(XD_{\mathbf{w}}\right)^T\left(XD_{\mathbf{w}}\right)\Big) = \|XD_{\mathbf{w}}\|_F^2 \\
&= \left\|\begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1d} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2d} \\ x_{30} & x_{31} & x_{32} & \cdots & x_{3d} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{N0} & x_{N1} & x_{N2} & \cdots & x_{Nd}\end{pmatrix}\begin{pmatrix} w_0 & 0 & 0 & \cdots & 0 \\ 0 & w_1 & 0 & \cdots & 0 \\ 0 & 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & w_d\end{pmatrix}\right\|_F^2 \\
&= \left\|\begin{pmatrix} w_0x_{10} & w_1x_{11} & w_2x_{12} & \cdots & w_dx_{1d} \\ w_0x_{20} & w_1x_{21} & w_2x_{22} & \cdots & w_dx_{2d} \\ w_0x_{30} & w_1x_{31} & w_2x_{32} & \cdots & w_dx_{3d} \\ \vdots & \vdots & \vdots & & \vdots \\ w_0x_{N0} & w_1x_{N1} & w_2x_{N2} & \cdots & w_dx_{Nd}\end{pmatrix}\right\|_F^2 \\
&= \sum_{i=0}^{d}\sum_{j=1}^{N}\left(w_ix_{ji}\right)^2 = \sum_{i=0}^{d}\sum_{j=1}^{N}w_i^2x_{ji}^2 = \sum_{i=0}^{d}\left(\sum_{j=1}^{N}x_{ji}^2\right)w_i^2
\end{aligned}
$$

接著，利用 quadratic form，可得

$$
\begin{aligned}
&tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) = \sum_{i=0}^{d}\left(\sum_{j=1}^{N} x_{ji}^2\right) w_i^2 \\
&= \mathbf{w}^T \begin{pmatrix} \sum_{j=1}^N x_{j0}^2 & 0 & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^N x_{j1}^2 & 0 & \cdots & 0 \\ 0 & 0 & \sum_{j=1}^N x_{j2}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \sum_{j=1}^N x_{jd}^2 \end{pmatrix} \mathbf{w} \\
&= \mathbf{w}^T \left(\begin{pmatrix} \sum_{j=1}^N x_{j0}^2 & \sum_{j=1}^N x_{j0}x_{j1} & \sum_{j=1}^N x_{j0}x_{j2} & \cdots & \sum_{j=1}^N x_{j0}x_{jd} \\ \sum_{j=1}^N x_{j1}x_{j0} & \sum_{j=1}^N x_{j1}^2 & \sum_{j=1}^N x_{j1}x_{j2} & \cdots & \sum_{j=1}^N x_{j1}x_{jd} \\ \sum_{j=1}^N x_{j2}x_{j0} & \sum_{j=1}^N x_{j2}x_{j1} & \sum_{j=1}^N x_{j2}^2 & \cdots & \sum_{j=1}^N x_{j2}x_{jd} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{j=1}^N x_{jd}x_{j0} & \sum_{j=1}^N x_{jd}x_{j1} & \sum_{j=1}^N x_{jd}x_{j2} & \cdots & \sum_{j=1}^N x_{jd}^2 \end{pmatrix} \odot \right. \\
&\qquad \left. \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \right) \mathbf{w} \\
&= \mathbf{w}^T \left(\left(\begin{pmatrix} x_{00} & x_{10} & x_{20} & \cdots & x_{N0} \\ x_{01} & x_{11} & x_{21} & \cdots & x_{N1} \\ x_{02} & x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{0d} & x_{1d} & x_{2d} & \cdots & x_{Nd} \end{pmatrix} \begin{pmatrix} x_{00} & x_{01} & x_{02} & \cdots & x_{0d} \\ x_{10} & x_{11} & x_{12} & \cdots & x_{1d} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{N0} & x_{N1} & x_{N2} & \cdots & x_{Nd} \end{pmatrix} \right) \odot \right. \\
&\qquad \left. \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \right) \mathbf{w} \\
&= \mathbf{w}^T \left(\left(X^T X \right) \odot I \right) \mathbf{w}
\end{aligned}
$$

因此可得

$$
\begin{aligned}
&E_{\mathbf{p}}\left[\|\mathbf{y} - X(\mathbf{w}\odot\mathbf{p})\|^2\right] \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T \mathbf{y} + \frac{1}{4}\mathbf{w}^T X^T X \mathbf{w} + \frac{1}{4} tr\left(D_{\mathbf{w}}^T X^T X D_{\mathbf{w}}\right) \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T \mathbf{y} + \frac{1}{4}\mathbf{w}^T X^T X \mathbf{w} + \frac{1}{4} \mathbf{w}^T \left(\left(X^T X\right) \odot I\right) \mathbf{w}
\end{aligned}
$$

$$= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T \mathbf{y} + \frac{1}{4}\mathbf{w}^T \left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{w}$$

所以

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{w}} E_{\mathbf{p}} \left[\|\mathbf{y} - X(\mathbf{w} \odot \mathbf{p})\|^2\right] \\
=& \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T \mathbf{y} + \frac{1}{4}\mathbf{w}^T \left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{w}\right) \\
=& -X^T\mathbf{y} + \frac{1}{2}\left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{w}
\end{aligned}$$

若令

$$\frac{\partial}{\partial \mathbf{w}} E_{\mathbf{p}} \left[\|\mathbf{y} - X(\mathbf{w} \odot \mathbf{p})\|^2\right] = -X^T\mathbf{y} + \frac{1}{2}\left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{w} = \mathbf{0}$$

則當 $X^T X + \left(X^T X\right) \odot I$ 爲 invertible 時，可得

$$\mathbf{w} = 2\left(X^T X + \left(X^T X\right) \odot I\right)^{-1} X^T \mathbf{y}$$

而當 $X^T X + \left(X^T X\right) \odot I$ 不爲 invertible 時，因爲 $\forall\ \mathbf{u} \in \mathbb{R}^{d+1}$ 且 $\mathbf{u} \neq \mathbf{0}$，皆有

$$\begin{aligned}
&\mathbf{u}^T \left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{u} \\
=& \mathbf{u}^T X^T X \mathbf{u} + \mathbf{u}^T \left(\left(X^T X\right) \odot I\right) \mathbf{u} \\
=& (X\mathbf{u})^T (X\mathbf{u}) + \mathbf{u}^T \left(\left(X^T X\right) \odot I\right) \mathbf{u} \\
=& \|X\mathbf{u}\|^2 + \mathbf{u}^T \left(\left(X^T X\right) \odot I\right) \mathbf{u} \\
\geq& \mathbf{u}^T \left(\left(X^T X\right) \odot I\right) \mathbf{u}
\end{aligned}$$

其中，由上可知 $\mathbf{u}^T \left(\left(X^T X\right) \odot I\right) \mathbf{u} = \sum_{i=0}^{d}\sum_{j=1}^{N} u_i^2 x_{ji}^2$，因此可得

$$\begin{aligned}
&\mathbf{u}^T \left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{u} \\
\geq& \mathbf{u}^T \left(\left(X^T X\right) \odot I\right) \mathbf{u} = \sum_{i=0}^{d}\sum_{j=1}^{N} u_i^2 x_{ji}^2 \geq 0
\end{aligned}$$

所以 $X^T X + \left(X^T X\right) \odot I$ 爲 positive semi-definite，因此 $E_{\mathbf{p}} \left[\|\mathbf{y} - X(\mathbf{w} \odot \mathbf{p})\|^2\right]$ $= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T X^T \mathbf{y} + \frac{1}{4}\mathbf{w}^T \left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{w}$ 爲 convex quadratic function，故其必定存在 optimal solution，意即 $\frac{\partial}{\partial \mathbf{w}} E_{\mathbf{p}} \left[\|\mathbf{y} - X(\mathbf{w} \odot \mathbf{p})\|^2\right] =$ $-X^T\mathbf{y} + \frac{1}{2}\left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{w} = \mathbf{0}$ 必定有解，因此可取

$$\mathbf{w} = 2\left(X^T X + \left(X^T X\right) \odot I\right)^{\dagger} X^T \mathbf{y}$$

其爲 $\frac{\partial}{\partial \mathbf{w}} E_{\mathbf{p}} \left[\|\mathbf{y} - X(\mathbf{w} \odot \mathbf{p})\|^2\right] = -X^T\mathbf{y} + \frac{1}{2}\left(X^T X + \left(X^T X\right) \odot I\right) \mathbf{w} = \mathbf{0}$ 一個解。

Aggregation

11. 當 g_1、g_2、g_3 之中至少有兩者都對某一筆資料分類錯誤時，G 才會將該筆資料分類錯誤，因此，當 g_1、g_2、g_3 分類錯誤的資料皆不相同時 (如 **Figure. 1** 所示)，G 會將所有的資料皆分類正確，此時可得 $E_{out}(G)$ 的最小值爲 0，而因爲 g_3 的 error 最高，且 $E_{out}(g_3) > E_{out}(g_1) + E_{out}(g_2)$，因此若要讓 $E_{out}(G)$ 有最大值，則必須 g_1 和 g_2 分類錯誤的資料皆不同，且被 g_1 或 g_2 分類錯誤的資料也要被 g_3 分類錯誤 (如 **Figure. 2** 所示)，此時可得 $E_{out}(G)$ 的最大值爲 $E_{out}(g_1) + E_{out}(g_2) = 0.08 + 0.16 = 0.24$，故 $0 \leq E_{out}(G) \leq 0.24$。

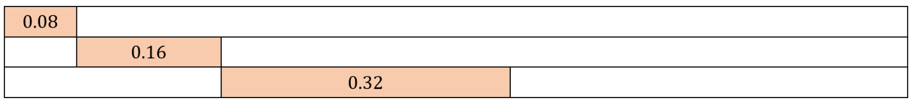


Figure. 1

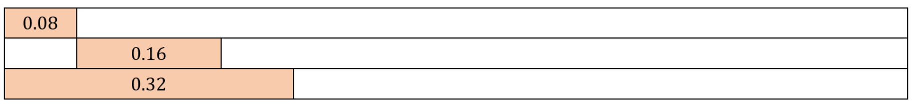


Figure. 2

12. 由於 K 爲奇數，因此當 g_1、g_2、$\cdots$、g_K 之中至少有 $\frac{K+1}{2}$ 個 classifier 都對某一筆資料分類錯誤時，G 才會將該筆資料分類錯誤。令資料的數量爲 N，則 g_k 所產生的分類錯誤總共有 $e_k N$ 個，因此 g_1、g_2、$\cdots$、g_K 所產生的分類錯誤總共有 $\sum_{k=1}^{K} e_k N$ 個，當被 G 分類正確的資料都會被 g_1、g_2、$\cdots$、g_K 分類正確，而被 G 分類錯誤的資料，其僅會被 g_1、g_2、$\cdots$、g_K 之中的 $\frac{K+1}{2}$ 個 classifier 分類錯誤，此時可得會被 G 分類錯誤的資料數目上限爲

$$\frac{\sum_{k=1}^{K} e_k N}{\frac{K+1}{2}} = \frac{2}{K+1} \sum_{k=1}^{K} e_k N$$

因此可得 $E_{out}(G)$ 的一個上限爲

$$\frac{\frac{2}{K+1} \sum_{k=1}^{K} e_k N}{N} = \frac{2}{K+1} \sum_{k=1}^{K} e_k$$

13. 首先，$\forall\ a>0$，因爲

$$
\begin{aligned}
ln\left(\lim_{N\to\infty}\left(1-\frac{a}{N}\right)^{pN}\right) &= \lim_{N\to\infty} ln\left(1-\frac{a}{N}\right)^{pN}\ \ (since\ ln(x)\ is\ continuous)\\
&= \lim_{n\to\infty} pNln\left(1-\frac{a}{N}\right)\\
&= \lim_{N\to\infty}\frac{ln\left(1-\frac{a}{N}\right)}{\frac{1}{pN}}\ \left(indeterminate\ form\ \frac{0}{0}\right)\\
&= \lim_{N\to\infty}\frac{\frac{1}{1-\frac{a}{N}}\cdot\frac{a}{N^2}}{-\frac{1}{pN^2}}\ \ (L'Hopital's\ rule)\\
&= \lim_{N\to\infty}\frac{-ap}{1-\frac{a}{N}} = -ap
\end{aligned}
$$

所以

$$\lim_{N\to\infty}\left(1-\frac{a}{N}\right)^{pN} = e^{-ap}$$

因此，$\forall\ a>0$，當 N 足夠大時，便有 $\left(1-\frac{a}{N}\right)^{pN}\approx e^{-ap}$。接著開始說明第 13 題。若令每一次 sample 的結果之間互爲 independent，則對每一筆資料而言，其在 sample $N'=pN$ 次之後完全沒有被 sample 到的機率爲 $\left(\frac{N-1}{N}\right)^{pN}=\left(1-\frac{1}{N}\right)^{pN}\approx e^{-p}$，因此該筆資料至少有被 sample 到一次的機率大約爲 $1-e^{-p}$。接著，$\forall\ i\in\{1,2,\cdots,N\}$，令 X_i 爲第 i 筆資料是否有被 sample 到的 random variable，$X_i=1$ (成功) 代表第 i 筆資料至少有被 sample 到一次，而 $X_i=0$ (失敗) 代表第 i 筆資料完全沒有被 sample 到，由上可知，$P(X_i=1)\approx 1-e^{-p}$，$P\left(X_i=0\right)\approx e^{-p}$，注意 $\forall\ k\in\mathbb{N}$ 且 $2\le k\le N$，考慮任意 k 個 random variable X_{i_1}、X_{i_2}、$\cdots$、X_{i_k}，令 x_{i_1}、x_{i_2}、$\cdots$、$x_{i_k}\in\{0,1\}$，並令 s 爲其中成功的次數，t 爲其中失敗的次數，因爲

$$P(X_{i_1}=x_{i_1})P(X_{i_2}=x_{i_2})\cdots P(X_{i_k}=x_{i_k})\approx(1-e^{-p})^s(e^{-p})^t$$

而由 inclusion-exclusion principle，可得

$$
\begin{aligned}
&P(X_{i_1}=x_{i_1},X_{i_2}=x_{i_2},\cdots,X_{i_k}=x_{i_k})\\
&= \left(\frac{N-t}{N}\right)^{pN}-\binom{s}{1}\left(\frac{N-(t+1)}{N}\right)^{pN}+\binom{s}{2}\left(\frac{N-(t+2)}{N}\right)^{pN}-\cdots+\\
&\quad(-1)^s\binom{s}{s}\left(\frac{N-(t+s)}{N}\right)^{pN}\\
&= \left(1-\frac{t}{N}\right)^{pN}-\binom{s}{1}\left(1-\frac{t+1}{N}\right)^{pN}+\binom{s}{2}\left(1-\frac{t+2}{N}\right)^{pN}-\cdots+\\
&\quad(-1)^s\binom{s}{s}\left(1-\frac{t+s}{N}\right)^{pN}\\
&\approx e^{-tp}-\binom{s}{1}e^{-(t+1)p}+\binom{s}{2}e^{-(t+2)p}-\cdots+(-1)^s\binom{s}{s}e^{-(t+s)p}
\end{aligned}
$$

$$
\begin{aligned}
&= e^{-tp}\left(1-\begin{pmatrix} s \\ 1 \end{pmatrix} e^{-p}+\begin{pmatrix} s \\ 2 \end{pmatrix} e^{-2p}-\cdots+(-1)^s\begin{pmatrix} s \\ s \end{pmatrix} e^{-sp}\right) \\
&= e^{-tp}(1-e^{-p})^s
\end{aligned}
$$

故 $P(X_{i_1}=x_{i_1}, X_{i_2}=x_{i_2}, \cdots, X_{i_k}=x_{i_k}) \approx P(X_{i_1}=x_{i_1})P(X_{i_2}=x_{i_2})\cdots P(X_{i_k}=x_{i_k})$，意即當 N 足夠大時，X_1、X_2、$\cdots$、X_N 爲 independent random variable，因此此時可以將檢驗每一筆資料是否有被 sample 過的過程視爲 binomial experiment，而平均來說至少有被 sample 過一次的資料數量，即成功次數的期望值，爲 $(1-e^{-p})N = N-(e^{-p}\cdot N)$。

Kernel for Decision Stumps

14. 首先，考慮在 s 和 i 固定的情況下，由於 $\forall\, \mathbf{x}\in\mathcal{X}$，$x_i$ 皆爲 integer，因此當 θ 和 $\tilde{\theta}$ 皆在 $(-\infty, L]$、$(L, L+1]$、$(L+1, L+2]$、$\cdots$、$(R-2, R-1]$、$(R-1, R]$、(R, ∞) 之中某一個相同的 interval 時，$g_{s,i,\theta}(\mathbf{x}) = s\cdot sign(x_i-\theta)$ 和 $g_{s,i,\tilde{\theta}}(\mathbf{x}) = s\cdot sign(x_i-\tilde{\theta})$ 的值便會相同，即 $g_{s,i,\theta}$ 和 $g_{s,i,\tilde{\theta}}$ 爲相同的 decision stump，而當 θ 和 $\tilde{\theta}$ 在 $(-\infty, L]$、$(L, L+1]$、$(L+1, L+2]$、$\cdots$、$(R-2, R-1]$、$(R-1, R]$、(R, ∞) 之中不同的 interval 時，便 $\exists\, \mathbf{x}\in\mathcal{X}$ 且 $x_i\in(min(\theta,\tilde{\theta}), max(\theta,\tilde{\theta}))$，使得 $g_{s,i,\theta}(\mathbf{x}) = s\cdot sign(x_i-\theta)$ 和 $g_{s,i,\tilde{\theta}}(\mathbf{x}) = s\cdot sign(x_i-\tilde{\theta})$ 的值不同，即 $g_{s,i,\theta}$ 和 $g_{s,i,\tilde{\theta}}$ 爲不同的 decision stump，由此可知，在 s 和 i 固定的情況下，decision stump 的數量會和 interval $(-\infty, L]$、$(L, L+1]$、$(L+1, L+2]$、$\cdots$、$(R-2, R-1]$、$(R-1, R]$、(R, ∞) 的數量相同，爲 $R-L+2$ 個。接著，由於 $s\in\{+1,-1\}$ 有 2 個不同的值，$i\in\{1,2,\cdots,d\}$ 有 d 個不同的值，因此不同的 decision stump 數量至多爲 $2d(R-L+2)$ 個，但是當 $s=+1$ 且 $\theta\in(-\infty, L]$，以及 $s=-1$ 且 $\theta\in(R,\infty)$ 時，$\forall\, \mathbf{x}\in\mathcal{X}$，皆有 $g_{s,i,\theta}(\mathbf{x}) = s\cdot sign(x_i-\theta) = +1$，因此這 $2d$ 個 decision stump 事實上皆爲相同的 decision stump，故多算了 $2d-1$ 個 decision stump，同理，當 $s=-1$ 且 $\theta\in(-\infty, L]$，以及 $s=+1$ 且 $\theta\in(R,\infty)$ 時，$\forall\, \mathbf{x}\in\mathcal{X}$，皆有 $g_{s,i,\theta}(\mathbf{x}) = s\cdot sign(x_i-\theta) = -1$，因此這 $2d$ 個 decision stump 事實上皆爲相同的 decision stump，故多算了 $2d-1$ 個 decision stump，因此可得不同的 decision stump 數量應爲

$$2d(R-L+2) - 2\times(2d-1) = 2d(R-L)+2$$

故當 $d=4$、$L=0$、$R=5$ 時，不同的 decision stump 數量爲

$$2\times 4\times(5-0)+2 = 42$$

15.

$$K_{ds}(\mathbf{x},\mathbf{x}') = (\phi_{ds}(\mathbf{x}))^T(\phi_{ds}(\mathbf{x}')) = \sum_{g_{s,i,\theta}\in\mathcal{G}} g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}')$$

令 $m_i = min(x_i, x'_i)$，$M_i = max(x_i, x'_i)$。首先，在 s 和 i 固定的情況下，由第 14 題可知，此時有 $R-L+2$ 個不同的 $g_{s,i,\theta}$，其中，當 θ 屬於 $(m_i, m_i+1]$、$(m_i+1, m_i+2]$、$\cdots$、$(M_i-2, M_i-1]$、$(M_i-1, M_i]$ 這 $M_i - m_i$ 個 interval 之中的其中一個時，$g_{s,i,\theta}(\mathbf{x}) = s \cdot sign(x_i - \theta)$ 和 $g_{s,i,\theta}(\mathbf{x}') = s \cdot sign(x'_i - \theta)$ 異號，意即會有 $M_i - m_i$ 個不同的 $g_{s,i,\theta}$ 使得 $g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}') = -1$，而當 θ 不屬於 $(m_i, m_i+1]$、$(m_i+1, m_i+2]$、$\cdots$、$(M_i-2, M_i-1]$、$(M_i-1, M_i]$ 這 $M_i - m_i$ 個 interval 之中的任何一個時，$g_{s,i,\theta}(\mathbf{x}) = s \cdot sign(x_i - \theta)$ 和 $g_{s,i,\theta}(\mathbf{x}') = s \cdot sign(x'_i - \theta)$ 同號，意即會有 $(R-L+2) - (M_i - m_i)$ 個不同的 $g_{s,i,\theta}$ 使得 $g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}') = +1$，因此在 s 和 i 固定的情況下，將 $g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}')$ 對 θ 作 summmation 後可得

$$\begin{aligned}&(M_i - m_i) \times (-1) + ((R-L+2) - (M_i - m_i)) \times (+1)\\ =\ &(R-L+2) - 2(M_i - m_i) = (R-L+2) - 2|x_i - x'_i|\end{aligned}$$

接著，將上式對 s 和 i 作 summation，可得

$$\begin{aligned}&\sum_{s\in\{+1,-1\}} \sum_{i=1}^{d} ((R-L+2) - 2|x_i - x'_i|)\\ =\ &\sum_{s\in\{+1,-1\}} \left(d(R-L+2) - 2\sum_{i=1}^{d} |x_i - x'_i| \right)\\ =\ &2\left(d(R-L+2) - 2\sum_{i=1}^{d} |x_i - x'_i| \right)\\ =\ &2d(R-L+2) - 4\sum_{i=1}^{d} |x_i - x'_i|\end{aligned}$$

令 $\|\cdot\|_1$ 爲 L1-norm，則上式可以寫爲

$$2d(R-L+2) - 4\|\mathbf{x} - \mathbf{x}'\|$$

但是當 $s = +1$ 且 $\theta \in (-\infty, L]$，以及 $s = -1$ 且 $\theta \in (R, \infty)$ 時，$\forall\ \mathbf{x} \in \mathcal{X}$，皆有 $g_{s,i,\theta}(\mathbf{x}) = s \cdot sign(x_i - \theta) = +1$，因此這 $2d$ 個 $g_{s,i,\theta}$ 事實上皆爲相同的 decision stump，故上式多算了 $2d-1$ 個 $g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}') = (+1) \times (+1) = 1$，同理，當 $s = -1$ 且 $\theta \in (-\infty, L]$，以及 $s = +1$ 且 $\theta \in (R, \infty)$ 時，$\forall\ \mathbf{x} \in \mathcal{X}$，皆有 $g_{s,i,\theta}(\mathbf{x}) = s \cdot sign(x_i - \theta) = -1$，因此這 $2d$ 個 $g_{s,i,\theta}$ 事實上皆爲相同的 decision stump，故上式多算了 $2d-1$ 個 $g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}') = (-1) \times (-1) = 1$ ，因此可得

$$\begin{aligned}K_{ds}(\mathbf{x}, \mathbf{x}') &= \sum_{g_{s,i,\theta}\in\mathcal{G}} g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}')\\ &= 2d(R-L+2) - 4\|\mathbf{x} - \mathbf{x}'\| - 2 \times (2d-1)\\ &= 2d(R-L) + 2 - 4\|\mathbf{x} - \mathbf{x}'\|\end{aligned}$$

16. 首先，在 s 和 i 固定的情況下，若 θ、$\tilde{\theta} \in [L, R]$，則當 $\theta \neq \tilde{\theta}$ 時，$\exists\, \mathbf{x} \in \mathcal{X}$ 且 x_i 介於 θ 和 $\tilde{\theta}$ 之間，使得 $g_{s,i,\theta}(\mathbf{x}) = s \cdot sign(x_i - \theta)$ 和 $g_{s,i,\tilde{\theta}}(\mathbf{x}) = s \cdot sign(x_i - \tilde{\theta})$ 的值不同，即 $g_{s,i,\theta}$ 和 $g_{s,i,\tilde{\theta}}$ 爲不同的 decision stump，而若 $\theta \in (-\infty, L)$，則 $\forall\, \mathbf{x} \in \mathcal{X}$，$g_{s,i,\theta}(\mathbf{x})$ 和 $g_{s,i,L}(\mathbf{x})$ 的值皆爲 $+s$，即 $g_{s,i,\theta}$ 和 $g_{s,i,L}$ 爲相同的 decision stump，而若 $\theta \in (R, \infty)$，則 $\forall\, \mathbf{x} \in \mathcal{X}$，$g_{s,i,\theta}(\mathbf{x})$ 和 $g_{s,i,R+1}(\mathbf{x})$ 的值皆爲 $-s$，即 $g_{s,i,\theta}$ 和 $g_{s,i,R+1}$ 爲相同的 decision stump，由上可知，在 s 和 i 固定的情況下，所有的 decision stump 皆可以由 $\theta \in [L, R] \cup \{R+1\}$ 來唯一決定，因此，將 $g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}')$ 對 θ 作 integral 時，integral region 爲 $[L, R] \cup \{R+1\}$ (注意 $\{R+1\}$ 在 $\mathbb{R}$ 中爲 measure zero，因此在 $\theta = R+1$ 上的 integral 爲 0，可以不用考慮在 $\theta = R+1$ 上的 integral)，可得

$$
\begin{aligned}
& \int_L^R g_{s,i,\theta}(\mathbf{x})g_{s,i,\theta}(\mathbf{x}')d\theta \\
= & \int_L^R \left(s \cdot sign(x_i - \theta)\right)\left(s \cdot sign(x_i' - \theta)\right)d\theta \\
= & \int_L^R sign(x_i - \theta)sign(x_i' - \theta)d\theta
\end{aligned}
$$

其中，令 $m_i = min(x_i, x_i')$、$M_i = max(x_i, x_i')$，當 $\theta \in [m_i, M_i)$ 時，可得 $sign(x_i - \theta)$ 和 $sign(x_i' - \theta)$ 同號，即 $sign(x_i - \theta)sign(x_i' - \theta) = +1$，而當 $\theta \notin [m_i, M_i)$ 時，可得 $sign(x_i - \theta)$ 和 $sign(x_i' - \theta)$ 異號，即 $sign(x_i - \theta)sign(x_i' - \theta) = -1$，因此上式等於

$$
\begin{aligned}
& \int_L^R sign(x_i - \theta)sign(x_i' - \theta)d\theta \\
= & \int_L^{m_i} sign(x_i - \theta)sign(x_i' - \theta)d\theta + \\
& \int_{m_i}^{M_i} sign(x_i - \theta)sign(x_i' - \theta)d\theta + \\
& \int_{M_i}^R sign(x_i - \theta)sign(x_i' - \theta)d\theta
\end{aligned}
$$

$$
\begin{aligned}
= & \int_L^{m_i}(+1)d\theta + \int_{m_i}^{M_i}(-1)d\theta + \int_{M_i}^R(+1)d\theta \\
= & (m_i - L) - (M_i - m_i) + (R - M_i) \\
= & (R - L) - 2(M_i - m_i) \\
= & (R - L) - 2|x_i - x_i'|
\end{aligned}
$$

接著，將上式對 s 和 i 作 summation，可得

$$\begin{aligned}
&\sum_{s\in\{+1,-1\}}\sum_{i=1}^{d}((R-L)-2|x_i-x_i'|)\\
&=\sum_{s\in\{+1,-1\}}\left(d(R-L)-2\sum_{i=1}^{d}|x_i-x_i'|\right)\\
&=2d(R-L)-4\sum_{i=1}^{d}|x_i-x_i'|
\end{aligned}$$

令 $\|\cdot\|_1$ 爲 L1-norm，則上式可以寫爲

$$2d(R-L)-4\|\mathbf{x}-\mathbf{x}'\|_1$$

因此可得

$$K_{ds}(\mathbf{x},\mathbf{x}')=(\phi(\mathbf{x}))^T(\phi(\mathbf{x}'))=2d(R-L)-4\|\mathbf{x}-\mathbf{x}'\|_1$$

(注意雖然以上過程中會和第 14、15 題一樣考慮到重複的 decision stump，但由於這些重複的 decision stump 個數爲 finite，因此在 integral 時其爲 measure zero set，也因此不會影響到之後的 summation，故對以上的結果沒有影響)

Yes, A Lighter Homework :-)

17. 其實老師講授的每一個 Lecture 我都很喜歡，因爲雖然我之前已經有修過其它的機器學習課程，可能有些內容之前已經學過了，但是老師所講授的內容與方式卻是最深入淺出的，老師儘可能地用淺顯易懂而清晰的方式來授課，讓人很能掌握並深入課程所講授的知識，而且在機器學習基石與技法的課程中，老師有講到許多其它課程中所沒有講授或不同的觀點，因此讓我在這兩個課程之中收穫頗豐，尤其是 Lecture 7 : Blending and Bagging 之中，我更是從老師的課程中學到了和以往不同的解釋與觀點，因此是我最喜歡的一個 Lecture。

18. 其實老師講授的每一個 Lecture 我都很喜歡，硬要說的話大概是今年新增的課程內容 Activation in Deep Learning 和 Initialization/Optimization in Deep Learning 等部分吧，雖然老師有上傳手寫的 note 到課程網站方便同學複習，但個人還是比較喜歡看課程投影片來複習，因此有些遺憾，不過整體而言這學期的課程眞的非常充實，眞的萬分感謝老師的授課！