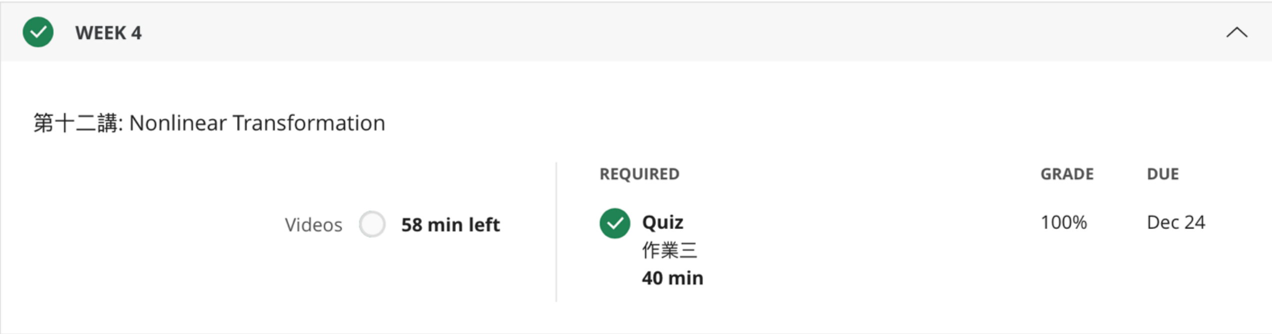
1.



2.

SGD:

$$err(W_t)\begin{cases} = 0 & (y_n W_t^T x_n > 0) \\ = -y_n W_t^T x_n & (y_n W_t^T x_n \leq 0) \end{cases}$$

$$\nabla err(W_t)\begin{cases} = 0 & (sign(W_t^T x_n) = y_n) \\ = -y_n x_n & (sign(W_t^T x_n) \neq y_n) \end{cases}$$

$$\therefore\ W_{t+1} = W_t - \nabla err(W_t)$$
$$= W_t - [\![sign(W_t^T x_n) \neq y_n]\!](-y_n x_n)$$
$$= W_t + [\![sign(W_t^T x_n) \neq y_n]\!](y_n x_n)$$

PLA:

$$W_{t+1} = W_t + [\![sign(W_t^T x_n) \neq y_n]\!](y_n x_n)$$

$\therefore\ err(W_t) = max(0, -y_n W_t^T x_n)$ results in PLA

3.

$$\max_h\ likelihood\,(h) \propto \prod_{n=1}^{N} h_{y_n}(x_n)$$

$$= \max_h\ ln \textstyle\prod_{n=1}^{N} h_{y_n}(x_n)$$

$$= \min_h \frac{1}{N}\sum_{n=1}^{N} -ln\big(h_{y_n}(x_n)\big)$$

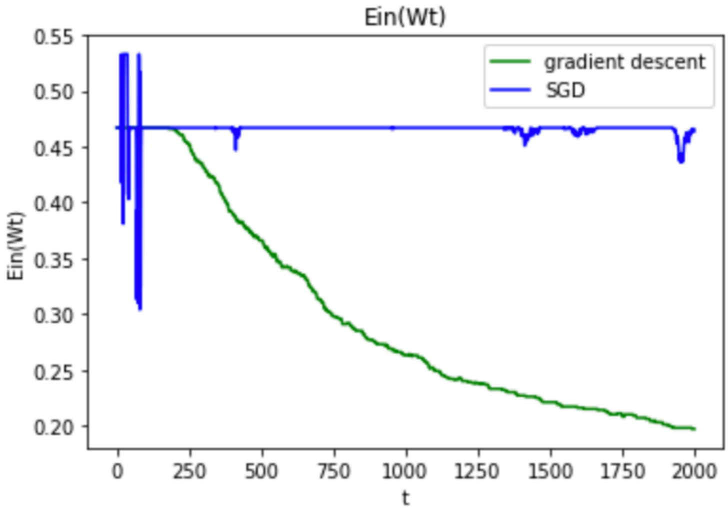
$$\therefore\ E_{in} = \tfrac{1}{N}\textstyle\sum_{n=1}^{N} -ln\big(h_{y_n}(x_n)\big)$$
$$= \tfrac{1}{N}\textstyle\sum_{n=1}^{N} -ln\big(exp(W_{y_n}^T x_n)/(\sum_{k=1}^{K} exp(W_k^T x_n))\big)$$
$$= \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\left(ln(\sum_{k=1}^{K} exp(W_k^T x_n)) - ln\left(exp(W_{y_n}^T x_n)\right)\right)$$
$$= \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\big(ln(\sum_{k=1}^{K} exp(W_k^T x_n)) - W_{y_n}^T x_n\big)$$

$$\therefore\ \frac{\partial E_{in}}{\partial W_i} = \frac{1}{N}\sum_{n=1}^{N}\left[\left(x_n \cdot exp(W_i^T x_n)\middle/\left(\sum_{k=1}^{K} exp(W_k^T x_n)\right)\right) - [\![y_n = i]\!]x_n\right]$$

$$\because\ h_i(x_n) = exp(W_i^T x_n)\middle/\left(\sum_{k=1}^{K} exp(W_k^T x_n)\right)$$

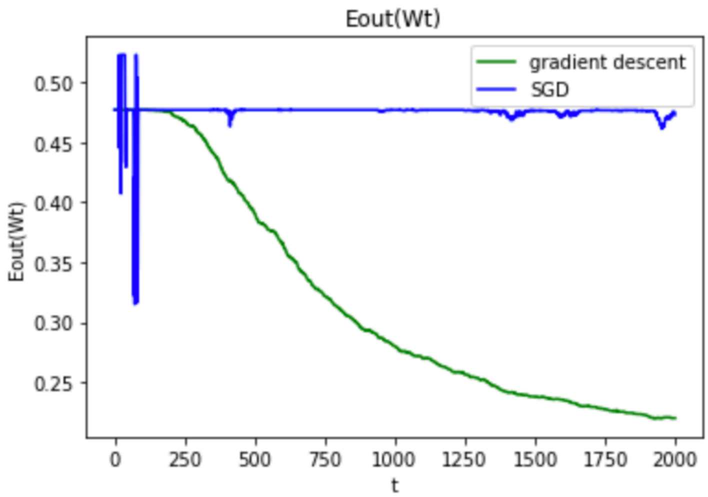
$$\therefore \frac{\partial E_{in}}{\partial W_i} = \frac{1}{N}\sum_{n=1}^{N}[x_n \cdot h_i(x_n) - [\![y_n = i]\!]x_n]$$
$$= \frac{1}{N}\sum_{n=1}^{N}[(h_i(x_n) - [\![y_n = i]\!])x_n]$$

4.



上圖中藍色和綠色的曲線分別為 gradient descent(GD)和 SGD 的 Ein，GD 的 Ein 隨 t 逐步下降，但是 SGD 的 Ein 卻有明顯的上下震盪。SGD 每次僅利用一個 data point 的梯度更新參數，梯度方向不一定是 loss function 最小的方向，這讓 SGD 的梯度稍稍偏離「正軌」，導致 Ein 並不會隨著 t 穩定下降。因為 GD 和 SGD 的 learning rate 分別是 0.01 和 0.001，learning rate 不同導致他們最後的 Ein 也不同，較大的 learning rate 可以讓最後的 Ein 更小。

5.



上圖中藍色和綠色的曲線分別為 gradient descent(GD)和 SGD 的 Eout, 在 VC bound 的保證下，無論 SGD 還是 GD，與之對應的 Ein 和 Eout 都很相近。GD 的 Eout 隨 t 逐步下降，最後的 Eout 約為 0.22，略高於其 Ein。SGD 的 Eout 有明顯的上下震盪，與其 Ein 的振幅相似。

6.

設定：

$$h_k(X) = \begin{bmatrix} h_k(x_1) \\ h_k(x_2) \\ \vdots \\ h_k(x_N) \end{bmatrix} \quad W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$A = \begin{bmatrix} h_1(x_1) & \dots & h_K(x_1) \\ h_1(x_2) & \dots & h_K(x_2) \\ \vdots & & \vdots \\ h_1(x_N) & \dots & h_K(x_N) \end{bmatrix} = [h_1(X) \quad h_2(X) \quad \dots \quad h_K(X)]$$

$$H(X) = [h_1(X) \quad h_2(X) \quad \dots \quad h_K(X)] \cdot W = AW$$

由題意可知：

① $e_0{}^2 = \frac{1}{N}\|Y\|^2$

$$= \frac{1}{N} Y^T Y$$

② $e_k{}^2 = \frac{1}{N}\|Y - h_k(X)\|^2$

$$= \frac{1}{N}\left(Y^T Y + h_k{}^T(X) h_k(X) - 2h_k{}^T(X) Y\right)$$

③ $RMSE^2(H) = \mathrm{E}^2 = \frac{1}{N}\|Y - H(X)\|^2$

$$= \frac{1}{N}(Y^T Y + H^T(X)H(X) - 2H^T(X)Y)$$

將①帶入②得：

$$Ne_k{}^2 = Ne_0{}^2 + h_k{}^T(X)h_k(X) - 2h_k{}^T(X)Y$$

$$h_k{}^T(X)Y = \frac{1}{2}\left(Ne_0{}^2 + h_k{}^T(X)h_k(X) - Ne_k{}^2\right) \cdots\cdots\cdots\cdots ④$$

$\because$ ③$\mathrm{E}^2 = \frac{1}{N}(Y^T Y + H^T(X)H(X) - 2H^T(X)Y)$

$$= \frac{1}{N}\left(Y^T Y + H^T(X)H(X) - 2W^T \begin{bmatrix} —— & h_1{}^T(X) & —— \\ & \vdots & \\ —— & h_K{}^T(X) & —— \end{bmatrix} Y\right)$$

$$= \frac{1}{N}\left(Y^T Y + H^T(X)H(X) - 2W^T \begin{bmatrix} h_1{}^T(X)Y \\ \vdots \\ h_K{}^T(X)Y \end{bmatrix}\right)$$

代入④得：

$$\mathrm{E}^2 = \frac{1}{N}\left(Y^T Y + H^T(X)H(X) - 2W^T \begin{bmatrix} \frac{1}{2}\left(Ne_0{}^2 + h_1{}^T(X)h_1(X) - Ne_1{}^2\right) \\ \vdots \\ \frac{1}{2}\left(Ne_0{}^2 + h_K{}^T(X)h_K(X) - Ne_K{}^2\right) \end{bmatrix}\right)$$

$$= \frac{1}{N}\left(Y^T Y + W^T A^T A W - W^T \begin{bmatrix} Ne_0{}^2 + h_1{}^T(X)h_1(X) - Ne_1{}^2 \\ \vdots \\ Ne_0{}^2 + h_K{}^T(X)h_K(X) - Ne_K{}^2 \end{bmatrix}\right)$$

$$\therefore\ \nabla \mathrm{E}^2(\mathrm{W}) = \frac{1}{N}\left(2A^TAW - \begin{bmatrix} N{e_0}^2 + {h_1}^T(X)h_1(X) - N{e_1}^2 \\ \vdots \\ N{e_0}^2 + {h_K}^T(X)h_K(X) - N{e_K}^2 \end{bmatrix}\right)$$

$$\because\ \mathrm{W}_{min} = \operatorname*{argmin}_W \mathrm{E}^2 = \operatorname*{argmin}_W \mathrm{E}$$

$$\nabla \mathrm{E}^2(\mathrm{W}_{min}) = 0$$

$$\therefore 2A^TA\mathrm{W}_{min} - \begin{bmatrix} N{e_0}^2 + {h_1}^T(X)h_1(X) - N{e_1}^2 \\ \vdots \\ N{e_0}^2 + {h_K}^T(X)h_K(X) - N{e_K}^2 \end{bmatrix} = 0$$

$$\mathrm{W}_{min} = \tfrac{1}{2}(A^TA)^{-1} \begin{bmatrix} N{e_0}^2 + {h_1}^T(X)h_1(X) - N{e_1}^2 \\ \vdots \\ N{e_0}^2 + {h_K}^T(X)h_K(X) - N{e_K}^2 \end{bmatrix}$$

$$= \frac{1}{2}\left(\begin{bmatrix} -- & {h_1}^T(X) & -- \\ & \vdots & \\ -- & {h_K}^T(X) & -- \end{bmatrix} \begin{bmatrix} h_1(X) & \dots & h_K(X) \end{bmatrix}\right)^{-1} \begin{bmatrix} N{e_0}^2 + {h_1}^T(X)h_1(X) - N{e_1}^2 \\ \vdots \\ N{e_0}^2 + {h_K}^T(X)h_K(X) - N{e_K}^2 \end{bmatrix}$$

$$= \frac{1}{2}\left(\begin{bmatrix} {h_1}^T(X)h_1(X) & {h_1}^T(X)h_2(X) & \dots & {h_1}^T(X)h_K(X) \\ {h_2}^T(X)h_1(X) & {h_2}^T(X)h_2(X) & \dots & {h_2}^T(X)h_K(X) \\ \vdots & \vdots & \dots & \vdots \\ {h_K}^T(X)h_1(X) & {h_K}^T(X)h_2(X) & \dots & {h_K}^T(X)h_K(X) \end{bmatrix}\right)^{-1} \left(\begin{bmatrix} {h_1}^T(X)h_1(X) \\ {h_2}^T(X)h_2(X) \\ \vdots \\ {h_K}^T(X)h_K(X) \end{bmatrix} + N\begin{bmatrix} {e_0}^2 - {e_1}^2 \\ {e_0}^2 - {e_2}^2 \\ \vdots \\ {e_0}^2 - {e_K}^2 \end{bmatrix}\right)$$