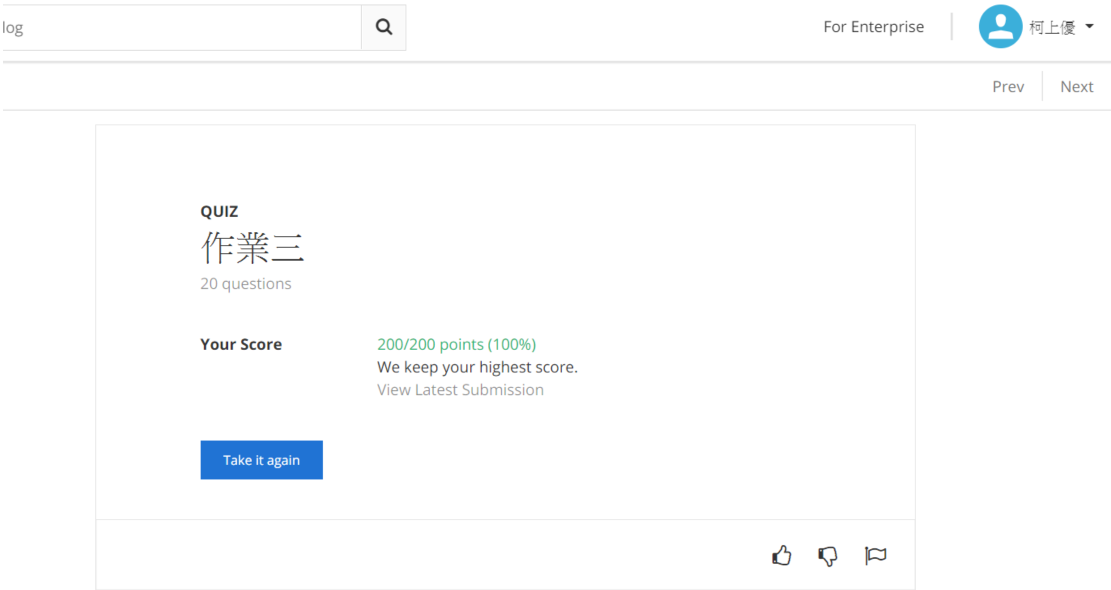
B05902109 資工二 柯上優

1.



2.

首先，證明$H^2 = H$

Pf>

$$\begin{aligned} & H^2 \\ &= (X(X^TX)^{-1}X^T)^2 \\ &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T = H \end{aligned}$$

再來，證明$(I-H)^2 = I - H$

Pf>

$$(I-H)^2 = I^2 - IH - HI + H^2 = I - 2H + H = I - H$$

3.

For PLA:

當 wx 與 y 同號，即 ywx > 0，不更新，即梯度為 0

當 wx 與 y 異號，即 ywx < 0，需要更新，此時梯度為-yx

For max(0, -ywx):

討論 y = +1，

wx > 0，即 ywx > 0，max(0, -ywx)=0，梯度為 0

wx < 0，即 ywx < 0，max(0, -ywx)= -ywx，梯度為-yx

討論 y = -1，

wx > 0，即 ywx < 0，max(0, -ywx)= -ywx，梯度為-yx

wx < 0，即 ywx > 0，max(0, -ywx)=0，梯度為 0

此外，對兩種方法，當 wx = 0，梯度必為 0

得證，err(w) = max(0, −ywx) results in PLA

4.

由 E 的二階泰勒展開式得到 E_2

$$E_2 = \mathrm{E} + \Delta\mathrm{u}\frac{\partial E}{\partial u} + \Delta\mathrm{v}\frac{\partial E}{\partial v} + \frac{1}{2!}(\Delta u^2\frac{\partial^2 E}{\partial u^2} + \Delta\mathrm{u}\Delta\mathrm{v}\frac{\partial^2 E}{\partial u\partial v} + \Delta v^2\frac{\partial^2 E}{\partial v^2})$$

計算其 Δu 與 Δv 的最小值。因為 E_2 對 Δu 與 Δv 微分為零時有最小值，得到

$$\frac{\partial E_2}{\partial \Delta u} = \frac{\partial E}{\partial u} + \Delta\mathrm{u}\frac{\partial^2 E}{\partial u^2} + \Delta\mathrm{v}\frac{\partial^2 E}{\partial u\partial v} = 0$$

$$\frac{\partial E_2}{\partial \Delta v} = \frac{\partial E}{\partial v} + \Delta\mathrm{v}\frac{\partial^2 E}{\partial v^2} + \Delta\mathrm{u}\frac{\partial^2 E}{\partial u\partial v} = 0$$

以矩陣表達

$$\begin{bmatrix} \frac{\partial^2 E}{\partial u^2} & \frac{\partial^2 E}{\partial u\partial v} \\ \frac{\partial^2 E}{\partial u\partial v} & \frac{\partial^2 E}{\partial v^2} \end{bmatrix}\begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -[\begin{matrix} \frac{\partial E}{\partial u} \\ \frac{\partial E}{\partial v} \end{matrix}]$$

也可寫作

$$\nabla^2 E(u,v) * E_2(\Delta u, \Delta v) = -\nabla\mathrm{E}(\mathrm{u},\mathrm{v})$$

可得所求

$$-(\nabla^2 E(u,v))^{-1}\nabla\mathrm{E}(\mathrm{u},\mathrm{v})$$

5.

如同上課所述，先從同樣的地方開始，依次化簡

$$\max\frac{1}{N}\prod_{n=1}^{N} h(x) = min - \frac{1}{N}\prod_{n=1}^{N} h(x) = min - \frac{1}{N}\prod_{n=1}^{N}\ln\big(h(x)\big)$$

$$= min\frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{1}{h(x)}\right) = min\frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{\sum_{i=1}^{K}\exp(w_i^T x_n)}{\exp\big(w_y^T x_n\big)}\right)$$

$$= min\frac{1}{N}\sum_{n=1}^{N}(\ln(\sum_{i=1}^{K}\exp(w_i^T x_n)) - \big(w_y^T x_n\big))$$

所求即

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N}(\ln(\sum_{i=1}^{K}\exp(w_i^T x_n)) - \left(w_y^T x_n\right))$$

6.

對前一題的答案微分

$$\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n \exp(w_i^T x_n)}{\sum_{i=1}^{K}\exp(w_i^T x_n)} - [\![y=i]\!]x_n\right) = \frac{1}{N}\sum_{n=1}^{N}\left(h_i(x_n) - [\![y=i]\!]x_n\right)$$

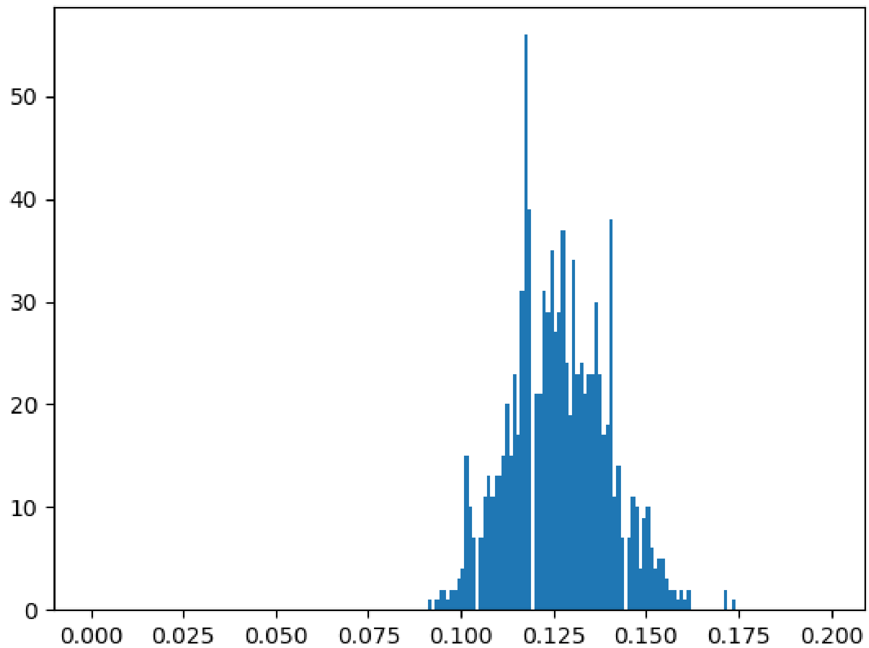
其中，括弧內前半項，由於分子中只有當 w_i^T 是該偏微分項時才會保留，其他項皆會捨去，化簡後以 h_i 取代。

括弧內後半項同理，只有當 w_i^T 是該偏微分項時才會保留，其他項皆會捨去，因此改以判斷式的雙括弧(值為 1 或 0)取代。

7.

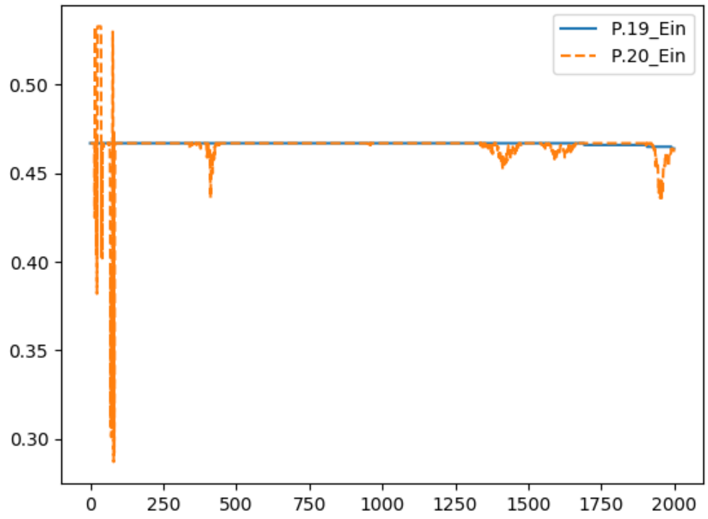
W: -0.9928 0.0013 0.0013 0.0007 1.5623 1.5589

E_{out}: 0.126056

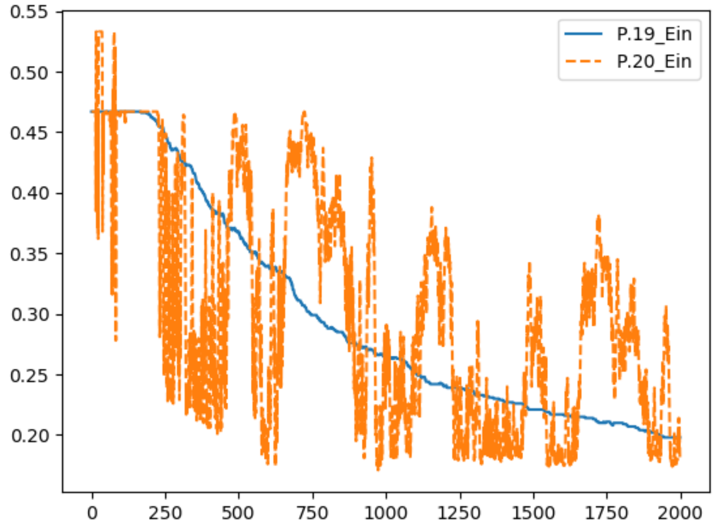


8.

Ita = 0.001



Ita = 0.01



比較 ita = 0.001 和 ita = 0.01，由於 0.001 移動一步的距離較短，變化非常不明顯

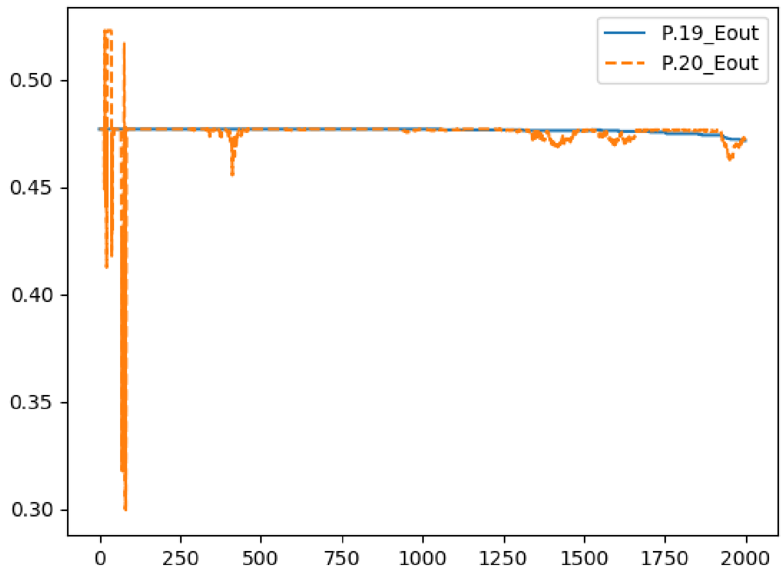
比較 GD 和 SGD，兩者都有往同樣的結果移動，但是過程中

the gradient descent version 藉著取平均梯度，穩定下降

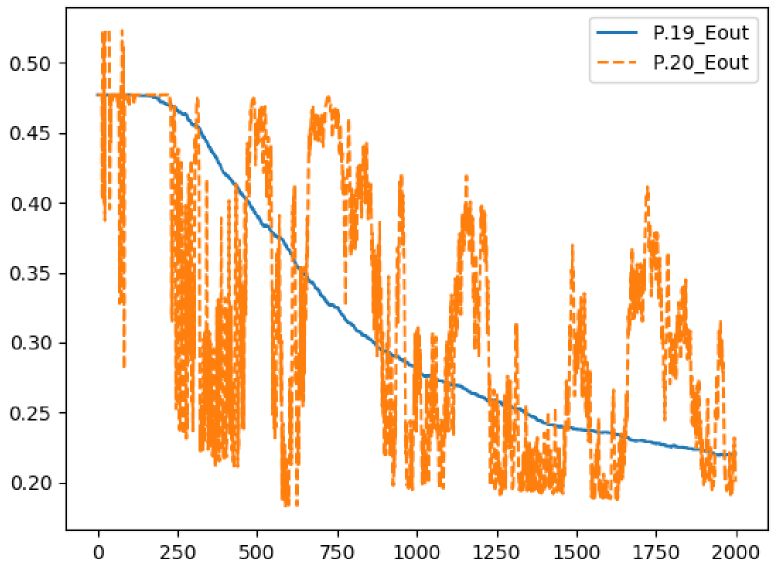
the stochastic gradient descent version 取其中一筆計算梯度，容易在下降時走太多甚至遠離 minima。

9.

Ita = 0.001



Ita = 0.01



和前一題一樣，ita=0.001 時移動較慢變化較不明顯

比較 GD 與 SGD，

the gradient descent version 藉著取平均梯度，穩定下降，Eout 也會穩定下降

the stochastic gradient descent version 取其中一筆計算梯度，容易在下降時走太多甚至遠離 minima，對於 Eout 也會發生不穩定偏移與突然的巨大變動。