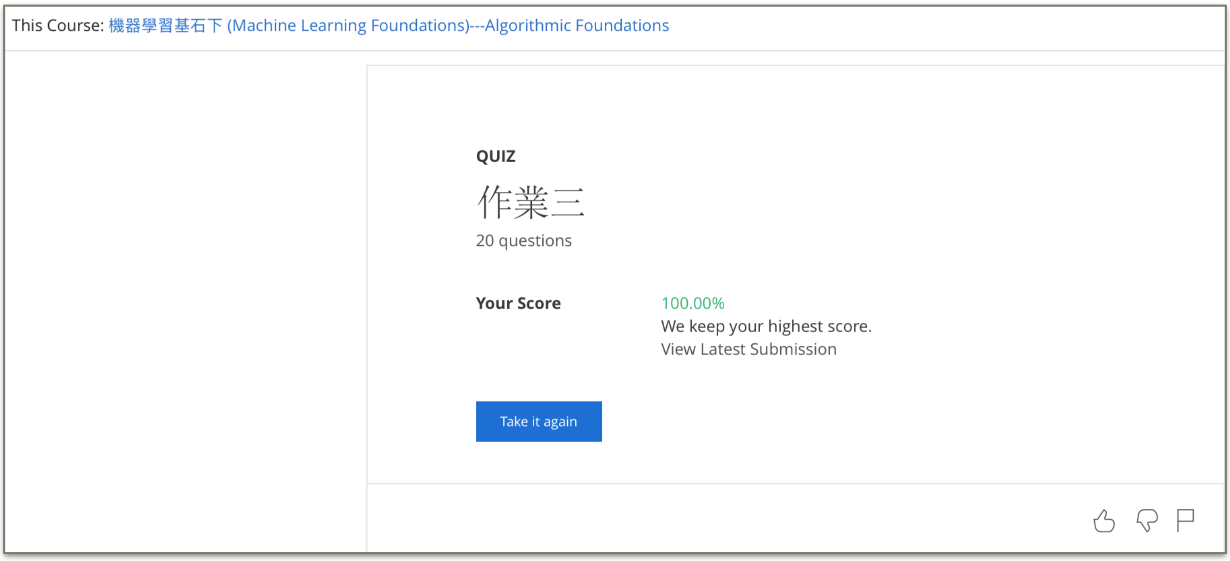
Machine Learning Foundations - *Homework #3*

B05902118 資工三 陳盈如

Problem 1.



Problem 2.

When using SGD on the following error function, prove that $err(w) = \max(0, -yw^Tx)$ results in PLA.

$\because$ In SGD: $w \leftarrow w - \eta \dfrac{\partial\, err(w)}{\partial w}$ / In PLA: $w \leftarrow w + 1 \cdot [\![y_n \neq sign(w^T x_n)]\!] (y_n x_n)$

if $y_n = sign(w^T x_n)$, SGD: $w \leftarrow w - \eta \dfrac{\partial\, 0}{\partial w} = w$

$$\text{PLA: } w \leftarrow w + 1 \cdot [\![y_n \neq sign(w^T x_n)]\!] (y_n x_n) = w + 0(y_n x_n) = w$$

else if $y_n \neq sign(w^T x_n)$, SGD: $w \leftarrow w - \eta \dfrac{\partial\, err(w)}{\partial w} = w - \eta \dfrac{\partial(-yw^Tx)}{\partial w} = w - \eta(-yx) = w + \eta yx$

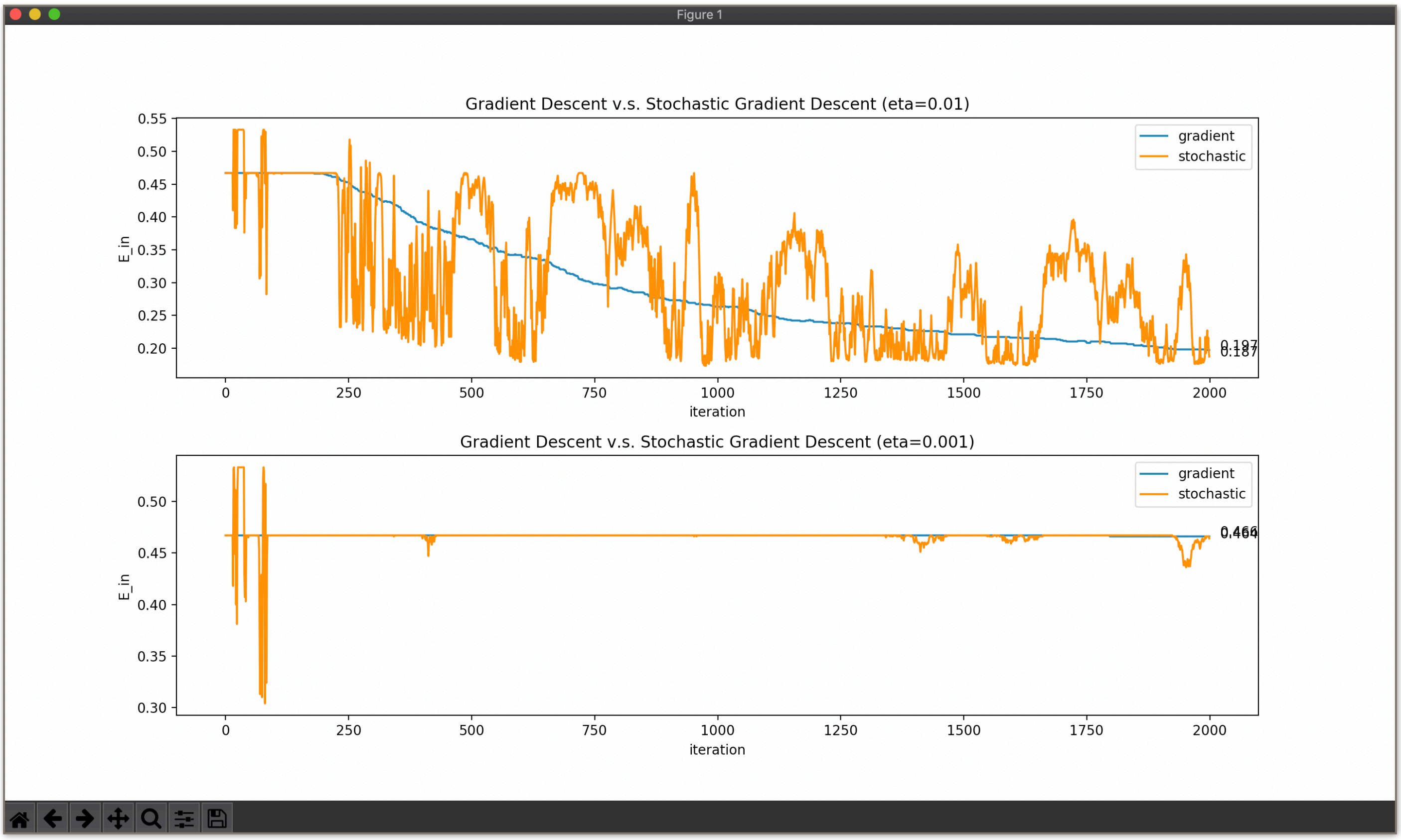
$$\text{PLA: } w \leftarrow w + 1 \cdot [\![y_n \neq sign(w^T x_n)]\!] (y_n x_n) = w + yx$$

$\therefore$ when $\eta = 1$, $err(w) = \max(0, -yw^Tx)$ results in PLA.

Problem 3.

$$
\begin{aligned}
\frac{\partial E_{in}}{\partial w_i} &= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \ln\left(\sum_{k=1}^{K}\exp(w_k^T x_n) - w_{y_n}^T x_n\right)}{\partial w_i}\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \ln\left(\sum_{k=1}^{K}\exp(w_k^T x_n)\right)}{\partial w_i} - \frac{\partial w_{y_n}^T x_n}{\partial w_i}\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \ln\left(\sum_{k=1}^{K}\exp(w_k^T x_n)\right)}{\partial\left(\sum_{k=1}^{K}\exp(w_k^T x_n)\right)}\frac{\partial\left(\sum_{k=1}^{K}\exp(w_k^T x_n)\right)}{\partial w_i} - \frac{\partial w_{y_n}^T x_n}{\partial w_i}\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\exp(w_i^T x_n)}{\sum_{k=1}^{K}\exp(w_k^T x_n)} - [[y_n == i]]x_n\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(h(x_n) - [[y_n == i]]x_n\right)
\end{aligned}
$$

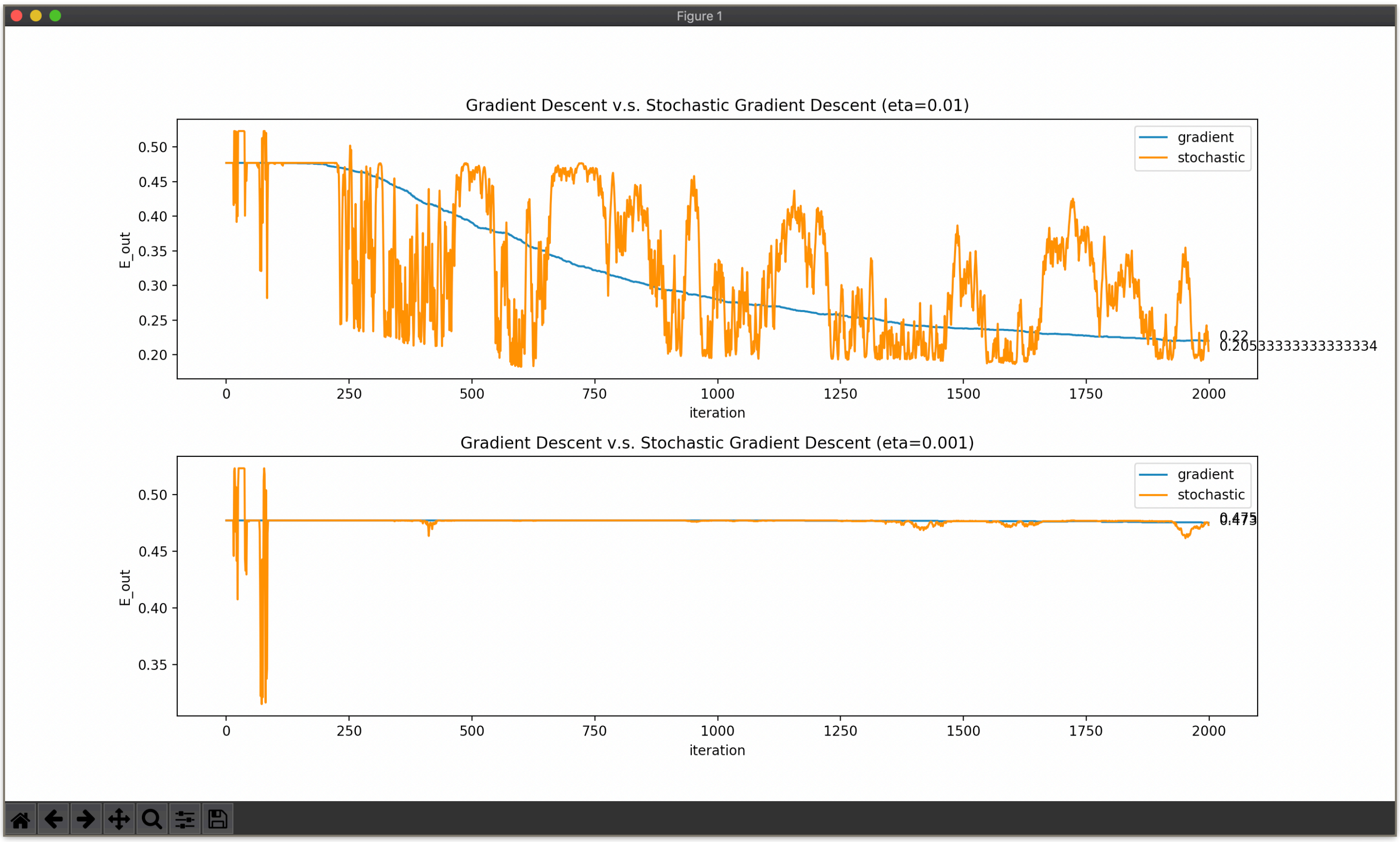
Problem 4.



Findings:

- No matter gradient descent or stochastic gradient descent, $\eta = 0.001$ is too small for both of them to run in 2000 iterations. In comparison, $\eta = 0.01$ is a better assumption.
- Gradient descent keeps descending after 2000 iterations.
- Stochastic gradient descent doesn't descend smoothly on the figure.

Problem 5.



Findings:

- The result of E_{out} is similar to the result of E_{in}. No matter which version we choose, $\eta = 0.001$ is too small for them to run in 2000 iterations, and $\eta = 0.01$ is a better choice.
- The result of E_{out} is greater than E_{in}.

Problem 6. (Bonus)

We can get $\min_{w_1, w_2, \ldots, w_K} RMSE(H) = \min_{w_1, w_2, \ldots, w_K} \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - h(x_n))^2}$ by $\min_{w_1, w_2, \ldots, w_K} \frac{1}{N}\sum_{n=1}^{N}(y_n - h(x_n))^2$.

We know that

$$e_k = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - h_k(x_n))^2} \rightarrow e_k^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - h_k(x_n))^2$$

$$e_0 = \sqrt{\frac{1}{N}\sum_{n=1}^{N}y_n^2} \rightarrow e_0^2 = \frac{1}{N}\sum_{n=1}^{N}y_n^2.$$

$$Let\ s_k = \sqrt{\frac{1}{N}\sum_{n=1}^{N}h_k(x_n)^2} \rightarrow s_k^2 = \frac{1}{N}\sum_{n=1}^{N}h_k(x_n)^2,$$

$$Then,\ \frac{1}{N}\sum_{n=1}^{N} h_i(w_n)y_n = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{-1}{2}\Big((y_n - h_k(x_n))^2 - y_n^2 - h_k(x_n)^2\Big)\right)$$

$$= \frac{-1}{2}\left(\frac{1}{N}\sum_{n=1}^{N}(y_n - h_k(x_n))^2 - \frac{1}{N}\sum_{n=1}^{N} y_n^2 - \frac{1}{N}\sum_{n=1}^{N} h_k(x_n)^2\right)$$

$$= \frac{-1}{2}(e_k^2 - e_0^2 - s_k^2)$$

$$\frac{\partial \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{k=1}^{K} w_k h_k(x_n) - y_n\right)^2}{\partial w_i} = 0$$

$$\Longrightarrow \frac{1}{N}\sum_{n=1}^{N}\frac{\partial\left(\sum_{k=1}^{K} w_k h_k(x_n) - y_n\right)^2}{\partial w_i} = 0$$

$$\Longrightarrow \frac{1}{N}\sum_{n=1}^{N}\frac{\partial\left(\sum_{k=1}^{K} w_k h_k(x_n) - y_n\right)^2}{\partial \sum_{k=1}^{K} w_k h_k(x_n)}\frac{\partial \sum_{k=1}^{K} w_k h_k(x_n)}{\partial w_i} = 0$$

$$\Longrightarrow \frac{1}{N}\sum_{n=1}^{N} 2\left(\sum_{k=1}^{K} w_k h_k(x_n) - y_n\right)h_i(x_n) = 0$$

$$\Longrightarrow \frac{1}{N}\sum_{n=1}^{N} 2\sum_{k=1}^{K} w_k h_k(x_n)h_i(x_n) = \frac{1}{N}\sum_{n=1}^{N} 2h_i(x_n)y_n$$

$$\Longrightarrow \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} w_k h_k(x_n)h_i(x_n) = s_k^2 + e_0^2 - e_k^2$$

$$\frac{2}{N}\begin{bmatrix} h_1(x_1) & h_1(x_2) & \cdots & h_1(x_N) \\ h_2(x_1) & h_2(x_2) & \cdots & h_2(x_N) \\ \vdots & \vdots & \ddots & \vdots \\ h_K(x_1) & h_K(x_2) & \cdots & h_K(x_N) \end{bmatrix}\begin{bmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_K(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_K(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x_N) & h_2(x_N) & \cdots & h_K(x_N) \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} s_0^2 + e_0^2 - e_0^2 \\ s_1^2 + e_0^2 - e_1^2 \\ \vdots \\ s_K^2 + e_0^2 - e_K^2 \end{bmatrix}$$

$$\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{bmatrix} = \underline{\frac{N}{2}\left(\begin{bmatrix} h_1(x_1) & h_1(x_2) & \cdots & h_1(x_N) \\ h_2(x_1) & h_2(x_2) & \cdots & h_2(x_N) \\ \vdots & \vdots & \ddots & \vdots \\ h_K(x_1) & h_K(x_2) & \cdots & h_K(x_N) \end{bmatrix}\begin{bmatrix} h_1(x_1) & h_2(x_1) & \cdots & h_K(x_1) \\ h_1(x_2) & h_2(x_2) & \cdots & h_K(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x_N) & h_2(x_N) & \cdots & h_K(x_N) \end{bmatrix}\right)^{-1}\begin{bmatrix} s_0^2 + e_0^2 - e_0^2 \\ s_1^2 + e_0^2 - e_1^2 \\ \vdots \\ s_K^2 + e_0^2 - e_K^2 \end{bmatrix}} \#$$

p.s. Suppose that inversion exists.