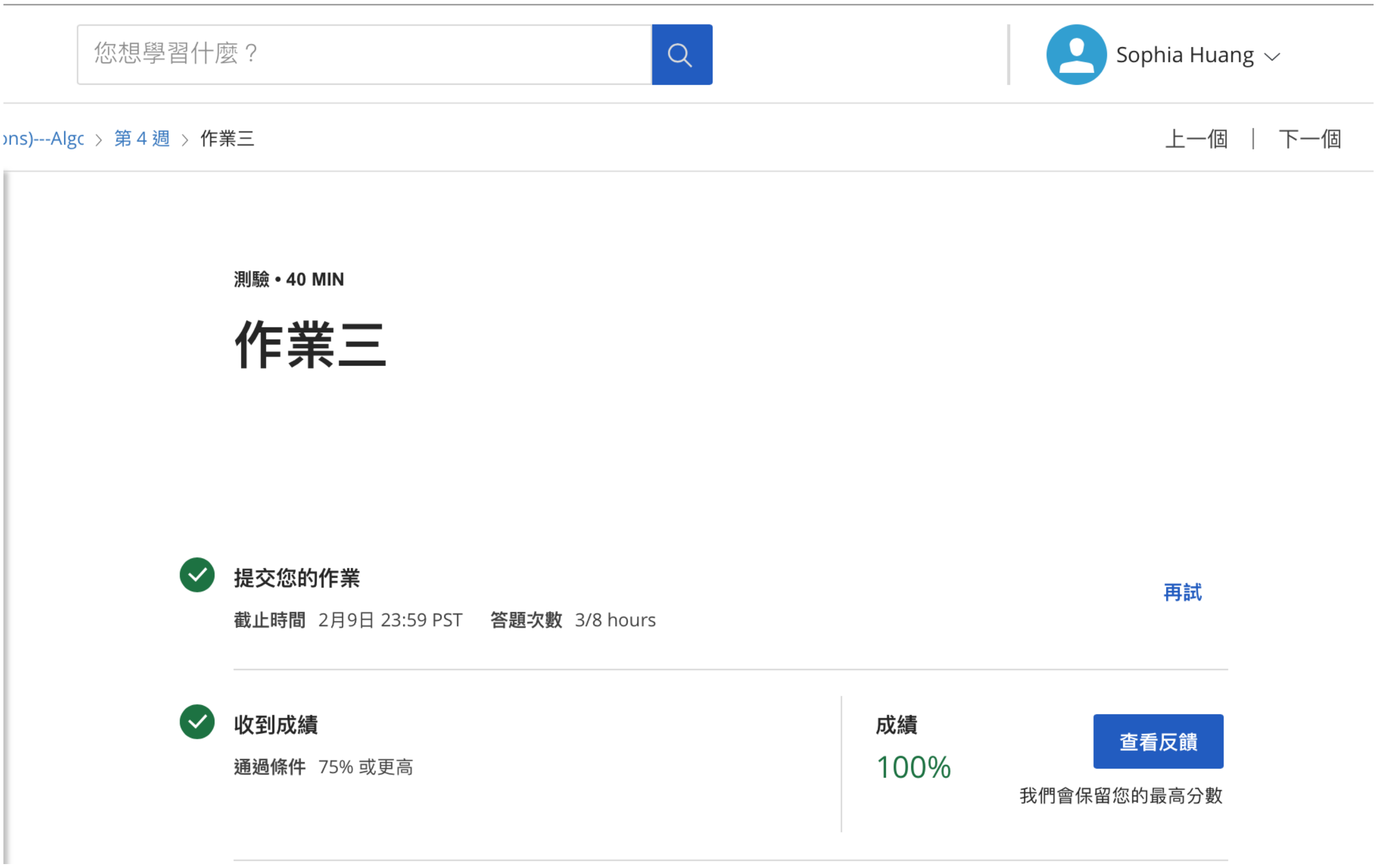
MLF homework3

b07902031 資工二 黃永雯

Problem 1



測驗 • 40 MIN

作業三

提交您的作業

截止時間 2月9日 23:59 PST 答題次數 3/8 hours

收到成績

通過條件 75% 或更高

成績

100%

我們會保留您的最高分數

Problem 2

For SGD, the update will be as follows:

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta \times$ (-∇err($\mathbf{w}_t$, $\mathbf{x}_n$, y_n))

(where err($\mathbf{w}_t$, $\mathbf{x}_n$, y_n) = err($\mathbf{w}$) = max(0, -y$\mathbf{w}^T$ $\mathbf{x}$))

For PLA, the update will be as follows:

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + 1 \times [[\text{sign}(\mathbf{w}_t^T\ \mathbf{x}_n) \neq y_n]] \times y_n\ \mathbf{x}_n$

There are two cases to consider:

(1) If sign(y) = sign($\mathbf{w}^T$ $\mathbf{x}$), then err($\mathbf{w}$) = max(0, -y$\mathbf{w}^T$ $\mathbf{x}$)) = 0 which means that $\mathbf{w}$ remains the same.

(2) If sign(y) $\neq$ sign($\mathbf{w}^T$ $\mathbf{x}$), then err($\mathbf{w}$) = -y$\mathbf{w}^T$ $\mathbf{x}$ and we can change the sentence for update as follows:

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta \times$ (-∇err($\mathbf{w}_t$, $\mathbf{x}_n$, y_n))

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta \times$ (-∇err($\mathbf{w}$))

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta \times$ (y_n $\mathbf{x}_n$))

If $\eta = 1$, then it can be represented by PLA.

From (1), (2) can know that err($\mathbf{w}$) = max(0, -y$\mathbf{w}^T$ $\mathbf{x}$) results in PLA.

Problem 3

The definition of Newton Direction is as follows:

f(x+Δx) = f(x) + f'(x)Δx + $\frac{1}{2}$ f''(x)Δx^2

When Δx is close to 0, the sentence will turned into:

0 = $\frac{d}{dx}$ (f(x) + f'(x)Δx + $\frac{1}{2}$ f''(x)Δx^2).

It can be changed to finding an x and let the following sentence correct:

f'(x) + f''(x) Δx = 0

Thus, it is the same as solving Δx = -$\frac{f'(x_n)}{f''(x_n)}$.

By the method mentioned above, we can get the sentence as follows:

∇E(u, v) + ∇^{2}E(u, v)Δ(u, v) = 0

Then we can get (Δu, Δv) = -$\frac{\nabla E(u,v)}{\nabla^2 E(u,v)}$, and it can also be represented as:

-($\nabla^2 E(u,v)$)$^{-1}$ ∇E(u, v).

Problem 4

As mentioned in class, we can calculate the likelihood as follows:

likelihood = $\prod_{n=1}^{N}$ $\frac{exp(w_{yn}^T x_n)}{\sum_{k=1}^{K} exp(w_k^T x_n)}$

And ln(likelihood) = $ln(\prod_{n=1}^{N}$ $\frac{exp(w_{yn}^T x_n)}{\sum_{k=1}^{K} exp(w_k^T x_n)})$

= $\sum_{n=1}^{N}[ln(exp(w_{y_n}^T x_n)) - ln(\sum_{k=1}^{K} exp(w_k^T x_n))]$

Since E_{in} is to minimize the negative log likelihood, the sentence will be as follows:

$\frac{1}{N}$ $\sum_{n=1}^{N}[ln(\sum_{k=1}^{K} exp(w_k^T x_n)) - ln(exp(w_{y_n}^T x_n))]$

= $\frac{1}{N}$ $\sum_{n=1}^{N}[ln(\sum_{k=1}^{K} exp(w_k^T x_n)) - w_{y_n}^T x_n]$

Problem 5

The definition is as follows:

X = $[x_1,\ x_2,...\ x_N]^T$

y = $[y_1,\ y_2,...\ y_N]^T$

$\tilde{X}$ = $[\tilde{x_1},\ \tilde{x_2},...\ \tilde{x_K}]^T$

y = $[\tilde{y_1},\ \tilde{y_2},...\ \tilde{y_K}]^T$

Then we can calculate E_{in} as follows:

$E_{in}(\mathbf{w}) = \frac{1}{N+K}(||$y - X$\mathbf{w}$ $||^2$ + || $\tilde{y}$ - $\tilde{X}$ $\mathbf{w}$ $||^2) = \frac{1}{N+K}(||$ X$\mathbf{w}$ - y $||^2$ + || $\tilde{X}$ $\mathbf{w}$ - $\tilde{y}$ $||^2)$

We can calculate $\frac{\partial E_{in}(\mathbf{W})}{\partial \mathbf{W}}$ = [[0]] to obtain the optimal $\mathbf{w}$.

Since $\frac{\partial E_{in}(\mathbf{W})}{\partial \mathbf{W}} = \frac{1}{N+K}(2((X^T$Xw - X^Ty) + ($\tilde{X}^T$Xw - $\tilde{X}^T$y))), w can be represented as $\frac{X^Ty+\tilde{X}^T\tilde{y}}{X^TX+\tilde{X}^T\tilde{X}} = (X^TX + \tilde{X}^T\tilde{X})^{-1}\ (X^Ty + \tilde{X}^T\tilde{y})$

Problem 6

From the definition mentioned in class can know that

$W_{REG} \leftarrow (X^TX + \lambda I)^{-1}\ X^T$y.

And from problem 5 can know that w can be represented as

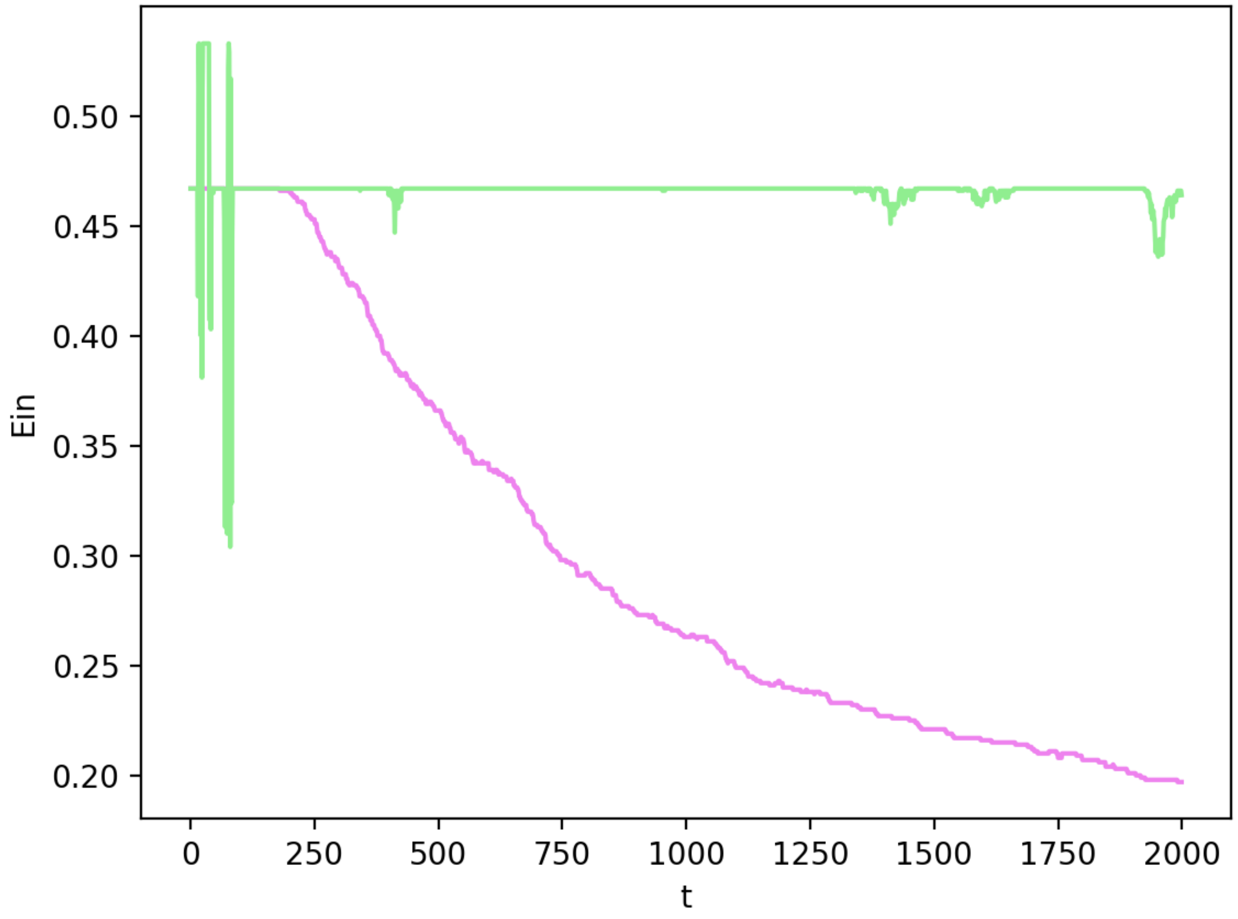
$(X^TX + \tilde{X}^T\tilde{X})^{-1}\ (X^Ty + \tilde{X}^T\tilde{y})$

If we take $\tilde{X} = \sqrt{\lambda}I$ and $\tilde{y} = 0$, the sentence above can be changed into

$(X^TX + (\sqrt{\lambda}I)^T\sqrt{\lambda}I)^{-1}\ (X^Ty + \tilde{X}^T\tilde{y})$ and will be the same as

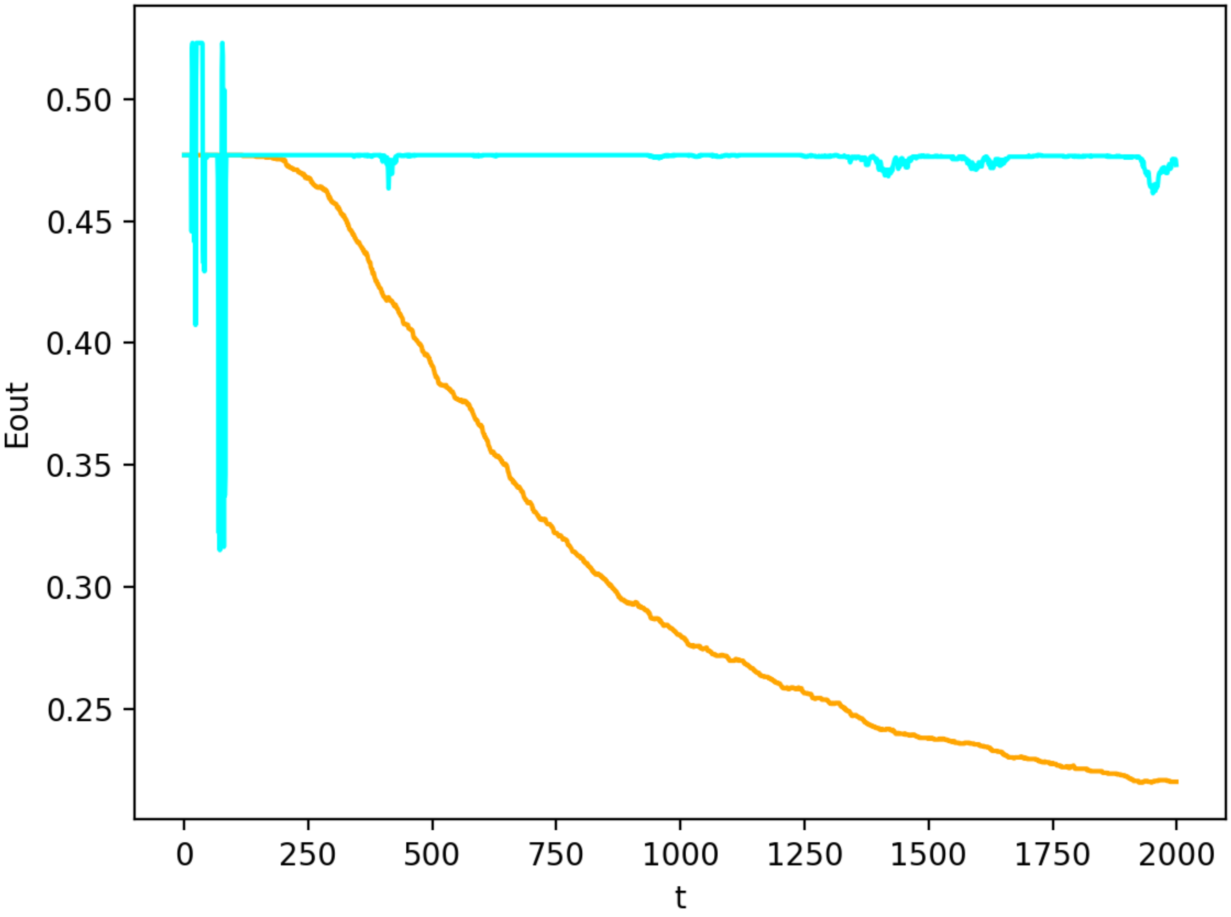
$W_{REG} \leftarrow (X^TX + \lambda I)^{-1}\ X^T$y.

Problem 7



In the picture, the purple line is the E_{in} for gradient descent version, and the green line is the E_{in} for stochastic gradient version. From the picture can see that for gradient descent version, E_{in} will be lower with the number of iteration increases. On the other hand, for stochastic gradient version, since it is chosen randomly (in this problem it is not totally chosen randomly), E_{in} will not continually decrease with the increase of iteration, but it still becomes more stable with the increase of iteration. Moreover, the gradient descent version will let E_{in} be close to 0.197 after 2000 times of iteration and stochastic gradient version will let E_{in} be close to 0.464 after 2000 times of iteration. The gradient descent version will change more substantially since the η we use is larger than that we choose for stochastic gradient version.

Problem 8



In the picture, the orange line is the E_{out} for gradient descent version, and the blue line is the E_{out} for stochastic gradient version. The changing is just like that in problem 7. However, E_{out} of both version are larger than E_{in}. For gradient descent version, E_{out} will be close to 0.22 after 2000 times of iteration. On the other hand, for stochastic gradient version, E_{out} will be close to 0.473 after 2000 times of iteration.

Problem 9

(a)

Since $V\Gamma^{-1}U^T$ is to define pseudo inverse, we can get $W_{lin} = V\Gamma^{-1}U^T$ y.

Therefore, we can change the sentence into:

$X^TXW_{lin} = X^TXV\Gamma^{-1}U^Ty$

$X^TXW_{lin} = X^TXX^{\dagger}y$

Since $XX^{\dagger} = (XX^{\dagger})^T$, the sentence can be changed into

$X^TXW_{lin} = X^T(XX^{\dagger})^Ty$

$X^TXW_{lin} = (XX^{\dagger}X)^Ty$

Moreover, since $XX^{\dagger}X =$ X, the sentence now becomes:

$X^TXW_{lin} = X^Ty$

(b)

(1) If there exists a solution that satisfies $X^TXw = X^Ty$, which means that we can project w onto range($(X^TX)^{\dagger}$) and $X^TXw = X^TXw_{lin}$.

The sentence $X^TXw = X^TXw_{lin}$ can be changed as follows:

$(X^TX)^{\dagger}X^TXw = (X^TX)^{\dagger}X^TXw_{lin}$

$(X^TX)^{\dagger}X^TXw = (X^TX)^{\dagger}X^Ty$ (known from problem 9 (a))

$(X^TX)^{\dagger}X^TXw = (V\Gamma^TU^TU\Gamma V^T)^{\dagger}V\Gamma^TU^Ty$ (since X = $U\Gamma V^T$)

$(X^TX)^{\dagger}X^TXw = V\Gamma^{-1}U^Ty = w_{lin}$

(2) From linear algebra can know that:

If W_2 is a subspace of inner product space W_1. x_3 is a vector of W_1, $x_1 \in W_2$, $x_2 \in (W_2)^{\perp}$ and it satisfies $x_1 + x_2 = x_3$. If $x_1 \neq x_3$, then $||x_3|| > ||x_1||$.

Therefore, from (1), (2) can know that $||w_{lin}|| \leq ||w||$