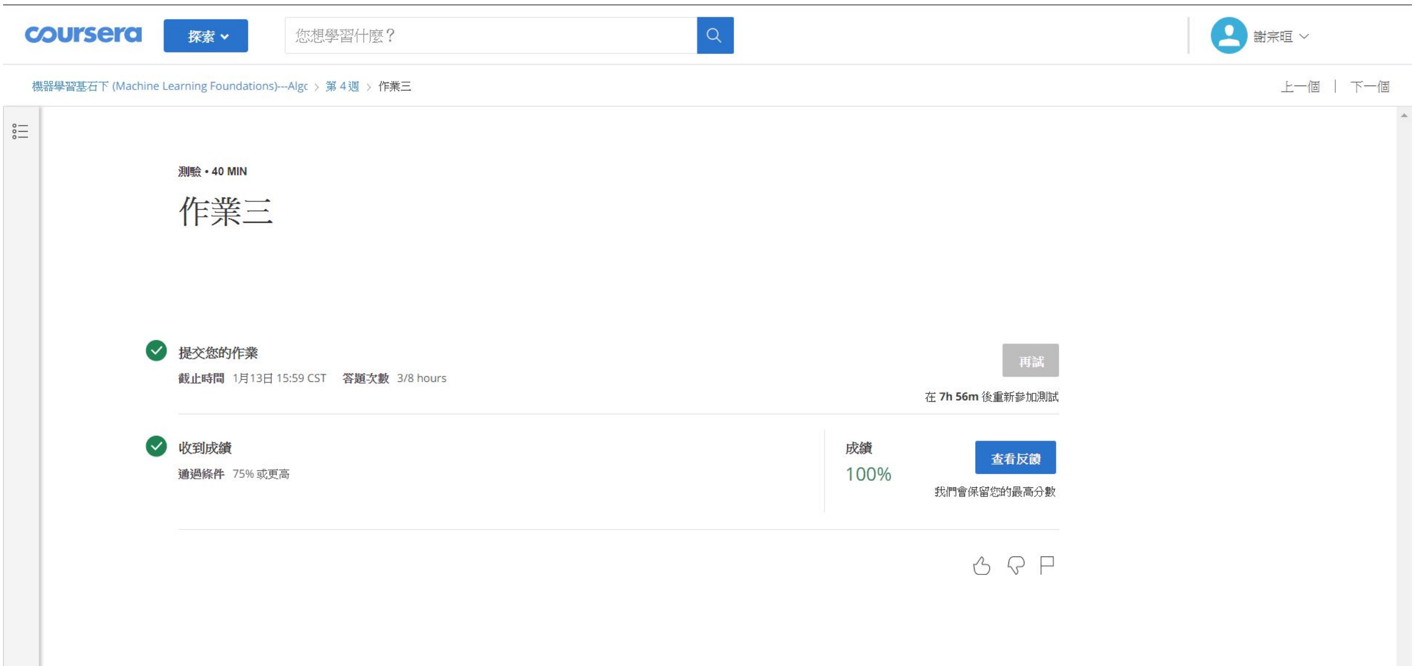
ML HW#3

B07902055 謝宗晅

1.



2.

Prove

分成兩個情況來討論，分別是 $y = 1$ 和 $y = -1$ 兩種情況：

- $y = 1$：
  - $\mathbf{w}^T\mathbf{x} > 0$
    PLA：正確，所以不用修正 $(\mathbf{w}_{t+1} = \mathbf{w}_t)$
    SGD：因為 $\max(0, -y\mathbf{w}^T\mathbf{x}) = 0$ 時的梯度 = 0，所以一樣是不用修正 $(\mathbf{w}_{t+1} = \mathbf{w}_t)$
  - $\mathbf{w}^T\mathbf{x} < 0$
    PLA：錯誤，所以要修正，$\mathbf{w}_{t+1} = \mathbf{w}_t + y\mathbf{x}$
    SGD：因為 $\max(0, -y\mathbf{w}^T\mathbf{x}) = -y\mathbf{w}^T\mathbf{x}$ 時的梯度是 $-y\mathbf{x}$，所以要修正，$\mathbf{w}_{t+1} = \mathbf{w}_t - (-y\mathbf{x}) = \mathbf{w}_t + y\mathbf{x}$
- $y = -1$
  - $\mathbf{w}^T\mathbf{x} < 0$
    PLA：正確，所以不用修正 $(\mathbf{w}_{t+1} = \mathbf{w}_t)$
    SGD：因為 $\max(0, -y\mathbf{w}^T\mathbf{x}) = 0$ 時的梯度 = 0，所以一樣是不用修正 $(\mathbf{w}_{t+1} = \mathbf{w}_t)$
  - $\mathbf{w}^T\mathbf{x} > 0$
    PLA：錯誤，所以要修正，$\mathbf{w}_{t+1} = \mathbf{w}_t + y\mathbf{x}$

SGD：因為 $\max(0, -y\mathbf{w}^T\mathbf{x}) = -y\mathbf{w}^T\mathbf{x}$ 時的梯度是 $-y\mathbf{x}$，所以要修正，$\mathbf{w}_{t+1} = \mathbf{w}_t - (-y\mathbf{x}) = \mathbf{w}_t + y\mathbf{x}$

因為 SGD 每次也是選取單一個點來更新 $\mathbf{w}$，因此可以證明以 $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$ 為 error function 的 SGD 做出來的結果會和 PLA 一樣。

3.

要最小化 $\hat{E}_2(\Delta u, \Delta v)$ 的話，就要讓 $\nabla\hat{E}_2(\Delta u, \Delta v) = 0$。為了方便表示，我將 (u, v) 用 $\mathbf{a}$ 表示，也就是說 $\hat{E}_2(u, v) = \hat{E}_2(\mathbf{a})$。因此當 $\mathbf{x} = (u + \Delta u, v + \Delta v)$ 很靠近 $\mathbf{a}$ 時 (Δu 和 Δv 很小)，就可以把式子寫成：

$$E(\mathbf{x}) \approx \hat{E}_2(\mathbf{x})$$

$$= E(\mathbf{a}) + (\nabla E(\mathbf{a}))^T(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T(H)(\mathbf{x} - \mathbf{a})$$

$$= E(\mathbf{a}) + (\nabla E(\mathbf{a}))^T(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x}^T H\mathbf{x} - 2\mathbf{a}^T H\mathbf{x} + \mathbf{a}^T H\mathbf{a})$$

$$\Rightarrow \nabla\hat{E}_2(\mathbf{x}) = H\mathbf{x} + (\nabla E(a) - H\mathbf{a}) = 0$$

$$\Rightarrow \mathbf{x} = -(H^{-1})(\nabla E(\mathbf{a}) - H\mathbf{a})$$

$$\Rightarrow \Delta\mathbf{x} = \mathbf{x} - \mathbf{a} = -(H^{-1})(\nabla E(\mathbf{a})) + (H^{-1})(H\mathbf{a}) - \mathbf{a}$$

$$= -(H^{-1})(\nabla E(\mathbf{a}))$$

因為 Hessian 矩陣是 positive definite 的，因此他的反矩陣必定存在。所以我們要找的 $(\Delta u, \Delta v)$ 就是上方的 $\Delta\mathbf{x} = -(H^{-1})(\nabla E(\mathbf{a}))$。

4.

和上課的內容是幾乎一樣的，只是要將 target function 想成有 K 個 (雖然實際上只有一個)，其中每一個想像出來的 target function，$f_k(\mathbf{x})$，代表的是這組 $\mathbf{x}$ 被分類到的結果是 k 的機率，所以一組 $\mathbf{x}$ 是 k 的 likelihood 就可以正比於

$$\prod_{i=1}^{K} h_i(y_i\mathbf{x})$$

其中 $y_i \in \{+1, -1\}$，+1 表示 output 是 i，-1 表示 output 不是 i。所以全部的 $\mathbf{x}$ 的 likelihood 就會有下列這個式子：

$$\max_h \text{likelihood } (h) \propto \prod_{n=1}^{N}\left(\prod_{i=1}^{K} h_i(y_{i,n}\mathbf{x}_n)\right) = \prod_{n=1}^{N}\left(\prod_{i=1}^{K} \frac{\exp(y_{i,n}\mathbf{w}_i^T\mathbf{x}_n)}{\sum_{k=1}^{K}\exp(y_{i,n}\mathbf{w}_k^T\mathbf{x}_n)}\right)$$

再運用課堂上提過的取 $\ln$ 技巧，就會變成：

$$\max_{h} \sum_{n=1}^{N}\left(\sum_{i=1}^{K}\ln\left(\frac{\exp(y_{i,n}\mathbf{w}_i^T\mathbf{x}_n)}{\sum_{k=1}^{K}\exp(y_{i,n}\mathbf{w}_k^T\mathbf{x}_n)}\right)\right)$$

加上負號，把前面的 $\max$ 變成 $\min$：

$$\min_{h} \sum_{n=1}^{N}\left(\sum_{i=1}^{K}\ln\left(\frac{\sum_{k=1}^{K}\exp(y_{i,n}\mathbf{w}_k^T\mathbf{x}_n)}{\exp(y_{i,n}\mathbf{w}_i^T\mathbf{x}_n)}\right)\right)$$

其中，$\ln$ 分母的部分可以分離開來，並且又因為之前的定義，$y_{i,n}=1$ 若且唯若對於 $\mathbf{x}_n$ 這筆資料的類別是 i，所以：

$$\min_{h} \sum_{n=1}^{N}\left(\ln\left(\sum_{k=1}^{K}\exp(\mathbf{w}_k^T\mathbf{x}_n)\right)-(\mathbf{w}_{y_n}^T\mathbf{x}_n)\right)$$

加上常數 $\frac{1}{N}$：

$$\frac{1}{N}\sum_{n=1}^{N}\left(\ln\left(\sum_{k=1}^{K}\exp(\mathbf{w}_k^T\mathbf{x}_n)\right)-(\mathbf{w}_{y_n}^T\mathbf{x}_n)\right)$$

5.

我們的目標是找到使得梯度為 0 的那個 $\mathbf{w}$。
題目的式子可以被化成以下這樣子 (參考自 Linear regression 的 slide)：

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

展開之後得到 (前面的 $\frac{1}{N+K}$ 對於求梯度 = 0 時的 $\mathbf{w}$ 不會有影響，所以暫時省略)：

$$\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}-2\mathbf{w}^T\mathbf{X}^T\mathbf{y}+\mathbf{y}^T\mathbf{y}\right)+\left(\mathbf{w}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{w}-2\mathbf{w}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{y}}+\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}\right)$$

把上面的式子對於 $\mathbf{w}$ 偏微分就會變成梯度：

$$\left(2\mathbf{X}^T\mathbf{X}\mathbf{w}-2\mathbf{X}^T\mathbf{y}\right)+\left(2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{w}-2\tilde{\mathbf{X}}^T\mathbf{y}\right)$$

因此，當梯度 = 0 的時候，題目的式子就會是最小值，所以：

$$\left(2\mathbf{X}^T\mathbf{X}\mathbf{w}-2\mathbf{X}^T\mathbf{y}\right)+\left(2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{w}-2\tilde{\mathbf{X}}^T\mathbf{y}\right)=0$$

$$\Rightarrow \mathbf{w}=\left(\mathbf{X}^T\mathbf{X}+\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}+\tilde{\mathbf{X}}^T\tilde{\mathbf{y}}\right)$$

6.

由第 5 題的第一個式子：

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

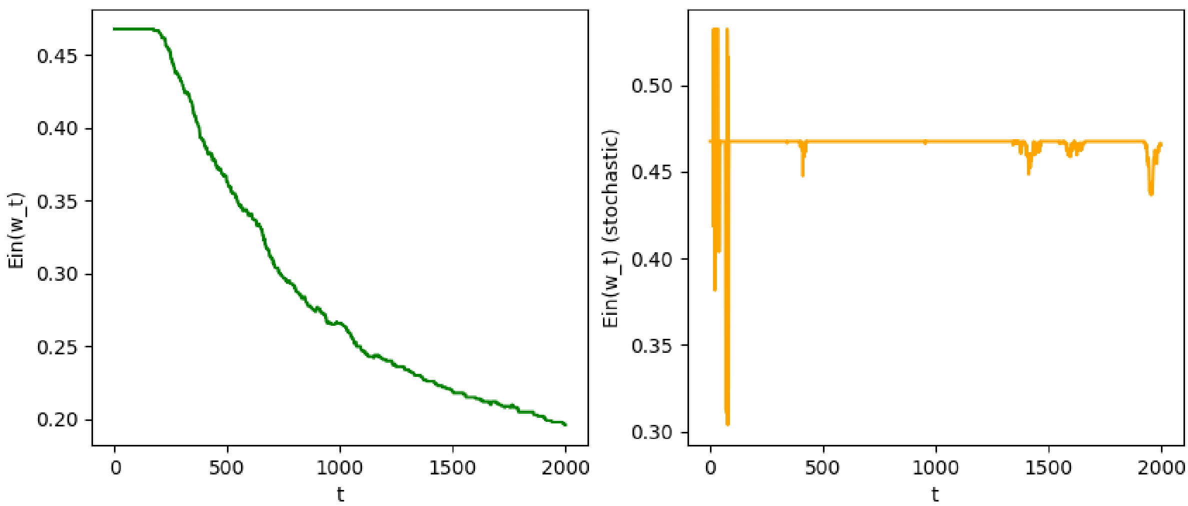
可以得到令：

$$\tilde{\mathbf{X}}=\sqrt{\lambda}\mathbf{I},\quad \tilde{\mathbf{y}}=\mathbf{0}$$

就可以得到：

$$\begin{aligned}&\underset{\mathbf{w}}{\arg\min}\ \frac{1}{N+d}\left(\left\|\mathbf{X}\mathbf{w}-\mathbf{y}\right\|^2+\left\|\sqrt{\lambda}\mathbf{w}\right\|^2\right)\\ &=\underset{\mathbf{w}}{\arg\min}\ \frac{1}{N+d}\left(\left\|\mathbf{X}\mathbf{w}-\mathbf{y}\right\|^2+\lambda\left\|\mathbf{w}\right\|^2\right)\\ &\approx\underset{\mathbf{w}}{\arg\min}\ \frac{\lambda}{N}\left\|\mathbf{w}\right\|^2+\frac{1}{N}\left\|\mathbf{X}\mathbf{w}-\mathbf{y}\right\|^2\end{aligned}$$

7.

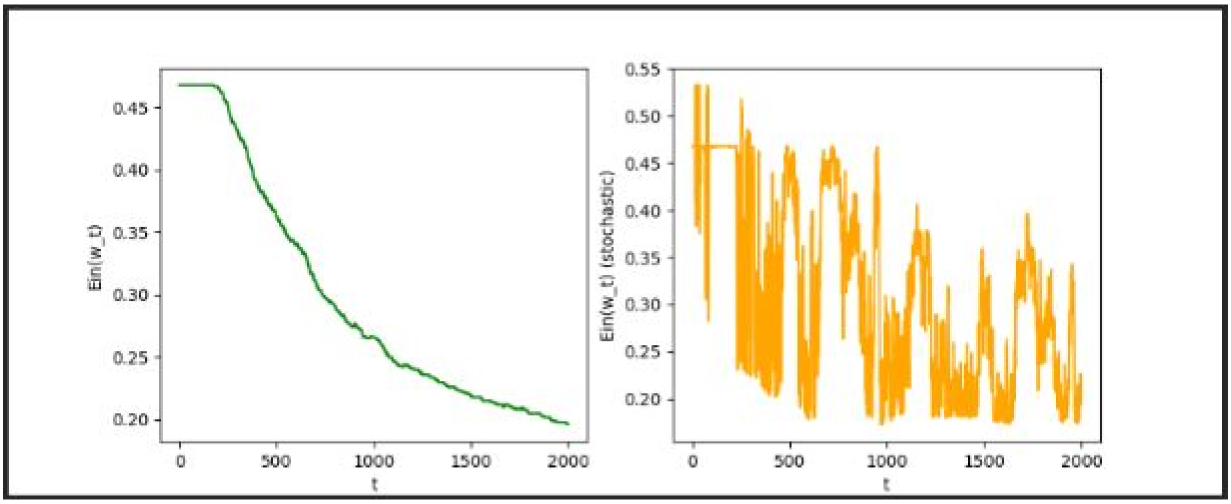


Findings

左邊的是用一般的 gradient descent，learning rate = 0.01 畫的圖

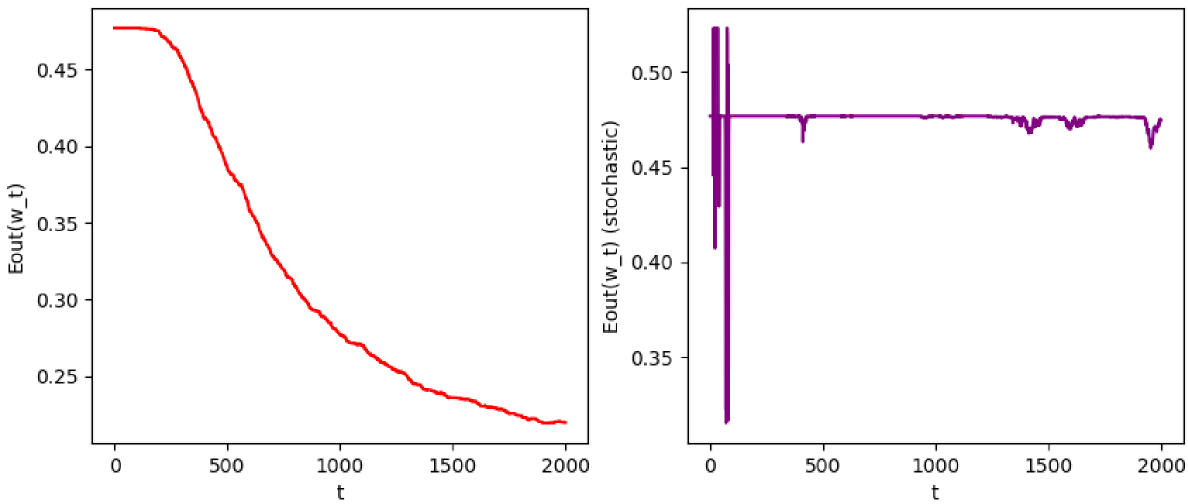
右邊的是用 stochastic gradient descent，learning rate = 0.001 畫的圖

可以看到左邊的 $E_{in}(\mathbf{w}_t)$ 有在下降的趨勢，而且是相當明顯的趨勢，但是右邊的圖看起來幾乎沒有在學習的樣子，我想應該是因為 learning rate 的差距造成了這樣的結果，右邊的圖因為 learning rate 太小了，所以幾乎沒有學習的跡象，而且我也有做了額外的實驗，是將 stochastic gradient descent 的 learning rate 調成 0.01，結果輸出的樣子會是很大的上下起伏，就像是 learning rate = 0.001 時，前 50 次更新的樣子，但是隨著很大的上下起伏，整體的 E_{in} 還是有在逐漸下降的趨勢，就像是下面這樣：



E_{in} 最終還是下降到了接近 0.2 的地方。

8.



Findings

左邊的是用一般的 gradient descent，learning rate = 0.01 畫的圖
右邊的是用 stochastic gradient descent，learning rate = 0.001 畫的圖
其實 E_{out} 的趨勢和上方的 E_{in} 的趨勢是很接近的，如果不說是哪一張圖的話其實看不太出來哪一張是 E_{in}，哪一張是 E_{out}，而我也有對於不同的 learning rate 做額外的實驗，和第 7 題一樣，結果也是和第 7 題很接近。

Bonus

(a)

根據題目的定義，我們可以寫出：

$$\begin{aligned} X^T X \mathbf{w}_{\text{lin}} &= X^T X \left(V \Gamma^{-1} U^T \mathbf{y}\right) = X^T \left(U \Gamma V^T\right)\left(V \Gamma^{-1} U^T \mathbf{y}\right) \\ &= X^T \left(U \Gamma \Gamma^{-1} U^T \mathbf{y}\right) = X^T \left(U U^T \mathbf{y}\right) = X^T \mathbf{y} \end{aligned}$$

因此 $\mathbf{w}_{\text{lin}} = \left(V \Gamma^{-1} U^T \mathbf{y}\right)$ 是一個解。

(b)

先證明一個小引理：
令 U 是內積空間 W 的子空間，令 $z \in W$，以及 $x \in U$ 和

$$y \in V = U^{\perp} = \{v \in W \mid v \cdot u = 0,\ \forall u \in U\}$$

使得 $z = x + y$，則

$$\|z\| \geq \|x\|$$

證明：
因為

$$\|z\|^2 = \|x\|^2 + 2(x \cdot y) + \|y\|^2 = \|x\|^2 + \|y\|^2$$

又 $\|y\| \geq 0$，得證。

令一個 subspace $W = \text{span}(\text{col}(X))$，其維度是 ρ，且題目中的 $\mathbf{w}_{\text{lin}} = X^{\dagger}\mathbf{y}$，令另外一個 $\mathbf{w}$ 滿足 $X^T X \mathbf{w} = X^T \mathbf{y}$，則將 $\mathbf{w}$ 垂直投影到 $\text{span}(\text{col}(X^{\dagger})) = \text{Null}(\text{col}(X)) = W^{\perp}$ 的結果就會是 $\mathbf{w}_{\text{lin}}$：

$$X^{\dagger} X \mathbf{w} = X^{\dagger} X \mathbf{w}_{\text{lin}} = X^{\dagger} X X^{\dagger} \mathbf{y} = X^{\dagger} \mathbf{y} = \mathbf{w}_{\text{lin}}$$

再根據我們的小引理，因此有了 $\|\mathbf{w}_{\text{lin}}\| \leq \|\mathbf{w}\|$ 的結論。