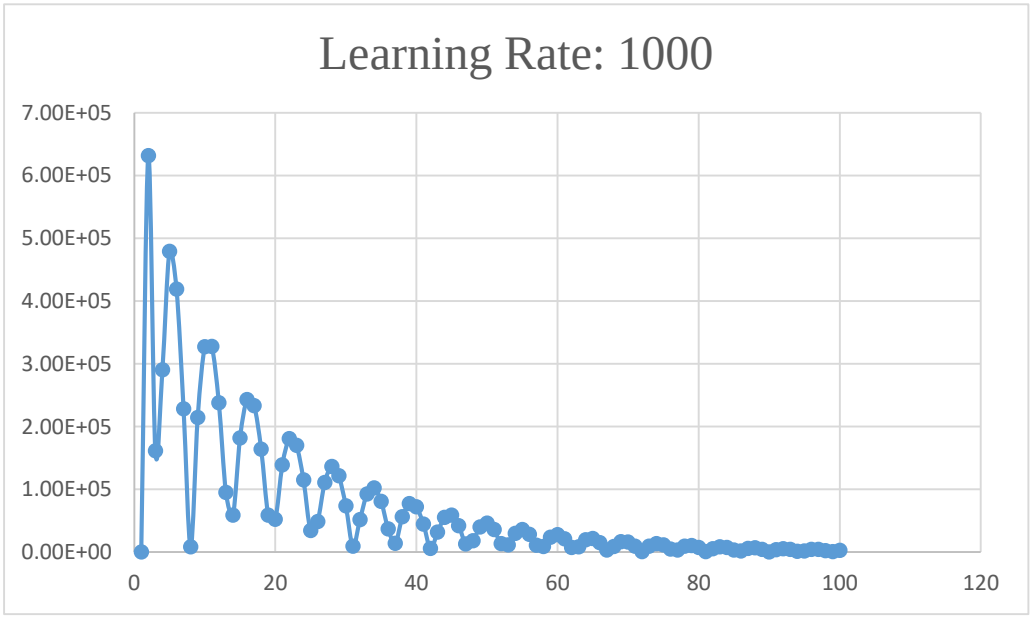
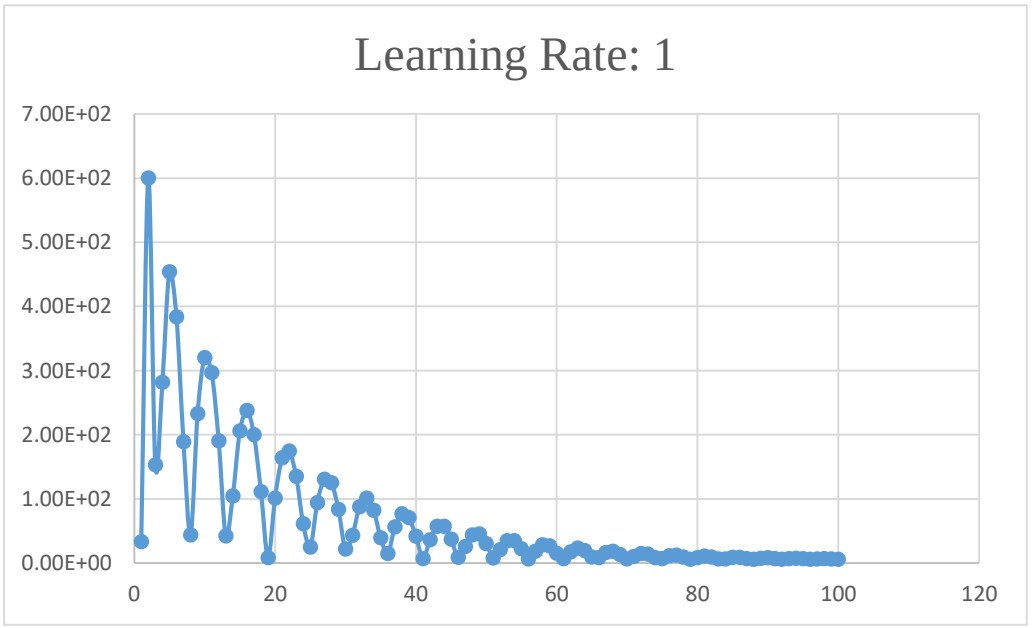
Homework 1 Report - PM2.5 Prediction

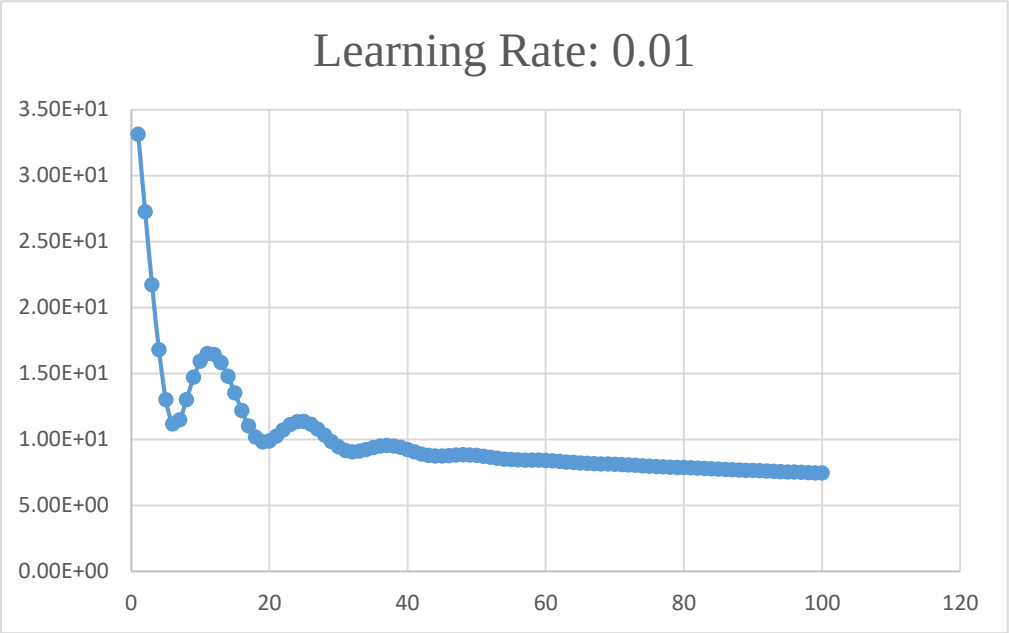
學號：R07922120　系級：資工碩一　姓名：陳禹達

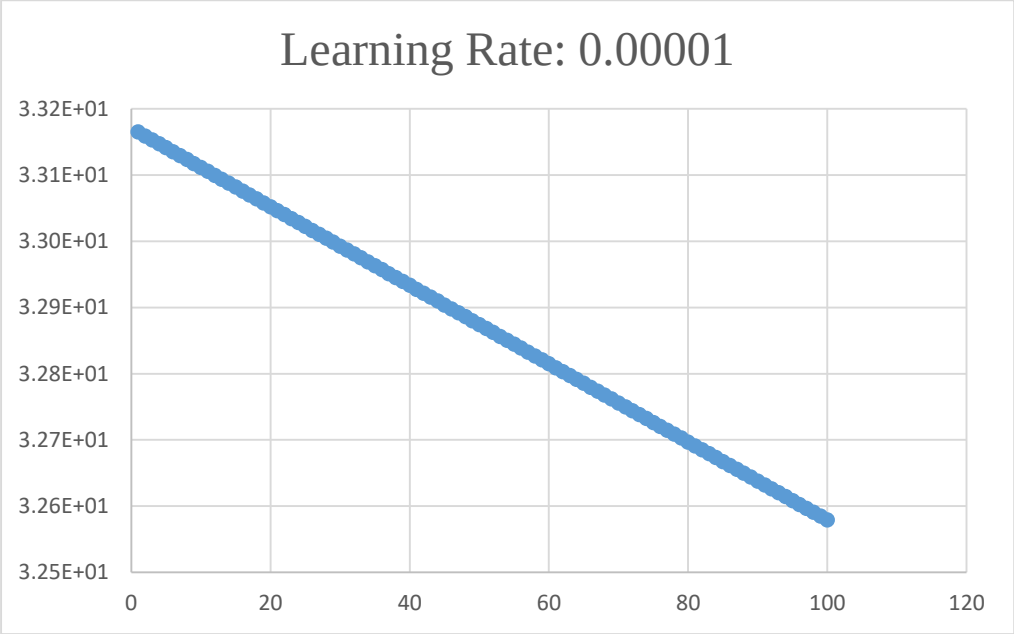
1. (1%) 請分別使用至少 **4** 種不同數值的 learning rate 進行 training（其他參數需一致），對其作圖，並且討論其收斂過程差異。

A:









這邊另外說明我用來降低 Learning Rate 的方式是 Adam，我使用得 Learning Rate 分別為: 1000、1、0.01、0.00001，從上圖可以看出，當 Initial Learning Rate 過大時，如:1000 或者 1，在剛開始時 RMSE 的震盪幅度會很大，這是因為當 Learning Rate 過大時，會使得在找尋參數時”找過頭”，使得 Loss 無法穩定的下降，但相反的 Loss 也會較為快速的下降至低點；而當 Initial Learning Rate 過小時，如:0.00001,0.01，則會使 Loss 下降得太慢，造成 iteration 的次數必須相對的提高，但是可以從圖中看出，Loss 是穩定的下降的，若 iteration 的次數夠大，仍然可以找到最佳解，因此選擇適中的 Learning Rate 非常重要。

2. (1%) 請分別使用每筆 data9 小時內所有 feature 的一次項（含 bias 項）以及每筆 data9 小時內 PM2.5 的一次項（含 bias 項）進行 training，比較並討論這兩種模型的 root mean-square error（根據 kaggle 上的 public/private score）。

   A:

Feature/RMSE	Public Error	Private Error
All	6.54	6.64
Only PM2.5	6.79	7.16

   從兩種模型所得到的 root mean-square error 去進行比較，可以看出使用所有的 Feature 所得到的模型較佳，這是因為影響 PM2.5 的因素有許多，例如 NOx 可能會在幾個小時後經過化學作用變成 PM2.5 的懸浮粒子，只要不會有 Overfitting 的問題，基本上多用了幾個，相較於只用一個 PM2.5 預測，精確度會高出許多。

3. (1%)請分別使用至少四種不同數值的 regularization parameter λ 進行 training（其他參數需一至），討論及討論其 RMSE(training, testing)（testing 根據 kaggle 上的 public/private score）以及參數 weight 的 L2 norm。

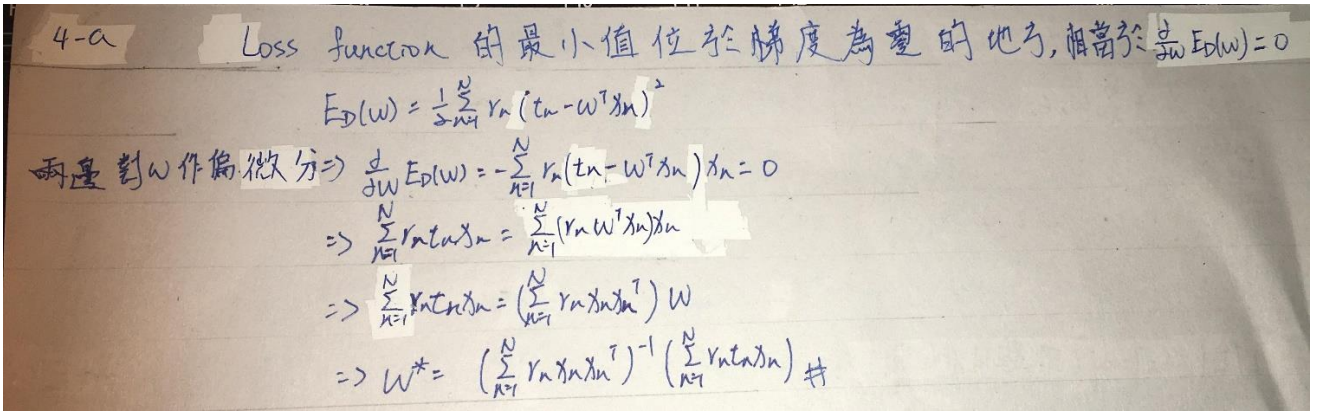
   A:

regularization parameter /RMSE	Training Error	Public Error	Private Error
0.01	5.52	6.15	6.65
0.1	5.40	6.14	6.65
10	5.82	6.11	6.65
1000	164.54	144.04	69.39

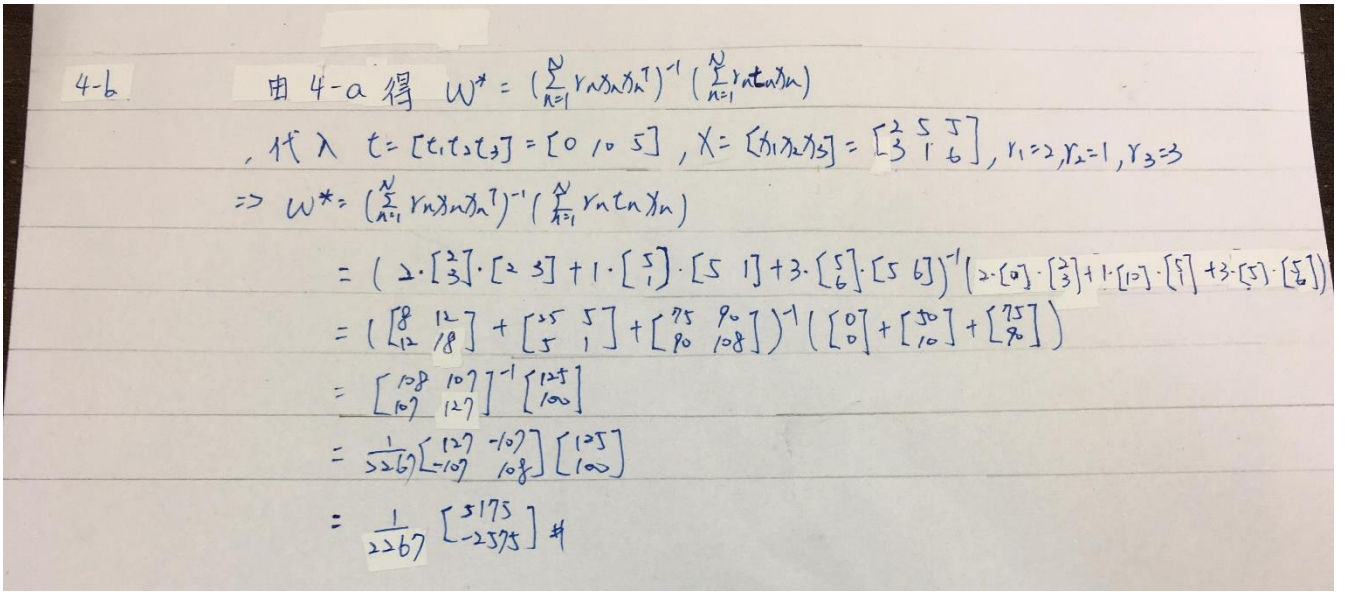
   由上表可以看出當 λ 調的過高時，Error 會特別高，這是因為當 λ 調太高時，會造成我們可用的 Feature 過少，這時就會發生 Underfitting，因此我們的預測誤差就會過大，另外因為這次實作過程中我使用的 Feature 並沒有這麼多，因此 regularization 並沒有什麼效果。

4.

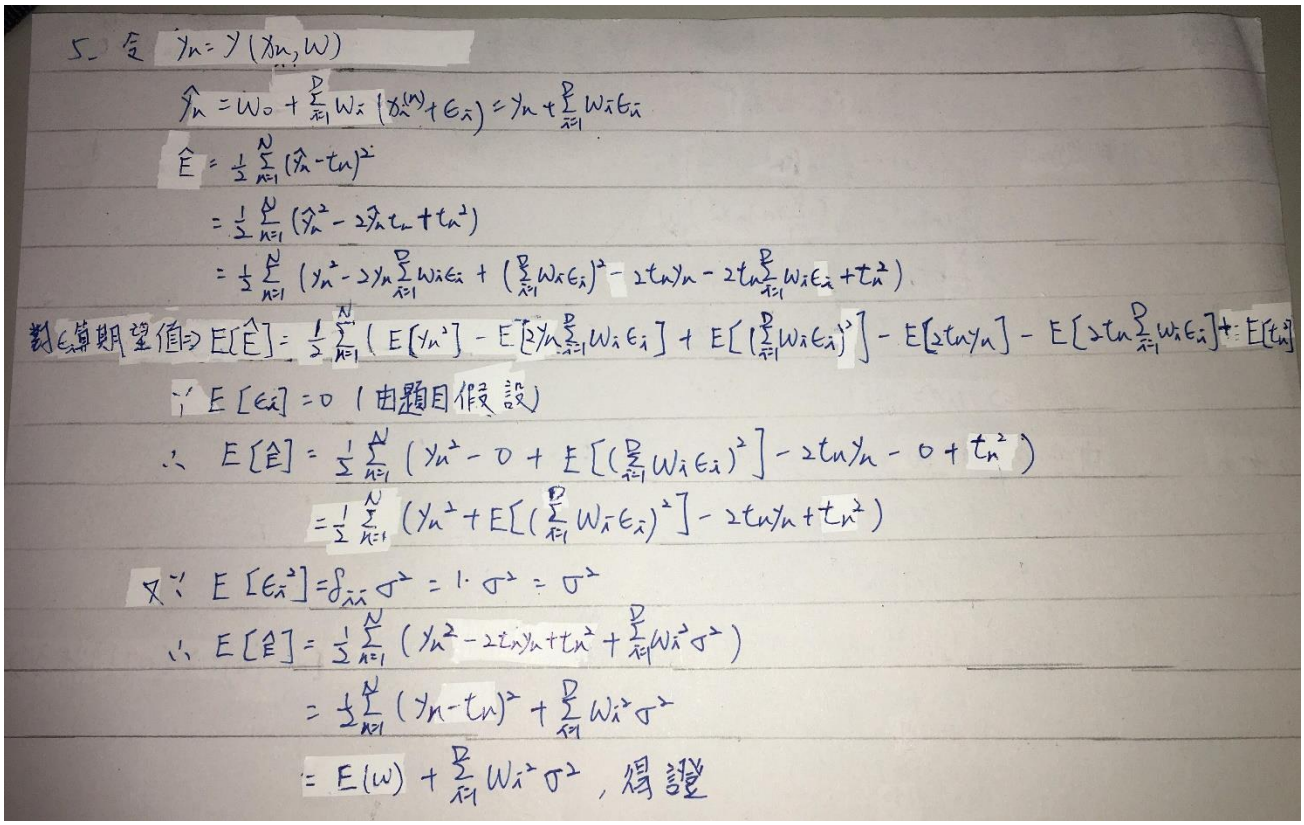
4-a.


4-a
Loss function 的最小值位於梯度為零的地方，相當於 $\nabla E_D(w)=0$
$E_D(w)=\frac{1}{2}\sum_{n=1}^{N} r_n(t_n-w^T x_n)^2$
兩邊對 w 作偏微分 $\Rightarrow \frac{d}{dw}E_D(w)=-\sum_{n=1}^{N} r_n(t_n-w^T x_n)x_n=0$
$\Rightarrow \sum_{n=1}^{N} r_n t_n x_n=\sum_{n=1}^{N}(r_n w^T x_n)x_n$
$\Rightarrow \sum_{n=1}^{N} r_n t_n x_n=\left(\sum_{n=1}^{N} r_n x_n x_n^T\right)w$
$\Rightarrow w^*=\left(\sum_{n=1}^{N} r_n x_n x_n^T\right)^{-1}\left(\sum_{n=1}^{N} r_n t_n x_n\right)$ #

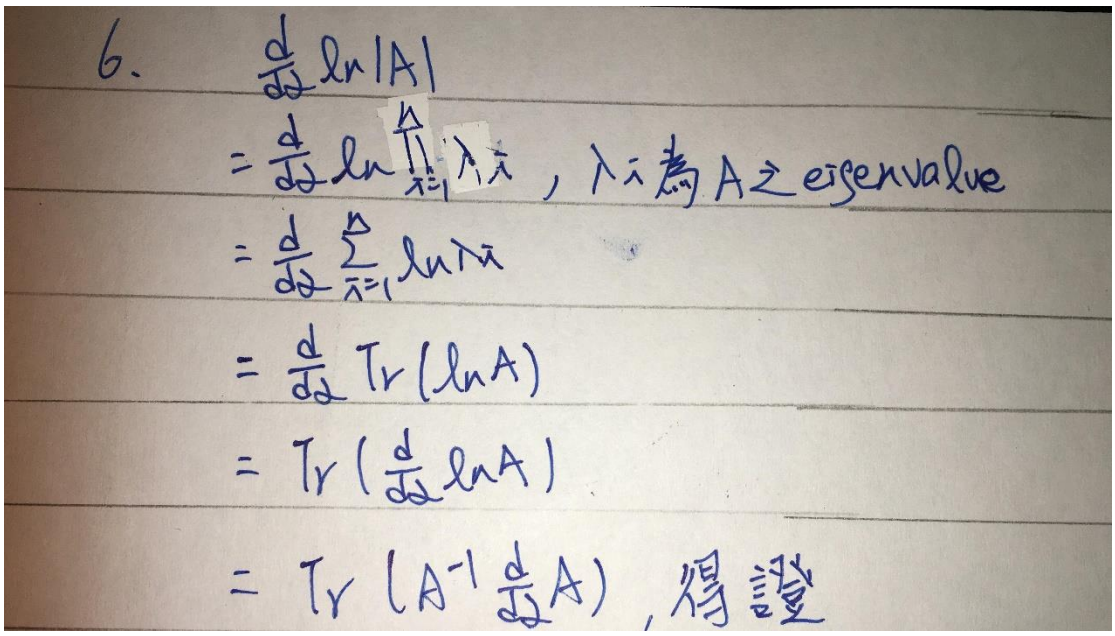
4-b.


4-b
由 4-a 得 $w^*=\left(\sum_{n=1}^{N} r_n x_n x_n^T\right)^{-1}\left(\sum_{n=1}^{N} r_n t_n x_n\right)$
代入 $t=[t_1 t_2 t_3]=[0\ 10\ 5]$, $X=[x_1 x_2 x_3]=\begin{bmatrix}2&5&1\\3&1&6\end{bmatrix}$, $r_1=2, r_2=1, r_3=3$
$\Rightarrow w^*=\left(\sum_{n=1}^{N} r_n x_n x_n^T\right)^{-1}\left(\sum_{n=1}^{N} r_n t_n x_n\right)$
$=\left(2\cdot\begin{bmatrix}2\\3\end{bmatrix}\cdot[2\ 3]+1\cdot\begin{bmatrix}5\\1\end{bmatrix}\cdot[5\ 1]+3\cdot\begin{bmatrix}1\\6\end{bmatrix}\cdot[1\ 6]\right)^{-1}\left(2\cdot[0]\cdot\begin{bmatrix}2\\3\end{bmatrix}+1\cdot[10]\cdot\begin{bmatrix}5\\1\end{bmatrix}+3\cdot[5]\cdot\begin{bmatrix}1\\6\end{bmatrix}\right)$
$=\left(\begin{bmatrix}8&12\\12&18\end{bmatrix}+\begin{bmatrix}25&5\\5&1\end{bmatrix}+\begin{bmatrix}3&18\\18&108\end{bmatrix}\right)^{-1}\left(\begin{bmatrix}0\\0\end{bmatrix}+\begin{bmatrix}50\\10\end{bmatrix}+\begin{bmatrix}15\\90\end{bmatrix}\right)$
$=\begin{bmatrix}136&107\\107&127\end{bmatrix}^{-1}\begin{bmatrix}125\\100\end{bmatrix}$
$=\frac{1}{2267}\begin{bmatrix}127&-107\\-107&136\end{bmatrix}\begin{bmatrix}125\\100\end{bmatrix}$
$=\frac{1}{2267}\begin{bmatrix}5175\\-2375\end{bmatrix}$ #

5.


5. 令 $y_n = y(x_n, w)$
$\hat{y}_n = w_0 + \sum_{i=1}^{D} w_i (x_n^{(i)} + \epsilon_i) = y_n + \sum_{i=1}^{D} w_i \epsilon_i$
$\hat{E} = \frac{1}{2} \sum_{n=1}^{N} (\hat{y}_n - t_n)^2$
$= \frac{1}{2} \sum_{n=1}^{N} (\hat{y}_n^2 - 2\hat{y}_n t_n + t_n^2)$
$= \frac{1}{2} \sum_{n=1}^{N} (y_n^2 - 2y_n \sum_{i=1}^{D} w_i \epsilon_i + (\sum_{i=1}^{D} w_i \epsilon_i)^2 - 2t_n y_n - 2t_n \sum_{i=1}^{D} w_i \epsilon_i + t_n^2)$
對 ϵ 算期望值 $E[\hat{E}] = \frac{1}{2} \sum_{n=1}^{N} (E[y_n^2] - E[2y_n \sum_{i=1}^{D} w_i \epsilon_i] + E[(\sum_{i=1}^{D} w_i \epsilon_i)^2] - E[2t_n y_n] - E[2t_n \sum_{i=1}^{D} w_i \epsilon_i] + E[t_n^2]$
$\because E[\epsilon_i] = 0$ (由題目假設)
$\therefore E[\hat{E}] = \frac{1}{2} \sum_{n=1}^{N} (y_n^2 - 0 + E[(\sum_{i=1}^{D} w_i \epsilon_i)^2] - 2t_n y_n - 0 + t_n^2)$
$= \frac{1}{2} \sum_{n=1}^{N} (y_n^2 + E[(\sum_{i=1}^{D} w_i \epsilon_i)^2] - 2t_n y_n + t_n^2)$
又 $\because E[\epsilon_i^2] = \delta_{ii} \sigma^2 = 1 \cdot \sigma^2 = \sigma^2$
$\therefore E[\hat{E}] = \frac{1}{2} \sum_{n=1}^{N} (y_n^2 - 2t_n y_n + t_n^2 + \sum_{i=1}^{D} w_i^2 \sigma^2)$
$= \frac{1}{2} \sum_{n=1}^{N} (y_n - t_n)^2 + \sum_{i=1}^{D} w_i^2 \sigma^2$
$= E(w) + \sum_{i=1}^{D} w_i^2 \sigma^2$，得證

6.


6. $\frac{d}{dx} \ln |A|$
$= \frac{d}{dx} \ln \prod_{i=1}^{D} \lambda_i$，$\lambda_i$ 為 A 之 eigenvalue
$= \frac{d}{dx} \sum_{i=1}^{D} \ln \lambda_i$
$= \frac{d}{dx} Tr(\ln A)$
$= Tr(\frac{d}{dx} \ln A)$
$= Tr(A^{-1} \frac{d}{dx} A)$，得證