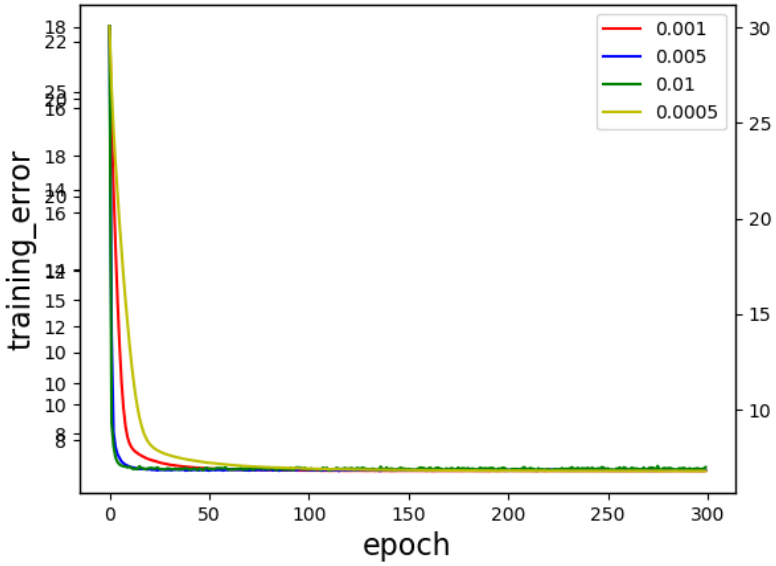
Homework 1 Report - PM2.5 Prediction

學號：b05901009 系級：電機三 姓名:高瑋聰

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。



使用 Adam 為 optimizer，由於是一種 Adaptive learning rate 的方法，因此起始 learning rate 的大小影響不大，由圖中也可看出只有一開始的收斂速度略有影響，但最後的 loss/error 則降到差不多的地方。

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）。

	Public	Private
使用所有 feature :	7.35981	7.51149
只使用 PM2.5 :	9.43819	9.34912

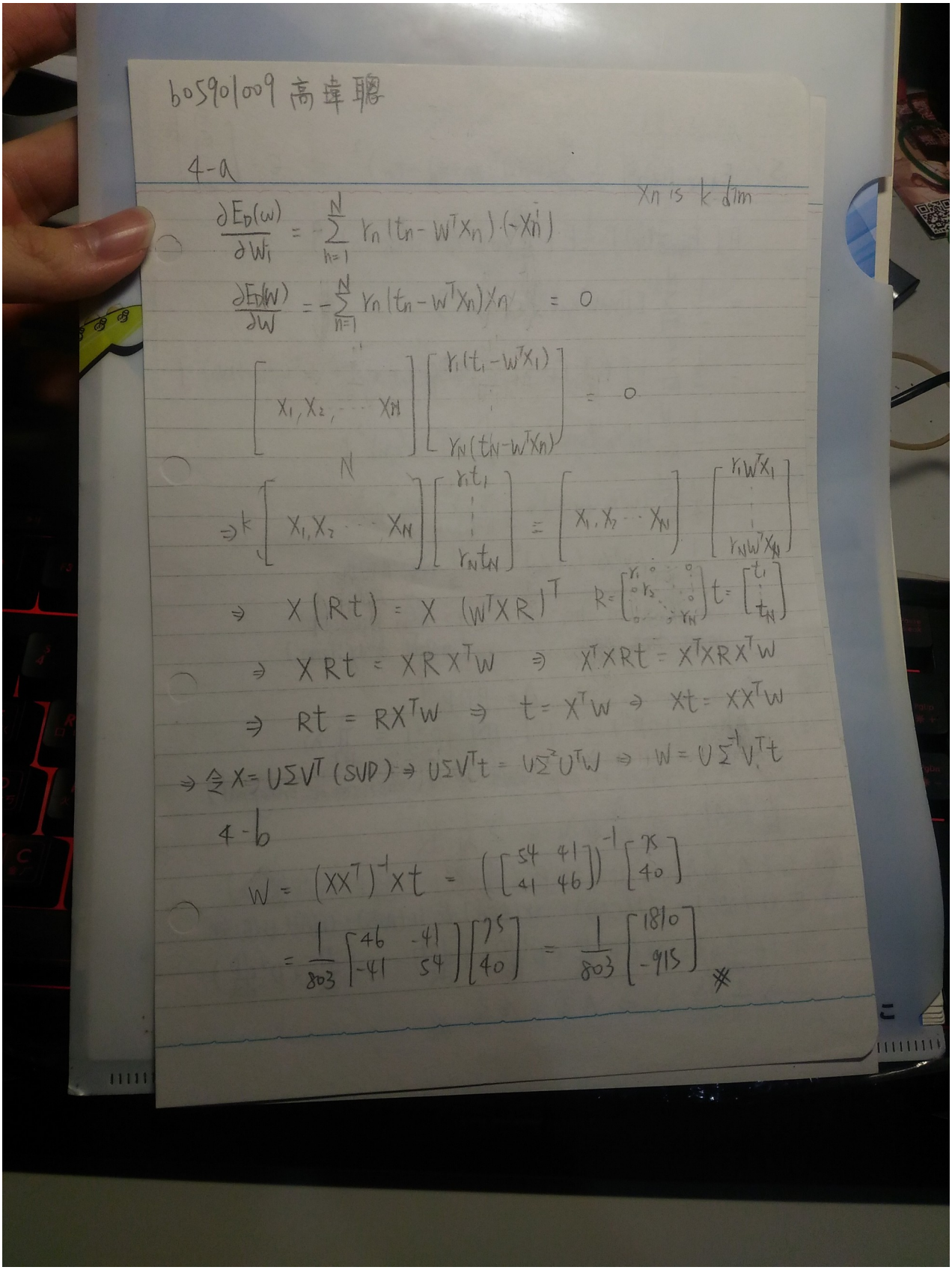
由此可知使用其他 feature 有助於提昇 PM2.5 的預測，在只使用 PM2.5 的資料時 RMSE 有明顯的上升。換言之，PM2.5 數值與空氣中的其他成份有一定的關聯性。

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

	Train	Public	Private	L2norm
λ = 0 :	6.1609	7.36817	7.53199	44.1842
λ = 0.001 :	6.2420	7.50029	7.61390	43.3508
λ = 0.01 :	6.8251	7.48780	7.51978	39.3967
λ = 0.1 :	9.9944	8.08249	7.97652	32.1635

根據以上數據，可看出 L2 regularization 大時確實可以讓 weight 的 norm 變小，同時適當的 regularization 也有助於減少 training error 與 testing error 的差，但是太大的 regularization 會導致 training 和 testing 一起變差，應是 regularization 的影響太大導致無法 fit training data。

4~6.請見下兩頁照片


b05901009 高韋聰
4-a
X_n is k dim
$\frac{\partial E_D(w)}{\partial w_i} = -\sum_{n=1}^{N} r_n(t_n - w^T x_n)(-x_n^i)$
$\frac{\partial E_D(w)}{\partial w} = -\sum_{n=1}^{N} r_n(t_n - w^T x_n)x_n = 0$
$\begin{bmatrix} x_1, x_2, \cdots x_N \end{bmatrix}\begin{bmatrix} r_1(t_1 - w^T x_1) \\ \vdots \\ r_N(t_N - w^T x_N) \end{bmatrix} = 0$
$\Rightarrow k\begin{bmatrix} x_1, x_2 \cdots x_N \end{bmatrix}\begin{bmatrix} r_1 t_1 \\ \vdots \\ r_N t_N \end{bmatrix} = \begin{bmatrix} x_1, x_2 \cdots x_N \end{bmatrix}\begin{bmatrix} r_1 w^T x_1 \\ \vdots \\ r_N w^T x_N \end{bmatrix}$
$\Rightarrow X(Rt) = X(w^T X R)^T \quad R = \begin{bmatrix} r_1 & & \\ & r_2 & \\ & & r_N \end{bmatrix} \quad t = \begin{bmatrix} t_1 \\ t_N \end{bmatrix}$
$\Rightarrow XRt = XRX^T w \Rightarrow X^T XRt = X^T XRX^T w$
$\Rightarrow Rt = RX^T w \Rightarrow t = X^T w \Rightarrow Xt = XX^T w$
$\Rightarrow \because X = U\Sigma V^T$ (SVD) $\Rightarrow U\Sigma V^T t = U\Sigma^2 U^T w \Rightarrow w = U\Sigma^{-1}V^T t$
4-b
$w = (XX^T)^{-1}Xt = \left(\begin{bmatrix} 54 & 41 \\ 41 & 46 \end{bmatrix}\right)^{-1}\begin{bmatrix} 75 \\ 40 \end{bmatrix}$
$= \frac{1}{803}\begin{bmatrix} 46 & -41 \\ -41 & 54 \end{bmatrix}\begin{bmatrix} 75 \\ 40 \end{bmatrix} = \frac{1}{803}\begin{bmatrix} 1810 \\ -915 \end{bmatrix}$ ※

5. $E_{noise}(w) = \frac{1}{2}\sum_{n=1}^{N}(w_0 + w^T(x_n+\epsilon_n) - t_n)^2$, $\epsilon = \begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \vdots\\ \epsilon_D\end{bmatrix}$

$$E[E_{noise}(w)] = \frac{1}{2}E\left[\sum_{n=1}^{N}(w_0 + w^Tx_n + w^T\epsilon_n - t_n)^2\right]$$

$$= \frac{1}{2}\sum_{n=1}^{N}E[(w_0 + w^Tx_n + w^T\epsilon_n - t_n)^2]$$

$$= \frac{1}{2}\sum_{n=1}^{N}E[(w^T\epsilon_n)^2 + 2w^T\epsilon_n(w^Tx_n - t_n + w_0) + (w^Tx_n - t_n + w_0)^2]$$

$$= \frac{1}{2}\sum_{n=1}^{N}(w^Tx_n - t_n + w_0)^2 + \frac{1}{2}\sum_{n=1}^{N}E[(w^T\epsilon_n)^2]$$

$$= \frac{1}{2}\sum_{n=1}^{N}(w^Tx_n - t_n + w_0)^2 + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{D}w_k^2\sigma^2$$

$$= \underbrace{\frac{1}{2}\sum_{n=1}^{N}(w^Tx_n - t_n + w_0)^2}_{\text{sum-of-square error for noise free input}} + \underbrace{\frac{1}{2}N\sigma^2\|w\|_2}_{\text{L2-norm regularization (weight-decay term)}}$$

6. A is symmetric $\rightarrow$ $A = PDP^{-1}$

(左) $|A| = |PDP^{-1}| = |P|\cdot|D|\,|P^{-1}| = \prod_{i=1}^{n}\lambda_i$

(∵ P is composed of eigenvectors of A, which can be orthogonal)

$$\frac{d}{dx}\ln|A| = \frac{d}{dx}\sum_{i=1}^{n}\ln\lambda_i = \sum\frac{1}{\lambda_i}\frac{d\lambda_i}{dx}$$

(右) $tr\left(A^{-1}\frac{d}{dx}A\right) = tr\left(PD^{-1}P^{-1}\left(\frac{dP}{dx}DP^{-1} + P\frac{dD}{dx}P^{-1} + PD\frac{dP^{-1}}{dx}\right)\right)$

由 $tr(ABC) = tr(CAB) = tr(BCA)$ 及 $tr(A+B) = tr(A) + tr(B)$ 知

$$tr\left(A^{-1}\frac{d}{dx}A\right) = tr\left(P^{-1}\frac{dP}{dx} + D^{-1}\frac{dD}{dx} + P\frac{dP^{-1}}{dx}\right) = tr\left(\frac{dI}{dx} + D^{-1}\frac{dD}{dx}\right)$$

$$= \sum\frac{1}{\lambda_i}\frac{d\lambda_i}{dx} \Rightarrow \text{左式} = \text{右式}\ \text{※}$$