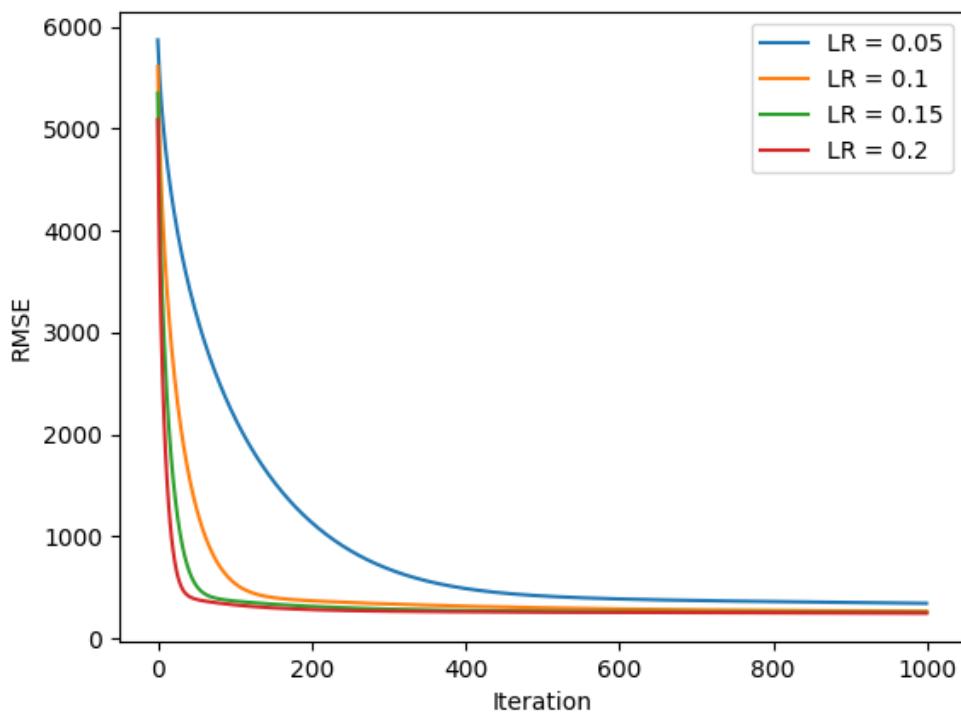


# ML HW1 REPORT

b04504042 電機三 劉家豪

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。

以下為示意圖：



learning rate 較大的，收斂速度較快，RMSE 驟降，很快地loss就趨於平穩。而 learning rate較小的，曲線較圓滑，收斂較慢，花較多時間才趨於平穩。

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
所有feature	8.80762	8.82028
PM2.5	9.03261	8.74818

所有feature都加，做出來的效果略好一些，推測應該是因為pm2.5裡面有些叫偏差的數據，沒有被處理掉，而全部feature都下讓loss function 更複雜，且各種數據綜合下，能彌補pm2.5數據的一些偏差。

3. (1%) 請分別使用至少四種不同數值的regularization parameter  $\lambda$ 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。

lambda	Public	Private	norm
0.001	7.54265	7.45600	1.87276
0.01	7.51033	7.51576	1.88265
0.05	7.55539	7.52132	1.88374
0.1	7.61134	7.84042	1.92568

feature 使用 pm2.5 , pm10 ,co , so2的一次項，取前九小時  
四組數據差異並沒有太多，可能是我iteration做比較多次，最後都收斂在差不多地方，norm的大小也沒有太大的差異。

4~6 (3%) 請參考數學題目 (連結：)，將作答過程以各種形式 (latex尤佳) 清楚地呈現在pdf檔中 (手寫再拍照也可以，但請注意解析度)。

collaborator : b04203058 蘇軒 , b04505025 陳在賢

(4-a)

$$\begin{aligned}
 E_D(w) &= \frac{1}{2} (\hat{y}^T - \hat{w}^T \hat{x}) \hat{R} (\hat{y}^T - \hat{w}^T \hat{x})^T \\
 &= \frac{1}{2} (\hat{y}^T - \hat{w}^T \hat{x}) \hat{R} (\hat{y} - \hat{x}^T \hat{w}) \\
 &= \frac{1}{2} (\hat{y}^T \hat{R} \hat{y} - \hat{w}^T \hat{x} \hat{R} \hat{y} - \hat{y}^T \hat{R} \hat{x}^T \hat{w} + \hat{w}^T \hat{x} \hat{R} \hat{x}^T \hat{w})
 \end{aligned}$$

$$\hat{y} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

$$\hat{x} = \begin{bmatrix} \vec{x}_1, \vec{x}_2, \dots, \vec{x}_N \end{bmatrix}$$

minimum  $\Rightarrow \nabla_w E_D(w) = \hat{x} \hat{R} \hat{x}^T \hat{w} - \hat{x} \hat{R} \hat{y} = 0$

$$\hat{x} \hat{R} \hat{x}^T \hat{w} = \hat{x} \hat{R} \hat{y}$$

$$w^* = (\hat{x} \hat{R} \hat{x}^T)^{-1} \hat{x} \hat{R} \hat{y}$$

$$\hat{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}$$

(4-b)

$$\begin{aligned}
 \hat{x} \hat{R} \hat{x}^T &= \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 5 & 1 \\ 5 & 6 \end{bmatrix} \\
 &= \begin{bmatrix} 108 & 107 \\ 107 & 127 \end{bmatrix}
 \end{aligned}$$

$$\hat{R} = \begin{bmatrix} r_{11} & 0 & 0 & \cdots & 0 \\ 0 & r_{22} & \cdots & 0 \\ 0 & 0 & r_3 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r_n \end{bmatrix}$$

$$(\hat{x} \hat{R} \hat{x}^T)^{-1} = \begin{bmatrix} \frac{127}{2267} & -\frac{107}{2267} \\ -\frac{107}{2267} & \frac{108}{2267} \end{bmatrix}$$

$$\hat{x} \hat{R} \hat{y} = \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix} = \begin{bmatrix} 125 \\ 100 \end{bmatrix}$$

$$w^* = \begin{bmatrix} 2.2827 \\ -1.1358 \end{bmatrix}$$

5.

$$y(x, w) = w_0 + \sum_{i=1}^D w_i x_i$$

$$\text{noise-free} \Rightarrow E(w) = \frac{1}{2} \sum_{n=1}^N (y(x_n, w) - t_n)^2$$

$$E(w) = \frac{1}{2} \sum_{n=1}^N \left( \sum_{i=1}^D \sum_{j=1}^D w_i w_j x_{ni} x_{nj} - 2(w_0 - t_n) \sum_{i=1}^D x_{ni} + (w_0 - t_n)^2 \right)$$

$$\text{with noise} \Rightarrow \bar{E}(w) = \frac{1}{2} \sum_{n=1}^N \left( \sum_{i=1}^D \sum_{j=1}^D w_i w_j (x_{ni} + \epsilon_{ni})(x_{nj} + \epsilon_{nj}) - 2(w_0 - t_n) \sum_{i=1}^D (x_{ni} + \epsilon_{ni}) + (w_0 - t_n)^2 \right)$$

$$\begin{aligned} E[\bar{E}(w)] &= \frac{1}{2} \sum_{n=1}^N \left( \sum_{i=1}^D \sum_{j=1}^D E[w_i w_j (x_{ni} + \epsilon_{ni})(x_{nj} + \epsilon_{nj})] - 2(w_0 - t_n) \sum_{i=1}^D E[x_{ni} + \epsilon_{ni}] + (w_0 - t_n)^2 \right) \\ &= \frac{1}{2} \sum_{n=1}^N \left[ \sum_{i=1}^D w_i^2 D \sigma^2 + \sum_{i=1}^D \sum_{j=1}^D w_i w_j x_{ni} x_{nj} - 2(w_0 - t_n) \sum_{i=1}^D x_{ni} + (w_0 - t_n)^2 \right] \\ &= \underbrace{E[E(w)]}_{\text{期望值}} + \underbrace{\frac{NP}{2} \sigma^2 \|w\|^2}_{\text{方差}} \end{aligned}$$

6.

$$\frac{d}{dx} \ln(|A|) = \frac{1}{|A|} \frac{d}{dx} |A| = \frac{1}{|A|} \left( \frac{d}{dx} |A| \right) = \text{Tr} \left( \frac{d}{dx} A \right)$$

$\frac{d}{dx} |A| = \text{Tr}(\text{adj}(A) \cdot \frac{dA}{dx})$ , for which  $\text{adj}(A)$  is  $A$ 's adjoint matrix

$$\text{from } A' = \frac{1}{|A|} \text{adj}(A)$$

$$\text{we get } \frac{d}{dx} |A| = \text{Tr}(\text{adj}(A) \cdot \frac{dA}{dx})$$

$$\Rightarrow \frac{1}{|A|} \frac{d}{dx} |A| = \frac{1}{|A|} |A| \text{Tr} \left( A^{-1} \cdot \frac{dA}{dx} \right)$$

$$= \underbrace{\text{Tr} \left( A^{-1} \frac{dA}{dx} \right)}_{\text{迹}}$$