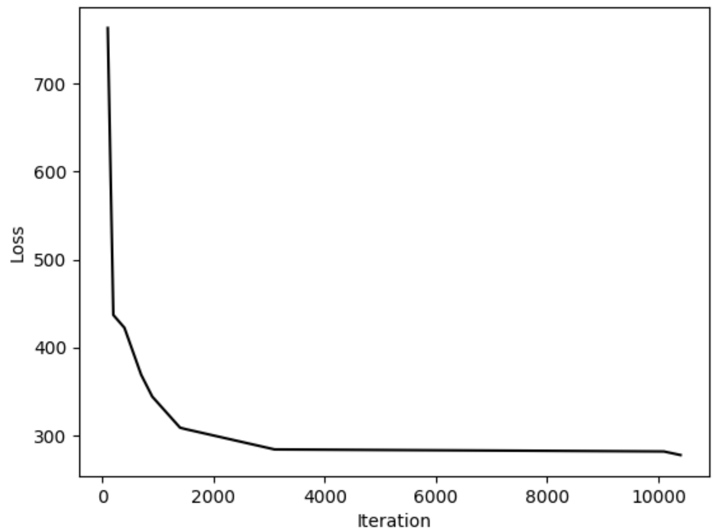
Homework 1 Report - PM2.5 Prediction

學號：B05901043　系級：電機三　姓名: 莊鎧爾

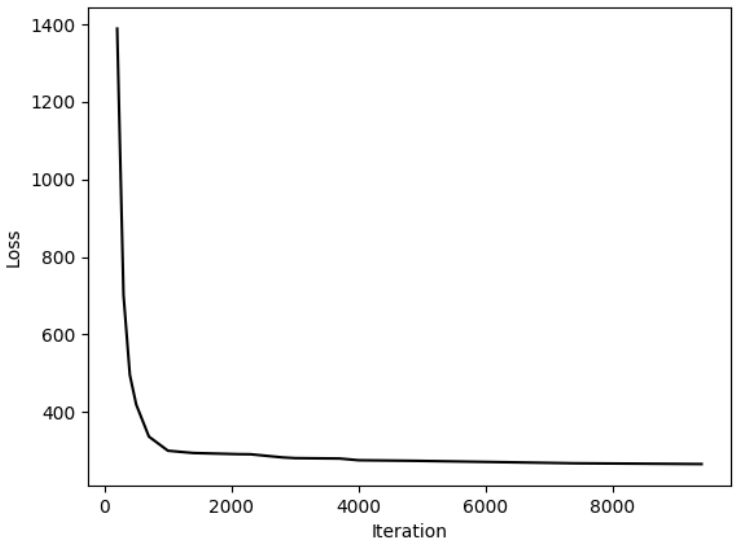
1. (1%) 請分別使用至少 **4** 種不同數值的 learning rate 進行 training（其他參數需一致），對其作圖，並且討論其收斂過程差異。

我將 Loss function 定義為 $Loss = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{2\times n}$

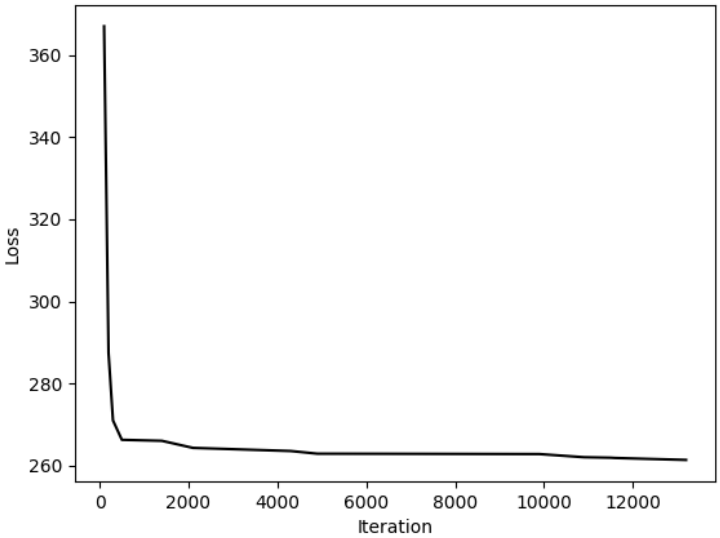
Training rate = 500



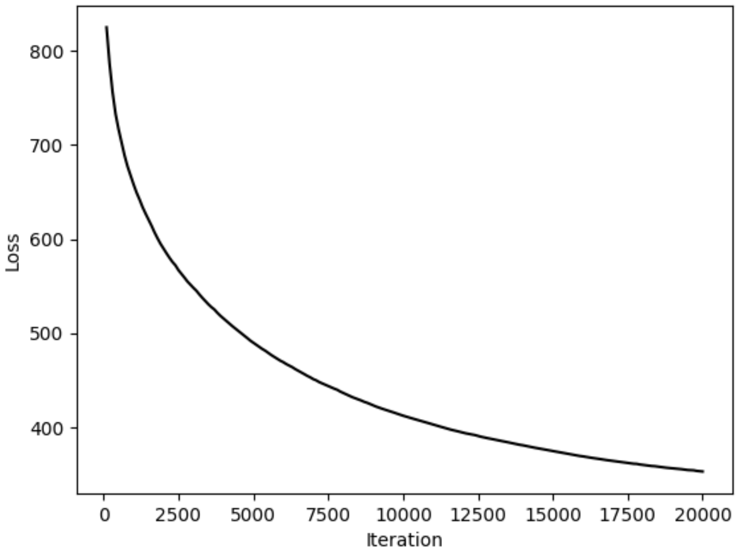
Training rate = 100



Training rate = 10



Training rate = 0.1



可以知道，training rate 介於 10 至 100 這個範圍時，會有比較快的收斂結果，尤其是 training rate = 10 時，不到 1000 次就幾乎要收斂到最小的 loss，而當 training rate 提高到 500 時，gradient descent 執行時會走的太大步，而走過最低點，因此收斂速度反而不如 learning rate = 10 和 100 時快，當 training rate 下降到 0.1 時，gradient descent 執行時每一次參數更新太少，所以收斂速度非常緩慢，但可以看到 loss 呈現穩定下降，因此只要等待的時間夠久，也可以走到最小值。

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）。

結果如下（根據 Kaggle public score）：

i. 每筆 data9 小時內所有 feature 的一次項（含 bias 項）：9.49520

ii. 每筆 data9 小時內 PM2.5 的一次項（含 bias 項）：9.56416

可以知道若只使用 PM2.5 的 data，資料量太少，得到的結果反而不如資料量多時好。

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。

regularization parameter λ	0.1	1.0	10	50	100
RMSE (training)	22.87	22.89	23.01	24.53	25.88
RMSE (testing) Kaggle public score	8.99193	9.65006	9.26648	9.23758	9.63401
L2 norm	2.957×10^6	2.961×10^6	3.012×10^6	3.400×10^6	3.785×10^6

當 λ 變大時，training set 上表現的結果比較差，所以 RMSE 和 L2 norm 都比較差，但在 testing set 上，可以看到，若 λ 不要設太大，可以得到比較好的結果，但是如果設太大，反而過度的降低 weight，而在 testing set 上也表現的不好。

4.

(4-a)

$$let\ \boldsymbol{X} = [\boldsymbol{x_1}\ \boldsymbol{x_2} \cdots\ \boldsymbol{x_n}], \qquad \boldsymbol{t} = [t_1\ t_2 \cdots t_n]^T, \qquad \boldsymbol{R} = [c_{ij}]\ where\ c_{ij} = \begin{cases} 0\ if\ i \neq j \\ r_i\ if\ i = j \end{cases}$$

$$Therefore, E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n(t_n - \boldsymbol{w^T x_n})^2 = \frac{1}{2}(\boldsymbol{t} - \boldsymbol{X^T w})^T \boldsymbol{R}(\boldsymbol{t} - \boldsymbol{X^T w})$$

$Find\ \nabla_{\boldsymbol{w}} E_D(\boldsymbol{w})$

$E_D(\boldsymbol{w} + \boldsymbol{\Delta w}) - E_D(\boldsymbol{w})$

$$= \frac{1}{2}(\boldsymbol{t} - \boldsymbol{X^T}(\boldsymbol{w} + \boldsymbol{\Delta w}))^T \boldsymbol{R}\Big(\boldsymbol{t} - \boldsymbol{X^T}(\boldsymbol{w} + \boldsymbol{\Delta w})\Big) - \frac{1}{2}(\boldsymbol{t} - \boldsymbol{X^T w})^T \boldsymbol{R}(\boldsymbol{t} - \boldsymbol{X^T w})$$

$$= \frac{1}{2}[-(\boldsymbol{X^T \Delta w})^T \boldsymbol{R}(\boldsymbol{t} - \boldsymbol{X^T w}) - (\boldsymbol{t} - \boldsymbol{X^T w})^T \boldsymbol{R}(\boldsymbol{X^T \Delta w}) + (\boldsymbol{X^T \Delta w})^T \boldsymbol{R}(\boldsymbol{X^T \Delta w})]$$

$$= \frac{1}{2}[-2(\boldsymbol{X^T \Delta w})^T \boldsymbol{R}(\boldsymbol{t} - \boldsymbol{X^T w}) + (\boldsymbol{X^T \Delta w})^T \boldsymbol{R}(\boldsymbol{X^T \Delta w})]$$

$$= \frac{1}{2}\boldsymbol{\Delta w^T}[-2\boldsymbol{XR}(\boldsymbol{t} - \boldsymbol{X^T w}) + \boldsymbol{XRX^T \Delta w}] \xRightarrow{\boldsymbol{\Delta w}\ is\ small} -\boldsymbol{\Delta w^T XR}(\boldsymbol{t} - \boldsymbol{X^T w})$$

$\because \boldsymbol{\Delta w^T}\ \nabla_{\boldsymbol{w}} E_D(\boldsymbol{w}) = E_D(\boldsymbol{w} + \boldsymbol{\Delta w}) - E_D(\boldsymbol{w}) = -\boldsymbol{\Delta w^T XR}(\boldsymbol{t} - \boldsymbol{X^T w})$

$\therefore\ \nabla_{\boldsymbol{w}} E_D(\boldsymbol{w}) = -\boldsymbol{XR}(\boldsymbol{t} - \boldsymbol{X^T w})$

$Find\ \boldsymbol{w}^*\ that\ minimizes\ E_D(\boldsymbol{w}), we\ solve\ \nabla_{\boldsymbol{w}} E_D(\boldsymbol{w}^*) = -\boldsymbol{XR}(\boldsymbol{t} - \boldsymbol{X^T w^*}) = 0$

$\boldsymbol{XRt} = \boldsymbol{XRX^T w^*}$

$\boldsymbol{w}^* = (\boldsymbol{XRX^T})^{-1}(\boldsymbol{XRt})$

(4-b)

$$\boldsymbol{X} = \begin{bmatrix} 2 & 5 & 5 \\ 3 & 1 & 6 \end{bmatrix}, \boldsymbol{R} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \boldsymbol{t} = \begin{bmatrix} 0 \\ 10 \\ 5 \end{bmatrix}$$

$$\boldsymbol{w}^* = (\boldsymbol{XRX^T})^{-1}(\boldsymbol{XRt}) = \begin{bmatrix} 2.283 \\ -1.136 \end{bmatrix}$$

5.

Because the syntax in question is misleading, I let

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

Now, I define E averaged over the noise distribution is E'

$$E'(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(\boldsymbol{x_n}', \boldsymbol{w}) - t_n)^2 = \frac{1}{2}\sum_{n=1}^{N}\left(w_0 + \sum_{i=1}^{D} w_i(x_i^n + \epsilon_i^n) - t_n\right)^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left(\left(w_0 + \sum_{i=1}^{D} w_i x_i^n - t_n\right) + \sum_{i=1}^{D} w_i \epsilon_i^n\right)^2$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left(w_0 + \sum_{i=1}^{D} w_i x_i^n - t_n\right)^2 + \sum_{n=1}^{N}\left[\left(w_0 + \sum_{i=1}^{D} w_i x_i^n - t_n\right)\left(\sum_{i=1}^{D} w_i \epsilon_i^n\right)\right]$$

$$+ \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{i=1}^{D} w_i \epsilon_i^n\right)^2$$

$$= E(\boldsymbol{w}) + \sum_{n=1}^{N}\left[(y(\boldsymbol{x_n}, \boldsymbol{w}) - t_n)\left(\sum_{i=1}^{D} w_i \epsilon_i^n\right)\right] + \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{i=1}^{D} w_i \epsilon_i^n\right)^2$$

Because of $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ and $\mathbb{E}[\epsilon_i] = 0$, we know $\sum_{n=1}^{N} \epsilon_i \epsilon_j = \delta_{ij} N \sigma^2$ and $\sum_{n=1}^{N} \epsilon_j = 0$

$$\sum_{n=1}^{N}\left[(y(\boldsymbol{x_n}, \boldsymbol{w}) - t_n)\left(\sum_{i=1}^{D} w_i \epsilon_i^n\right)\right] = 0$$

$$\sum_{n=1}^{N}\left(\sum_{i=1}^{D} w_i \epsilon_i^n\right)^2 = \sum_{n=1}^{N}\sum_{i=1}^{D}(w_i \epsilon_i)^2 = N\sigma^2 \sum_{i=1}^{D} w_i^2$$

Therefore,

$$E'(\boldsymbol{w}) = E(\boldsymbol{w}) + N\sigma^2 \sum_{i=1}^{D} w_i^2$$

The noise distribution is equivalent to sum-of-squares error for noise-free input variables with the addition of a weight -decay regularization term, in which the bias parameter w_0 is omitted.

6.

Let $\mathbf{A}$ with eigenvalue $\lambda_1, \lambda_2, \cdots, \lambda_n$, then

Because $\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i$

$$\frac{d}{d\alpha}\ln(|\boldsymbol{A}|) = \frac{d}{d\alpha}\ln\left(\prod_{i=1}^{n}\lambda_i\right) = \frac{d}{d\alpha}\sum_{i=1}^{n}\ln(\lambda_i) = \sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln(\lambda_i) = \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\mathrm{d}}{\mathrm{d}\alpha}\lambda_i$$

And we know that

$$A^{-1} \text{ with eigenvalue } \frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \cdots, \frac{1}{\lambda_n}$$

$$\frac{d}{d\alpha}A \text{ with eigenvalue } \frac{d}{d\alpha}\lambda_1, \frac{d}{d\alpha}\lambda_2, \cdots, \frac{d}{d\alpha}\lambda_n$$

$$\boldsymbol{A}^{-1}\frac{d}{d\alpha}\boldsymbol{A} \text{ with eigenvalue } \frac{1}{\lambda_1}\frac{d}{d\alpha}\lambda_1, \frac{1}{\lambda_2}\frac{d}{d\alpha}\lambda_2, \cdots, \frac{1}{\lambda_n}\frac{d}{d\alpha}\lambda_n$$

Because $\text{trace}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$

$$Tr\left(\boldsymbol{A}^{-1}\frac{d}{d\alpha}\boldsymbol{A}\right) = \sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{d}{d\alpha}\lambda_i$$

Therefore,

$$\frac{d}{d\alpha}\ln(|\boldsymbol{A}|) = Tr\left(\boldsymbol{A}^{-1}\frac{d}{d\alpha}\boldsymbol{A}\right)$$