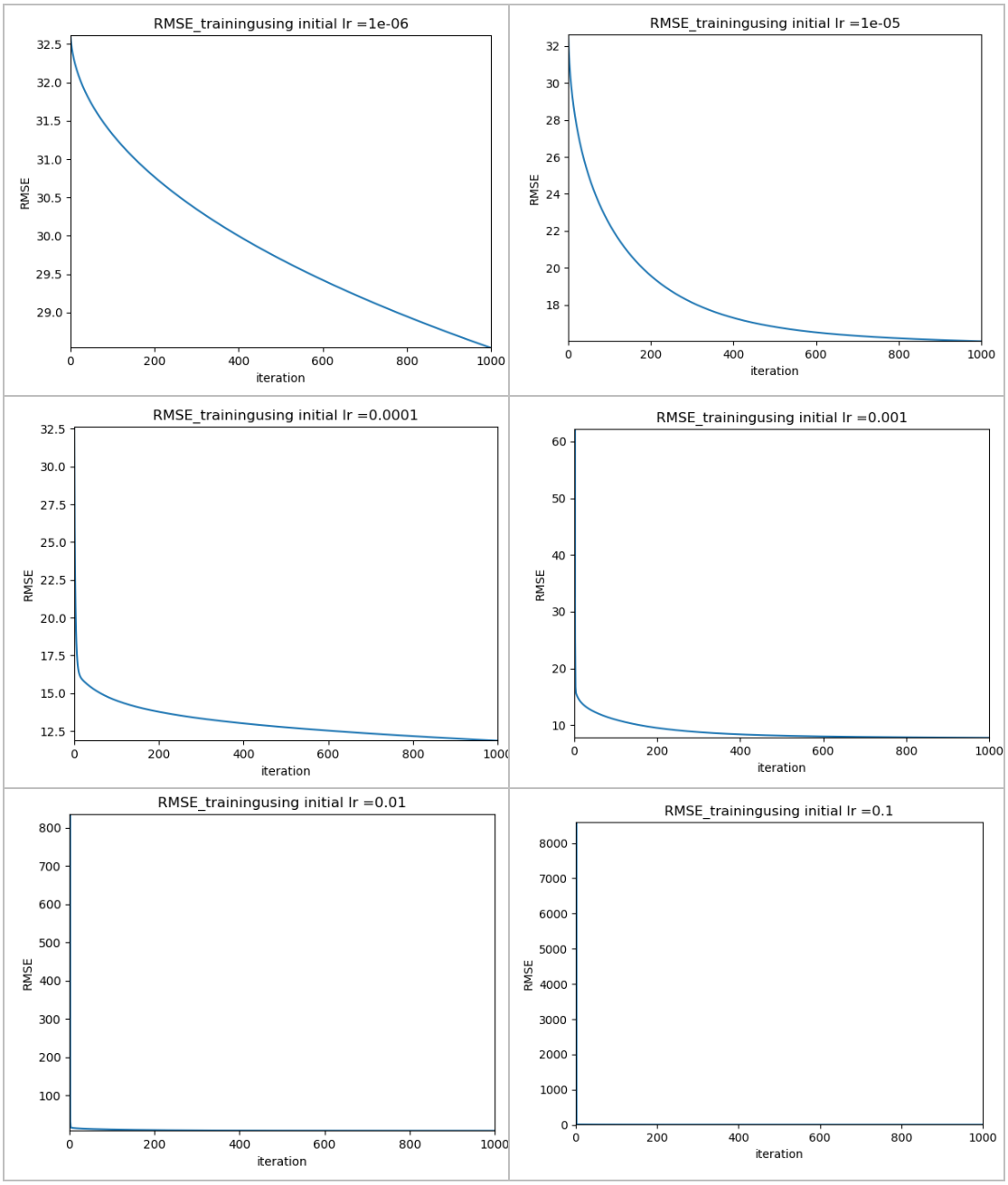
Homework 1 Report - PM2.5 Prediction

學號：r06631035　系級：生機所碩二　姓名: 王凱陞

1.



由左至右，上至下的順序 learning rate 分別為 0.000001、0.00001、0.0001、0.001、0.01、0.1 的 RMSE-iteration 變化圖，固定 iteration 為 1000 且加入 regularization 固定參數為 0.05，可從上述順序看出，learning rate 的上升會使整個 training 的收斂加快，也就是如同課堂所說的 learning rate 的提高會愈快趨近到低處，但這部分看不到的是 learning rate 過大產生的震盪效果，是由於這部分我已經採用 adagrad，所以實際 training 進去的 learning rate 會隨著權重和 iteration 的進行而被拉低。

2.

Submission and Description	Private Score	Public Score
submission.csv a minute ago by r06631035_王凱陞 For Question 2, all feature.	8.79604	9.07504

上圖為所有 feature 的一次項所得到的結果，iteration 為 5000，learning rate 為 0.01，未加入 regularization，僅做將小於零的資料轉為 0 的資料處理所跑出的結果。

Submission and Description	Private Score	Public Score
submission.csv just now by r06631035_王凱陞 For Question 2, only PM2.5 as feature.	10.10747	10.03291

此為僅有 PM2.5 的一次項，其餘參數和資料處理皆和上圖一致。

可由此二結果看到，若單一只用 PM2.5 的一次項作為 feature，明顯會造成結果不理想，雖然並非所有 feature 都跟預測 PM2.5 具一定的相關性，但使用全部的 feature 會增加更多的資料，且有部分 feature 對 PM2.5 的預測有其相關性，故預測的結果會較好一些，但 PM2.5 的一次項在 public/private score 上較為接近一些，這有可能是資料分佈的情形不同，而使用單一使用 PM2.5 預測結果較差以外，也僅用單一 feature，故受到其餘 feature 的資料分部不均影響較小，所以才導致此情況。

3.

	Private	Public	Weight L2 norm	RMSE for training
$\lambda = 0.001$	7.63446	7.21778	14.05656	6.38875
$\lambda = 0.01$	7.64086	7.20220	13.73623	6.38943
$\lambda = 0.05$	7.66619	7.14815	12.32904	6.40561
$\lambda = 0.1$	7.74193	7.11676	10.62326	6.45660
$\lambda = 0.5$	10.54464	9.09383	8.236014	8.40138

上表為 regularization parameter λ 由上到下分別為 0.001、0.01、0.05、0.1、0.5 所輸出的結果，此結果有經資料處理且 iteration 固定為 6617，learning rate 設定在 0.001，可由結果看到當 λ 從 0.001 遞減至 0.1 間，public score 都是呈現變好的情況，但 private score 則是有持續變差，到了 0.5 則是兩個都變得很差，而 weight 都是下降的情況，表示 regularization 對權重確實有 bound 住的情況，這點在 training 的 RMSE 也可以看到，結果都是愈來愈差，在合理的情況下確實可以避免 overfitting 的狀況產生，在 testing score 上呼應了老師上課所說的 regularization 的參數設定上，需要經過 validation set 多方的嘗試，不可以太過相信在 public 的結果，不然很容易在 private 產生過多的限制或是 overfitting，像上表的結果明顯 regularization 太大之後就會產生 private 比較差的情況。

下圖為 kaggle 輸出結果，僅供參考:

Submission and Description	Private Score	Public Score
submission.csv just now by r06631035_王凱陞 Question 3, lambda = 0.5, L2 norm: 8.23601370871	10.54464	9.09383
submission.csv 4 minutes ago by r06631035_王凱陞 Question 3, lambda = 0.1, L2 norm: 10.6232589836	7.74193	7.11676
submission.csv 6 minutes ago by r06631035_王凱陞 Question 3, lambda = 0.05, L2 norm: 12.3290443083	7.66619	7.14815
submission.csv 8 minutes ago by r06631035_王凱陞 Question 3, lambda = 0.01, L2 norm: 13.7360522893	7.64086	7.20220
submission.csv 9 minutes ago by r06631035_王凱陞 Question 3, lambda = 0.001, L2 norm: 14.05656328	7.63446	7.21778

4 (a). 根據題意:

$$\text{SSE}(\mathbf{w}) = E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\,(t_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2$$

Let

$$\mathbf{x} = [\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{x}_N] = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{(M+1)1} & \cdots & x_{(M+1)N} \end{bmatrix}, \qquad \boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_{M+1} \end{bmatrix}$$

, $\mathbf{r} = \begin{bmatrix} r_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_N \end{bmatrix}$ (主對角線為r_1 、r_2 、 … 、r_N)

→則 $\mathbf{t} - \boldsymbol{x}^T\boldsymbol{w} = [t_1 \;\; t_2 \;\; \cdots \;\; t_N] - [\boldsymbol{w}^T\boldsymbol{x}_1 \;\; \boldsymbol{w}^T\boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{w}^T\boldsymbol{x}_N]$

$$\rightarrow \text{SSE} = [t_1 - \boldsymbol{w}^T\boldsymbol{x}_1 \;\; t_2 - \boldsymbol{w}^T\boldsymbol{x}_2 \;\; \cdots \;\; t_N - \boldsymbol{w}^T\boldsymbol{x}_N] \begin{bmatrix} r_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_N \end{bmatrix} \begin{bmatrix} t_1 - \boldsymbol{w}^T\boldsymbol{x}_1 \\ \vdots \\ t_N - \boldsymbol{w}^T\boldsymbol{x}_N \end{bmatrix}$$

$$= (\mathbf{t} - \boldsymbol{w}^T\boldsymbol{x})\mathbf{r}(\mathbf{t}^T - \boldsymbol{x}^T\mathbf{w}) = \mathbf{tr}\boldsymbol{t}^T - \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{t}^T + \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\boldsymbol{w}$$

再解$\text{SSE}(\mathbf{w} + \Delta\mathbf{w}) - \text{SSE}(\mathbf{w})$

$$= \frac{1}{2}[\mathbf{tr}\boldsymbol{t}^T - \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T(\boldsymbol{w} + \Delta\boldsymbol{w}) - (\boldsymbol{w} + \Delta\boldsymbol{w})^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{t}^T + (\boldsymbol{w} + \Delta\boldsymbol{w})^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T(\boldsymbol{w} + \Delta\boldsymbol{w}) -$$

$$\mathbf{tr}\boldsymbol{t}^T - \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{t}^T + \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\boldsymbol{w}]$$

$$= \frac{1}{2}[-\boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T\Delta\mathbf{w} - \Delta\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{t}^T + \Delta\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\Delta\boldsymbol{w} + \Delta\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\Delta\boldsymbol{w}]$$

由於$\boldsymbol{x}\boldsymbol{r}\boldsymbol{t}^T$、$\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\boldsymbol{w}$為 scalar，則

$$\rightarrow \frac{1}{2}[2\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\Delta\boldsymbol{w} - 2\boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T\Delta\mathbf{w} + \Delta\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\Delta\boldsymbol{w}]$$

又因為$\Delta\boldsymbol{w}$為 infinitesimal 故可消掉，則

$$= \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T\Delta\boldsymbol{w} - \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T\Delta\mathbf{w} = (\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T - \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T)\Delta\mathbf{w}$$

此 form 即符合 $\text{f(w} + \Delta\text{w)} - \text{f(w)} = \nabla_w f(w)(\Delta w)$，

故$\nabla_{\boldsymbol{w}} E_D(\boldsymbol{w}) = (\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T - \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T)$，而最佳解$\boldsymbol{w}^*$，即為$\nabla_{\boldsymbol{w}} E_D(\boldsymbol{w})$=0 時的 $\mathbf{w}$，故

$$\nabla_{\boldsymbol{w}} E_D(\boldsymbol{w}) = (\boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T - \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T) = \mathbf{0}$$

$$\rightarrow \boldsymbol{w}^T\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T = \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T$$

$$\rightarrow \boldsymbol{w}^T = \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T(\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T)^{-1}$$

$$\rightarrow \boldsymbol{w}^T = \boldsymbol{t}\boldsymbol{r}\boldsymbol{x}^T(\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T)^{-1} \qquad (\because ((\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T)^{-1})^T = (\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T)^{-1} \text{ ，} \boldsymbol{r}^T = \boldsymbol{r})$$

$$\rightarrow \boldsymbol{w}^* = (\boldsymbol{x}\boldsymbol{r}\boldsymbol{x}^T)^{-1}\boldsymbol{x}\boldsymbol{r}\boldsymbol{t}^T$$

4 (b).

根據上述，$\boldsymbol{w}^* = (\boldsymbol{xrx}^T)^{-1}\boldsymbol{xrt}^T$

代入題目參數:

$$\boldsymbol{w}^* = (\begin{bmatrix}2 & 5 & 5\\ 3 & 1 & 6\end{bmatrix}\begin{bmatrix}2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 3\end{bmatrix}\begin{bmatrix}2 & 3\\ 5 & 1\\ 5 & 6\end{bmatrix})^{-1}\begin{bmatrix}2 & 5 & 5\\ 3 & 1 & 6\end{bmatrix}\begin{bmatrix}2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 3\end{bmatrix}[0 \quad 10 \quad 5]^{\mathrm{T}}$$

$$= (\begin{bmatrix}108 & 107\\ 107 & 127\end{bmatrix})^{-1}\begin{bmatrix}2 & 5 & 5\\ 3 & 1 & 6\end{bmatrix}\begin{bmatrix}2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 3\end{bmatrix}[0 \quad 10 \quad 5]^{\mathrm{T}}$$

$$= (\begin{bmatrix}\frac{127}{2267} & \frac{-107}{2267}\\ \frac{-107}{2267} & \frac{108}{2267}\end{bmatrix}\begin{bmatrix}2 & 5 & 5\\ 3 & 1 & 6\end{bmatrix}\begin{bmatrix}2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 3\end{bmatrix}[0 \quad 10 \quad 5]^{\mathrm{T}}$$

$$= (\begin{bmatrix}\frac{127}{2267} & \frac{-107}{2267}\\ \frac{-107}{2267} & \frac{108}{2267}\end{bmatrix}\begin{bmatrix}2 & 5 & 5\\ 3 & 1 & 6\end{bmatrix}\begin{bmatrix}2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 3\end{bmatrix}[0 \quad 10 \quad 5]^{\mathrm{T}}$$

$$= \begin{bmatrix}\frac{5175}{2267}\\ \frac{-2575}{2267}\end{bmatrix}$$

5. 根據題意，加入高斯雜訊，則y可表示為$\mathrm{y}(\mathrm{x}, \mathbf{w}, \epsilon) = w_o + \sum_{i=1}^{D} w_i(x_i + \epsilon_i)$
則 SSE 為

$\mathrm{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \boldsymbol{w},) - t_n)^2$

$= \frac{1}{2}\sum_{n=1}^{N}(w_o + \sum_{i=1}^{D} w_i(x_{n,i} + \epsilon_i) - t_n)^2$

$= \frac{1}{2}\sum_{n=1}^{N}(w_o + \sum_{i=1}^{D} w_i \boldsymbol{x}_{n,i} - t_n + \sum_{i=1}^{D} w_i\epsilon_i)^2$

$= \frac{1}{2}\sum_{n=1}^{N}(w_o + \sum_{i=1}^{D} w_i x_{n,i} - t_n)^2 + \sum_{n=1}^{N}(w_o + \sum_{i=1}^{D} w_i x_{n,i} - t_n)\,(\sum_{i=1}^{D} w_i\epsilon_i) +$

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

則$\frac{1}{2}\sum_{n=1}^{N}(w_o + \sum_{i=1}^{D} w_i x_{n,i} - t_n)^2$，即為加高斯雜訊的 Loss function，

根據題意取期望值處理其他項 (即題意的平均):
$\sum_{n=1}^{N}(w_o + \sum_{i=1}^{D} w_i x_{n,i} - t_n)(\sum_{i=1}^{D} w_i\epsilon_i)$的期望值，會因為$\sum_{i=1}^{D} w_i\epsilon_i$項為線性函數，其期望值可表示為$\sum_{i=1}^{D} w_i\mathbb{E}[\epsilon_i]$，又$\mathbb{E}[\epsilon_i] = 0$，故此項為 0。

$\frac{1}{2}\sum_{n=1}^{N}(\sum_{i=1}^{D} w_i\epsilon_i)^2$項，$(\sum_{i=1}^{D} w_i\epsilon_i)^2$期望值，由題意可知每一項加入的高斯雜訊$\epsilon_i$皆為獨立，且$\sum_{i=1}^{D} w_i\epsilon_i$為線性函數，$w_i$可做為 scalar 提出，則可寫為

$(\sum_{i=1}^{D} w_i\mathbb{E}[\epsilon_i])\big(\sum_{j=1}^{D} w_j\mathbb{E}[\epsilon_j]\big) = \sum_{i=1}^{D}\sum_{j=1}^{D} w_i\, w_j\mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j] = \sum_{i=1}^{D}\sum_{j=1}^{D} w_i\, w_j\mathbb{E}[\epsilon_i\epsilon_j]$

又$\mathbb{E}[\epsilon_i\epsilon_j]$=$\delta_{ij}\sigma^2$，故

$\sum_{i=1}^{D}\sum_{j=1}^{D} w_i\, w_j\mathbb{E}[\epsilon_i\epsilon_j] = \sum_{i=1}^{D}\sum_{j=1}^{D} w_i\, w_j\delta_{ij}\sigma^2 = \sigma^2\sum_{i=1}^{D} {w_i}^2$

此即為 regularization 的形式，令$\sigma^2 = \lambda$，和前面的 Loss function 做結合，則可得下式，如同課堂所教的 regularization:

$$\frac{1}{2}\sum_{n=1}^{N}(w_o + \sum_{i=1}^{D} w_i x_{n,i} - t_n)^2 + \lambda\sum_{i=1}^{D} {w_i}^2$$

6.

先解等號左邊部分，已知|A|即為A的特徵值λ_1、λ_2、…、λ_n的乘積，取 ln 後即為連加故可寫為$\frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{i=1}^{n}\ln(\lambda_i)$，則

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{i=1}^{n}\ln(\lambda_i)=\sum_{i=1}^{n}\frac{1}{\lambda_i}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}$$

等號右邊部分，已知特徵值系統$\mathrm{Au}=\lambda\,\mathrm{u}$，A為可逆矩陣，則$\mathrm{A}^{-1}\mathrm{u}=\frac{1}{\lambda}\mathrm{u}$:

$\mathrm{Au}=\lambda\,\mathrm{u}$，對α微分

$\rightarrow\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathrm{Au}=\frac{\mathrm{d}}{\mathrm{d}\alpha}\lambda\,\mathrm{u}$ ，左乘A^{-1}

$\rightarrow\mathrm{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathrm{Au}=\mathrm{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\lambda\,\mathrm{u}=\frac{\mathrm{d}\lambda}{\mathrm{d}\alpha}\mathrm{A}^{-1}\mathrm{u}=\frac{\mathrm{d}\lambda}{\mathrm{d}\alpha}\frac{1}{\lambda}\mathrm{u}$ ，由於 A 為實對稱矩陣，可利用 eigenvector 進行對角化，故α會和 λ 有關，故可寫為$\frac{\mathrm{d}\lambda}{\mathrm{d}\alpha}$，

則可得$\mathrm{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathrm{A}=\frac{\mathrm{d}\lambda}{\mathrm{d}\alpha}\frac{1}{\lambda}\mathrm{I}$,I為$\mathrm{n}\times\mathrm{n}$的單位矩陣，

根據題意求$\mathrm{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathrm{A}$的 trace，即為$\mathrm{Trace}(\frac{\mathrm{d}\lambda}{\mathrm{d}\alpha}\frac{1}{\lambda}\mathrm{I})$

故$\mathrm{Trace}\left(\frac{\mathrm{d}\lambda}{\mathrm{d}\alpha}\frac{1}{\lambda}\mathrm{I}\right)=\sum_{i=1}^{n}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}\frac{1}{\lambda_i}$

則左式等於右式，等號成立，故得證。