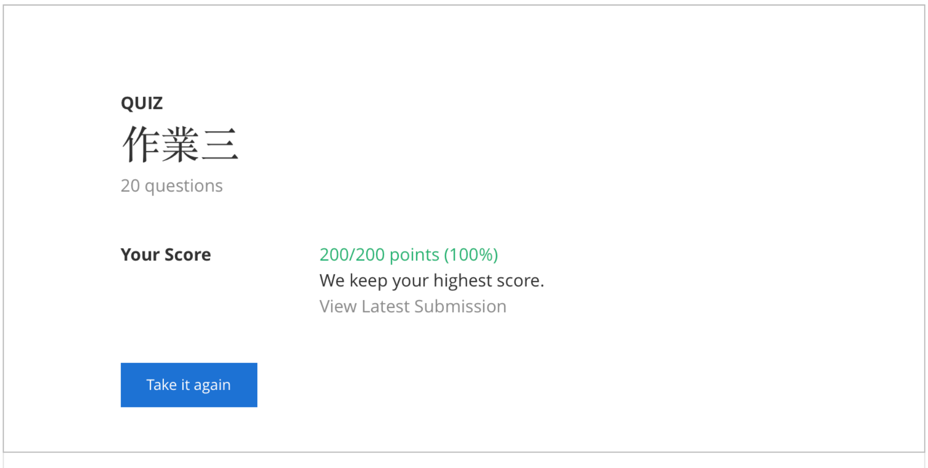
Machine Learning Foundation HW3

B04902004 王佑安

1.



2. $\mathrm{H}^2 = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}(\mathbf{X^{T}X(X^{T}X)^{-1}})\mathrm{X}^{\mathrm{T}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})\mathrm{X}^{-1} = \mathrm{H}$
$\Rightarrow (\mathrm{I}-\mathrm{H})^2 = \mathrm{I} - 2\mathrm{H} + \mathrm{H}^2 = \mathrm{I} - \mathrm{H}$

3. $sign(w^Tx) = y$
$err(w) = 0, \nabla err(w) = 0 \Rightarrow w_{t+1} \leftarrow w_t$ (same as PLA)
$sign(w^Tx) \neq y$
$err(w) = -yw^Tx, \nabla err(w) = -yx \Rightarrow w_{t+1} \leftarrow w_t + yx$ (same as PLA)

4. To minimize $\hat{E}_2(\Delta u, \Delta v)$, $\nabla\hat{E}_2(\Delta u, \Delta v)$
$= \nabla(E(u,v) + (\Delta u, \Delta v)\nabla E(u,v) + \frac{1}{2}((\Delta u, \Delta v)(\nabla E(u,v)))^2)$
$= \nabla E(u,v) + (\Delta u, \Delta v)(\nabla^2(u,v)) = 0$
$\Rightarrow (\Delta u, \Delta v) = -(\nabla^2 E(u,v))^{-1}\nabla E(u,v)$

5. $\max\limits_{w} likelihood(w) \propto \prod\limits_{n=1}^{N} h_{y_n}(x_n)$
$\Rightarrow \min\limits_{w} \frac{1}{N}\sum\limits_{n=1}^{N} -ln(h_{y_n}(x_n))$
$= \min\limits_{w} \frac{1}{N}\sum\limits_{n=1}^{N} ln(\sum\limits_{i=1}^{K} exp(w_i^Tx_n)) - ln(exp(w_{y_n}^Tx_n))$
$= \min\limits_{w} \frac{1}{N}\sum\limits_{n=1}^{N} ln(\sum\limits_{i=1}^{K} exp(w_i^Tx_n)) - w_{y_n}^Tx_n$

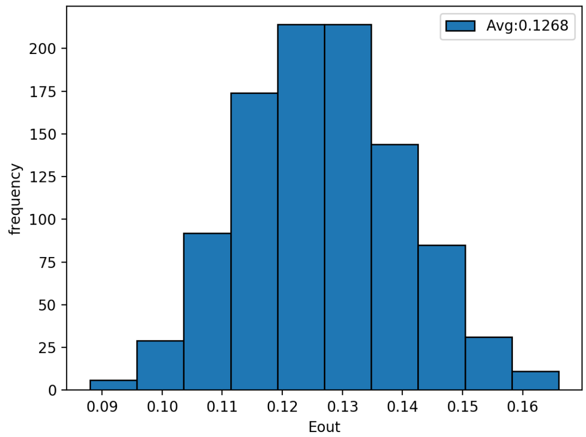
6. $\frac{\partial}{\partial w_i}\frac{1}{N}\sum_{n=1}^{N} ln(\sum_{i=1}^{K} exp(w_i^T x_n)) - w_{y_n}^T x_n$

$= \frac{1}{N}\sum_{n=1}^{N}\frac{x_n exp(w_i^T x_n)}{\sum_{i=1}^{K} exp(w_i^T x_n)} - [y_n = i]x_n$

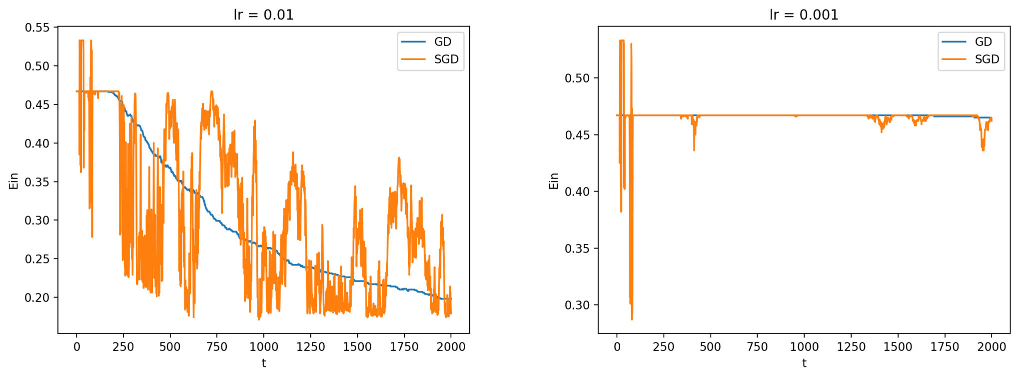
$= \frac{1}{N}\sum_{n=1}^{N} x_n h_{y_n}(x_n) - [y_n = i]x_n$

$= \frac{1}{N}\sum_{n=1}^{N} (h_{y_n}(x_n) - [y_n = i])x_n$

7.

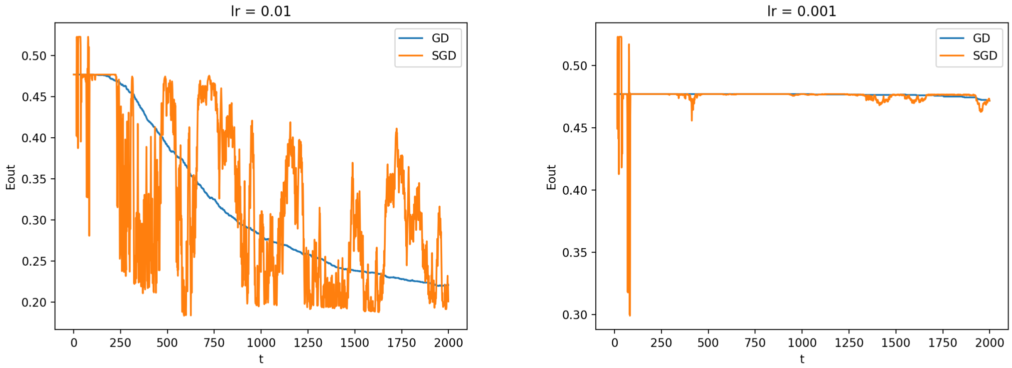


8.



首先可以發現 SGD 震盪的幅度比 GD 大很多，可以推論是因為 SGD 每次只看一筆 data 所以每次 update 不一定能讓整體的 error 降低。再來觀察 lr 的影響，可以發現 0.01 的 error 降低得比 0.001 快，更仔細觀察甚至會發現 0.001 的曲線跟 0.01 的前 200 形狀很接近，可以推論初始的 w 跟 minimal 差很多，每次 update 在大部分維度上的 gradient 方向都相同，所以把 lr 開大可以讓 error 更快降下去。

9.



跟上一題比較後發現，Eout 跟 Ein 變化的趨勢幾乎一樣，差別只在於 Eout 稍微高的了一點，因此可以推論這份 data 的 training set 跟 testing set 的 correlation 很高，在相同的 hypothesis 下 Ein 跟 Eout 成正比。

10. (a) $\mathrm{X}^T\mathrm{X}\mathbf{w}_{lin} = \mathrm{X}^T(\mathrm{U}\Gamma\mathrm{V}^T)(\mathrm{V}\Gamma^{-1}\mathrm{U}^T)y$

$= \mathrm{X}^T\mathrm{X}\mathbf{w}_{lin} = \mathrm{X}^T(\mathrm{U}(\Gamma(\mathrm{V}^T\mathrm{V})\Gamma^{-1})\mathrm{U}^T)y$

$= \mathrm{X}^T\mathrm{X}\mathbf{w}_{lin} = \mathrm{X}^T(\mathrm{U}(\Gamma\Gamma^{-1})\mathrm{U}^T)y$

$= \mathrm{X}^T\mathrm{X}\mathbf{w}_{lin} = \mathrm{X}^T(\mathrm{U}\mathrm{U}^T)y$

$= \mathrm{X}^T\mathrm{X}\mathbf{w}_{lin} = \mathrm{X}^Ty$