Machine Learning Foundation Hw3

資工四 B05902006 蕭縈瀅

Problem 1

Score：100 %

Problem 2

Prove $err(w) = max(0, -yw^Tx)$ results in PLA .

The update on **PLA** algorithm：$w_{t+1} \leftarrow w_t + [y \neq sign(w^Tx)]yx$

$$\nabla_w err(w) = \begin{cases} \frac{\delta - yw^Tx}{\delta w}, & if \ \ -yw^Tx > 0 \\ 0, & if \ \ -yw^Tx < 0 \end{cases}$$

$$= \begin{cases} -yx, & if \ yw^Tx < 0 \\ 0, & if \ yw^Tx > 0 \end{cases}$$

$$= -[y \neq sign(w^Tx)]yx$$

It's obvious to see that $w_{t+1} \leftarrow w_t + [y \neq sign(w^Tx)]yx$

$= w_t + \eta(-\nabla_w err(w))$, when $\eta = 1$

Problem 3

To minimize $\hat{E}_2(\Delta u, \Delta v)$, we would like to find $\Delta u, \Delta v$ such that $\nabla \hat{E}_2(\Delta u, \Delta v) = 0$

$\nabla \hat{E}_2(\Delta u, \Delta v) = \nabla(E(u,v) + (\Delta u, \Delta v)\nabla E(u,v) + \frac{1}{2}((\Delta u, \Delta v)(\nabla E(u,v)))^2)$

$= \nabla E(u,v) + (\Delta u, \Delta v)(\nabla^2(u,v)) = 0$

Therefore, $(\Delta u, \Delta v) = -(\nabla^2 E(u,v))^{-1}\nabla E(u,v)$

Problem 4

$likelihood(w) \propto \prod_{n=1}^{N} h_{y_n}(x_n) \propto \ln \prod_{n=1}^{N} h_{y_n}(x_n)$

$\max_w likelihood(w) \propto \prod_{n=1}^{N} h_{y_n}(x_n)$

$\max_w likelihood(w) \rightarrow \min_w \frac{1}{N}\Sigma_{n=1}^{N} - \ln(h_{y_n}(x_n))$

$\rightarrow \min_w \frac{1}{N}\Sigma_{n=1}^{N} \ln(\Sigma_{i=1}^{k} e^{w_i^T x_n}) - ln(e^{w_{y_n}^T x_n})$

$\rightarrow \min_w \frac{1}{N}\Sigma_{n=1}^{N} \ln(\Sigma_{i=1}^{k} e^{w_i^T x_n}) - w_{y_n}^T x_n$

Therefore, $E_{in} = \frac{1}{N}\Sigma_{n=1}^{N} \ln(\Sigma_{i=1}^{k} e^{w_i^T x_n}) - w_{y_n}^T x_n$

Problem 5

$E_{in}(w_{lin}) = \min_w \frac{1}{N+K}(\Sigma_{n=1}^{N}(y_n - w^T x_n) + \Sigma_{k=1}^{K}(\tilde{y}_k - w^T \tilde{x}_k))$

$= \min_w \frac{1}{N+K}[(w^T X^T X w + 2w^T X^T y + y^T y) + (w^T \tilde{X}^T \tilde{X} w + 2w^T \tilde{X}^T \tilde{y} + \tilde{y}^T \tilde{y})]$

$\Rightarrow \nabla E_{in}(w_{lin}) = \frac{2}{N+K}(x^T X w_{lin} - X^T y + \tilde{X}^T \tilde{X} w_{lin} - \tilde{X}^T \tilde{y}) = 0$

$\Rightarrow (X^T X + \tilde{X}^T \tilde{X}) w_{lin} = X^T y + \tilde{X}^T \tilde{y}$

$\Rightarrow w_{lin} = (X^T X + \tilde{X}^T \tilde{X})^{-1}(X^T y + \tilde{X}^T \tilde{y})$

Problem 6

From above, we know the optimal solution $w_{reg} \leftarrow (Z^T Z + \lambda I)^{-1} Z^T y$

That is, $\tilde{X}^T \tilde{X} = \lambda I, \tilde{X}^T \tilde{y} = 0$

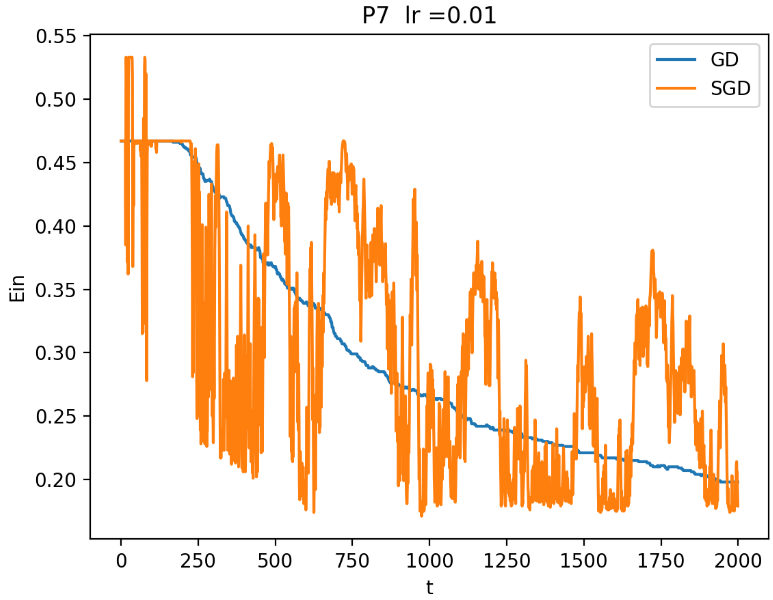
$\Rightarrow \tilde{X} = \sqrt{\lambda} I, \tilde{y} = 0$

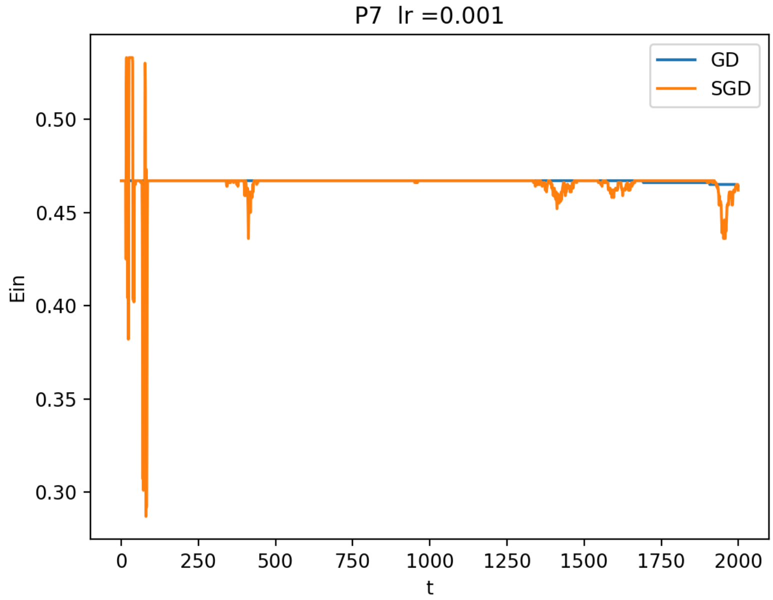
Problem 7

Finding：

- **vibration**：It is clearly to see the vibration of SGD is much bigger than GD. We can infer that it is owing to SGD only trains 1 data each time, so each update can't lower the total error effectively.

- η : Big η $(lr = 0.01)$ can lower total error rate faster than small one $lr = 0.001$.

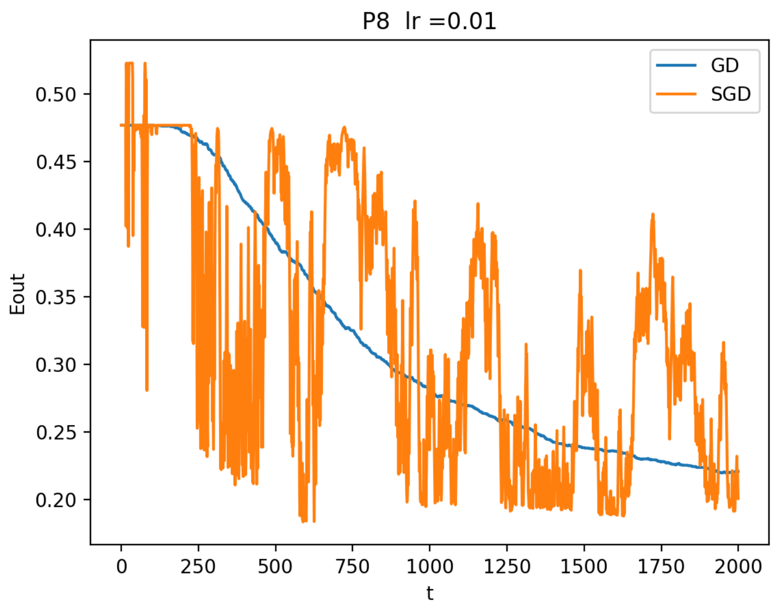


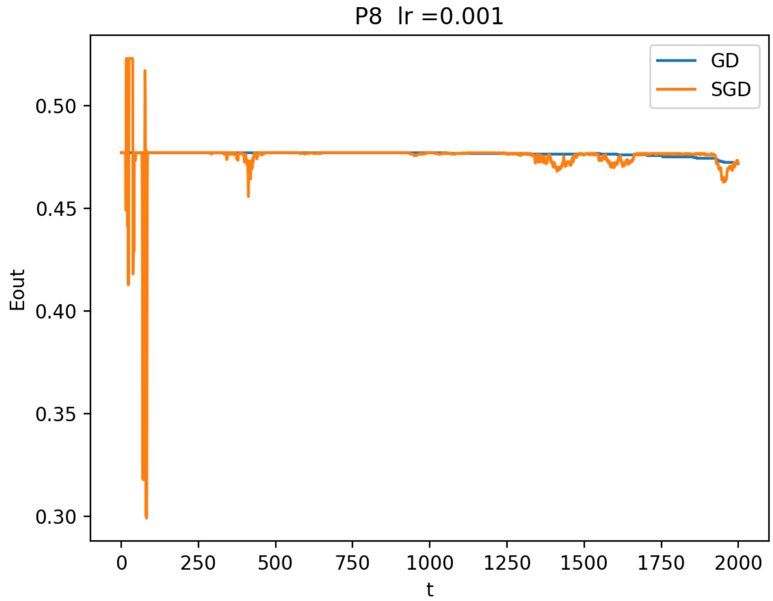


Problem 8

Finding :

- **pattern** : The pattern of $Eout$ is similar to Ein, while the value of $Eout$ seems to be a little bit higher than Ein. We can infer that $hw3\_train.dat$ and $hw3\_test.dat$ has high correlation. Because Ein is propotional to $Eout$ under the same hypothesis.





Bonus

(a)

$X^T X w_{lin} = X^T (U \Gamma V^T)(V \Gamma^{-1} U^T y)$

$= X^T U \Gamma (V^T V) \Gamma^{-1} U^T y \; (\because \textit{commutation law})$

$= X^T U (\Gamma \Gamma^{-1}) U^T y \; (\because V^T V = I \rho)$

$= X^T (U U^T) y \; (\because \Gamma \Gamma^{-1} = I \rho)$

$= X^T y \; (\because U^T U = I \rho)$