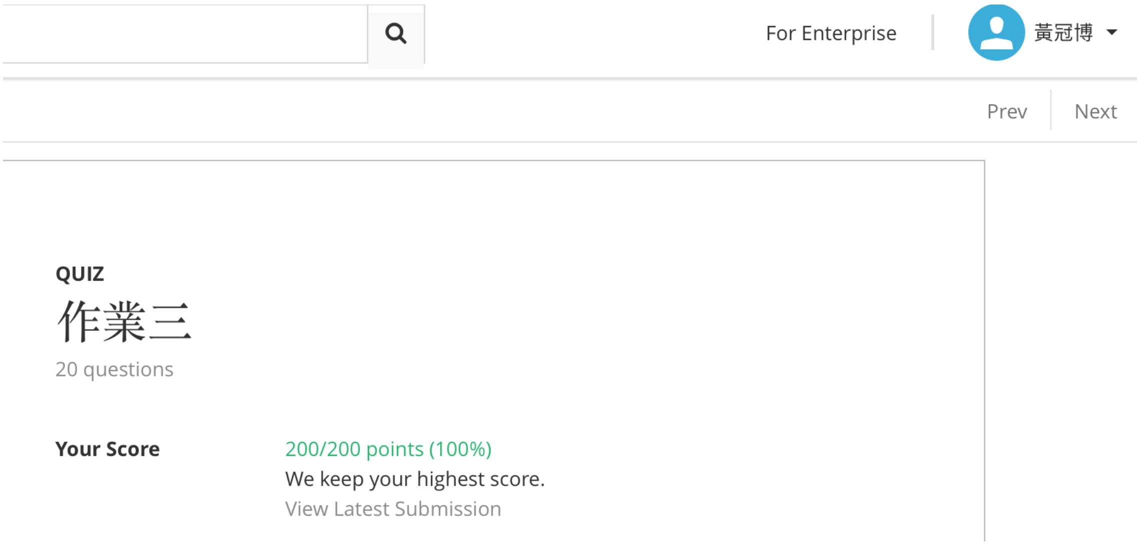
Machine Learning Foundations
Homework 2

b05902121 黃冠博

January 1, 2018

1.



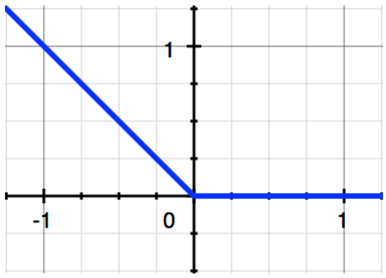
2.

$$
\begin{aligned}
H \cdot H &= X(X^TX)^{-1}X^T \cdot X(X^TX)^{-1}X^T \\
&= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T \\
&= XI(X^TX)^{-1}X^T \\
&= X(X^TX)^{-1}X^T \\
&= H
\end{aligned}
$$

$$
\begin{aligned}
(I-H)^2 &= (I-H)\cdot(I-H) \\
&= I - 2H + H \cdot H \\
&= I - 2H + H \\
&= I - H
\end{aligned}
$$

3.

For $y = 1$: Let the Y axis be $err(w)$ and the X axis be $\mathbf{w^T x}$, the graph of $max(0, -y\mathbf{w^T x})$ is shown below.



When $\mathbf{w^T x} < 0$, by PLA, the point (x, y) is wrong and needs to be corrected by the method below.

$$w_{t+1} \leftarrow w_t + y\mathbf{x}$$

The gradient is $\nabla(-\mathbf{w^T x}) = -\mathbf{x}$. Correct it by SGD as shown below.

$$w_{t+1} \leftarrow w_t - \nabla err(w) = w_t + \mathbf{x} = w_t + y\mathbf{x}$$

When $\mathbf{w^T x} > 0$, by PLA, the point (x, y) is correct and does not need to be corrected.
The gradient is $\nabla(0) = 0$.

$$w_{t+1} \leftarrow w_t - \nabla err(w) = w_t + 0 = w_t$$

By the result above, we know that the error function results in PLA when $y = 1$ and can easily verify when $y = -1$.

For $y = -1$:
When $\mathbf{w^T x} < 0$:
PLA: the point (x, y) is correct and does not need to be corrected
SGD:

$$w_{t+1} \leftarrow w_t - \nabla err(w) = w_t + 0 = w_t$$

When $\mathbf{w^T x} > 0$:
PLA: the point (x, y) is wrong and needs to be corrected

$$w_{t+1} \leftarrow w_t + y\mathbf{x} = w_t - \mathbf{x}$$

SGD:

$$w_{t+1} \leftarrow w_t - \nabla err(w) = w_t - \mathbf{x}$$

Hence, we know that the error function $max(0, -y\mathbf{w^T x})$ results in PLA.

4.

Two variable Taylor series second order:
Let f be an infinitely differentiable function in some open neighborhood around (x, y) = (a, b).

$$f(x,y) = f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) + \frac{1}{2!}[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(y-b)^2]$$

We can derive

$$\begin{aligned}\hat{E}_2(\Delta u, \Delta v) &= E(u,v) + E_u(u,v)\Delta u + E_v(u,v)\Delta v \\ &\quad + \frac{1}{2!}(E_{uu}(u,v)(\Delta u)^2 + 2E_{uv}(u,v)\Delta u\Delta v + E_{vv}(u,v)(\Delta v)^2) \\ &= E(u,v) + \nabla E(u,v)\begin{bmatrix}\Delta u \\ \Delta v\end{bmatrix} + \frac{1}{2!}(\Delta u, \Delta v)H(u,v)\begin{bmatrix}\Delta u \\ \Delta v\end{bmatrix}\end{aligned}$$

with $H(u,v)$ being the Hessian matrix

$$\begin{bmatrix}E_{uu}(u,v) & E_{uv}(u,v) \\ E_{vu}(u,v) & E_{vv}(u,v)\end{bmatrix}$$

To minimize $\hat{E}_2(\Delta u, \Delta v)$, set its gradient to 0.

$$\begin{aligned}\nabla\hat{E}_2(\Delta u, \Delta v) = 0 &\Rightarrow \nabla(E(u,v)) + \nabla(E_u(u,v)\Delta u + E_v(u,v)\Delta v) + \nabla(\frac{1}{2}[(\Delta u)^2 + (\Delta v)^2]H(u,v)) \\ &\Rightarrow \nabla(E(u,v)) + (E_u(u,v), E_v(u,v)) + H(u,v)(\Delta u, \Delta v) \\ &\Rightarrow 0 + \nabla E(u,v) + H(u,v)(\Delta u, \Delta v) = 0 \\ &\Rightarrow (\Delta u, \Delta v) = -[H(u,v)]^{-1}\nabla E(u,v) = -(\nabla^2 E(u,v))^{-1}\nabla E(u,v)\end{aligned}$$

5.

$$h_y(\mathbf{x}) = \frac{e^{\mathbf{w_y^T x}}}{\sum_{k=1}^{K} e^{\mathbf{w_k^T x}}}$$

Apply the method of minimizing likelihood (logistic h).

$$\begin{aligned}max\frac{1}{N}\prod_{n=1}^{N} h_y(x) &\Rightarrow min -\frac{1}{N}\prod_{n=1}^{N} h_y(\mathbf{x_n}) \\ &\Rightarrow min -\frac{1}{N}\sum_{i=1}^{N} \ln h_y(\mathbf{x_n}) \\ &\Rightarrow min -\frac{1}{N}\sum_{i=1}^{N}(\ln(e^{\mathbf{w_y^T x_n}}) - \ln\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}) \\ &\Rightarrow min\frac{1}{N}\sum_{n=1}^{N}(\ln\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}} - \ln e^{\mathbf{w_y^T x_n}}) \\ &\Rightarrow min\frac{1}{N}\sum_{n=1}^{N}(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}) - \mathbf{w_y^T x_n})\end{aligned}$$

6.

$$\mathbf{E}_{in} = \frac{1}{N}\sum_{n=1}^{N}(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}) - \mathbf{w_y^T x_n})$$

$$\frac{\partial \mathbf{E}_{in}}{\partial \mathbf{w}_i} = \frac{\partial}{\partial \mathbf{w}_i}[\frac{1}{N}\sum_{n=1}^{N}(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}) - \mathbf{w_y^T x_n})] = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}_i}(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}) - \mathbf{w_y^T x_n})$$

$$\frac{\partial}{\partial \mathbf{w}_i}(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}) - \mathbf{w_y^T x_n}) = \frac{\partial}{\partial \mathbf{w}_i}(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}) - \frac{\partial \mathbf{w_y^T x_n}}{\partial \mathbf{w}_i}$$

Be aware that:

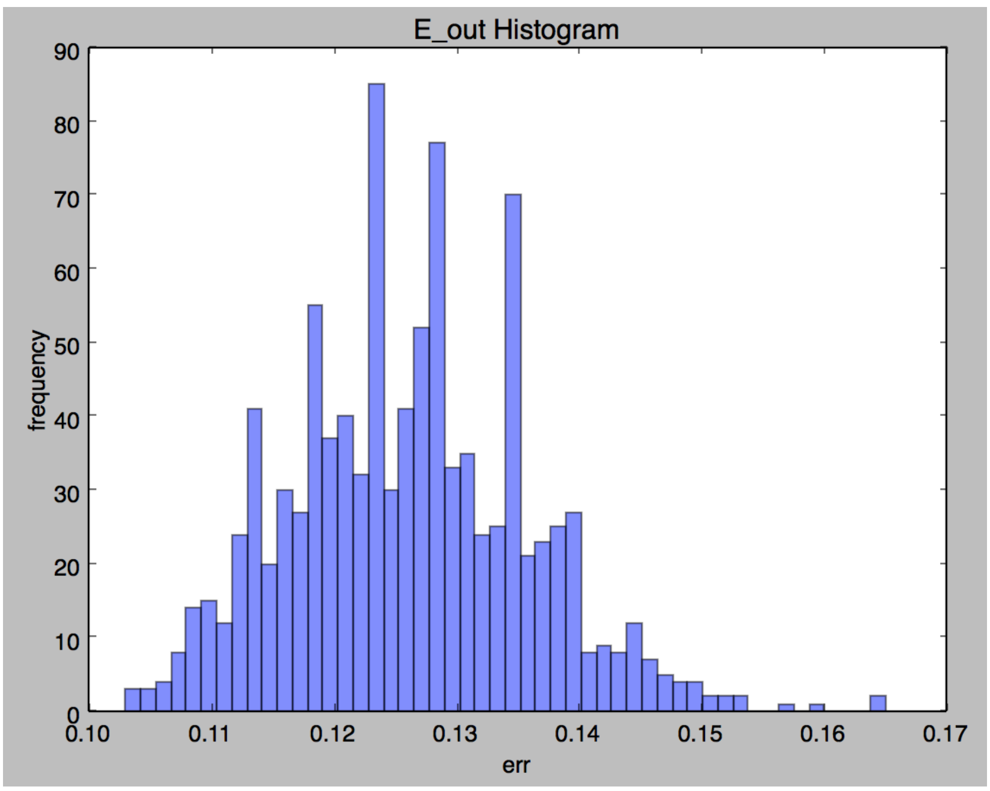
$$\frac{\partial \mathbf{w_y^T x_n}}{\partial \mathbf{w}_i} = \begin{cases} 0 & \text{if } y \neq \text{i} \\ \mathbf{x_n} & \text{if y = i} \end{cases}$$

$$\begin{aligned}
\frac{\partial(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}))}{\partial \mathbf{w}_i} &= \frac{\partial(\ln(\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}))}{\partial \sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}} \cdot \frac{\partial \sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}}{\partial \mathbf{w_i^T x_n}} \cdot \frac{\partial \mathbf{w_i^T x_n}}{\partial \mathbf{w_i}} \qquad \text{by chain rule} \\
&= \frac{1}{\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}} \cdot \frac{\partial(e^{\mathbf{w_1^T x_n}} + \cdots + e^{\mathbf{w_i^T x_n}} + \cdots + e^{\mathbf{w_K^T x_n}})}{\partial \mathbf{w_i^T x_n}} \cdot \mathbf{x_n} \\
&= \frac{1}{\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}} \cdot \frac{e^{\mathbf{w_i^T x_n}}}{\partial \mathbf{w_i^T x_n}} \cdot \mathbf{x_n} \\
&= \frac{e^{\mathbf{w_i^T x_n}}}{\sum_{k=1}^{K} e^{\mathbf{w_k^T x_n}}} \cdot \mathbf{x_n} \\
&= h_i(\mathbf{x_n}) \cdot \mathbf{x_n}
\end{aligned}$$

Hence,

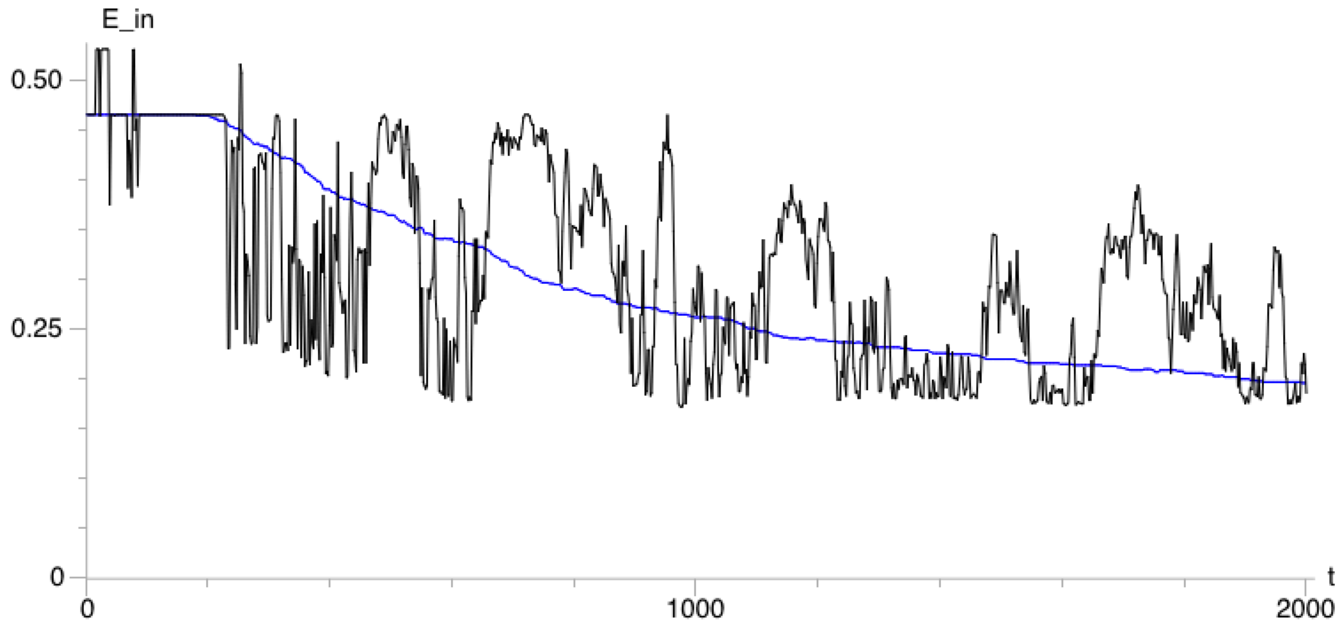
$$\frac{\partial \mathbf{E}_{in}}{\partial \mathbf{w}_i} = \frac{1}{N}\sum_{n=1}^{N}((h_i(\mathbf{x_n}) - [[y_n = i]]))\mathbf{x_n}$$

7. Average E_out: 0.126089

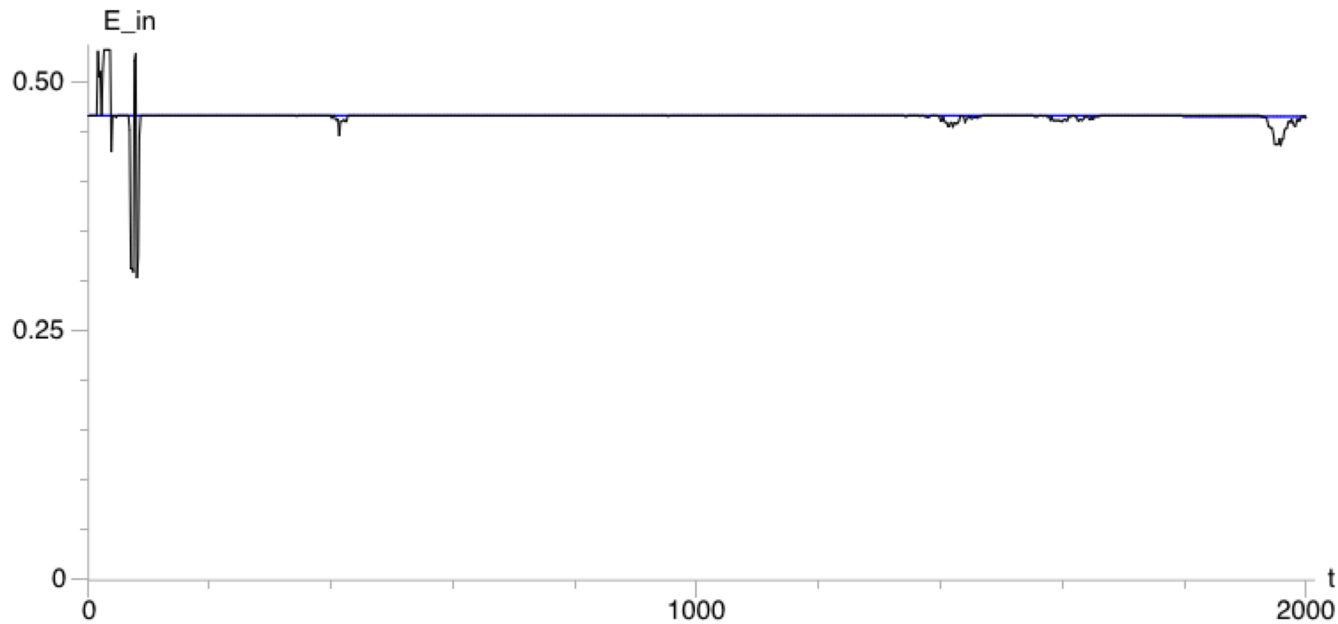


8. E_in

$\eta = 0.01$, GD: E_in = 0.197000, SGD: E_in = 0.187000



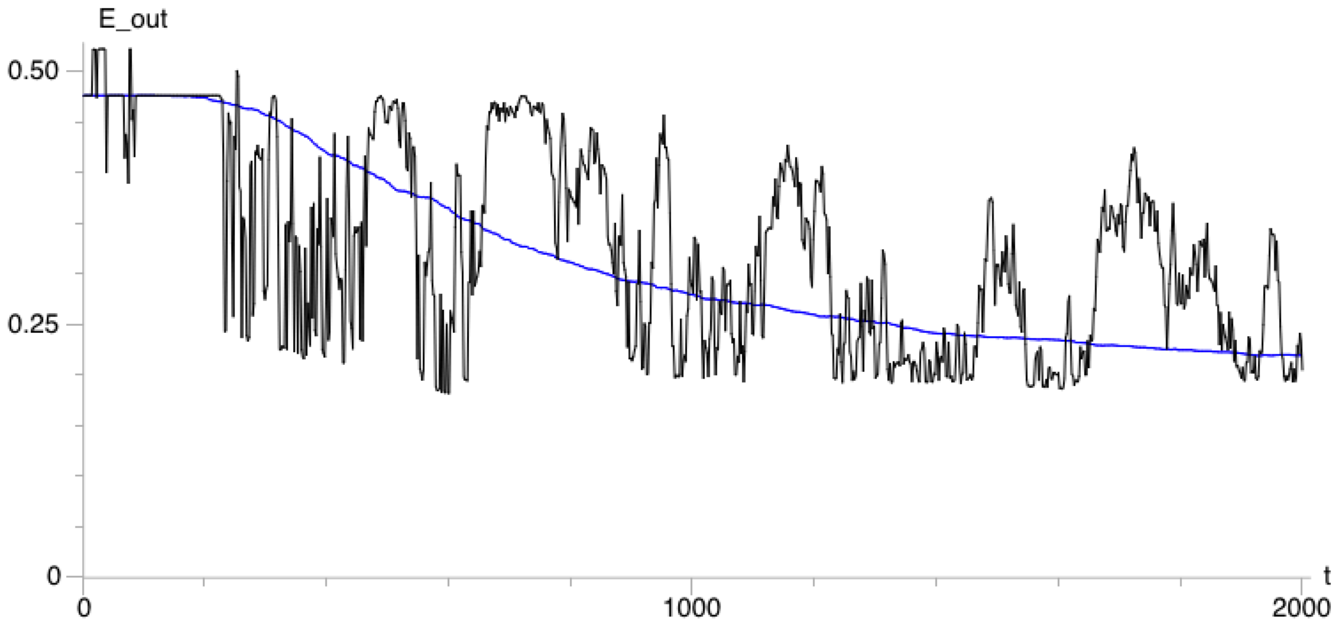
$\eta = 0.001$, GD: E_in = 0.466000, SGD: E_in = 0.464000



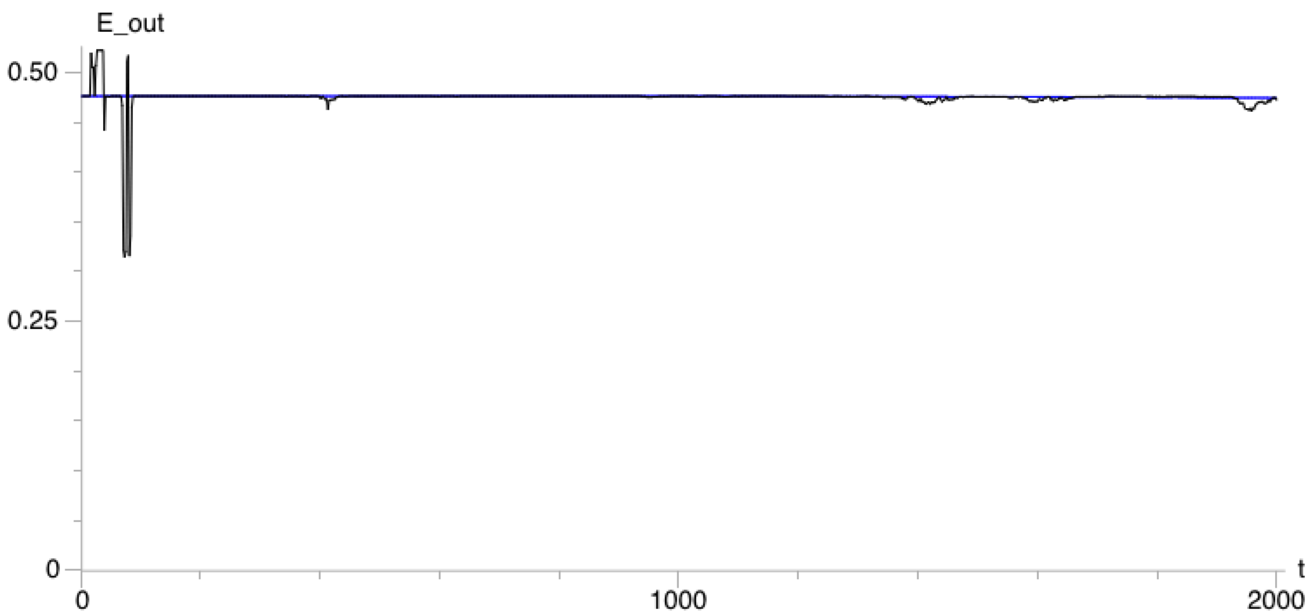
The blue curve is GD, the black curve is SGD.
When η is 0.01, E_{in} curves of GD is smooth and monotonic.
When η is 0.01, E_{in} curves of SGD has drastic jumping patterns.
However, the two curves decline after more rounds.
E_{in} curves of GD SGD are similar when η is 0.001. There is no decline after more rounds (almost horizontal).

9. E_out

$\eta = 0.01$, GD: E_out = 0.220000, SGD: E_out = 0.205333



$\eta = 0.001$, GD: E_out = 0.475000, SGD: E_out = 0.473000



The blue curve is GD, the black curve is SGD.
When η is 0.01, E_{out} curves of GD is smooth and monotonic.
When η is 0.01, E_{out} curves of SGD has drastic jumping patterns.
However, the two curves decline after more rounds.
E_{out} curves of GD SGD are similar when η is 0.001. There is no decline after more rounds (almost horizontal). However, there is a little jumping pattern during the first few rounds.
The curves of E_out are similar to E_in, and this is what we want.

10.

(a)

$$
\begin{aligned}
X^T X \mathbf{w_{lin}} &= X^T (U\Gamma V^T)(V\Gamma^{-1}U^T y) \\
&= X^T U\Gamma(V^T V)\Gamma^{-1}U^T y \\
&= X^T U\Gamma I_\rho \Gamma^{-1}U^T y \\
&= X^T U(\Gamma\Gamma^{-1})U^T y \\
&= X^T U I_\rho U^T y \\
&= X^T (UU^T) y \\
&= X^T I_N y \\
&= X^T y
\end{aligned}
$$

(b)

notations:

A^+ is called the Moore-Penrose inverse of A

A^* is the Hermitian transpose of A

Hermitian transpose:

Taking the transpose and then taking the complex conjugate of each entry. ($A^* = \overline{A^T}$)

some properties:

1. $AA^+A = A$,
2. $A^+AA^+ = A^+$,
3. $(AA^+)^* = AA^+$,
4. $(A^+A)^* = A^+A$.

Denote X^TX as A, $\mathbf{w}$ as x, X^Ty as b and $\mathbf{w_{lin}}$ as z.
If $Ax = b$ has a solution, then $z = A^+b$ is a solution. We show that z is the smallest such solution.
Define Q as A^+A. Note that $Qz = A^+AA^+b = (A^+AA^+)b = A^+b = z$ and $Q* = Q$. And we have:

$$
\begin{aligned}
z^*(x-z) &= (Qz)^*(x-z) \\
&= z^*Q(x-z) \\
&= z^*(Qx - Qz) \\
&= z^*(A^+Ax - z) \\
&= z^*(A^+b - z) \\
&= z^*(z-z) = 0
\end{aligned}
$$

$$
\begin{aligned}
x = z + (x-z) \Rightarrow \|x\|^2 &= \|z\|^2 + 2z^*(x-z) + \|x-z\|^2 \\
&= \|z\|^2 + \|x-z\|^2 \\
&\geq \|z\|^2
\end{aligned}
$$

Hence, $\|\mathbf{w_{lin}}\| \leq \|\mathbf{w}\|$