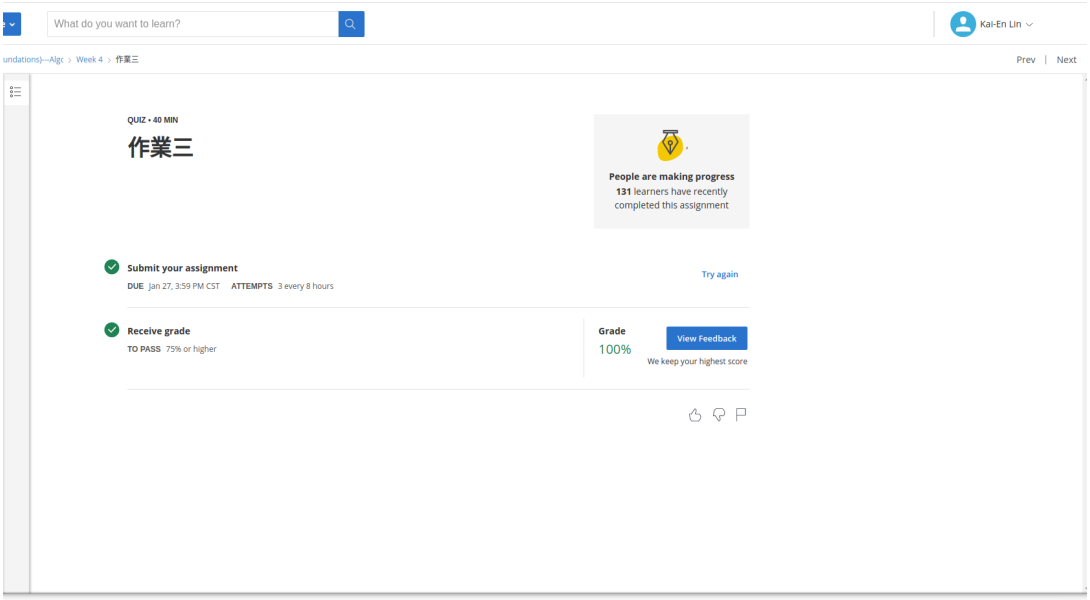
Homework #3

Student Name: 林楷恩
Student ID: b07902075

1



2

- **SGD:** $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla err(\mathbf{w}_t, \mathbf{x}_n, y_n)$
- **PLA:** $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + [\![y_n \neq sign(\mathbf{w}_t^T \mathbf{x}_n)]\!] (y_n \mathbf{x}_n)$

If $err(w) = \max(0, -y\mathbf{w}^T\mathbf{x})$, since **SGD** computes ∇err by one single $(\mathbf{x}_n, y_n)$, the error function can be rewritten as $err(\mathbf{w}_t, \mathbf{x}_n, y_n) = \max(0, -y_n \mathbf{w}_t^T \mathbf{x}_n)$. Then, I derive $\nabla err(\mathbf{w}_t, \mathbf{x}_n, y_n)$ in 2 cases:

- If $y_n = sign(\mathbf{w}_t^T \mathbf{x}_n)$: This implies $-y_n \mathbf{w}_t^T \mathbf{x}_n < 0$, so $err(\mathbf{w}_t, \mathbf{x}_n, y_n) = 0$ and this point is not differentiable. By the problem description, we just ignore it. Thus w_t is not updated, which is the same as the behavior of **PLA**.
- If $y_n \neq sign(\mathbf{w}_t^T \mathbf{x}_n)$: This implies $-y_n \mathbf{w}_t^T \mathbf{x}_n > 0$, so $err(\mathbf{w}_t, \mathbf{x}_n, y_n) = -y_n \mathbf{w}_t^T \mathbf{x}_n$ and its gradient is derived as follow:

$$\frac{\partial err(\mathbf{w}_t, \mathbf{x}_n, y_n)}{\partial \mathbf{w}_{t,i}} = -y_n \mathbf{x}_{n,i} \implies \nabla err(\mathbf{w}_t, \mathbf{x}_n, y_n) = -y_n \mathbf{x}_n$$

Hence, in this case, if $\eta = 1$, then the formula of **SGD** becomes $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_n \mathbf{x}_n$, which is the same as how **PLA** updates its $\mathbf{w}_t$.

3

The second-order Taylor's expansion of E, $\hat{E}_2(\Delta u, \Delta v)$, is as follow:

$$\hat{E}_2(\Delta u, \Delta v) = \frac{E_{uu}(u,v)(\Delta u)^2}{2} + \frac{E_{vv}(u,v)(\Delta v)^2}{2} + \frac{E_{uv}(u,v)(\Delta u)(\Delta v)}{2} + E_u(u,v)\Delta u + E_v(u,v)\Delta v + E(u,v)$$

Partial differentiate $\hat{E}_2(\Delta u, \Delta v)$ to find the critical point. Since it is given that the Hessian matrix is positive definite, we always have minimum at the critical point:

$$\frac{\partial \hat{E}_2}{\partial \Delta u} = E_{uu}(u,v)\Delta u + E_{uv}(u,v)(\Delta v) + E_u(u,v) = 0$$
$$\frac{\partial \hat{E}_2}{\partial \Delta v} = E_{vv}(u,v)\Delta v + E_{uv}(u,v)(\Delta u) + E_v(u,v) = 0$$

$$\Longrightarrow \begin{bmatrix} E_{uu}(u,v) & E_{uv}(u,v) \\ E_{uv}(u,v) & E_{vv}(u,v) \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \begin{bmatrix} -E_u(u,v) \\ -E_v(u,v) \end{bmatrix} \text{ (rewrite in matrix form)}$$

$$\Longrightarrow \nabla^2 E(u,v) \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -\nabla E(u,v)$$

$$\Longrightarrow \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -(\nabla^2 E(u,v))^{-1} \nabla E(u,v) \ \#$$

4

The likelyhood that h generates $\mathcal{D}$ is the product of the probability that h outputs the correct $\mathbf{y}$ for each $\mathbf{x}_n$, and our goal is to maximize this likelyhood.

$$\max_{\mathbf{w}} \quad \prod_{n=1}^{N} \frac{exp(\mathbf{w}_{y_n}^T x_n)}{\sum_{k=1}^{K} exp(\mathbf{w}_k^T x_n)}$$
$$\Longrightarrow \max_{\mathbf{w}} \quad ln \prod_{n=1}^{N} \frac{exp(\mathbf{w}_{y_n}^T x_n)}{\sum_{k=1}^{K} exp(\mathbf{w}_k^T x_n)}$$
$$\Longrightarrow \max_{\mathbf{w}} \quad \sum_{n=1}^{N} \left(\mathbf{w}_{y_n}^T x_n - ln(\sum_{k=1}^{K} exp(\mathbf{w}_k^T x_n)) \right)$$
$$\Longrightarrow \min_{\mathbf{w}} \quad \sum_{n=1}^{N} \left(ln(\sum_{k=1}^{K} exp(\mathbf{w}_k^T x_n)) - \mathbf{w}_{y_n}^T x_n \right) \text{ (negative maximum is minimum)}$$
$$\Longrightarrow \min_{\mathbf{w}} \quad \frac{1}{N} \sum_{n=1}^{N} \left(ln(\sum_{k=1}^{K} exp(\mathbf{w}_k^T x_n)) - \mathbf{w}_{y_n}^T x_n \right) \text{ (multiply a constant)}$$

5

Rewrite the problem in matrix form:

$$\min_{\mathbf{w}} \frac{1}{N+K}\left(\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2+\|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2\right) \implies \min_{\mathbf{w}} E_{in}(\mathbf{w})$$

To minimize $E_{in}(\mathbf{w})$, we should find the $\mathbf{w}_{\text{reg}}$ such that $\nabla E_{in}(\mathbf{w}_{\text{reg}}) = \vec{0}$. The derivation of $\nabla E_{in}(\mathbf{w})$ is as follow:

$$\begin{aligned}
E_{in}(\mathbf{w}) =& \frac{1}{N+K}\left(\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2+\|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2\right) \\
=& \frac{1}{N+K}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}-2\mathbf{w}^T\mathbf{X}^T\mathbf{y}+\mathbf{y}^T\mathbf{y}+\mathbf{w}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{w}-2\mathbf{w}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{y}}+\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}\right) \\
=& \frac{1}{N+K}\left(\mathbf{w}^T(\mathbf{X}^T\mathbf{X}+\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\mathbf{w}-2\mathbf{w}^T(\mathbf{X}^T\mathbf{y}+\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})+(\mathbf{y}^T\mathbf{y}+\tilde{\mathbf{y}}^T\tilde{\mathbf{y}})\right) \\
\implies \nabla E_{in}(\mathbf{w}) =& \frac{1}{N+K}\left(2(\mathbf{X}^T\mathbf{X}+\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\mathbf{w}-2(\mathbf{X}^T\mathbf{y}+\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})\right)
\end{aligned}$$

Then solve the equation $\nabla E_{in}(\mathbf{w}_{\text{reg}}) = \vec{0}$:

$$\begin{aligned}
&\frac{1}{N+K}\left(2(\mathbf{X}^T\mathbf{X}+\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\mathbf{w}_{\text{reg}}-2(\mathbf{X}^T\mathbf{y}+\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})\right) = \vec{0} \\
\implies& 2(\mathbf{X}^T\mathbf{X}+\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\mathbf{w}_{\text{reg}}-2(\mathbf{X}^T\mathbf{y}+\tilde{\mathbf{X}}^T\tilde{\mathbf{y}}) = \vec{0} \\
\implies& (\mathbf{X}^T\mathbf{X}+\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})\mathbf{w}_{\text{reg}} = \mathbf{X}^T\mathbf{y}+\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} \\
\implies& \mathbf{w}_{\text{reg}} = (\mathbf{X}^T\mathbf{X}+\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{X}^T\mathbf{y}+\tilde{\mathbf{X}}^T\tilde{\mathbf{y}}) \ \#
\end{aligned}$$

6

By works in **Problem 5**:

$$\begin{aligned}\mathbf{w}_{\text{reg}} =& \text{argmin}_{\mathbf{w}} \frac{1}{N+K}\left(\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 + \|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2\right) \\ =& \text{argmin}_{\mathbf{w}} \frac{1}{N+K}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 + \frac{1}{N+K}\|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2\end{aligned}$$

Compare with the equation in problem description:

$$\mathbf{w}_{\text{reg}} = \text{argmin}_{\mathbf{w}} \frac{\lambda}{N}\|\mathbf{w}\|^2 + \frac{1}{N}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2$$

First, since both $\frac{1}{N+K}$ and $\frac{1}{N}$ are just positive constants that multiply the whole function, they do not affect the comparison of different $\mathbf{w}$, and thus do not affect the solution $\mathbf{w}_{\text{reg}}$ that minimizes the function. So, we can just take out $\frac{1}{N+K}$ and $\frac{1}{N}$ from the two equation:

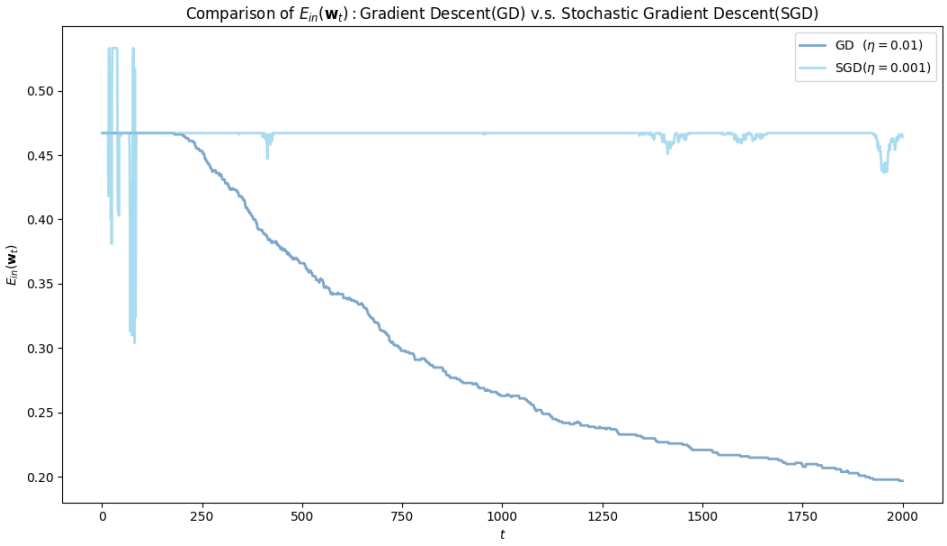
$$\begin{aligned}\mathbf{w}_{\text{reg}} &= \text{argmin}_{\mathbf{w}} \lambda\|\mathbf{w}\|^2 + \|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 \\ \mathbf{w}_{\text{reg}} &= \text{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 + \|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2\end{aligned}$$

By observation, both equations have $\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2$, so if $\lambda\|\mathbf{w}\|^2 = \|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2$, the two equations will be exactly the same.

$$\begin{aligned}&\text{Let } \tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0} \\ \Longrightarrow\ &\|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2 = \|\sqrt{\lambda}\mathbf{w}\|^2 = (\sqrt{\lambda}\mathbf{w})^T(\sqrt{\lambda}\mathbf{w}) = \lambda\mathbf{w}^T\mathbf{w} = \lambda\|\mathbf{w}\|^2\end{aligned}$$

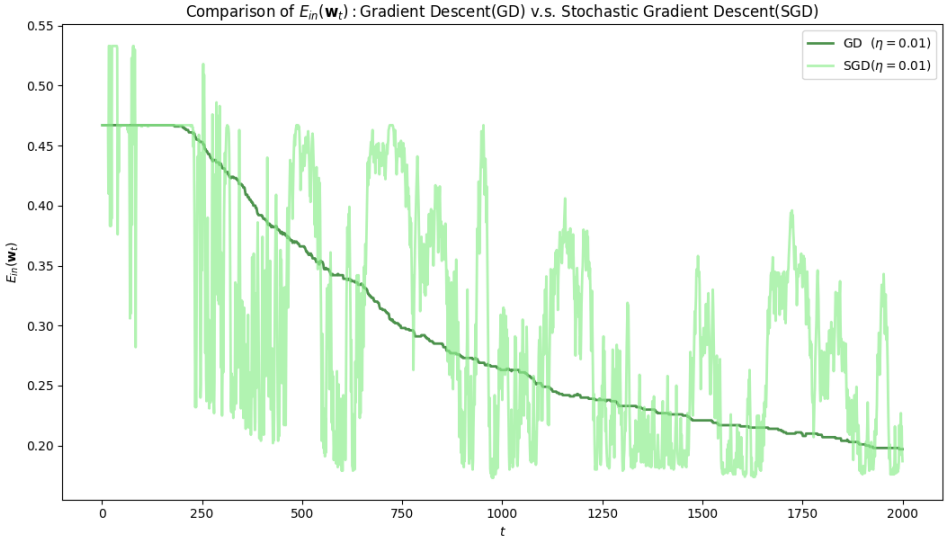
When $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$, the two equations are exactly the same (except the constant multiplier), which implies their solutions will also be the same.

7



My findings:

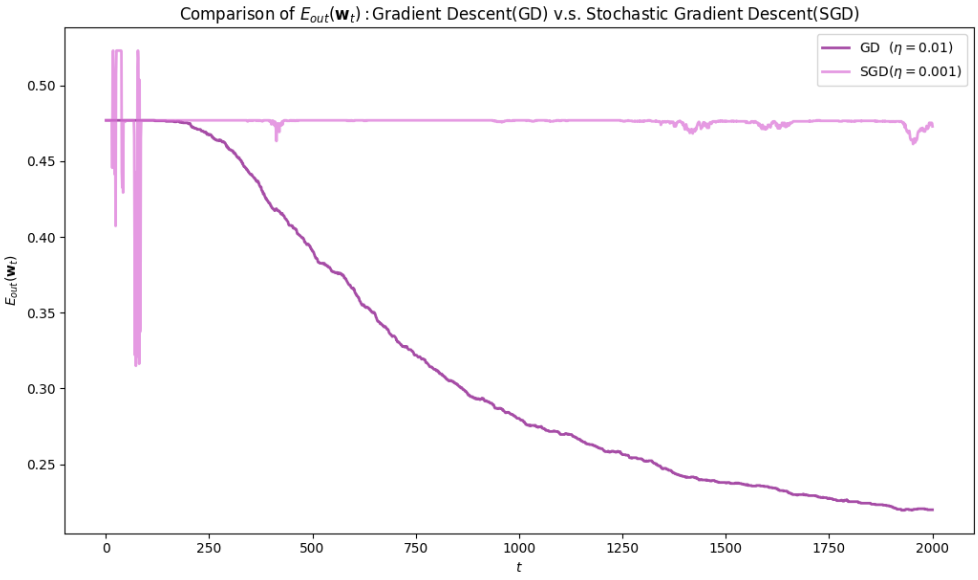
- The E_{in} of **GD** is stably decreasing, which shows a better accuracy. This is because normal gradient descent always finds the "true" gradient, which guarantees the improvement of function. However, computing true gradient takes much more time, thus the execution time of **GD** is considerably longer than the one of **SGD**.
- The E_{in} of **SGD** keeps fluctuating and seems to not have a tendency of improving. I thinks the fluctuation is due to the unstable nature of **SGD**, while the poor accuracy is because the learning rate(η) is too small, so it takes longer time to move to the "valley". Thus, if we apply the same learning rate as **GD**(0.01) to **SGD**, as the following figure shows, we can see that it still fluctuates but has a tendency of decrease, converging to a E_{in} similar to the result of **GD**, which shows that **SGD** can actually do as well as **GD**.



8

My findings:

- The evolution of E_{out} is similar to E_{in} ($E_{in} \approx E_{out}$)
- The E_{out} of **GD** is stably decreasing, while the E_{out} of **SGD** keeps fluctuating and decreases quite slowly.



9

(a) Substitute $X = U\Gamma V^T$ and $\mathbf{w}_{\text{lin}} = V\Gamma^{-1}U^T\mathbf{y}$ into $X^TX\mathbf{w}_{\text{lin}}$

$$
\begin{aligned}
X^TX\mathbf{w}_{\text{lin}} =& (U\Gamma V^T)^T(U\Gamma V^T)(V\Gamma^{-1}U^T\mathbf{y}) \\
=& (V\Gamma^TU^T)(U\Gamma V^T)(V\Gamma^{-1}U^T\mathbf{y}) \\
=& V\Gamma^T\Gamma V^TV\Gamma^{-1}U^T\mathbf{y} && (\text{by } U^TU = I_\rho) \\
=& V\Gamma^T\Gamma\Gamma^{-1}U^T\mathbf{y} && (\text{by } V^TV = I_\rho) \\
=& V\Gamma^TU^T\mathbf{y} && (\text{by } \Gamma\Gamma^{-1} = I_\rho) \\
=& (U\Gamma V^T)^T\mathbf{y} \\
=& X^T\mathbf{y} \ \#
\end{aligned}
$$

$\implies$ $\mathbf{w}_{\text{lin}} = V\Gamma^{-1}U^T\mathbf{y}$ satisfies the equation and thus is a solution of $X^TX\mathbf{w}_{\text{lin}} = X^T\mathbf{y}$.

(b) **I.** Let $P = X^\dagger X = V\Gamma^{-1}U^T U\Gamma V^T$. By $U^T U = I_\rho$ and $\Gamma^{-1}\Gamma = I_\rho$, P can be simplified to VV^T. Then by observation, we can see that P has 2 property:

1° symmetric: $P^T = (VV^T)^T = (V^T)^T V^T = VV^T = P$

2° $PX^\dagger = X^\dagger$: $PX^\dagger = VV^T V\Gamma^{-1}U^T$, by $V^T V = I_\rho$, $PX^\dagger = V\Gamma^{-1}U^T = X^\dagger$

II. Let $\mathbf{w}$ be an arbitrary solution of $X^T X\mathbf{w} = X^T\mathbf{y}$, and let $\mathbf{z} = \mathbf{w} - \mathbf{w}_{\text{lin}}$,

$$\begin{aligned} X^T X\mathbf{z} =& X^T X(\mathbf{w} - \mathbf{w}_{\text{lin}}) \\ =& X^T X\mathbf{w} - X^T X\mathbf{w}_{\text{lin}} \\ =& X^T\mathbf{y} - X^T\mathbf{y} \text{ (since both } \mathbf{w}_{\text{lin}} \text{ and } \mathbf{w} \text{ are solutions of } X^T X\mathbf{w} = X^T\mathbf{y}) \\ =& \vec{\mathbf{0}} \end{aligned}$$

Pre-multiply with $\mathbf{z}^T$, then we can derive:

$$\begin{aligned} & \mathbf{z}^T X^T X\mathbf{z} = \vec{\mathbf{0}} \\ \Longrightarrow\ & (X\mathbf{z})^T (X\mathbf{z}) = \vec{\mathbf{0}} \\ \Longrightarrow\ & \|X\mathbf{z}\|^2 = 0 \\ \Longrightarrow\ & \|X\mathbf{z}\| = 0 \\ \Longrightarrow\ & X\mathbf{z} = \vec{\mathbf{0}} \end{aligned}$$

III. Prove $\mathbf{z}^T\mathbf{w}_{\text{lin}} = \vec{\mathbf{0}}$:

$$\begin{aligned} \mathbf{z}^T\mathbf{w}_{\text{lin}} &= \mathbf{z}^T X^\dagger\mathbf{y} \\ &= \mathbf{z}^T PX^\dagger\mathbf{y} && \text{.....by } \textbf{I.}\ 2^\circ\text{: } PX^\dagger = X^\dagger \\ &= \mathbf{z}^T P^T X^\dagger\mathbf{y} && \text{.....by } \textbf{I.}\ 1^\circ\text{: } P = P^T \\ &= (P\mathbf{z})^T X^\dagger\mathbf{y} \\ &= (X^\dagger X\mathbf{z})^T X^\dagger\mathbf{y} && \text{.....by } P = X^\dagger X \\ &= \vec{\mathbf{0}} && \text{.....by } \textbf{II.},\ X\mathbf{z} = \vec{\mathbf{0}} \end{aligned}$$

By **I. II. III.**, we can show that $\|\mathbf{w}_{\text{lin}}\| \leq \|\mathbf{w}\|$ for all $\mathbf{w}$ that satisfies $X^T X\mathbf{w} = X^T\mathbf{y}$:

$$\begin{aligned} \|\mathbf{w}\|^2 &= \|\mathbf{z} + \mathbf{w}_{\text{lin}}\|^2 && \text{.....by definition in } \textbf{II.},\ \mathbf{z} = \mathbf{w} - \mathbf{w}_{\text{lin}} \\ &= \|\mathbf{z}\|^2 + 2\mathbf{z}^T\mathbf{w}_{\text{lin}} + \|\mathbf{w}_{\text{lin}}\|^2 \\ &= \|\mathbf{z}\|^2 + \|\mathbf{w}_{\text{lin}}\|^2 && \text{.....by } \textbf{III.},\ \mathbf{z}^T\mathbf{w}_{\text{lin}} = \vec{\mathbf{0}} \\ &\geq \|\mathbf{w}_{\text{lin}}\|^2 && \text{.....} \|\mathbf{z}\|^2 \geq 0 \\ \Longrightarrow\ & \|\mathbf{w}_{\text{lin}}\| \leq \|\mathbf{w}\| \end{aligned}$$

This completes my proof that $\mathbf{w}_{\text{lin}}$ is the shortest weight vector that minimize E_{in}.

* reference: https://math.stackexchange.com/questions/2026901/explain-why-x-ab-is-the-shortest-possible-solution-to-ata-hatx-atb