

Exercise 1

(a) In order to find the minimum of $L(\theta)$ for $\theta = \theta_0$
We need to have θ_0 satisfy $\frac{\partial L(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} = 0$.

$$\rightarrow L(\theta) = \sum (y_n - \theta^T x_n)^2 + \lambda \|\theta\|^2$$

$$\rightarrow \frac{\partial L(\theta)}{\partial \theta} = \sum (y_n - x_n^T \theta) (-2x_n) + 2\lambda \theta$$

We want:

$$\frac{\partial L(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} = 0 \Rightarrow -2 \sum x_n (y_n - x_n^T \theta_0) + 2\lambda \theta_0 = 0$$

~~$$\Rightarrow \sum (x_n y_n - x_n x_n^T \theta_0) - \lambda \theta_0 = 0$$~~

$$\Rightarrow \left(\sum x_n x_n^T + \lambda I \right) \theta_0 = \sum x_n y_n$$

So the solution is $\left(\sum_{n=1}^N x_n x_n^T + \lambda I \right) \theta_0 = \sum_{n=1}^N x_n y_n$

(b) Let $x \in \mathbb{R}^l$ and dataset X contains N points $X = \{(x_1, y_1), (x_2, y_2) \dots, (x_N, y_N)\}$

then as known let θ be $\theta = \begin{bmatrix} \theta_0 \\ \theta \end{bmatrix} \in \mathbb{R}^{l+1}$

$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$ where $x_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{il} \end{bmatrix} \in \mathbb{R}^{l+1}$, the leading 1 after the l -dim data-point.

So: $\sum_{n=1}^N x_n x_n^T$ is nothing more than the point wise

function of $X^T X$, as $X^T = \begin{bmatrix} x_1^T & x_2^T & \dots & x_N^T \end{bmatrix} (l+1) \times N$

and $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} N \times (l+1)$ so $X^T X = [x_1, x_2, \dots, x_n]$ $\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$

meaning that $X^T X = x_1 x_1^T + x_2 x_2^T + \dots + x_N x_N^T = \sum_{n=1}^N x_n x_n^T$

Also $X^T y$ is easily proven

$$X^T y = [x_1, x_2, \dots, x_N] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \sum_{n=1}^N x_n y_n \quad (b)$$

From (a), (b) we have that:

$$\left(\sum_{n=1}^N x_n x_n^T + \lambda I \right) \theta = \sum_{n=1}^N y_n x_n \Rightarrow \underbrace{\left(X^T X + \lambda I \right)}_{(a)} \cdot \underbrace{\theta}_{(b)} = \underbrace{X^T y}_{(b)}$$

so $\theta = (X^T X + \lambda I)^{-1} X^T y$

Exercise 2

Let 1-dim problem $y = \theta x + \eta$ (1)

where true value of $\theta = \theta_0 \stackrel{(1)}{\Rightarrow} y = \theta_0 x + \eta$ (2)

Also $\hat{\theta}_{Mvu}$ is the minimum variance ^{unbiased} estimator of θ_0 .

and F the parametric set that

$$\hat{\theta}_b = (1+a)\hat{\theta}_{Mvu}, \quad a \in \mathbb{R}$$

First of all let's see what we know:

If $\hat{\theta}_{Mvu}$ is MVEstimator then.

$$MSE(\hat{\theta}_{Mvu}) = \text{Var}[\hat{\theta}_{Mvu}] + \text{Bias}[\hat{\theta}_{Mvu}]^2$$

$\hat{\theta}_{Mvu}$ is unbiased $E[\hat{\theta}_{Mvu}] = \theta_0 \Rightarrow \text{Bias}[\hat{\theta}_{Mvu}] = 0$. (3)

Then

$$\begin{aligned} MSE(\hat{\theta}_{Mvu}) &= \text{Var}[\hat{\theta}_{Mvu}] = E[(\hat{\theta}_{Mvu} - E[\hat{\theta}_{Mvu}])^2] \\ &= E[(\hat{\theta}_{Mvu} - \theta_0)^2] = E[\hat{\theta}_{Mvu}^2] - \theta_0^2 \end{aligned}$$

(a). From $\hat{\theta}_{Mvu}$ being an unbiased estimator of θ_0 we know that $E[\hat{\theta}_{Mvu}] = \theta_0$ and $MSE[\hat{\theta}_{Mvu}] = \text{Var}[\hat{\theta}_{Mvu}]$

• From the fact that is the minimum unbiased estimator we know that if MVU exists, it is unique.

That may be biased estimators with lower MSE than $\hat{\theta}_{Mvu}$. And due to Cramer-RAO the MVU in LS for linear regression.

(b) Prove that all θ_b 's of F with $a \neq 0$ are biased.

• Let $\hat{\theta}_b = (1+a)\hat{\theta}_{Mvu}$ then $E[\hat{\theta}_b] = E[(1+a)\hat{\theta}_{Mvu}]$

$E[\hat{\theta}_b] = (1+a)E[\hat{\theta}_{Mvu}] = (1+a)\theta_0$ that for $a \neq 0$ is ~~biased~~ biased.

$$(c) \text{MSE}[\hat{\theta}_{Mvu}] = E[(\hat{\theta}_{Mvu} - \theta_0)^2] = \text{Var}[\hat{\theta}_{Mvu}]$$

this term can be 0 when $\text{Var}[\hat{\theta}_{Mvu}] = 0$.

Due to the fact that $\hat{\theta}_{Mvu}$ is an estimator deriving from N points of a dataset, its variance could ~~theoretically~~ theoretically approach 0 at $N \rightarrow \infty$. However, that is not feasible. The best approach would be just a very large number of points.

$$\begin{aligned} (d) \text{MSE}(\hat{\theta}_b) &= E[(1+a)\hat{\theta}_{Mvu} - E[(1+a)\hat{\theta}_{Mvu}]]^2 + (E[(1+a)\hat{\theta}_{Mvu}] - \theta_0)^2 \\ &= (1+a)^2 E[(\hat{\theta}_{Mvu} - E[\hat{\theta}_{Mvu}])^2] + (a E[\hat{\theta}_{Mvu}])^2 \\ &= (1+a)^2 \text{MSE}[\hat{\theta}_{Mvu}] + (a\theta_0)^2 \end{aligned}$$

(e) We know that there are biased estimators ($\hat{\theta}_b$)
 where

$$MSE(\hat{\theta}_b) < MSE(\hat{\theta}_{Mvu})$$

$$(1+a)^2 MSE(\hat{\theta}_{Mvu}) + a^2 \theta_0^2 < MSE(\hat{\theta}_{Mvu})$$

Let $MSE(\hat{\theta}_{Mvu})$ be e :

$$(1+a)^2 e + a^2 \theta_0^2 < e$$

$$(a^2 + 2a + 1)e + a^2 \theta_0^2 - e < 0$$

$$a^2 e + 2ae + a^2 \theta_0^2 < 0$$

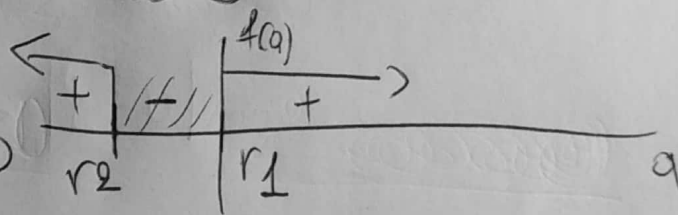
$$f(a) = (e + \theta_0^2) \cdot a^2 + 2e \cdot a + 0 = a(2e + (e + \theta_0^2)a)$$

$$\text{Dis: } \Delta(a) = 4e^2 > 0$$

In order to be negative ~~we need~~ we need:

Root 1: $a = 0$

Root 2: $a = \frac{-2e}{e + \theta_0^2} < 0$



$$-\frac{2e}{e + \theta_0^2} < a < 0$$

(4) We know that $-\frac{2e}{e+\theta_0^2} < a < 0$

Adding 1 to each side gives us:

$$1 - \frac{2e}{e+\theta_0^2} < a+1 < 1$$

$$\frac{e+\theta_0^2-2e}{e+\theta_0^2} < a+1 < 1$$

$$\frac{\theta_0^2 - e}{\theta_0^2 + e} < a+1 < 1 \Rightarrow a+1 < 1$$

Absolute on each side:

$$|a+1| < |1|$$

Then since $\hat{\theta}_b = (1+a)\hat{\theta}_{MVO} \Rightarrow |\hat{\theta}_b| = |1+a||\hat{\theta}_{MVO}|$

$$|\hat{\theta}_b| < |\hat{\theta}_{MVO}| \cdot |1| \Rightarrow |\hat{\theta}_b| < |\hat{\theta}_{MVO}|$$

(g) As found in ex (d)

$$MSE(\hat{\theta}_b) = (1+a)^2 MSE(\hat{\theta}_{Mvu}) + a^2 \theta_0^2$$

We need to Minimize $MSE(\hat{\theta}_b)$ wrt a thus,

$$\frac{\partial MSE(\hat{\theta}_b)}{\partial a} = 2(1+a) \cdot MSE(\hat{\theta}_{Mvu}) + 2a\theta_0^2$$

Find a^* such as:

$$\left. \frac{\partial MSE(\hat{\theta}_b)}{\partial a} \right|_{a=a^*} = 0 \Rightarrow 2(1+a^*) MSE(\hat{\theta}_{Mvu}) + 2a^* \theta_0^2 = 0$$

$$\Rightarrow (1+a^*) MSE(\hat{\theta}_{Mvu}) + a^* \theta_0^2 = 0$$

$$\Rightarrow MSE(\hat{\theta}_{Mvu}) + a^* MSE(\hat{\theta}_{Mvu}) + a^* \theta_0^2 = 0$$

$$a^* (MSE(\hat{\theta}_{Mvu}) + \theta_0^2) = -MSE(\hat{\theta}_{Mvu}) \Rightarrow$$

$$\Rightarrow a^* = \frac{-MSE(\hat{\theta}_{Mvu})}{MSE(\hat{\theta}_{Mvu}) + \theta_0^2}$$

(h) Lets see again the components of a^*

$$\left. \begin{aligned} -MSE[\hat{\theta}_{Mvu}] &= \text{Var}[\hat{\theta}_{Mvu}] = E[\hat{\theta}_{Mvu}^2] - \theta_0^2 \\ -\theta_0^2 \end{aligned} \right\}$$

Both contain the θ_0 which is not known, it is always to be determined! Thus its impossible to find a^*

Exercise 3

We have N pairs satisfying the equation

$$y_n = \theta_0^T x_n + \eta_n, \quad \eta_n \in \mathcal{N}(0, \sigma_n^2)$$

For the LS estimator we know that:

$$\left(\sum_{n=1}^N x_n x_n^T \right) \theta = \sum_{n=1}^N y_n x_n \quad (4)$$

Now $\theta \in \mathbb{R}$ and $x_n = 1 \quad \forall n \in N$. We have the scalar problem where $y_n = \theta_0 \cdot 1 + \eta_n \Rightarrow y_n = \theta_0 + \eta_n$

RV y_n models ~~$y_n = \theta_0 + \eta_n$~~ $y_n = \theta_0 + \eta_n$
and $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$.

(a) In eq 4 we have $\left(\sum_{n=1}^N x_n x_n^T \right) \theta = \sum_{n=1}^N y_n x_n$.

Substituting for $x_n = 1 \Rightarrow x_n^T = 1$ we have:

$$\left(\sum_{n=1}^N 1 \cdot 1 \right) \theta = \sum_{n=1}^N y_n \cdot 1 \Rightarrow N\theta = \sum_{n=1}^N y_n \Rightarrow \theta_{LS} = \frac{1}{N} \sum_{n=1}^N y_n$$

$$\Rightarrow \hat{\theta}_{LS} = \bar{y}$$

(b) We have that $E[y_n] = E[\theta_0 + \eta_n]$

$$= E[\theta_0] + E[\eta_n] = \theta_0 + 0 = \theta_0, \text{ thus } y_n \text{ is unbiased estimator of } \theta_0$$

(c) For \bar{y} we have that $\hat{\theta}_{LS} = \bar{y} = \frac{1}{N} \sum_{n=1}^N y_n$.

$$\begin{aligned} E[\hat{\theta}_{LS}] &= E[\bar{y}] = E\left[\frac{1}{N} \sum y_n\right] = \frac{1}{N} E\left[\sum y_n\right] = \frac{1}{N} \sum E[y_n] \\ &= \frac{1}{N} \sum \theta_0 = \frac{1}{N} N \theta_0 = \theta_0, \text{ meaning that } \hat{\theta}_{LS} = \bar{y} \text{ is also} \\ &\text{an unbiased estimator of } \theta_0. \end{aligned}$$

(d) $\bar{y} = \hat{\theta}_{LS} = \hat{\theta}_{MVO}$

(e) Starting from $\left(\sum_{n=1}^N x_n x_n^T + \lambda I\right) \hat{\theta} = \sum_{n=1}^N y_n x_n$
we have that:

$$\left(\sum_{n=1}^N 1 \cdot 1 + \lambda\right) \hat{\theta} = \sum_{n=1}^N y_n \cdot 1$$

$$(N + \lambda) \hat{\theta} = \sum_{n=1}^N y_n \Rightarrow \hat{\theta}_{\text{ridge}} = \frac{1}{(\lambda + N)} \sum_{n=1}^N y_n$$

~~$\hat{\theta}_{\text{ridge}} = \frac{1}{N + \lambda} \sum y_n = \frac{1}{N + \lambda} \sum y_n$~~ $\Rightarrow \hat{\theta}_{\text{ridge}} = \frac{\sum y_n}{N + \lambda}$

(f) $\hat{\theta}_{\text{ridge}} = \frac{\sum y_n}{\lambda + N} = \frac{\frac{1}{N} \sum y_n}{\frac{1}{N}(\lambda + N)} = \frac{1}{\frac{\lambda}{N} + 1} \hat{\theta}_{MVO} = \frac{N}{\lambda + N} \hat{\theta}_{MVO}$

$$(g) E[\hat{\theta}_{\text{ridge}}] = E\left[\frac{N}{\lambda+N} \hat{\theta}_{\text{Mvu}}\right] = \frac{N}{\lambda+N} E[\hat{\theta}_{\text{Mvu}}] = \frac{N}{\lambda+N} \theta_0 \neq \theta_0, \text{ thus}$$

$\hat{\theta}_{\text{ridge}}$ is a biased estimator.

$$(h) \text{ We have that } |\hat{\theta}_{\text{ridge}}| = \left| \frac{N}{\lambda+N} \hat{\theta}_{\text{Mvu}} \right| \Rightarrow$$

$$\Rightarrow |\hat{\theta}_{\text{ridge}}| = \left| \frac{N}{\lambda+N} \right| |\hat{\theta}_{\text{Mvu}}| \Rightarrow |\hat{\theta}_{\text{ridge}}| = |\hat{\theta}_{\text{Mvu}}| \left| \frac{N}{\lambda+N} \right|$$

$$\text{but } N, \lambda \in \mathbb{R}^+ \text{ and } \lambda > 0 \text{ thus } N < N+\lambda \Rightarrow \left| \frac{N}{\lambda+N} \right| < 1$$

$$\text{so } |\hat{\theta}_{\text{ridge}}| < |\hat{\theta}_{\text{Mvu}}|$$

$$(i) \text{ We have that } \hat{\theta}_{\text{biased}} = (1+a) \hat{\theta}_{\text{Mvu}}$$

$$\text{In our case } \hat{\theta}_{\text{biased}} = \hat{\theta}_{\text{ridge}} \text{ and } \hat{\theta}_{\text{Mvu}} = \hat{\theta}_{\text{LS}} = \bar{y}$$

So:

$$\hat{\theta}_{\text{ridge}} = (1+a) \hat{\theta}_{\text{Mvu}} \Rightarrow \frac{N}{N+\lambda} \hat{\theta}_{\text{Mvu}} = (1+a) \hat{\theta}_{\text{Mvu}}$$

$$\Rightarrow 1+a = \frac{N}{N+\lambda} \Rightarrow \boxed{a = -\frac{\lambda}{N+\lambda}}$$

$$\text{We had that the range was } \left(-\frac{2 \text{MSE}(\hat{\theta}_{\text{Mvu}})}{\text{MSE}(\hat{\theta}_{\text{Mvu}}) + \theta_0^2}, 0 \right)$$

$$\text{so } -\frac{2 \text{MSE}(\hat{\theta}_{\text{Mvu}})}{\text{MSE}(\hat{\theta}_{\text{Mvu}}) + \theta_0^2} < -\frac{\lambda}{N+\lambda} < 0 \Rightarrow 0 < \lambda < \frac{2N \cdot \text{MSE}(\hat{\theta}_{\text{Mvu}})}{\theta_0^2 - \text{MSE}(\hat{\theta}_{\text{Mvu}})}$$