# Exersize 1

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

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

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

We want:

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

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

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

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 $\theta$ be $\theta = \begin{bmatrix} \theta_0 \\ \theta_l \end{bmatrix} \in \mathbb{R}^{l+1}$

$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$ 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 $X^T = \begin{bmatrix} x_1 & x_2 & \dots & x_N \end{bmatrix}_{(l+1) \times N}$

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 $X^T X = x_1 x_1^T + x_2 x_2^T + \dots x_N x_N^T = \sum_{n=1}^{N} x_n x_n^T$ (a)

---

Also $X^T y$ is easily proven

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

---

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

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

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

# Exersize 2

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

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

Also $\hat{\theta}_{MVU}$ is the minimum variance unbiased estimator of $\theta_0$.

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.

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

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

Then

$$MSE(\hat{\theta}_{MVU}) = Var[\hat{\theta}_{MVU}] = E\left[\left(\hat{\theta}_{MVU} - E[\hat{\theta}_{MVU}]\right)^2\right]$$

$$= E\left[\left(\hat{\theta}_{MVU} - \theta_0\right)^2\right] = E[\hat{\theta}_{MVU}^2] - \theta_0^2$$

---

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

• From the fact that is the minimum unbiased estimator we know that if MVU exists, it is unique. That may are biased estimators with lower MSE than $\hat{\theta}_{MVU}$. And due to Cramer-RAO the MVU in LS for lin.reg is effi.

(b) Prove that all $\theta_b$'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.

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

this term can be 0 when $Var[\hat{\theta}_{MVU}] = 0$.
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 $N \to +\infty$. However, that is not feasible. The best approach would be just a very lage number of points.

(d) $MSE(\hat{\theta}_b) = E[((1+a)\hat{\theta}_{MVU} - E[(1+a)\hat{\theta}_{MVU}])^2] + (E[(1+a)\theta_{MVU}] - \theta_0)^2$

$= (1+a)^2 E[(\hat{\theta}_{MVU} - E[\hat{\theta}_{MVU}])^2] + (a E[\hat{\theta}_{MVU}])^2$

$= (1+a)^2 MSE[\hat{\theta}_{MVU}] + (a\theta_0)^2$

(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 $e$:

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

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

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

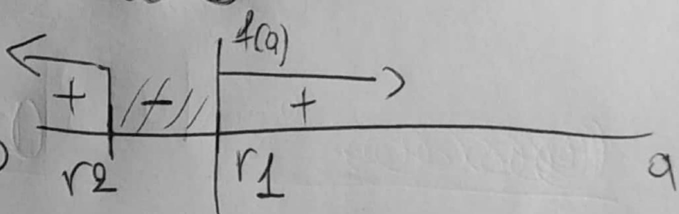
$$f(a) = (e+\theta_0^2)\cdot a^2 + 2e\cdot a + 0 = a(2e+(e+\theta_0^2)a)$$

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

In order to be negative we need:

root 1: $a=0$

root 2: $a = \dfrac{-2e}{e+\theta_0^2} < 0$




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

(f) 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 $\hat{\theta_b} = (1+a)\hat{\theta}_{MVU} \Rightarrow |\hat{\theta_b}| = |1+a|\,|\hat{\theta}_{MVU}|$

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

(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 <u>a</u> thus,

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

<u>Find $a^*$ such as:</u>

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

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

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

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

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

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

$$\left.\begin{array}{l} - MSE[\hat{\theta}_{MVU}] = Var[\hat{\theta}_{MVU}] = E[\hat{\theta}_{MVU}^2] - \theta_0^2 \\ - \theta_0^2 \end{array}\right\}$$

Both contain the $\theta_0$ which is not known, it is always to be determinded ! Thus its impossible to find $a^*$

## Exersize 3

We have N pairs satisfying the equation

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

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 $x_n = 1 \; \forall n \in N$. 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 $y_n = \theta_0 + \eta_n$
and $\bar{y} = \frac{1}{N} \sum_{1=1}^{N} y_n$.

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

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

$E[\hat{\theta}_{LS}] = E[\bar{y}] = E[\frac{1}{N}\sum y_n] = \frac{1}{N}E[\sum y_n] = \frac{1}{N}\sum E[y_n]$

$= \frac{1}{N}\sum \theta_0 = \frac{1}{N} N\theta_0 = \theta_0$, meaning that $\hat{\theta}_{LS} = \bar{y}$ is also an unbiased estimator of $\theta_0$.

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

(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:

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

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

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

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

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

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

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

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

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

So $|\hat{\theta}_{ridge}| < |\hat{\theta}_{MVU}|$

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

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

so:

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

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

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

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

## Exersize 4

(e) Discuss briefly on the results

Let's see the results of our fit on the data

-> LS estimator: Fairly good fit, with weights above $\theta_3$ ~~being~~ fluctuating between positive and negative values at large values. This is a sign of overfit meaning our model is heavily unconfident about new data and presents great variance as an estimator.

-> Ridge Estimator: Here we maintain a fairly good fit for values $10^{-4}, 10^{-3}, 10^{-2}$ and we see a great penalization on the terms above the 2nd degree. Theta (3+) values are relatively small compared to the lesser terms. Howerer at $l_2 = 0,1$ we see that the penalization interferes with the quality of the fit, ledding to worse results.

→ **Lasso Estimation**: In our lasso regression estimation we see a good fit at $l_1 = 5 \cdot 10^{-4}$ and $l_1 = 10^{-4}$ with a strict penalization of terms other than $\theta_0$ and $\theta_2$, a hard assumption that our model is of $y = \theta_2 x_2^2 + \theta_0$ form. At values greater than that our model seems to over-penalize the fit, leading from a poor-fitted model to a complete 0-variance straight line at $l_1 = 0,1$.

# Exersize 6

(ii) Comment briefly on the results.

In exersize 5 we proved by example that the fit of a LSE on a random dataset, after some trials approaches the true ~~estimated~~ values, under of course the assumption of a proper fit. That means our $\hat{\theta}$ estimator ~~follows a distribution~~ eventually converges to its dataset-generating value at high N (a lot of trials). Some fits may estimate higher, some lower but their mean "cancels-out" the deviation between them, leading to a better estimate. Also we proved that $l_2$ penalization (Ridge Reg.) is not always a solution. We need to pick a sensible value of $l_2$ in order to get good results. As I plotted in my Jupyter Notebook, a relatively low range of such values (2-40) gives the best results.