3nd Homework

Thomas Saltos

Exercise 1

We want to minimize the L(θ), so we have to set the derivative to 0:

$$\frac{\partial \mathrm{L}(\theta)}{\partial \theta} = 0 \Rightarrow 2\sum_{n=1}^{N}\left((-x_n)(y_n - x_n^T\theta) + \lambda\theta\right) = 0 \Rightarrow \sum_{n=1}^{N}(x_n x_n^T + \lambda\theta) = \sum_{n=1}^{N} x_n\, y_n$$

Let's define the matrix $\mathrm{X} = \begin{pmatrix} \mathrm{x}_1^{\mathrm{T}} \\ \mathrm{x}_2^{\mathrm{T}} \\ . \\ . \\ \mathrm{x}_{\mathrm{N}}^{\mathrm{T}} \end{pmatrix}$. Now the $\sum_{n=1}^{N} x_n\, x_n^T$ can be written in matrix form as $X^T X$.

Moreover, the term $\sum_{n=1}^{N} x_n\, y_n$ equals $X^T y$ where $\mathrm{y} = \begin{pmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ . \\ . \\ \mathrm{y}_{\mathrm{N}} \end{pmatrix}$.

The ridge-regression solution can be expressed as $\theta = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda\mathrm{I}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y}$

Exercise 2

a) The second term (biased term) equals to 0, since θ_{MVU} is unbiased estimator of θ_0, thus $E(\theta_{MVU}) = \theta_0$

b) $\mathrm{E}(\theta_b) = \mathrm{E}\left((1+a)\theta_{MVU}\right) = (1+a)\mathrm{E}(\theta_{MVU}) = (1+a)\theta_0$. For a $\neq 0$, θ_b is biased estimator of θ_0.

c) $MSE(\theta_{MVU}) = E[(\theta_{MVU} - E[\theta_{MVU}])^2]$. The mean square error cannot be zero unless θ_{MVU} has zero variance. This cannot be achieved in practice since we have a finite data set of points N. Zero variance is achieved for infinite size of data samples.

d) $MSE(\theta_b) = E\left[\left((1+a)\theta_{MVU} - E[(1+a)\theta_{MVU}]\right)^2\right] + (E[(1+a)\theta_{MVU}] - \theta_0)^2 =$
$(1+a)^2 E\left[\theta_{MVU} - E[\theta_{MVU}]\right]^2 + \left(E\left[\theta_{MVU} - \theta_0 + aE[\theta_{MVU}]\right]\right)^2 =$
$MSE(\theta_{MVU}) + (E[\theta_{MVU}])^2 \Rightarrow$
$MSE(\theta_b) = (1+a)^2 MSE(\theta_{MVU}) + (a\theta_0)^2$

e)

$$\mathrm{MSE}(\theta_b) < \mathrm{MSE}(theta_{MVU}) \Rightarrow (1+a)^2\mathrm{MSE}(\theta_{MVU}) + a^2\theta_0^2 < \mathrm{MSE}(\theta_{MVU})$$
$$\Rightarrow (a^2+2a)\mathrm{MSE}(\theta_0^2) + 2\mathrm{aMSE}(\theta_{MVU}) < 0$$
$$\Rightarrow a^2(MSE(\theta_{MVU}) + \theta_0^2) + 2aMSE(\theta_{MVU}) < 0$$

The above polynomial is negative when $-\frac{2MSE(\theta_{MVU})}{MSE(\theta_{MVU})+\theta_0^2} < a < 0$.

f) If we add 1 to the last equality, we have:

$$1 - \frac{2MSE(\theta_{MVU})}{MSE(\theta_{MVU}) + \theta_0^2} < a + 1 < 1 \Rightarrow |a+1| < 1$$

We conclude that:

$$|\theta_b| < |1+a||\theta_{MVU}|$$

g)

$$\frac{\partial MSE(\theta_b)}{\partial a} = 0 \Rightarrow (1+a)MSE(\theta_{MVU}) + a\theta_0^2 = 0 \Rightarrow a^* = -\frac{MSE(\theta_{MVU})}{MSE(\theta_{MVU}) + \theta_0^2}$$

h) a^* cannot be determined because θ_0 is unknown.

Exercise 3

a) From the equation of the LS we have that:

$$\mathrm{N}\theta = \sum_{n=1}^{N} y_n \Rightarrow \theta = \frac{\sum_{n=1}^{N} y_n}{\mathrm{N}}$$

b) $\mathrm{E}[y_n] = \mathrm{E}[\theta_0] + \mathrm{E}[\eta_n] = \theta_0 + 0 = \theta_0,\ \mathrm{E}[\eta_n] = 0$ since the noise is zero mean.

c) $\mathrm{E}[\bar{y}] = E\left[\frac{\sum_{n=1}^{N} y_n}{\mathrm{N}}\right] = \frac{N\theta_0}{N} = \theta_0$.

d)

e) Using the formula from 1a, we have that:

$$(N+\lambda)\acute{\theta} = \sum_{n=1}^{N} y_n \Rightarrow \acute{\theta} = \frac{\sum_{n=1}^{N} y_n}{\mathrm{N}+\lambda}$$

f) From (a) we have that the ridge regression estimator $\acute{\theta}$ is related to the LS estimator denoted as $\theta_{\acute{\mathrm{MVU}}}$ as follows:

$$\theta = \frac{N}{N+\lambda}\theta_{\acute{\mathrm{MVU}}}$$

g)

$$\mathrm{E}[\acute{\theta}] = \mathrm{E}\left[\frac{N}{N+\lambda}\theta_{\acute{MVU}}\right] = \frac{N}{N+\lambda}\mathrm{E}[\theta_{\acute{MVU}}] = \frac{N}{N+\lambda}\theta_0$$

Thus, $\acute{\theta}$ is a biased estimator.

h) From (f) we have that $|\theta| = |\frac{N}{N+\lambda}||\theta_{\acute{\mathrm{MVU}}}|$ and $\left|\frac{N}{N+\lambda}\right| < 1 for \lambda > 0 \Rightarrow |\acute{\theta}| < |\theta_{\acute{MVU}}|$

i) Recalling from exercise 2 that $\acute{\theta_b} = (1+a)\theta_{\acute{MVU}}$ we can get that $(1+a) = \frac{N}{N+\lambda}$

Then from (2e):

$$\frac{\lambda}{N+\lambda} < \frac{2MSE(\theta_{\acute{MVU}})}{MSE(\theta_{\acute{MVU}}) + \theta_0^2} \Rightarrow \lambda < \frac{2MSE(\theta_{\acute{MVU}})N}{\theta_0^2 - MSE(\theta_{\acute{MVU}})}$$

Exercise 4

d) It is noticed the 8th degree polynomial model has high complexity for this problem, which is linear, thus our model overfit the data. It is worth noticing that the coefficients of higher orders are estimated to large values. On the contrary, in the regularized LS the higher orders coefficient takes values close to zero. Regularization is a useful tool when the model which generate the data is unknown.