02471 Machine Learning for Signal Processing

Solution

Exercise 2: Parameter Estimation

2.1 Linear Models

Exercise 2.1.2

The book uses column vectors, so we get the following dimensions for the vectors: $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{\theta} \in \mathbb{R}^{(l+1)\times 1}$ where l is the number of dimensions in the input data. X then neeeds to be a $\mathbb{R}^{N\times(l+1)}$ matrix. We define these as

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,l} & 1 \\ x_{2,1} & \cdots & x_{2,l} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{N,1} & \cdots & x_{N,l} & 1 \end{bmatrix}$$

If we write the sum for the first data-point we get:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{n=1}^{N=1} (y_n - \boldsymbol{\theta}_b^T \boldsymbol{x}_n - \theta_0)^2 \\ &= (y_1 - \theta_1 x_{1,1} - \cdots - \theta_l x_{1,l} - \theta_0)^2 \end{aligned}$$

Similarly, the first component of the vector $(\boldsymbol{y} - X\boldsymbol{\theta})$ is

$$y_1 - \theta_1 x_{1,1} - \cdots - \theta_l x_{1,l} - \theta_0 \cdot 1$$

Since the inner product of a vector is the sum of all the components squared, we have shown the relation.

Exercise 2.1.3

We first make the following rewrites:

$$\begin{aligned} J(\boldsymbol{\theta}) &= (\boldsymbol{y} - X\boldsymbol{\theta})^T (\boldsymbol{y} - X\boldsymbol{\theta}) \\ &= (\boldsymbol{y}^T - (X\boldsymbol{\theta})^T)(\boldsymbol{y} - X\boldsymbol{\theta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\theta} - (X\boldsymbol{\theta})^T\boldsymbol{y} + (X\boldsymbol{\theta})^T X\boldsymbol{\theta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - (X\boldsymbol{\theta})^T\boldsymbol{y} - (X\boldsymbol{\theta})^T\boldsymbol{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2(X\boldsymbol{\theta})^T\boldsymbol{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta} \end{aligned}$$

We can now use the following rules from appendix A

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^T\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{\partial \boldsymbol{x}^T A\boldsymbol{x}}{\partial \boldsymbol{x}} = \left(A + A^T\right)\boldsymbol{x}$$

We get

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}) &= -\frac{\partial}{\partial \boldsymbol{\theta}} 2(X\boldsymbol{\theta})^T \boldsymbol{y} + \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T X^T X \boldsymbol{\theta} \\
&= -2\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T X^T \boldsymbol{y} + (X^T X + (X^T X)^T)\boldsymbol{\theta} \\
&= -2X^T \boldsymbol{y} + 2X^T X \boldsymbol{\theta}
\end{aligned}
$$

Note that $X^T\boldsymbol{y}$ results in a $(l+1 \times 1)$ sized vector, so $\boldsymbol{\theta}^T X^T \boldsymbol{y}$ is indeed an inner product between two vectors.

2.2 Estimation

Exercise 2.2.1

Assume the model $\mathrm{y} = g(\mathbf{x}) + \eta$, then we have (since η is zero-mean, and $\boldsymbol{x}$ is observed, that is $\mathbf{x} = \boldsymbol{x}$), then

$$
\begin{aligned}
\mathbb{E}[\mathrm{y}|\boldsymbol{x}] &= \mathbb{E}[g(\boldsymbol{x}) + \eta] \\
&= g(\boldsymbol{x}) + \mathbb{E}[\eta] \\
&= g(\boldsymbol{x}) \\
\mathrm{MSE} &= \mathbb{E}\big[(\mathrm{y} - \mathbb{E}[\mathrm{y}|\boldsymbol{x}])^2\big] \\
&= \mathbb{E}\big[(g(\boldsymbol{x}) + \eta - \mathbb{E}[\mathrm{y}|\boldsymbol{x}])^2\big] \\
&= \mathbb{E}\big[(g(\boldsymbol{x}) + \eta - g(\boldsymbol{x})^2\big] \\
&= \mathbb{E}\big[\eta^2\big]
\end{aligned}
$$

The variance is defined as $\mathrm{var}[\eta] = \mathbb{E}[(\eta - \mathbb{E}[\eta])^2]$, so for a zero mean variable we have $\mathrm{var}[\eta] = \mathbb{E}[\eta^2]$. Hence we get the result

$$
\mathrm{MSE} = \mathbb{E}\big[\eta^2\big] = \sigma_\eta^2
$$

Exercise 2.2.2

Since we are dealing with unbiased estimator, we know that:

$$
\mathbb{E}[\hat{\theta}_i] = \boldsymbol{\theta}
$$

We also know that estimators are uncorrelated and all have the variance:

$$
\sigma^2 = \mathbb{E}\big[(\theta_i - \boldsymbol{\theta}_o)^T(\theta_i - \boldsymbol{\theta}_o)\big]
$$

$$
\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i
$$

Putting this together yields:

$$
\mathbb{E}[\hat{\theta}] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\hat{\theta}_i] = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\theta} = \boldsymbol{\theta}
$$

Next, assuming that estimators are uncorrelated, meaning:

$$
\mathbb{E}\Big[(\hat{\theta}_i - \boldsymbol{\theta}_o)^T(\hat{\theta}_j - \boldsymbol{\theta}_o)\Big] = \sigma^2\delta_{ij}, \text{ where } \delta_{ij} = 1 \text{ when } i = j \text{ and zero otherwise.}
$$

Use substitution to obtain:

$$\begin{aligned}
\sigma_c^2 &= \mathbb{E}\Big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)\Big] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o\right)^T\left(\frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o\right)\right] \\
&= \mathbb{E}\left[\frac{1}{m^2}\left(\sum_{i=1}^{m}\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o\right)^T\left(\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o\right)\right] \\
&= \mathbb{E}\left[\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\Big(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o\Big)^T\Big(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o\Big)\right] \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbb{E}\Big[(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_o)^T(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_o)\Big] \\
&= \frac{1}{m^2}m\sigma^2 \\
&= \frac{1}{m}\sigma^2
\end{aligned}$$

Exercise 2.2.3

We know that $\hat{\theta}_u$ is an unbiased estimator, so that $\mathbb{E}[\hat{\theta}_u] = \theta_0$. For this exercise, the biased estimator is defined as $\hat{\theta}_b := (1+\alpha)\hat{\theta}_u$. Further, we assume that $\mathrm{MSE}(\hat{\theta}_b) > 0$ (this is only zero if we have zero noise and a perfect fit) and that $\theta_o > 0$ (since $\theta_o = 0$ is the trivial case of a null-fit).

First calculate the MSE of biased estimator:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}_b) &= \mathbb{E}\Big[(\hat{\theta}_b - \theta_o)^2\Big] \\
&= \mathbb{E}\Big[((1+\alpha)\hat{\theta}_u - \theta_o)^2\Big]
\end{aligned}$$

The trick is to add and subtract term $\alpha\theta_o$:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}_b) &= \mathbb{E}\Big[((1+\alpha)\hat{\theta}_u + \alpha\theta_o - \alpha\theta_o - \theta_o)^2\Big] \\
&= \mathbb{E}\Big[((1+\alpha)\hat{\theta}_u + \alpha\theta_o - \theta_o(\alpha+1))^2\Big] \\
&= \mathbb{E}\Big[((1+\alpha)(\hat{\theta}_u - \theta_o) + \alpha\theta_o)^2\Big] \\
&= \mathbb{E}\Big[(1+\alpha)^2(\hat{\theta}_u - \theta_o)^2 + \alpha^2\theta_o^2 + 2(1+\alpha)(\hat{\theta}_u - \theta_o)\alpha\theta_o\Big]
\end{aligned}$$

Since α and θ_o are deterministic, we can narrow the scope of the expectations (expectations only needs to be taken wrt random variables):

$$\mathrm{MSE}(\hat{\theta}_b) = (1+\alpha)^2\mathbb{E}\Big[(\hat{\theta}_u - \theta_o)^2\Big] + \alpha^2\theta_o^2 + 2\alpha(1+\alpha)(\mathbb{E}[\hat{\theta}_u] - \theta_o)\theta_o$$

Taking into account that $\mathbb{E}\Big[(\hat{\theta}_u - \theta_o)^2\Big] = \mathrm{MSE}(\hat{\theta}_u)$ and $\mathbb{E}[\hat{\theta}_u] = \theta_o$, we end up with:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}_b) &= (1+\alpha)^2\mathrm{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 + 2\alpha(1+\alpha)(\theta_o - \theta_o)\theta_o \\
&= (1+\alpha)^2\mathrm{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2
\end{aligned}$$

Now we have an expression for $\text{MSE}(\hat{\theta}_b)$. Next we seek the solution for α, so that

$$\text{MSE}(\hat{\theta}_b) < \text{MSE}(\hat{\theta}_u)$$

By substitution, we get:

$$\begin{aligned}
(1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< \text{MSE}(\hat{\theta}_u) \\
\Rightarrow (1+\alpha^2+2\alpha)\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< \text{MSE}(\hat{\theta}_u) \\
\Rightarrow \text{MSE}(\hat{\theta}_u) + \alpha^2\text{MSE}(\hat{\theta}_u) + 2\alpha\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< \text{MSE}(\hat{\theta}_u) \\
\Rightarrow \alpha^2\text{MSE}(\hat{\theta}_u) + 2\alpha\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_o^2 &< 0 \\
\Rightarrow \alpha\Big(\alpha\text{MSE}(\hat{\theta}_u) + 2\text{MSE}(\hat{\theta}_u) + \alpha\theta_o^2\Big) &< 0
\end{aligned}$$

If we multiply both sides by $\frac{1}{\theta_o^2+\text{MSE}(\hat{\theta}_u)}$ (which is a positive quantity), we get:

$$\begin{aligned}
\Rightarrow \frac{\alpha\Big(\alpha\text{MSE}(\hat{\theta}_u) + 2\text{MSE}(\hat{\theta}_u) + \alpha\theta_o^2\Big)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)} &< 0 \\
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)} + \alpha\frac{\theta_o^2+\text{MSE}(\hat{\theta}_u)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)}\right) &< 0 \\
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)} + \alpha\right) &< 0
\end{aligned}$$

To proceed, we need to handle the three cases of α: $\alpha < 0$, $\alpha > 0$, and $\alpha = 0$. The case where $\alpha = 0$, does not really make sense to consider, since this means our biased estimator will be defined as our unbiased estimator and no bias is then induced.

For the case $\alpha > 0$: here the biased estimator "expands" $\hat{\theta}_u$ (by noting the definition of $\hat{\theta}_b$), and we get:

$$\begin{aligned}
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)} + \alpha\right) &< 0 \\
\Rightarrow \frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)} + \alpha &< 0
\end{aligned}$$

Since the terms $\text{MSE}(\hat{\theta}_u)$ and θ_o^2 are both positive $\text{MSE}(\hat{\theta}_u)$, and α assumed positive, this inequality cannot hold, since there is no route to make the result on the left hand side below 0. Hence, $\alpha > 0$ will not result in a reduction in MSE.

For the case $\alpha < 0$: here the biased estimator "shrinks" $\hat{\theta}_u$, and we get

$$\begin{aligned}
\Rightarrow \alpha\left(\frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)} + \alpha\right) &< 0 \\
\Rightarrow \frac{2\text{MSE}(\hat{\theta}_u)}{\theta_o^2+\text{MSE}(\hat{\theta}_u)} + \alpha &> 0
\end{aligned}$$

The term on the left hand side, is only positive if

$$\frac{2\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u)+\theta_o^2} > -\alpha \quad \Leftrightarrow \quad -\frac{2\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u)+\theta_o^2} < \alpha$$

Additionally, we can establish a lower bound for α by analyzing $\frac{\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u)+\theta_o^2}$. Since all the individual terms in $\frac{\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u)+\theta_o^2}$ are assumed positive, we get:

$$0 < \frac{\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u) + \theta_o^2} < 1$$

That means:

$$-2 < \alpha < 0 \Rightarrow -1 < 1 + \alpha < 1 \Rightarrow |1 + \alpha| < 1$$

and additionally

$$|\hat{\theta}_b| = |(1 + \alpha)\hat{\theta}_u| = |1 + \alpha||\hat{\theta}_u| < |\hat{\theta}_u|$$

2.3 Bias-variance trade-off

Exercise 2.3.1

Reusing the rewrites we made in exercise 2.1, we can readily see we can write the cost-function as

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - X\boldsymbol{\theta})^T(\boldsymbol{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}$$

And using the same procedure for differentation we get

$$\frac{\partial}{\partial\boldsymbol{\theta}}J(\boldsymbol{\theta}) = -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\theta} + 2\lambda I\boldsymbol{\theta}$$

Equating the derivative to 0 and re-arrange yields

$$\begin{aligned} -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\theta} + 2\lambda I\boldsymbol{\theta} &= \mathbf{0} & \Leftrightarrow \\ (2X^TX + 2\lambda I)\boldsymbol{\theta} &= 2X^T\boldsymbol{y} & \Leftrightarrow \\ (X^TX + \lambda I)\boldsymbol{\theta} &= X^T\boldsymbol{y} \end{aligned}$$