1 Ex.6.2

Define the vector-valued function $b(x)^T = \left(1, x, x^2, \ldots, x^k\right)$ for $k \geq 0$. Let $\mathbf{B}$ be the $N \times (k+1)$ regression matrix with i th row $b\left(x_i\right)^T$, and $\mathbf{W}\left(x_0\right)$ the $N \times N$ diagonal matrix with i th diagonal element $K_\lambda\left(x_0, x_i\right)$. Then we have

$$b\left(x_0\right)^T = b\left(x_0\right)^T \left(\mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}. \quad (1)$$

Note the definition of $l_i\left(x_0\right)$ in (6.9) in text, from (1), we have

$$\begin{aligned}
&1 = b\left(x_0\right)^T \left(\mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{1} = \textstyle\sum_{i=1}^N l_i\left(x_0\right) \\
&x_0 = b\left(x_0\right)^T \left(\mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}_2 = \textstyle\sum_{i=1}^N l_i\left(x_0\right) x_i \\
&\ldots \\
&x_0^k = b\left(x_0\right)^T \left(\mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}_{k+1} = \textstyle\sum_{i=1}^N l_i\left(x_0\right) x_i^k
\end{aligned}$$

where $\mathbf{B}_i$ is the i th column of $\mathbf{B}$ (note that $\mathbf{B}_1 = \mathbf{1}$). Therefore we have $b_0\left(x_0\right) = \sum_{i=1}^N l_i\left(x_0\right) = 1$ and

$$b_1\left(x_0\right) = \sum_{i=1}^N \left(x_i - x_0\right) l_i\left(x_0\right) = \sum_{i=1}^N l_i\left(x_0\right) x_i - x_0 \sum_{i=1}^N l_i\left(x_0\right) = x_0 - x_0 \cdot 1 = 0.$$

For $j \geq 2$, we have

$$\begin{aligned}
b_j\left(x_0\right) &= \sum_{i=1}^N \left(x_i - x_0\right)^j l_i\left(x_0\right) \\
&= \sum_{i=1}^N \left(\sum_{b=0}^j C_j^b (-1)^b x_i^{j-b} x_0^b\right) l_i\left(x_0\right) \\
&= \sum_{b=0}^j C_j^b (-1)^b x_0^b \left(\sum_{i=1}^N l_i\left(x_0\right) x_i^{j-b}\right) \\
&= \sum_{b=0}^j C_j^b (-1)^b x_0^b x_0^{j-b} \\
&= \sum_{b=0}^j C_j^b (-1)^b x_0^j \\
&= (1-1)^j x_0^j \\
&= 0.
\end{aligned}$$

By Taylor expansion we have

$$\begin{aligned}
E\left[\hat{f}\left(x_0\right)\right] - f\left(x_0\right) =& \sum_{i=1}^N l_i\left(x_0\right) f\left(x_i\right) - f\left(x_0\right) \\
=& f\left(x_0\right) \sum_{i=1}^N l_i\left(x_0\right) - f\left(x_0\right) + f'\left(x_0\right) \sum_{i=1}^N \left(x_i - x_0\right) l_i\left(x_0\right) \\
&+ \frac{f''\left(x_0\right)}{2} \sum_{i=1}^N \left(x_i - x_0\right)^2 l_i\left(x_0\right) \\
&+ \ldots \\
&+ (-1)^k \frac{f^{(k)}}{k!} \sum_{i=1}^N \left(x_i - x_0\right)^k l_i\left(x_0\right) \\
&+ R \\
=& R,
\end{aligned}$$

where the remainder term R involves $(k+1)$ th and higher-order derivatives of f, on which the bias only depends.

2 Ex.6.3

Let's first introduce notations. Define the vector-valued function $b(x)^T = \left(1, x, x^2, \ldots, x^d\right)$ for $d \geq 1$. Let $\mathbf{B}$ be the $N \times (d+1)$ regression matrix with i th row $b\left(x_i\right)^T$.

$$\mathbf{B} = \left(\begin{array}{cccc} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \cdots & x_N^d \end{array}\right) = \left(\begin{array}{c} b\left(x_1\right)^T \\ b\left(x_2\right)^T \\ \vdots \\ b\left(x_N\right)^T \end{array}\right) \in R^{N \times (d+1)}$$

and

Let $\mathbf{W}(x)$ the $N \times N$ diagonal matrix with i th diagonal element $K_\lambda\left(x, x_i\right)$, that is,

$$\mathbf{W}(x) = \left(\begin{array}{cccc} K_\lambda\left(x, x_1\right) & 0 & \cdots & 0 \\ 0 & K_\lambda\left(x, x_2\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_{\lambda\left(x, x_N\right)} \end{array}\right) \in R^{N \times N}$$

Note that $\mathbf{W}(x) = \mathbf{W}^T(x)$ By definition of $l(x)$ (see, e.g... (6.9) in the text), we have

$$l\left(x_0\right)^T = b\left(x_0\right)^T \left(\mathbf{B}^T \mathbf{W}\left(x_0\right) \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{W}\left(x_0\right)$$

Denote $b = b\left(x_0\right)$ and $\mathbf{W} = \mathbf{W}\left(x_0\right)$ to simplify the notations from now on, we have

$$\begin{aligned} \left\|l\left(x_0\right)\right\|^2 &= l\left(x_0\right)^T l\left(x_0\right) \\ &= b^T \left(\mathbf{B}^T \mathbf{W B}\right)^{-1} \mathbf{B}^T \mathbf{W} \mathbf{W}^T \mathbf{B} \left(\mathbf{B}^T \mathbf{W B}\right)^{-1} b \end{aligned} \tag{1}$$

We need to show $\left\|l\left(x_0\right)\right\|^2$ is increasing in d. The expression involves with the weighted kernel matrix $\mathbf{W}$, however it turns out $\left\|l\left(x_0\right)\right\|^2$ does not depend on $\mathbf{W}$. Note that we could plug $\mathbf{I} = \mathbf{B B}^T \left(\mathbf{B B}^T\right)^{-1} = \left(\mathbf{B B}^T\right)^{-1} \mathbf{B B}^T$ between $\mathbf{W}$ and $\mathbf{W}^T$ in (1), we obtain

$$\begin{aligned} & \left\|l\left(x_0\right)\right\|^2 \\ = & b^T \left(\mathbf{B}^T \mathbf{W B}\right)^{-1} \mathbf{B}^T \mathbf{W B B}^T \left(\mathbf{B B}^T\right)^{-1} \left(\mathbf{B B}^T\right)^{-1} \mathbf{B B}^T \mathbf{W}^T \mathbf{B} \left(\mathbf{B}^T \mathbf{W B}\right)^{-1} b \\ = & b^T \mathbf{B}^T \left(\mathbf{B B}^T\right)^{-1} \left(\mathbf{B B}^T\right)^{-1} \mathbf{B} b, \end{aligned}$$

therefore we see $\left\|l\left(x_0\right)\right\|^2$ is independent of $\mathbf{W}$ So we can take $\mathbf{W} = \mathbf{I}$ in (1), which gives

$$\left\|l\left(x_0\right)\right\|^2 = b^T \left(\mathbf{B}^T \mathbf{B}\right)^{-1} b \tag{2}$$

Now consider the case for $d+1$, we denote

$$\hat{\mathbf{B}} = \left(\begin{array}{ccccc} 1 & x_1 & \cdots & x_1^d & x_1^{d+1} \\ 1 & x_2 & \cdots & x_2^d & x_2^{d+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_N & \cdots & x_N^d & x_N^{d+1} \end{array}\right) = \left(\begin{array}{cc} \mathbf{B} & c \end{array}\right) \in R^{N \times (d+2)}$$

where

$$c = \left(\begin{array}{c} x_1^{d+1} \\ x_2^{d+1} \\ \vdots \\ x_N^{d+1} \end{array}\right) \in R^{N \times 1}$$

Similarly, denote $\hat{b}^T = \left(b^T, x_0^{d+1}\right) = \left(1, x_0, x_0^2, \ldots, x_0^d, x_0^{d+1}\right) \in R^{1 \times (d+2)}$. In that case, (2) becomes

$$\left\|\hat{l}\left(x_0\right)\right\|^2=\hat{b}^T\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{b} \tag{3}$$

Further, we have

$$\hat{\mathbf{B}}^T\hat{\mathbf{B}}=\begin{pmatrix}\mathbf{B}^T\\ c^T\end{pmatrix}\begin{pmatrix}\mathbf{B} & c\end{pmatrix}=\begin{pmatrix}\mathbf{B}^T\mathbf{B} & \mathbf{B}^Tc\\ c^T\mathbf{B} & c^Tc\end{pmatrix}\in R^{(d+2)\times(d+2)}.$$

Note that $c^Tc\in R^1$ is a scalar. Recall the formula for block matrix inverse, (e.g., Schur complement), we have

$$\begin{aligned}&\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\\ =&\begin{pmatrix}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}+\frac{1}{k}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^Tcc^T\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1} & -\frac{1}{k}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^Tc\\ -\frac{1}{k}c^T\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1} & \frac{1}{k}\end{pmatrix}\end{aligned}$$

where

$$k=c^Tc-c^T\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^Tc \tag{4}$$

Denote $\beta=\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^Tc\in R^{(d+2)\times 1}$, plug $\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}$ into (3), we get

$$\begin{aligned}&\left\|\hat{l}\left(x_0\right)\right\|^2\\ =&\begin{pmatrix}b^T & x_0^{d+1}\end{pmatrix}\begin{pmatrix}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}+\frac{1}{k}\beta\beta^T & -\frac{1}{k}\beta\\ -\frac{1}{k}\beta^T & \frac{1}{k}\end{pmatrix}\begin{pmatrix}b\\ x_0^{d+1}\end{pmatrix}\\ =&b^T\left(\mathbf{B}^T\mathbf{B}\right)^{-1}b+\frac{1}{k}\left[b^T\beta\beta^Tb-x_0^{d+1}\beta^Tb-x_0^{d+1}b^T\beta+\left(x_0^{d+1}\right)^2\right]\\ =&b^T\left(\mathbf{B}^T\mathbf{B}\right)^{-1}b+\frac{1}{k}\left(x_0^{d+1}-b^T\beta\right)^2\quad\left(\text{ note } b^T\beta\in R\right)\\ =&\left\|l\left(x_0\right)\right\|^2+\frac{1}{k}\left(x_0^{d+1}-b^T\beta\right)^2.\end{aligned}$$

Therefore, it suffices to show that $k>0$ for k defined in (4). To do that, we only need to show

$$\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\preceq\mathbf{I}_N.$$

Consider the QR decomposition of $\mathbf{B}$

$$\mathbf{B}=\mathbf{Q}\mathbf{R}$$

where $\mathbf{Q}$ is an $N\times(d+1)$ orthogonal matrix, $\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_N$, and $\mathbf{R}\in R^{(d+1)\times(d+1)}$ is an upper triangular matrix. Then

$$\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T=\mathbf{Q}\mathbf{R}\left(\mathbf{R}^T\mathbf{R}\right)^{-1}\mathbf{R}^T\mathbf{Q}^T=\mathbf{Q}\mathbf{Q}^T. \tag{5}$$

Let $(\mathbf{Q}\mathbf{Q}_1)$ be an $N\times N$ orthogonal matrix, we have

$$\mathbf{I}_N=\begin{pmatrix}\mathbf{Q} & \mathbf{Q}_1\end{pmatrix}\begin{pmatrix}\mathbf{Q}^T\\ \mathbf{Q}_1^T\end{pmatrix}=\mathbf{Q}\mathbf{Q}^T+\mathbf{Q}_1\mathbf{Q}_1^T$$

The result (5) follows by noting $\mathbf{Q}_1\mathbf{Q}_1^T$ is positive semi-definite. The proof is now complete.

3 Ex.6.7

Note that local regression smoothers are linear estimators, and we can write

$$\hat{\mathbf{f}}=\mathbf{S}_{\lambda\mathbf{y}}$$

where $\{\mathbf{S}_\lambda\}_{ij}=l_i\left(x_j\right)$ for $l_i(x)$ defined by (6.8) in the text. Then we know

$$y_i-\hat{f}^{-i}\left(x_i\right)=\frac{y_i-\hat{f}\left(x_i\right)}{1-\{\mathbf{S}_\lambda\}_{ii}}$$

4 Ex.6.10

Consider the in-sample prediction error (7.18) and the training error $\overline{\mathrm{err}}$ in the case of squared-error loss:

$$\begin{aligned}
\mathrm{Err}_{\mathrm{in}} &= \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}\left(Y_i^0 - \hat{f}\left(x_i\right)\right)^2 \\
\overline{\mathrm{err}} &= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}\left(x_i\right)\right)^2 .
\end{aligned}$$

Add and subtract $f\left(x_i\right)$ and $E\hat{f}\left(x_i\right)$ in each expression and expand. Hence establish that the average optimism in the training error is

$$\frac{2}{N}\sum_{i=1}^{N}\operatorname{Cov}\left(\hat{y}_i, y_i\right),$$

. So we know that

$$\mathrm{PE}(\lambda) = \mathrm{ASR}(\lambda) + \frac{2}{N}\sum_{i=1}^{N}\operatorname{Cov}\left(\hat{y}_i, y_i\right)$$

and we have

$$\begin{aligned}
\sum_{i=1}^{N}\operatorname{Cov}\left(\hat{y}_i, y_i\right) &= \operatorname{trace}(\operatorname{Cov}(\hat{\mathbf{y}}, \mathbf{y})) \\
&= \operatorname{trace}(\operatorname{Cov}(\mathbf{S}\mathbf{y}, \mathbf{y})) \\
&= \operatorname{trace}(\mathbf{S}\mathbf{Cov}(\mathbf{y}, \mathbf{y})) \\
&= \operatorname{trace}(\mathbf{S}\mathbf{Var}(\mathbf{y})) \\
&= \operatorname{trace}(\mathbf{S})\sigma_\epsilon^2 .
\end{aligned}$$

Then the proof is straightforward.