7
Splitting the Sample

In the previous chapters the parameters of the estimates with the optimal rate of convergence depend on the unknown distribution of (X, Y), especially on the smoothness of the regression function. In this and in the following chapter we present data-dependent choices of the smoothing parameters. We show that for bounded Y the estimates with parameters chosen in such an adaptive way achieve the optimal rate of convergence.

7.1 Best Random Choice of a Parameter

Let $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be the sample as before. Assume a finite set $\mathcal{Q}_n$ of parameters such that for every parameter $h \in \mathcal{Q}_n$ there is a regression function estimate $m_n^{(h)}(\cdot) = m_n^{(h)}(\cdot, D_n)$. Let $\hat{h} = \hat{h}(D_n) \in \mathcal{Q}_n$ be such that

$$\int |m_n^{(\hat{h})}(x) - m(x)|^2 \mu(dx) = \min_{h \in \mathcal{Q}_n} \int |m_n^{(h)}(x) - m(x)|^2 \mu(dx),$$

where $\hat{h}$ is called the best random choice of the parameter. Obviously, $\hat{h}$ is not an estimate, it depends on the unknown m and μ.

This best random choice can be approximated by splitting the data. Let $D_{n_l} = \{(X_1, Y_1), \dots, (X_{n_l}, Y_{n_l})\}$ be the learning (training) data of size n_l and $D_n \backslash D_{n_l}$ the testing data of size n_t ($n = n_l + n_t \geq 2$). For every parameter $h \in \mathcal{Q}_n$ let $m_{n_l}^{(h)}(\cdot) = m_{n_l}^{(h)}(\cdot, D_{n_l})$ be an estimate of m depending only on the learning data D_{n_l} of the sample D_n. Use the testing data to

choose a parameter $H = H(D_n) \in \mathcal{Q}_n$:

$$\frac{1}{n_t}\sum_{i=n_l+1}^{n_l+n_t} |m_{n_l}^{(H)}(X_i) - Y_i|^2 = \min_{h\in\mathcal{Q}_n} \frac{1}{n_t}\sum_{i=n_l+1}^{n_l+n_t} |m_{n_l}^{(h)}(X_i) - Y_i|^2. \quad (7.1)$$

Define the estimate by

$$m_n(x) = m_n(x, D_n) = m_{n_l}^{(H)}(x, D_{n_l}). \quad (7.2)$$

We show that H approximates the best random choice $\hat{h}$ in the sense that $\mathbf{E}\int |m_n(x) - m(x)|^2\mu(dx)$ approximates $\mathbf{E}\int |m_{n_l}^{(\hat{h})}(x) - m(x)|^2\mu(dx)$.

Theorem 7.1. *Let $0 < L < \infty$. Assume*

$$|Y| \le L \quad a.s. \quad (7.3)$$

and

$$\max_{h\in\mathcal{Q}_n} \|m_{n_l}^{(h)}\|_\infty \le L \quad a.s. \quad (7.4)$$

Then, for any $\delta > 0$,

$$\mathbf{E}\int |m_n(x) - m(x)|^2\mu(dx)$$

$$\le (1+\delta)\mathbf{E}\int |m_{n_l}^{(\hat{h})}(x) - m(x)|^2\mu(dx) + c\frac{1+\log(|\mathcal{Q}_n|)}{n_t}, \quad (7.5)$$

where $\hat{h} = \hat{h}(D_{n_l})$ and $c = L^2(16/\delta + 35 + 19\delta)$.

The only assumption on the underlying distribution in Theorem 7.1 is the boundedness of $|Y|$ (cf. (7.3)). It can be applied to any estimate which is bounded in supremum norm by the same bound as the data (cf. (7.4)). We can always truncate an estimate at $\pm L$, which implies that (7.4) holds. If (7.3) holds, then the regression function will be bounded in absoulte value by L, too, and hence the L_2 error of the truncated estimate will be less than or equal to the L_2 error of the original estimate, so the truncation has no negative consequence in view of the error of the estimate.

In the next section we will apply this theorem to partitioning, kernel, and nearest neighbor estimates. We will choose $\mathcal{Q}_n$ and n_t such that the second term on the right-hand side of (7.5) is less than the first term. This implies that the expected L_2 error of the estimate is bounded by some constant times the expected L_2 error of an estimate, which is applied to a data set of size n_l (rather than n) and where the parameter is chosen in an optimal way for this data set. Observe that this is not only true asymptotically, but true for each finite sample size.

Proof of Theorem 7.1. An essential tool in the proof will be Bernstein's inequality together with majorization of a variance by some constant times the corresponding expectation. This will yield the denominator n_t in the

result instead of $\sqrt{n_t}$ attainable by the use of Hoeffding's inequality (cf. Problem 7.2).

We will use the error decomposition

$$\begin{aligned}
& \mathbf{E}\left\{\int |m_n(x)-m(x)|^2\mu(dx)\Big|D_{n_l}\right\} \\
= \;& \mathbf{E}\left\{\int |m_{n_l}^{(H)}(x)-m(x)|^2\mu(dx)\Big|D_{n_l}\right\} \\
= \;& \mathbf{E}\left\{|m_{n_l}^{(H)}(X)-Y|^2\Big|D_{n_l}\right\} - \mathbf{E}|m(X)-Y|^2 \\
=: \;& T_{1,n}+T_{2,n},
\end{aligned}$$

where

$$T_{1,n} = \mathbf{E}\left\{|m_{n_l}^{(H)}(X)-Y|^2\Big|D_{n_l}\right\} - \mathbf{E}|m(X)-Y|^2 - T_{2,n}$$

and

$$T_{2,n} = (1+\delta)\frac{1}{n_t}\sum_{i=n_l+1}^{n_l+n_t}\left(|m_{n_l}^{(H)}(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\right).$$

Because of (7.1),

$$T_{2,n} \le (1+\delta)\frac{1}{n_t}\sum_{i=n_l+1}^{n_l+n_t}\left(|m_{n_l}^{(\hat{h})}(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\}\right),$$

hence,

$$\begin{aligned}
\mathbf{E}\{T_{2,n}|D_{n_l}\} \;&\le\; (1+\delta)\left(\mathbf{E}\left\{|m_{n_l}^{(\hat{h})}(X)-Y|^2\Big|D_{n_l}\right\} - \mathbf{E}|m(X)-Y|^2\right) \\
&=\; (1+\delta)\int |m_{n_l}^{(\hat{h})}(x)-m(x)|^2\mu(dx).
\end{aligned}$$

In the sequel we will show

$$\mathbf{E}\left\{T_{1,n}|D_{n_l}\right\} \le c\frac{(1+\log|\mathcal{Q}_n|)}{n_t}, \tag{7.6}$$

which, together with the inequality above, implies the assertion. Let $s>0$ be arbitrary. Then

$$\begin{aligned}
& \mathbf{P}\{T_{1,n}\ge s|D_{n_l}\} \\
= \;& \mathbf{P}\Bigg\{(1+\delta)\Bigg(\mathbf{E}\{|m_{n_l}^{(H)}(X)-Y|^2|D_{n_l}\} - \mathbf{E}|m(X)-Y|^2 \\
& \qquad -\frac{1}{n_t}\sum_{i=n_l+1}^{n}\{|m_{n_l}^{(H)}(X_i)-Y_i|^2 - |m(X_i)-Y_i|^2\}\Bigg)
\end{aligned}$$

$$\geq s + \delta \left(\mathbf{E}\{|m_{n_l}^{(H)}(X) - Y|^2 | D_{n_l}\} - \mathbf{E}|m(X) - Y|^2 \right) \Big| D_{n_l} \Bigg\}$$

$$\begin{aligned} \leq \quad & \mathbf{P}\Bigg\{ \exists h \in \mathcal{Q}_n : \mathbf{E}\{|m_{n_l}^{(h)}(X) - Y|^2 | D_{n_l}\} - \mathbf{E}|m(X) - Y|^2 \\ & \qquad - \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{ |m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right\} \\ & \geq \frac{1}{1+\delta} \left(s + \delta \mathbf{E} \left\{ |m_{n_l}^{(h)}(X) - Y|^2 - |m(X) - Y|^2 \Big| D_{n_l} \right\} \right) \Big| D_{n_l} \Bigg\} \\ \leq \quad & |\mathcal{Q}_n| \max_{h \in \mathcal{Q}_n} \mathbf{P}\Bigg\{ \Bigg\{ \mathbf{E} \left\{ |m_{n_l}^{(h)}(X) - Y|^2 | D_{n_l} \right\} - \mathbf{E}|m(X) - Y|^2 \\ & \qquad - \frac{1}{n_t} \sum_{i=n_l+1}^{n} \left\{ |m_{n_l}^{(h)}(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right\} \Bigg\} \\ & \geq \frac{1}{1+\delta} \left(s + \delta \mathbf{E} \left\{ |m_{n_l}^{(h)}(X) - Y|^2 - |m(X) - Y|^2 \Big| D_{n_l} \right\} \right) \Big| D_{n_l} \Bigg\}. \end{aligned}$$

Fix $h \in \mathcal{Q}_n$. Set

$$Z = |m_{n_l}^{(h)}(X) - Y|^2 - |m(X) - Y|^2$$

and

$$Z_i = |m_{n_l}^{(h)}(X_{n_l+i}) - Y_{n_l+i}|^2 - |m(X_{n_l+i}) - Y_{n_l+i}|^2 \quad (i = 1, \ldots, n - n_l).$$

Using Bernstein's inequality (see Lemma A.2) and

$$\begin{aligned} \sigma^2 \quad := \quad & \mathbf{Var}\{Z | D_{n_l}\} \\ \leq \quad & \mathbf{E}\{Z^2 | D_{n_l}\} \\ = \quad & \mathbf{E} \Bigg\{ \left| (m_{n_l}^{(h)}(X) - Y) - (m(X) - Y) \right|^2 \\ & \qquad \times \left| (m_{n_l}^{(h)}(X) - Y) + (m(X) - Y) \right|^2 \Big| D_{n_l} \Bigg\} \\ \leq \quad & 16L^2 \int |m_{n_l}^{(h)}(x) - m(x)|^2 \mu(dx) \\ = \quad & 16L^2 \mathbf{E}\{Z | D_{n_l}\} \end{aligned} \tag{7.7}$$

we get

$$
\begin{aligned}
&\mathbf{P}\Bigg\{\Bigg\{\mathbf{E}\{|m_{n_l}^{(h)}(X)-Y|^2|D_{n_l}\}-\mathbf{E}|m(X)-Y|^2 \\
&\qquad\qquad -\frac{1}{n_t}\sum_{i=n_l+1}^{n}\{|m_{n_l}^{(h)}(X_i)-Y_i|^2-|m(X_i)-Y_i|^2\}\Bigg\} \\
&\quad\geq \frac{1}{1+\delta}\left(s+\delta\mathbf{E}\left\{|m_{n_l}^{(h)}(X)-Y|^2-|m(X)-Y|^2\middle|D_{n_l}\right\}\right)\middle|D_{n_l}\Bigg\} \\
&= \mathbf{P}\left\{\mathbf{E}\{Z|D_{n_l}\}-\frac{1}{n_t}\sum_{i=1}^{n_t}Z_i\geq\frac{1}{1+\delta}\left(s+\delta\cdot\mathbf{E}\{Z|D_{n_l}\}\right)\middle|D_{n_l}\right\} \\
&\leq \mathbf{P}\left\{\mathbf{E}\{Z|D_{n_l}\}-\frac{1}{n_t}\sum_{i=1}^{n_t}Z_i\geq\frac{1}{1+\delta}\left(s+\delta\cdot\frac{\sigma^2}{16L^2}\right)\middle|D_{n_l}\right\} \\
&\leq \exp\left(-n_t\frac{\frac{1}{(1+\delta)^2}\left(s+\delta\frac{\sigma^2}{16L^2}\right)^2}{2\sigma^2+\frac{2}{3}\frac{8L^2}{1+\delta}\left(s+\delta\frac{\sigma^2}{16L^2}\right)}\right).
\end{aligned}
$$

Here we don't need the factor 2 before the exponential term because we don't have absolute value inside the probability (cf. proof of Lemma A.2). Next we observe

$$
\begin{aligned}
&\frac{\frac{1}{(1+\delta)^2}\left(s+\delta\frac{\sigma^2}{16L^2}\right)^2}{2\sigma^2+\frac{2}{3}\frac{8L^2}{1+\delta}\left(s+\delta\frac{\sigma^2}{16L^2}\right)} \\
\geq\ &\frac{s^2+2s\delta\frac{\sigma^2}{16L^2}}{\frac{16}{3}L^2(1+\delta)s+\sigma^2\left(2(1+\delta)^2+\frac{1}{3}\delta(1+\delta)\right)}.
\end{aligned}
$$

An easy but tedious computation (cf. Problem 7.1) shows

$$
\frac{s^2+2s\delta\frac{\sigma^2}{16L^2}}{\frac{16}{3}L^2(1+\delta)s+\sigma^2\left(2(1+\delta)^2+\frac{1}{3}\delta(1+\delta)\right)}\geq\frac{s}{c}, \tag{7.8}
$$

where $c=L^2(16/\delta+35+19\delta)$. Using this we get that

$$
\mathbf{P}\left\{T_{1,n}\geq s|D_{n_l}\right\}\leq|\mathcal{Q}_n|\exp\left(-n_t\frac{s}{c}\right).
$$

It follows, for arbitrary $u>0$,

$$
\begin{aligned}
\mathbf{E}\left\{T_{1,n}|D_{n_l}\right\} &\leq u+\int_u^\infty\mathbf{P}\left\{T_{1,n}>s|D_{n_l}\right\}ds \\
&\leq u+\frac{|\mathcal{Q}_n|\,c}{n_t}\exp\left(-\frac{n_t\,u}{c}\right).
\end{aligned}
$$

Setting $u=\frac{c\log(|\mathcal{Q}_n|)}{n_t}$, this implies (7.6), which in turn implies the assertion. □