# Chapter 6

# Adaptive estimation

## 6.1 Adaptation and data-driven tuning

So far we have studied various high- or infinite-dimensional statistical problems and methods in some detail. For various problems we have derived minimax lower bounds for the risk that estimators can achieve uniformly over certain classes of underlying truths. Moreover, we have seen that different estimators that we have considered in Chapter 3 achieve these optimal rates.

An important observation is that for all methods that we have seen so far it holds that to achieve an optimal convergence rate uniformly over a certain class of truths, it is required to choose a specific setting of a tuning parameter that depends on some feature of the class. This feature however is an aspect of the unknown truth, and it is typically not very realistic to assume that we actually have access to it. The kernel estimators or projection estimator for instance can achieve optimal rates over Hölder or Sobolev balls, but the rate-optimal estimators use the knowledge of the degree of smoothness $\beta$ of the unknown function to set the appropriate bandwidth or point of truncation, see Examples 3.4.2, 3.4.3 and 3.4.4. Similarly, the rate-optimal estimators that we considered for the estimation of sparse vectors in the normal means model use the knowledge of the degree of sparsity $s$ of the unknown signal, cf. Example 3.4.5 and Exercise 3.18. These examples illustrate that the rate-optimality results we have proved so far should really be seen as oracle results: there exists rate-optimal estimators for various problems, but they can not be practically implemented unless you have access to an oracle that has information about certain properties of the unknown truth.

The question that we address in this chapter is whether it is possible to construct estimators that do *not* use knowledge of aspects of the truth such as its degree of smoothness or sparsity, but that *still* achieve optimal convergence rates over smoothness or sparsity classes, for instance. If an estimator has this property, it is said to be *adaptive* with respect to the specific feature under consideration. Adaptive methods somehow manage, either explicitly or implicitly, to automatically choose the correct value of their tuning parameters

in a purely data-driven way.

## 6.2 Lepski's method

The first method for adaptation that we consider tries to select the best estimator among a scale of non-adaptive estimators by trading off bias and variance in a purely data-driven manner. Of course this is not entirely straightforward, since we do not have direct access to these quantities in general. Still, it turns out that by considering certain cleverly chosen statistics, we can get sufficient information to make an optimal choice. For concreteness, we study the method in the context of the simple projection estimator of Example 3.2.1. However, as we will repeat at the end of this section, the method is much more general in nature and can be adapted to work in many more situations.

So we assume we are in the setting of the sequence formulation (2.4) of the white noise model. For every $k \in \mathbb{N}$ we can then consider the projection estimator $\hat{f}_k = (Y_1, Y_2, \ldots, Y_k, 0, 0, \ldots)$. Lepski's method for choosing the appropriate truncation point $k$ is based on an analysis of the statistics $\|\hat{f}_k - \hat{f}_l\|_2^2$ for different $k$ and $l$. For $l > k$ we can write

$$\begin{aligned} \|\hat{f}_k - \hat{f}_l\|_2^2 &= \|\hat{f}_k - \mathrm{E}_f \hat{f}_k - (\hat{f}_l - \mathrm{E}_f \hat{f}_l) + \mathrm{E}_f \hat{f}_l - \mathrm{E}_f \hat{f}_k\|_2^2 \\ &= \frac{1}{n} \sum_{k<j\le l} Z_j^2 + \sum_{k<j\le l} f_j^2 - 2\frac{1}{\sqrt{n}} \sum_{k<j\le l} f_j Z_j. \end{aligned} \tag{6.1}$$

Note that the first term in the last line of the display is of stochastic order $l/n$, the second term is a lower bound for the squared bias of $\hat{f}_k$ (see Example 3.2.1). If $f \in H_R^\beta[0,1]$ for some $\beta, R > 0$, then the third term is of lower order than the first two (see the proof of Lemma 6.2.1 ahead). Hence if $\|\hat{f}_k - \hat{f}_l\|_2^2 \gg l/n$ for some $l \ge k$, then the squared bias of $\hat{f}_k$ must be of larger order than $l/n \ge k/n$, hence the bias is too large relative to the variance. If, on the other hand, $\|\hat{f}_k - \hat{f}_l\|_2^2 \ll l/n$ for all $l \ge k$, then this holds in particular for $l$ a very large multiple of $k$, and hence the squared bias of $\hat{f}_k$, which is approximately the second term in that case, is of smaller order than $k/n$. So in that case the bias is too small. This reasoning indicates that the 'right' choice of $k$, which balances squared bias and variance, is the choice that ensures that $\|\hat{f}_k - \hat{f}_l\|_2^2$ is of the order $l/n$ for $l \ge k$.

These considerations motivate the following procedure. For $\tau > 0$, define

$$\hat{k} = \min\Big\{k \le n : \|\hat{f}_k - \hat{f}_l\|_2^2 \le \frac{\tau l}{n} \text{ for all } k \le l \le n\Big\}.$$

We use this data-driven truncation level and define the estimator for $f$ by $\hat{f} = \hat{f}_{\hat{k}} = (Y_1, Y_2, \ldots, Y_{\hat{k}}, 0, 0, \ldots)$. A crucial step in the proof of the fact that the estimator $\hat{f}$ is rate-optimal over Sobolev balls is the following lemma, which asserts that with probability tendinding to 1, the truncation level $\hat{k}$ is upper bounded by the oracle choice (see Example 3.4.2).

**Lemma 6.2.1.** *Suppose that $\beta, R > 0$ and $\tau \geq 4$. There exist universal constants $c_1, c_2 > 0$ such that*

$$\sup_{f \in H^\beta_R[0,1]} \mathrm{P}_f\Big(\hat{k} > (nR^2)^{1/(1+2\beta)}\Big) \leq c_1 e^{-c_2 \tau (nR^2)^{1/(1+2\beta)}}.$$

*Proof.* Set $k = (nR^2)^{1/(1+2\beta)}$. By definition of $\hat{k}$ and a union bound,

$$\mathrm{P}_f(\hat{k} > k) \leq \sum_{l \geq k} \mathrm{P}_f\Big(\|\hat{f}_k - \hat{f}_l\|_2^2 > \frac{\tau l}{n}\Big).$$

For $f \in H^\beta_R[0,1]$ we have the bound

$$\sum_{k<j\leq l} f_j^2 \leq R^2 k^{-2\beta} \leq l/n$$

for all $l \geq k$. In view of the expansion (6.1) it follows that

$$\mathrm{P}_f\Big(\|\hat{f}_k - \hat{f}_l\|_2^2 > \frac{\tau l}{n}\Big) \leq \mathrm{P}\Big(\frac{1}{n}\sum_{k<j\leq l} Z_j^2 > \frac{(\tau-1)l}{2n}\Big) + \mathrm{P}\Big(2\frac{1}{\sqrt{n}}\sum_{k<j\leq l} f_j Z_j > \frac{(\tau-1)l}{2n}\Big),$$

where the $Z_j$ are independent standard normal variabels. The variance of the random variable in the second probability on the right is

$$\frac{4}{n}\sum_{k<j\leq l} f_j^2 \leq \frac{4l}{n^2}.$$

Hence, the probability is bounded by

$$1 - \Phi\Big(\frac{(\tau-1)\sqrt{l}}{4}\Big) \leq e^{-\frac{(\tau-1)^2 l}{32}}$$

By Markov's inequality the first probability is bounded by

$$e^{-\frac{(\tau-1)l}{8}}\Big(\mathrm{E}e^{\frac{1}{4}Z_1^2}\Big)^l = e^{-(\frac{(\tau-1)}{8} - \log\sqrt{2})l}$$

(check!). For $\tau \geq 4$ it holds that $(\tau-1)/8 - \log\sqrt{2} > c_0\tau$ for some $c_0 > 0$. We conclude that there exists a constant $c_2 > 0$ such that for all $\tau \geq 6$,

$$\mathrm{P}_f(\hat{k} > k) \leq 2\sum_{l>k} e^{-c_2\tau l} = \frac{2e^{-c_2\tau}}{1-e^{-c_2\tau}} e^{-c_2\tau k}.$$

Since $x \mapsto x/(1-x)$ is increasing on $(0,1)$, taking

$$c_1 = \frac{2e^{-2c_2}}{1-e^{-2c_2}}$$

completes the proof. □

So with probability tending to 1 the data-driven truncation point $\hat{k}$ is not too large, i.e. we are not undersmoothing. The same can not be proved for the the potential problem of oversmoothing, but it turns out that we can have adaptive estimation nonetheless.

**Theorem 6.2.2.** *Suppose that $\beta, R > 0$ and $\tau \geq 4$. Then*

$$\sup_{f \in H^\beta_R[0,1]} \mathrm{E}_f \|\hat{f} - f\|_2 \lesssim n^{-\beta/(1+2\beta)}.$$

*Proof.* Again let $k = (nR^2)^{1/(1+2\beta)}$. By Cauchy-Schwarz,

$$\mathrm{E}_f \|\hat{f} - f\|_2 1_{\hat{k}>k} = \sum_{k<j\leq n} \mathrm{E}_f \|\hat{f}_j - f\|_2 1_{\hat{k}=j} \leq \sum_{k<j\leq n} \sqrt{\mathrm{E}_f \|\hat{f}_j - f\|_2^2} \sqrt{\mathrm{P}_f(\hat{k} = j)}.$$

For the risk of the projection estimator $\hat{f}_j$ we have, for $k < j \leq n$ and $f \in H^\beta_R[0,1]$,

$$\mathrm{E}_f \|\hat{f}_j - f\|_2^2 = \sum_{i>j} f_i^2 + \frac{j}{n} \leq R^2 j^{-2\beta} + 1 \leq R^2 + 1.$$

Hence, by the preceding lemma,

$$\sup_{f \in H^\beta_R[0,1]} \mathrm{E}_f \|\hat{f} - f\|_2 1_{\hat{k}>k} \leq n\sqrt{1+R^2} \sqrt{\mathrm{P}_f(\hat{k} > k)} = o\Big(n^{-\beta/(1+2\beta)}\Big).$$

Next, we note that

$$\mathrm{E}_f \|\hat{f} - f\|_2 1_{\hat{k}\leq k} \leq \mathrm{E}_f \|\hat{f} - \hat{f}_k\|_2 1_{\hat{k}\leq k} + \mathrm{E}_f \|\hat{f}_k - f\|_2 1_{\hat{k}\leq k}.$$

The second term is bounded by $(\mathrm{E}_f \|\hat{f}_k - f\|_2^2)^{1/2}$, which is bounded by a constant times $n^{-\beta/(1+2\beta)}$, uniformly for $f \in H^\beta_R[0,1]$ (see (3.14)). For the first term we note that by definition of $\hat{k}$,

$$\mathrm{E}_f \|\hat{f} - \hat{f}_k\|_2 1_{\hat{k}\leq k} \leq \sqrt{\frac{\tau k}{n}} \leq n^{-\beta/(1+2\beta)}.$$

This completes the proof. □

So indeed, we see that the estimator $\hat{f}$ attains the optimal rate $n^{-\beta/(1+2\beta)}$ uniformly over Sobolev balls with regularity $\beta$. Moreover, the estimator does not use knowledge of the parameter $\beta$. Hence, the procedure automatically adapts to the smoothness of the unkown signal!

For concreteness we have worked out the performance of Lepski's method in the context of the white noise model, using the projection estimators as class of non-adaptive estimators that serve as input for the method and obtaining

optimal rates relative to the $L^2$-norm. All of this can be generalized. The proofs show that the essence of the argument goes through as soon as we have a scale of estimators with a one-dimensional tuning parameter such that the bias and the variance in the relevant norm are tractable and depend in a monotone way on the tuning parameter.