$$\begin{aligned}
\int \Big(p^{(\ell)}(x+t) - p^{(\ell)}(x)\Big)^2 dx & \qquad (1.21)\\
&= \int \Big(t \int_0^1 p^{(\ell+1)}(x+\theta t) d\theta\Big)^2 dx \\
&\le t^2 \Big(\int_0^1 \Big[\int \Big(p^{(\ell+1)}(x+\theta t)\Big)^2 dx \Big]^{1/2} d\theta \Big)^2 \\
&= t^2 \int (p^{(\beta)}(x))^2 dx
\end{aligned}$$

in view of the generalized Minkowski inequality. ∎

1.2.4 Lack of asymptotic optimality for fixed density

How to choose the kernel K and the bandwidth h for the kernel density estimators in an optimal way? An old and still popular approach is based on minimization in K and h of the asymptotic MISE *for fixed density* p. However, this does not lead to a consistent concept of optimality, as we are going to explain now. Other methods for choosing h are discussed in Section 1.4.

The following result on asymptotics for fixed p or its versions are often considered.

Proposition 1.6 *Assume that:*

(i) the function K is a kernel of order 1 satisfying the conditions

$$\int K^2(u)du < \infty, \qquad \int u^2 |K(u)| du < \infty, \qquad S_K \triangleq \int u^2 K(u) du \neq 0;$$

(ii) the density p is differentiable on $\mathbf{R}$, the first derivative p' is absolutely continuous on $\mathbf{R}$ and the second derivative satisfies

$$\int (p''(x))^2 dx < \infty.$$

Then for all $n \geq 1$ the mean integrated squared error of the kernel estimator $\hat{p}_n$ satisfies

$$\begin{aligned}
\text{MISE} &\equiv \mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx \\
&= \left[\frac{1}{nh} \int K^2(u)du + \frac{h^4}{4} S_K^2 \int (p''(x))^2 dx \right] (1 + o(1)), \qquad (1.22)
\end{aligned}$$

where the term $o(1)$ is independent of n (but depends on p) and tends to 0 as $h \to 0$.

A proof of this proposition is given in the Appendix (Proposition A.1).

The main term of the MISE in (1.22) is

$$\frac{1}{nh}\int K^2(u)du + \frac{h^4}{4}\ S_K^2 \int (p''(x))^2 dx. \tag{1.23}$$

Note that if K is a nonnegative kernel, expression (1.23) coincides with the nonasymptotic upper bound for the MISE which holds for all n and h (cf. Theorem 1.3 with $\beta = 2$).

The approach to optimality that we are going to criticize here starts from the expression (1.23). This expression is then minimized in h and in nonnegative kernels K, which yields the "optimal" bandwidth for given K:

$$h^{MISE}(K) = \left(\frac{\int K^2}{nS_K^2 \int (p'')^2}\right)^{1/5} \tag{1.24}$$

and the "optimal" nonnegative kernel:

$$K^*(u) = \frac{3}{4}(1-u^2)_+ \tag{1.25}$$

(the Epanechnikov kernel; cf. bibliographic notes in Section 1.11). In particular,

$$h^{MISE}(K^*) = \left(\frac{15}{n\int (p'')^2}\right)^{1/5}. \tag{1.26}$$

Note that the choices of h as in (1.24), (1.26) are not feasible since they depend on the second derivative of the unknown density p. Thus, the basic formula (1.2) with kernel $K = K^*$ and bandwidth $h = h^{MISE}(K^*)$ as in (1.26) does not define a valid estimator, but rather a random variable that can be qualified as a pseudo-estimator or *oracle* (for a more detailed discussion of oracles see Section 1.8 below). Denote this random variable by $p_n^E(x)$ and call it the *Epanechnikov oracle*. Proposition 1.6 implies that

$$\lim_{n\to\infty} n^{4/5}\mathbf{E}_p \int (p_n^E(x) - p(x))^2 dx = \frac{3^{4/5}}{5^{1/5}4}\left(\int (p''(x))^2 dx\right)^{1/5}. \tag{1.27}$$

This argument is often exhibited as a benchmark for the optimal choice of kernel K and bandwidth h, whereas (1.27) is claimed to be the best achievable MISE. The Epanechnikov oracle is declared optimal and its feasible analogs (for which the integral $\int (p'')^2$ in (1.26) is estimated from the data) are put forward. We now explain why such an approach to optimality is misleading. The following proposition is sufficiently eloquent.

Proposition 1.7 *Let assumption (ii) of Proposition 1.6 be satisfied and let K be a kernel of order 2 (thus, $S_K = 0$), such that*

$$\int K^2(u)du < \infty.$$

Then for any $\varepsilon > 0$ the kernel estimator $\hat{p}_n$ with bandwidth

$$h = n^{-1/5}\varepsilon^{-1}\int K^2(u)du$$

satisfies

$$\limsup_{n\to\infty} n^{4/5}\mathbf{E}_p \int (\hat{p}_n(x) - p(x))^2 dx \le \varepsilon. \tag{1.28}$$

The same is true for the positive part estimator $\hat{p}_n^+ = \max(0, \hat{p}_n)$:

$$\limsup_{n\to\infty} n^{4/5}\mathbf{E}_p \int (\hat{p}_n^+(x) - p(x))^2 dx \le \varepsilon. \tag{1.29}$$

A proof of this proposition is given in the Appendix (Proposition A.2).

We see that for all $\varepsilon > 0$ small enough the estimators $\hat{p}_n$ and $\hat{p}_n^+$ of Proposition 1.7 have smaller asymptotic MISE than the Epanechnikov oracle, *under the same assumptions on* p. Note that $\hat{p}_n$, $\hat{p}_n^+$ are true estimators, not oracles. So, if the performance of estimators is measured by their asymptotic MISE *for fixed* p there is a multitude of estimators that are strictly better than the Epanechnikov oracle. Furthermore, Proposition 1.7 implies:

$$\inf_{T_n} \limsup_{n\to\infty} n^{4/5}\mathbf{E}_p \int (T_n(x) - p(x))^2 dx = 0, \tag{1.30}$$

where $\inf_{T_n}$ is the infimum over all the kernel estimators or over all the positive part kernel estimators.

The positive part estimator $\hat{p}_n^+$ is included in Proposition 1.7 on purpose. In fact, it is often argued that one should use nonnegative kernels because the density itself is nonnegative. This would support the "optimality" of the Epanechnikov kernel because it is obtained from minimization of the asymptotic MISE over nonnegative kernels. Note, however, that non-negativity of density estimators is not necessarily achieved via non-negativity of kernels. Proposition 1.7 presents an estimator $\hat{p}_n^+$ which is nonnegative, asymptotically equivalent to the kernel estimator $\hat{p}_n$, and has smaller asymptotic MISE than the Epanechnikov oracle.

Proposition 1.7 plays the role of counterexample. The estimators $\hat{p}_n$ and $\hat{p}_n^+$ of Proposition 1.7 are by no means advocated as being good. They can be rather counterintuitive. Indeed, their bandwidth h contains an arbitrarily large constant factor ε^{-1}. This factor serves to diminish the variance term, whereas, for fixed density p, the condition $\int u^2 K(u)du = 0$ eliminates the main bias term if n is large enough, that is, if $n \ge n_0$, starting from some n_0 that depends on p. This elimination of the bias is possible for fixed p but not uniformly over p in the Sobolev class of smoothness $\beta = 2$. The message of

Proposition 1.7 is that even such counterintuitive estimators outperform the Epanechnikov oracle as soon as the asymptotics of the MISE *for fixed* p is taken as a criterion.

To summarize, the approach based on fixed p asymptotics does not lead to a consistent concept of optimality. In particular, saying that "the choice of h and K as in (1.24) – (1.26) is optimal" does not make much sense.

This explains why, instead of studying the asymptotics for fixed density p, in this book we focus on the *uniform* bounds on the risk over classes of densities (Hölder, Sobolev, Nikol'ski classes). We compare the behavior of estimators in a minimax sense on these classes. This leads to a valid concept of optimality (*among all estimators*) that we develop in detail in Chapters 2 and 3.

Remarks.

(1) Sometimes asymptotics of the MSE (risk at a fixed point) *for fixed* p is used to derive "optimal" h and K, leading to expressions similar to (1.24) – (1.26). This is yet another version of the inconsistent approach to optimality. The above critical remarks remain valid when the MISE is replaced by the MSE.

(2) The result of Proposition 1.7 can be enhanced. It can be shown that, under the same assumptions on p as in Propositions 1.6 and 1.7, one can construct an estimator $\tilde{p}_n$ such that

$$\lim_{n\to\infty} n^{4/5}\mathbf{E}_p \int (\tilde{p}_n(x) - p(x))^2 dx = 0 \tag{1.31}$$

(cf. Proposition 3.3 where we prove an analogous fact for the Gaussian sequence model). Furthermore, under mild additional assumptions, for example, if the support of p is bounded, the result of Proposition 1.7 holds for the estimator $p_n^+ / \int p_n^+$, which itself is a probability density.

1.3 Fourier analysis of kernel density estimators

In Section 1.2.3 we studied the MISE of kernel density estimators under classical but restrictive assumptions. Indeed, the results were valid only for densities p whose derivatives of given order satisfy certain conditions. In this section we will show that more general and elegant results can be obtained using Fourier analysis. In particular, we will be able to analyze the MISE of kernel estimators with kernels K that do not belong to $L_1(\mathbf{R})$, such as the *sinc kernel*

$$K(u) = \begin{cases} \frac{\sin u}{\pi u}, & \text{if } u \neq 0, \\ \frac{1}{\pi}, & \text{if } u = 0, \end{cases} \tag{1.32}$$

and will see that this kernel is better than the Epanechnikov kernel, the latter being inadmissible in the sense to be defined below.