Probab. Theory Relat. Fields (2009) 143:569–596
DOI 10.1007/s00440-008-0137-y

An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation

Evarist Giné · Richard Nickl

Received: 13 August 2007 / Revised: 14 December 2007 / Published online: 29 January 2008

Abstract It is shown that the uniform distance between the distribution function $F_n^K(h)$ of the usual kernel density estimator (based on an i.i.d. sample from an absolutely continuous law on $\mathbb{R}$) with bandwidth h and the empirical distribution function F_n satisfies an exponential inequality. This inequality is used to obtain sharp almost sure rates of convergence of $\|F_n^K(h_n) - F_n\|_\infty$ under mild conditions on the range of bandwidths h_n, including the usual MISE-optimal choices. Another application is a Dvoretzky–Kiefer–Wolfowitz-type inequality for $\|F_n^K(h) - F\|_\infty$, where F is the true distribution function. The exponential bound is also applied to show that an adaptive estimator can be constructed that efficiently estimates the true distribution function F in sup-norm loss, and, at the same time, estimates the density of F—if it exists (but without assuming it does)—at the best possible rate of convergence over Hölder-balls, again in sup-norm loss.

Keywords Kernel density estimator · Exponential inequalities · Adaptive estimation · Sup-norm · Plug-in property

Mathematics Subject Classification (2000) Primary: 62G07; Secondary: 60F05

1 Introduction

Let $X_1, \ldots, X_n$ be independent random variables each having law P, and denote by P_n the usual empirical measure induced by the sample. Let F and F_n denote the distribu-

E. Giné (✉) · R. Nickl
Department of Mathematics, University of Connecticut, Storrs, CT 06269-3009, USA
e-mail: gine@math.uconn.edu

R. Nickl
e-mail: nickl@math.uconn.edu

tion functions of P and P_n. In the recent articles [2,17,18,29], it was shown that several nonparametric density estimators (such as maximum likelihood, wavelet or kernel estimators) are within a $\|\cdot\|_{\mathcal{F}}$-ball of probabilistic size $o(1/\sqrt{n})$ around the empirical measure, where $\|\mu\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\int f d\mu|$ is the usual supnorm of a measure μ over some (Donsker) class $\mathcal{F}$. The special case $\mathcal{F} = \{1_{(-\infty,x]} : x \in \mathbb{R}\}$ corresponds to the distribution function $\hat{F}_n$ of the corresponding density estimator, and in this case the quantity $\|\hat{F}_n - F_n\|_\infty$ can be analyzed with somewhat more precision. For example, in the case of the maximum likelihood estimator of a monotone density, a classical result by Kiefer and Wolfowitz [23] is that $\|\hat{F}_n - F_n\|_\infty = O_{a.s.}\left((n/\log n)^{-2/3}\right)$. Similar results were recently proved for other shape-constrained minimum contrast estimators, see [1,8,9].

The first goal of the present article is to give a more precise analysis of the stochastic behavior of $\|F_n^K(h) - F_n\|_\infty$ if $F_n^K(h)$ is the distribution function of the classical Rosenblatt–Parzen kernel density estimator with bandwidth h. In Theorem 1 we shall prove an exponential inequality for the tail probabilities of $\sqrt{n}\|F_n^K(h) - F_n\|_\infty$, under mild and general conditions on P and the bandwidth h. The proof consists in an application of Talagrand's inequality, together with expectation bounds for empirical processes over VC-classes. Although our theorem is confined to this particular class $\mathcal{F}$, the model of the proof extends easily to other Donsker class $\mathcal{F}$. The inequality implies a rate of convergence of $\|F_n^K(h_n) - F_n\|_\infty$ to zero, which—as we also show—cannot be improved upon in general. This is in the spirit of the above mentioned result by [23]. Another direct consequence is a Dvoretzky–Kiefer–Wolfowitz [10] type inequality (up to constants) for the distribution function of the classical kernel density estimator with mean-integrated-squared error (MISE) optimal (and other) bandwidths.

The second goal of this article is the following adaptation result, that uses Theorem 1: Without any assumptions on the underlying data generating process, we show that one can construct a purely data driven estimator which estimates the distribution function F efficiently in sup-norm loss and, at the same time, estimates the density of F (if it exists, but without apriori assuming it does), at the best possible rate of convergence over Hölder balls, again in sup-norm loss. To do this, we slightly modify Lepski's method for adaptive nonparametric density estimation (see [25], and further developments given, e.g., in [4,26,27,33]). The idea basically consists in applying the usual method, but confined to kernel density estimators whose distribution functions are contained in a $\|\cdot\|_\infty$-ball of size smaller than $1/\sqrt{n}$ around F_n—the exact size depending on the bandwidth of the estimator—and then using our exponential bound to control the probability of the event that $\sqrt{n}\|F_n^K(h) - F_n\|_\infty$ is "too large".

We think that this adaptation result is theoretically interesting in several ways, although it has practical drawbacks. First, the estimator we construct is robust with respect to the choice of loss function, as it is optimal in sup-norm loss both for the distribution function *and* the density. In particular this shows that—in a certain asymptotic sense—one can outperform the empirical distribution function as an estimator of F. Of course we do not advocate replacing F_n by our more complicated estimator—see also the last paragraph of the introduction—but we suggest that estimators other than F_n can be thought of even if nothing is known about F.

A second motivation requires a little more background. Density estimators that achieve the minimax rate of convergence in some loss function (we typically have L^p-loss in mind) *and simultaneously* satisfy central limit theorems over Donsker classes $\mathcal{F}$, are interesting in several statistical problems. Bickel and Ritov [2] label this the "plug-in property" of an estimator, which is useful, e.g., in the estimation of non- or semiparametric functionals (see also Sects. 3.3–4 in [29]): Suppose, e.g., $\Phi(\cdot)$ is a nonlinear functional defined on some set of bounded densities. In statistical applications, $\Phi(p)$ is often estimated by the plug-in estimator $\Phi(p_n)$, where p_n is some density estimator for the true density p. Suppose the approximation

$$\Phi(p_n) - \Phi(p) = D\Phi(p)[p_n - p] + O(\|p_n - p\|_\infty^2)$$

holds. The rate of convergence of the remainder term is of order $\|p_n - p\|_\infty^2$, which typically depends on unknown properties of p, and here adaptive p_n is desirable, see also the recent article [5] on higher order efficiency of semiparametric estimators. But one still has to prove asymptotic normality of the linear term $D\Phi(p)[p_n - p]$—which will often have the form of an integral functional—and this easily follows from general CLTs for the process $\sqrt{n}(\int(p_n - p)f)_{f\in\mathcal{F}}$. Theorem 2—within its limited context—provides both the CLT for the linear part and best possible rates for the remainder.

Third, while adaptive estimation of a density on the real line in L^p-loss, $1 \le p < \infty$, has been considered in several articles, e.g., [7,19,21,22], the case $p = \infty$ does not seem to have been treated in the literature, except in the context of the Gaussian white noise model: Using Lepski's method, Tsybakov [33] treated the case $p = \infty$ as well as pointwise density estimation in this Gaussian framework, and using an estimator different from ours. Building on [33], Butucea [4] pioneered the use of Lepski's method in the density model by treating the pointwise case. Our results show that adaptation by Lepski's method applied to regular kernel density estimators also works in the sup-norm case.

We emphasize in advance that Theorem 2 is asymptotic in nature, and that the applicability of the estimator we construct is limited, at least for the following two reasons. First, the constants involved in the "Lepski-type" tests we construct may be too large to give reasonable results for small or even moderate sample sizes. Clearly, the reason for these large constants relies on the fact that we try to implement Lepski's method in a density model rather than in the Gaussian white noise model: in Lemma 1, we use Talagrand's inequality combined with expectation bounds for empirical processes based on the entropy integral, instead of the direct Gaussian tail bounds that one can use in the Gaussian white noise model considered in most of the adaptive literature so far. Similar problems were encountered in [21], where constants even larger than ours were necessary. This points to the need for bounds with more reasonable constants for the moments of empirical processes (than those we obtain in the Appendix). A second limitation of Theorem 2 relates to the fact that higher order kernels are necessary if one wants to adapt to a wider range of smoothness, and the finite-sample performance of any kernel-based estimator might suffer from this fact, see Sect. 6 in [28] and also Remark 5.

2 Basic notation

For an arbitrary (non-empty) set M, $\ell^\infty(M)$ will denote the Banach space of bounded real-valued functions H on M normed by

$$\|H\|_M := \sup_{m\in M} |H(m)|,$$

but we will use the usual symbol $\|f\|_\infty$ to denote $\sup_{x\in\mathbb{R}} |f(x)|$ for $f : \mathbb{R} \to \mathbb{R}$. For Borel-measurable functions $h : \mathbb{R} \to \mathbb{R}$ and Borel measures μ on $\mathbb{R}$, we set $\mu h := \int_{\mathbb{R}} h d\mu$, and we denote by $\mathcal{L}^p(\mathbb{R}, \mu)$ the usual Lebesgue-spaces of Borel-measurable functions from $\mathbb{R}$ to $\mathbb{R}$. If $d\mu(x) = dx$ is Lebesgue measure, we set shorthand $\mathcal{L}^p(\mathbb{R}) := \mathcal{L}^p(\mathbb{R}, \mu)$, and, for $1 \le p < \infty$, we abbreviate the norm by $\|\cdot\|_p$. The convolution $f * g(x)$ of two measurable functions f, g on $\mathbb{R}$ is defined by $\int g(x - y) f(y) dy$ if the integral converges. Similarly, if μ is any finite signed measure and f is a measurable function, set $\mu * f(x) = \int f(x - y) d\mu(y)$ if the integral exists.

Given n independent random variables $X_1, \ldots, X_n$ identically distributed according to some Borel law P on $\mathbb{R}$, we denote by $P_n = n^{-1} \sum_{j=1}^n \delta_{X_j}$ the empirical measure. We assume throughout that the variables X_j are the coordinate projections of $(\mathbb{R}^\mathbb{N}, \mathcal{B}^\mathbb{N}, P^\mathbb{N})$, and we set $\Pr := P^\mathbb{N}$. The empirical process indexed by $\mathcal{F} \subseteq \mathcal{L}^2(\mathbb{R}, P)$ is given by

$$f \longmapsto \sqrt{n}\,(P_n - P)\, f = \frac{1}{\sqrt{n}} \sum_{j=1}^n (f(X_j) - Pf).$$

Convergence in law of random elements in $\ell^\infty(\mathcal{F})$ is defined in the usual way, see, e.g., 5.1.1 in [6], and will be denoted by the symbol $\rightsquigarrow_{\ell^\infty(\mathcal{F})}$. The class $\mathcal{F}$ is said to be *P-Donsker* if $\sqrt{n}\,(P_n - P) \rightsquigarrow_{\ell^\infty(\mathcal{F})} G_P$ where G_P is the (generalized) Brownian bridge process indexed by $\mathcal{F}$ with covariance $EG_P(f)G_P(g) = P[(f - Pf)(g - Pg)]$ and if G_P is sample-bounded and -continuous w.r.t. the covariance metric.

3 An exponential inequality for the distribution function of the kernel estimator

We will consider the usual kernel density estimator: If $X_1, \ldots, X_n$ are i.i.d. on the real line, then

$$p_n^K(h)(x) = P_n * K_h(x) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{x - X_j}{h}\right), \quad x \in \mathbb{R}, \tag{1}$$

where the kernel $K : \mathbb{R} \to \mathbb{R}$ is a symmetric, integrable function that integrates to 1, $K_h(x) := h^{-1}K(x/h)$, and $h := h_n \searrow 0$, $h_n > 0$. The dependence of h on n will be assumed without displaying. We use kernels of order $r > 0$,

$$\int_{\mathbb{R}} y^j K(y)dy = 0 \text{ for } j = 1, \ldots, \{r\}, \quad \text{and} \quad \int_{\mathbb{R}} |y|^r |K(y)|dy < \infty,$$

where $\{r\}$ is the largest integer strictly smaller than r.

Denote now by

$$F_n(x) = \int_{-\infty}^{x} dP_n(u), \quad x \in \mathbb{R},$$

the empirical distribution function and by

$$F_n^K(h)(x) = \int_{-\infty}^{x} p_n^K(h)(u)du, \quad x \in \mathbb{R},$$

the distribution function of the kernel density estimator (1). Also, define for any nonnegative integer s the spaces $\mathsf{C}^s(\mathbb{R})$ of all bounded continuous real-valued functions that are s-times continuously differentiable on $\mathbb{R}$, equipped with the norm

$$\|f\|_{s,\infty} = \sum_{0 \le \alpha \le s} \left\|D^\alpha f\right\|_\infty,$$

with the convention that $D^0 =: id$ and then $\mathsf{C}^0(\mathbb{R}) =: \mathsf{C}(\mathbb{R})$. For noninteger $s > 0$, set

$$\mathsf{C}^s(\mathbb{R}) = \left\{ f \in \mathsf{C}^{[s]}(\mathbb{R}) : \|f\|_{s,\infty} := \sum_{0 \le \alpha \le [s]} \left\|D^\alpha f\right\|_\infty + \sup_{x \ne y} \frac{\left|D^{[s]} f(x) - D^{[s]} f(y)\right|}{|x - y|^{s-[s]}} < \infty \right\}, \tag{2}$$

where $[s]$ denotes the integer part of s. We have the following theorem.

Theorem 1 *Suppose P has a density p_0 with respect to Lebesgue measure. Assume that p_0 is bounded, in which case we set $t = 0$ in what follows, or that $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some $t > 0$. Let $h := h_n \to 0$ as $n \to \infty$ satisfy $h \ge (\log n/n)$ and let K be a kernel of order $t + 1$. Then there exist finite positive constants $L := L(\|p_0\|_\infty, K)$ and $\Lambda_0 := \Lambda_0(\|p_0\|_{t,\infty}, K)$ such that for all $\lambda \ge \Lambda_0 \max(\sqrt{h \log(1/h)}, \sqrt{n}h^{t+1})$ and $n \in \mathbb{N}$,*

$$\Pr\left(\sqrt{n}\|F_n^K(h) - F_n\|_\infty > \lambda\right) \le 2\exp\left\{-L \min\left(h^{-1}\lambda^2, \sqrt{n}\lambda\right)\right\}.$$

Proof Note first that $F_n^K(h)(x) - F_n(x)$ is a random variable for each $x \in \mathbb{R}$ and hence so is $\|F_n^K(h) - F_n\|_\infty$ since, by right continuity, this is in fact a supremum

over a countable set. We will also use this observation when we apply Talagrand's inequality below.

We set $\mathcal{F} = \{1_{(-\infty,x]} : x \in \mathbb{R}\}$ throughout the proof, and note that $P_n 1_{(-\infty,x]} = F_n(x)$ as well as $(P_n * K_h)1_{(-\infty,x]} = F_n^K(x)$. We consider the decomposition:

$$P_n * K_h - P_n = P_n * K_h - P * K_h - P_n + P + P * K_h - P. \tag{3}$$

For the deterministic bias $P * K_h - P$, we have (as in Lemma 4 in [18]), for given $f \in \mathcal{F}$ with $\bar{f}(x) = f(-x)$

$$(P * K_h - P)f = \int_{\mathbb{R}} K(u)[p_0 * \bar{f}(hu) - p_0 * \bar{f}(0)]du.$$

First, if $t = 0$, we have for every $x \in \mathbb{R}$ and $u \geq 0$ that

$$\begin{aligned}\left|p_0 * \bar{f}(hu) - p_0 * \bar{f}(0)\right| &= \left|\int_{\mathbb{R}} \left(1_{(-\infty,x]}(y - hu) - 1_{(-\infty,x]}(y)\right) p_0(y)dy\right| \\ &\leq \left|\int_{\mathbb{R}} 1_{[x,x+hu]}(y)p_0(y)dy\right| \leq \|p_0\|_\infty hu\end{aligned}$$

and likewise for $u < 0$, and hence $\|P * K_h - P\|_{\mathcal{F}} \leq h\|p_0\|_\infty \int_{\mathbb{R}} |K(u)||u|du$. More generally, if $t > 0$, the distribution function F of p_0 is contained in $\mathsf{C}^{t+1}(\mathbb{R})$, so by standard Taylor expansions and since the kernel is of order $t + 1$ it follows that

$$\sup_{x\in\mathbb{R}} |(P * K_h - P)1_{(-\infty,x]}| = \sup_{x\in\mathbb{R}} \left|\int_{\mathbb{R}} K(u)[F(x + uh) - F(x)]du\right| \leq dh^{t+1} \tag{4}$$

for some constant d depending only on $\|p_0\|_{t,\infty}$ and $\int_{\mathbb{R}} |K(u)||u|^{t+1}du$.

For the remaining part of the decomposition, observe that, using the symmetry of the kernel,

$$(P_n * K_h - P * K_h - P_n + P)f = (P_n - P)(K_h * f - f)$$

for $f \in \mathcal{F}$. Consequently,

$$\begin{aligned}&\Pr\left(\sqrt{n}\|F_n^K(h) - F_n\|_\infty > \lambda\right) \\ &\leq \Pr\left(\sqrt{n}\sup_{f\in\mathcal{F}} |(P_n - P)(K_h * f - f)| > \lambda - d\sqrt{n}h^{t+1}\right) \\ &\leq \Pr\left(n\sup_{f\in\mathcal{F}} |(P_n - P)(K_h * f - f)| > \frac{\sqrt{n}\lambda}{2}\right),\end{aligned} \tag{5}$$

by assumption on λ, and we will apply Talagrand's inequality to the class

$$\tilde{\mathcal{F}} = \{K_h * f - f - P(K_h * f - f) : f \in \mathcal{F}\}$$

to bound the last probability, but first we need some other facts:

(a) First, we note that the class of functions $\{K_h * f - f : f \in \mathcal{F}\}$ is uniformly bounded by $2\|K\|_1$, and hence $\tilde{\mathcal{F}}$ has envelope $U = 4\|K\|_1$.

(b) Also,

$$\sup_{f\in\mathcal{F}} \|K_h * f - f\|_{2,P} \le Ch^{1/2} =: \sigma \tag{6}$$

for $C = \|p_0\|_\infty^{1/2} \int_{\mathbb{R}} |u|^{1/2}|K(u)|$, since

$$\begin{aligned} E(f(X+y) - f(X))^2 &\le E|f(X+y) - f(X)| \\ &= \int_{-\infty}^{\infty} 1_{[x-y,x)}(u)\, p_0(u)dx \\ &\le \|p_0\|_\infty y, \end{aligned}$$

if $y > 0$ and similarly if $y < 0$, and hence, using Minkowski's inequality for integrals

$$\begin{aligned} &\left(E\left(\int_{\mathbb{R}} (f(X+y) - f(X))K_h(y)dy \right)^2 \right)^{1/2} \\ &\le \int_{\mathbb{R}} \left(E(f(X+y) - f(X))^2 \right)^{1/2} |K_h(y)|dy \\ &\le \|p_0\|_\infty^{1/2} \int_{\mathbb{R}} |y|^{1/2}|K_h(y)|dy \\ &= \|p_0\|_\infty^{1/2} h^{1/2} \int_{\mathbb{R}} |u|^{1/2}|K(u)|. \end{aligned} \tag{7}$$

(c) Moreover, we will need the expectation bound

$$nE \sup_{f\in\mathcal{F}} |(P_n - P)(K_h * f - f)| \le d'\sqrt{nh\log(1/h)}, \tag{8}$$

for some constant $0 < d' < \infty$ depending only on $\|p_0\|_\infty$ and K, which is proved as follows: For each $h > 0$, the class $\{K_h * 1_{(-\infty,x]} : x \in \mathbb{R}\}$ is just $\{F^K(\frac{x-\cdot}{h}) : x \in \mathbb{R}\}$, where $F^K(t) = \int_{-\infty}^t K(s)ds$, since

$$K_h * 1_{(-\infty,x]}(u) = h^{-1} \int_{-\infty}^{x} K\left(\frac{y-u}{h}\right) dy = \int_{-\infty}^{\frac{x-u}{h}} K(t)dt = F^K\left(\frac{x-u}{h}\right),$$

and F^K is of bounded variation since it is the distribution function of a finite signed measure. Similarly, $\{1_{(-\infty,x]}(t) : x \in \mathbb{R}\} = \{1_{(-\infty,0]}(t-x) : x \in \mathbb{R}\}$, so $\{K_h * 1_{(-\infty,x]} - 1_{(-\infty,x]} : x \in \mathbb{R}\}$ is contained in the set of all translates of the function $F^K(\cdot/h) - 1_{(-\infty,0]}(\cdot)$, which is of bounded variation, and Lemma 3 hence gives an entropy bound for the class $\{K_h * f - f : f \in \mathcal{F}\}$ independent of h. This entropy bound and the bounds from (a) and (b) above now allow to apply Proposition 3, yielding (8) since $h \geq n^{-1}\log n$.

Finally, we apply Talagrand's inequality, see (37), with

$$x = L\min\left(\sqrt{n}\lambda, h^{-1}\lambda^2\right)$$

and σ, U as in (a) and (b), to the expression (5). For this we need the following bounds:

(I) First we have

$$nE\sup_{f\in\mathcal{F}}|(P_n - P)(K_h * f - f)| \leq d'\sqrt{nh\log(1/h)} \leq \frac{\sqrt{n}\lambda}{6}$$

by (8) and the assumption on λ.

(II) We also have, for V as defined in Section 5.1, that

$$V \leq C^2 nh + 8\|K\|_1 d'\sqrt{nh\log(1/h)} \leq C'nh$$

for some constant C' since $h \geq (\log n/n)$ and hence

$$\sqrt{2Vx} \leq \sqrt{2C'nhL\min\left(\sqrt{n}\lambda, h^{-1}\lambda^2\right)} \leq \sqrt{2C'L}\sqrt{n}\lambda \leq \frac{\sqrt{n}\lambda}{6}$$

for L small enough.

(III) Furthermore,

$$Ux/3 \leq (4/3)\|K\|_1 L\min\left(\sqrt{n}\lambda, h^{-1}\lambda^2\right) \leq \frac{\sqrt{n}\lambda}{6}.$$

Summarizing, the sum of the terms in I, II, III is smaller than $(\sqrt{n}\lambda/2)$ if L is chosen suitably small, and we obtain from (37) for the given choice of x that

$$\Pr\left(n\sup_{f\in\mathcal{F}}|(P_n - P)(K_h * f - f)| > \frac{\sqrt{n}\lambda}{2}\right) \leq 2\exp\{-x\},$$

which implies the theorem. □

We now discuss several probabilistic and statistical consequences of Theorem 1.

1. A DKW-type inequality. Theorem 1 implies a Dvoretzky-Kiefer-Wolfowitz [10] type exponential bound, up to constants; namely, there exist universal constants c_1, c_2 such that for $\Lambda_0 \max(\sqrt{h\log(1/h)}, \sqrt{n}h^{t+1}) \le \lambda \le \sqrt{n}$ we have, for $p_0 \in \mathsf{C}^t(\mathbb{R})$

$$\Pr(\sqrt{n}\|F_n^K(h) - F\|_\infty > \lambda) \le c_1 \exp\{-c_2\lambda^2\}. \tag{9}$$

The MISE-optimal choice $h^*(t) \simeq n^{-1/2t+1}$ is admissible for every $t > 0$, in which case the window for λ is $\Lambda_0 n^{-1/2(2t+1)}\sqrt{\log n} \le \lambda \le \sqrt{n}$.

2. LIL, invariance principles. Theorem 1 can also be used to transfer invariance principles and other limit theorems for $F_n - F$ to $F_n^K - F$. We only give the law of the iterated logarithm: Let

$$\mathcal{S} = \left\{ x \mapsto \int_{-\infty}^{x} f dP : \int_{\mathbb{R}} f dP = 0, \int_{\mathbb{R}} f^2 dP \le 1 \right\}$$

be the Strassen set. If P, K satisfy the conditions of Theorem 1 for some $t \ge 0$, and if $h_n \to 0$ as $n \to \infty$, $h_n \ge (\log n/n)$ and $\sup_n \sqrt{n}h_n^{t+1} = M < \infty$, then, almost surely, the sequence

$$\left\{ \sqrt{\frac{n}{2\log\log n}}(F_n^K(h_n) - F) \right\}_{n=3}^{\infty}$$

is relatively compact in $\ell^\infty(\mathbb{R})$ and its set of limit points coincides with the Strassen set $\mathcal{S}$.

3. Optimal choices of h. The most interesting bandwidths are

$$h^*(t) :\simeq n^{-1/(2t+1)}, \text{ and } h^{**}(t) \simeq (n/\log n)^{-1/(2t+1)},$$

since in the case of $h^*(t)$, the kernel estimator is optimal in mean integrated squared error, and in the case $h^{**}(t)$, it is optimal in sup-norm loss. Note that for these bandwidths and $p_0 \in \mathsf{C}^t(\mathbb{R})$, Theorem 1 implies

$$\|F_n^K(h^*(t)) - F_n\|_\infty = O_{a.s.}\left(n^{-(t+1)/(2t+1)}\sqrt{\log n}\right) = o_{a.s.}(n^{-1/2}), \tag{10}$$

as well as

$$\|F_n^K(h^{**}(t)) - F_n\|_\infty = O_{a.s.}\left((n/\log n)^{-(t+1)/(2t+1)}\right) = o_{a.s.}(n^{-1/2}). \tag{11}$$

We shall see below that these rates (including the logarithmic power) cannot be improved in general.

Remark 1 (*Comments on shape-constrained estimators.*) The case $t = 1$ in (10) and (11) can be (qualitatively) compared to the classical result of Kiefer and Wolfowitz [23], who considered the distribution function $\hat{F}_n$ of the maximum likelihood estimator of a monotone decreasing density p_0, and proved that, if p_0 is strictly monotone,

bounded from above and below and has a bounded continuous derivative (on its support), then $\|\hat{F}_n - F_n\|_\infty = O_{a.s.}\left((n/\log n)^{-2/3}\right)$. A similar result can be proved in a monotone regression framework, see [9], who also proved that this rate is best possible. Furthermore, Durot and Tocquet obtained a pointwise result similar to (12) (with a non-normal limiting distribution). An "updated" proof of the classical Kiefer–Wolfowitz result can be found in [1], who also considered an analog of it for the distribution function $\tilde{F}_n$ of the least squares estimator of a convex monotone density, and obtained $\|\tilde{F}_n - F_n\|_\infty = O_{a.s.}\left((n/\log n)^{-3/5}\right)$, which again can be compared to the case $t = 2$ in (10) and (11). Balabdaoui and Wellner [1] also conjectured that for k-monotone densities, the rate should be $(n/\log n)^{-(k+1)/(2k+1)}$, and Theorem 1 confirms the analog of this conjecture for the kernel density estimator (with $\sqrt{\log n}$ instead of $(\log n)^{(k+1)/(2k+1)}$ if the bandwidth $h^*(t)$ is considered).

In the remainder of this section, we show that Theorem 1 is optimal in several respects.

4. Rate of convergence of $\|F_n^K(h) - F_n\|_\infty$: lower bounds. First note that a simple application of the CLT (convergence of triangular arrays to the normal law), of (3), (4) and of (16), (20), gives that, if $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some integer $t > 0$ and K is a (e.g., compactly supported) kernel of order $t+1$, for every $x \in \mathbb{R}$,

$$n^{(t+1)/(2t+1)}\left(F_n^K(h^*(t))(x) - F_n(x)\right) \rightsquigarrow_{\mathbb{R}} N(D^t p_0(x)c(K), p_0(x)c'(K)), \quad (12)$$

and

$$(n/\log n)^{(t+1)/(2t+1)}\left(F_n^K(h^{**}(t))(x) - F_n(x)\right) \to D^t p_0(x)c(K) \quad \text{in probability}, \quad (13)$$

where $c(K) = \int_{\mathbb{R}} K(u)u^{t+1}du$, $c'(K) = \int(1_{[0,\infty)}(u) - F^K(u))^2 du$ and $F^K(u) = \int_{-\infty}^u K(v)dv$. Since no differentiable density p_0 on $\mathbb{R}$ has a t-th derivative that vanishes for all x, the second limit gives optimality of (11). (This is clear for $c(K) \neq 0$, and in the somewhat unnatural case $c(K) = 0$, the rate cannot improve uniformly in all densities in a $\|\cdot\|_{t,\infty}$-ball.) Moreover, if we ignore logarithmic terms, the limit (12) gives optimality of the convergence rate in (10), and the following lower bound for more general bandwidths will show, in particular, that the power of the logarithmic term in (10) cannot be improved in general, see Remark 2.

In the proof of the following proposition, we use a lower bound from [15], which is based on a Sudakov-type-inequality for Rademacher processes due to Talagrand (see [24, p. 114]).

Proposition 1 *Let $K(u) = 1_{[-1/2,1/2]}(u)$ and let h_n satisfy $h_n \to 0$ and $nh_n/\log n \to \infty$ as $n \to \infty$. Let p_0 be any density such that $p_0(x) = c > 0$ for all x on an interval of positive length. Then there exist positive constants c_1, c_2 such that*

$$\liminf_n \Pr\left\{\|F_n^K(h_n) - F_n\|_\infty > c_1\sqrt{h_n(\log h_n^{-1})/n}\right\} > c_2 > 0.$$

Proof As the proof will show, we can assume $c = 1$ w.l.o.g. Let $[a, b]$ be an interval where $p_0(x) = 1$ and let $I = [a + 2h, b - 2h]$ for h small enough. Then, by Taylor expansion and symmetry of K,

$$(P * K_h - P)1_{(-\infty,x]} = \int_{-1/2}^{1/2} [F(x + uh) - F(x)]du$$
$$= h^2 \int_{-1/2}^{1/2} u^2 Dp_0(x + uh\zeta)du = 0 \tag{14}$$

for all $x \in I$. Therefore, by (3) above,

$$\|F_n^K(h) - F_n\|_\infty \geq \sup_{x\in I} \left|(P_n - P)(K_h * 1_{(\infty,x]} - 1_{(-\infty,x]})\right|.$$

The main step is to show that there exist $c_3 > 0$ such that for all n large enough,

$$E \sup_{x\in I} \left|(P_n - P)(K_h * 1_{(-\infty,x]} - 1_{(-\infty,x]})\right| \geq c_3\sqrt{h(\log h^{-1})/n}. \tag{15}$$

Set

$$g_x = K_h * 1_{(-\infty,x]} - 1_{(-\infty,x]} \text{ and } \mathcal{F}' = \{g_x : x \in \mathbb{R}\}.$$

A computation along the lines of (7) gives

$$\sup_{x\in\mathbb{R}} |Pg_x| \leq c_4 h, \tag{16}$$

and therefore, by symmetrization (ε_j are the usual i.i.d. Rademacher variables, independent of the X_j's, all coordinates in a large probability space),

$$E \sup_{x\in I} |(P_n - P)g_x| \geq \frac{1}{2} E \sup_{x\in I} \left| n^{-1} \sum_{j=1}^n \varepsilon_j g_x(X_j) \right| - \frac{1}{2} E \left| n^{-1} \sum_{j=1}^n \varepsilon_j \right| \sup_{x\in I} |Pg_x|$$
$$= \frac{1}{2} E \sup_{x\in I} \left| n^{-1} \sum_{j=1}^n \varepsilon_j g_x(X_j) \right| - o\left(\sqrt{h \log h^{-1}/n}\right).$$

Hence inequality (15) will follow if we show

$$E \sup_{x\in I} \left| n^{-1} \sum_{j=1}^n \varepsilon_j g_x(X_j) \right| \geq c_5\sqrt{h(\log h^{-1})/n}. \tag{17}$$

To prove this we will apply Theorem 3.4 in [15], whose proof is actually given for the symmetrized process, and applies to this process even if the class of functions involved is not centered.

From the proof of Theorem 1 we already know that

$$\sup_Q \log N(\mathcal{F}', \mathcal{L}^2(Q), \epsilon) \leq c_6 \log(c_7/\epsilon) \tag{18}$$

for $\epsilon \leq 2 \sup_x \|g_x\|_\infty$, so inequality (17) will follow from Theorem 3.4 in [15] if we show that,

$$\log N(\mathcal{F}'/2, \mathcal{L}^2(P), \sigma/4) \geq c_8 \log(c_9/\sigma), \tag{19}$$

where $2 = \sup_x \|g_x\|_\infty > \sigma \geq \sup_{x\in I} \|g_x\|_{2,P}$. For the uniform kernel, we can sharpen the variance bound (6) to $\sigma = \sqrt{h/12}$, since, recalling that F^K is the cdf of K, we have, for $x \in I$,

$$\begin{aligned}
&E\left(K_h * 1_{(-\infty,x]} - 1_{(-\infty,x]}\right)^2 \\
&= \int \left(\int K(u) 1_{[(y-x)/h,\infty)}(u) du - 1_{(-\infty,x]}(y) \right)^2 p_0(y) dy \\
&= \int \left(1_{(x,\infty)}(y) - F^K((y-x)/h) \right)^2 p_0(y) dy \\
&= h \int \left(1_{[0,\infty)}(v) - F^K(v) \right)^2 p_0(vh + x) dv \\
&= h \int \left(1_{[0,\infty)}(v) - F^K(v) \right)^2 dv = h/12.
\end{aligned} \tag{20}$$

Note next that $g_x(y) = h^{-1}(x + (h/2)) - 1 - (y/h)$ if $x - (h/2) \leq y \leq x$, $g_x(y) = h^{-1}(x + (h/2)) - (y/h)$ if $x < y \leq x + (h/2)$ and $g_x(y)$ is 0 otherwise. Then, a simple computation shows that, for $\delta < h/2$ and $x \in I$, we have

$$\int (g_{x+\delta}(y) - g_x(y))^2 p_0(y) dy = \frac{2\delta^3}{3h^2} + 2\left(\frac{h}{2} - \delta\right)\frac{\delta^2}{h^2} + \delta\left(1 - \frac{\delta}{h}\right)^2$$

which, for $\delta = h/3$, equals $(17/3^4)h > \sigma^2 = h/12$. Since, for h and hence δ small enough, the interval I contains at least $(b-a)/(3\delta)$ points at distance at least δ from each other, it follows that

$$N\left(\mathcal{F}'/2, \mathcal{L}^2(P), \sigma/4\right) = N\left(\mathcal{F}', \mathcal{L}^2(P), \sigma/2\right) \geq (b-a)(3\delta)^{-1} = (b-a)(12\sigma^2)^{-1},$$

which proves (19) and therefore (15), via (17). The bound in probability follows from the lower bound (15), the upper bound

$$\left(E \sup_{x\in I} |(P_n - P)g_x|^2\right)^{1/2} \leq c_{10}\sqrt{h \log h^{-1}/n}$$

—which follows from Corollary 1 in Giné and Mason [16] and (18)—and the usual Paley–Zygmund argument: for any random variable ξ, one has $E|\xi| \le a + E(\xi^2)^{1/2} \times \Pr(|\xi| > a)^{1/2}$, which for $\xi = \sup_{t\in I} |(P_n - P)g_t|$ and $a = (c_3/2)\sqrt{h \log h^{-1}/n}$ gives, by virtue of the given bounds,

$$\Pr(\xi > a) \ge (c_3/(2c_{10}))^2 ,$$

proving the proposition. □

It is not difficult to devise kernels of order larger than one so that a result similar to this proposition holds as well.

This proposition shows that the requirement $\lambda \ge c\sqrt{h \log h^{-1}}$ in Theorem 1 is necessary for (some) densities that are contained in $\mathsf{C}^t(\mathbb{R})$ for any t. Regarding the second requirement in Theorem 1, namely that $\lambda \ge C\sqrt{n}h^{t+1}$ for some $0 < C < \infty$, one can distinguish three situations. If $\sqrt{n}h^{t+1} << \sqrt{h \log h^{-1}}$, this additional conditions is void. If $\sqrt{n}h^{t+1} >> \sqrt{h \log h^{-1}}$, necessity is easily seen from (4) (and (8)). If $\sqrt{n}h^{t+1} \simeq \sqrt{h \log h^{-1}}$, then $h \simeq h^{**}(t)$, and optimality of Theorem 1 follows from (13).

Remark 2 (*Optimality of* (10).) If one requires p_0 to be only uniformly continuous, one can find a nontrivial interval I around a point of maximum, where p_0 is almost constant, and the lower bound for the entropy of $\mathcal{F}'$ in the proof of Proposition 1 still holds. If in addition, p_0 is differentiable, the bias (14) is not necessarily zero, but is bounded by Ch^2 for some $C < \infty$. Hence, for $h \simeq h^*(1)$, the bias can be absorbed by the bound (17), and the proof of Proposition 1 shows that the order (10) is optimal for $t = 1$. Similar arguments imply that the bound (10) is sharp in general ($t > 1$), but we do not pursue this further.

5. The case of discrete P. Another question is whether the assumption that P has a bounded density can be relaxed. Inspection of the proof shows that boundedness of p_0 can be relaxed to $p_0 \in \mathcal{L}^p(\mathbb{R})$ for some $p > 1$, assuming $\sqrt{n}h^{1-1/p} \to 0$. On the other hand, if P has atoms, while F_n is still uniformly close (with large probability) to the discontinuous function F, the continuous function $F_n^K(h)$ is not:

Proposition 2 *Let P be a probability measure such that $P\{x_0\} = a > 0$ for some $x_0 \in \mathbb{R}$. If $K \in \mathcal{L}^1(\mathbb{R})$ and $h_n \to 0$ as $n \to \infty$ then*

$$\lim_n \Pr\left(\|F_n^K(h) - F_n\|_\infty > \frac{a}{3}\right) = 1,$$

in particular, the sequence $\sqrt{n}\|F_n^K(h) - F_n\|_\infty$ is not stochastically bounded.

Proof We adopt the notation from the proof of Theorem 1. By continuity of the measure $P * K_h$, we obviously have

$$\|P * K_h - P\|_{\mathcal{F}} = \|P * K_h(-\infty, \cdot] - P(-\infty, \cdot]\|_\infty \ge a/2.$$

On the other hand, even if P does not have a density, one still has

$$E\left(\int_{\mathbb{R}}(f(X+y)-f(X))K_h(y)dy\right)^2 \leq \|K\|_1^2,$$

which, by the same argument as in Part c) of the proof of Theorem 1, gives $\sqrt{n}E\|(P_n - P)(K_h * f - f)\|_{\mathcal{F}} = O(1)$, and hence by (3),

$$\sqrt{n}\|F_n^K(h) - F_n\|_\infty \geq \sqrt{n}a/2 - O_P(1),$$

which implies the proposition. □

6. The multivariate case. It is of interest to know whether a result similar to Theorem 1 can be proved in higher dimensions, i.e., for the distribution function of the kernel density estimator of a density p_0 in $\mathbb{R}^d$, $d > 1$. The first part of the decomposition (3) can be shown to satisfy $\sup_{f\in\mathcal{F}}|(P_n - P)(K_h * f - f)| = o_P(n^{-1/2})$ for $\mathcal{F} = \{1_{(-\infty,\mathbf{x}]} : \mathbf{x} \in \mathbb{R}^d\}$, see Theorem 2a and Corollary 1a in [18]. On the other hand, the "bias part"

$$\sqrt{n}\sup_{\mathbf{x}\in\mathbb{R}^d}|(P * K_h - P)(1_{(-\infty,\mathbf{x}]})| \tag{21}$$

can be shown not to be smaller than $\sqrt{n}h^{t+1}$ for most common kernels used in multivariate density estimation (and for p_0 t-times differentiable in $\mathbb{R}^d$). In particular, this is too large for the MISE-optimal bandwidths $n^{-1/(2t+d)}$ to be admissible (rather one would need a bias of order $o(\sqrt{n}h^{t+d/2})$). To see that (21) is not smaller than $\sqrt{n}h^{t+1}$, consider, for simplicity, $d = 2$, $t = 0$ and $K(x_1, x_2) = k(x_1)k(x_2)$ a positive kernel of order 2, supported in $(-1/2, 1/2)^2$, and let P be uniform on $(0, 1)$. It is easy to see that then the bias at the point $\mathbf{x} = (0, 1)$ equals $\sqrt{n}(ch+c'h^2)$, where c, c' are fixed positive constants. Hence, in dimension d, one can expect a similar exponential inequality as in Theorem 1 only in (the less interesting) range of h's where $\sqrt{n}h^{t+1} = o(1)$.

4 Adaptive Estimation of the distribution function and its density

Given independent $X_1, \ldots, X_n$ with common law P on the real line, we ask the following question in this section: If nothing at all is known about the probability measure P, can one efficiently estimate the cumulative distribution function $F(x)$ and, *simultaneously*, estimate the density p_0 of P, if it exists, at the minimax rate in sup-norm loss over Hölder balls? We recall that the best rate of convergence uniformly over balls of densities in $\mathsf{C}^t(\mathbb{R})$, is $(n/\log n)^{-t/(2t+1)}$, see [20].

The empirical distribution function $F_n(x)$, which estimates F efficiently, is *not* a candidate since it does not have a derivative. But the kernel density estimator $p_n^K(h)$ from (1) *is* a candidate: In fact, it is well known (e.g., [14]) that $p_n^K(h^{**}(t))$ estimates p_0, if the latter is known to exist and to have derivatives up to order $t > 0$, at the

optimal rate in sup-norm loss. In practice however, t is unknown, so that $h^{**}(t)$ is not available, and we have to find other ways to choose h, which we do in a purely data-driven way: Set

$$\mathcal{H} := \left\{ h_k = \rho^{-k} : k \in \mathbb{N} \cup \{0\}, \rho^{-k} > n^{-1} (\log n)^2 \right\}$$

where $\rho > 1$ is arbitrary. The number of elements in this grid is of order $\log n$, and we denote by $h_{\min}$ the last (i.e., smallest) element in the grid. We will construct a preliminary estimator $\hat{h}_n$ for the bandwidth as follows: First, as long as

$$\| F_n^K(h_{\min}) - F_n \|_\infty > \frac{\sqrt{h_{\min}} \, (\log(1/h_{\min}))^2}{\sqrt{n \log n}} \tag{22}$$

holds, we set $\hat{h}_n = 0$. This step can be interpreted as a test of whether P has a discrete part or not, cf. Proposition 2 and also Remark 3. If

$$\| F_n^K(h_{\min}) - F_n \|_\infty \leq \frac{\sqrt{h_{\min}} \, (\log(1/h_{\min}))^2}{\sqrt{n \log n}},$$

we proceed to use a modification of Lepski's method for adaptive estimation and check whether

$$\| p_n^K(h_{\min}^+) - p_n^K(h_{\min}) \|_\infty \leq \sqrt{\tilde{M} \frac{\log(1/h_{\min})}{n h_{\min}}} \quad AND \quad \| F_n^K(h_{\min}^+) - F_n \|_\infty \leq \frac{\sqrt{h_{\min}^+} \left(\log(1/h_{\min}^+) \right)^2}{\sqrt{n} \log n}$$

simultaneously hold, where $h_{\min}^+$ is the last but one element in the grid $\mathcal{H}$, and where

$$\tilde{M} := \tilde{M}_n := C \| p_n^K(h_{\min}) \|_\infty$$

with $\sqrt{C} := \sqrt{C(K)} = 384 \| K \|_2$. If this does not occur, we set $\hat{h}_n = h_{\min}$, otherwise, we define $\hat{h}_n$ as

$$\hat{h}_n = \max \left\{ h \in \mathcal{H} : \| p_n^K(h) - p_n^K(g) \|_\infty \leq \sqrt{\tilde{M} \frac{\log(1/g)}{ng}} \ \forall g < h, g \in \mathcal{H} \right.$$
$$\left. AND \quad \| F_n^K(h) - F_n \|_\infty \leq \frac{\sqrt{h} \, (\log(1/h))^2}{\sqrt{n} \log n} \right\}. \tag{23}$$

The adaptive estimator of F is now defined as $F_n^K(\hat{h}_n)$ with the convention that $F_n^K(0) = F_n$.

By Proposition 2, if P has a discrete part, then $\hat{h}_n = 0$—and therefore $F_n^K(\hat{h}_n) = F_n$—with probability approaching one. Consequently, if P has a discrete part, our estimator will be discrete with probability approaching one. If P has a bounded and uniformly continuous density, then the next theorem shows that $p_n^K(\hat{h}_n)$ exists and uniformly estimates the derivative p_0 of F with probability approaching 1. We recall that Pr is the product probability on $\mathbb{R}^{\mathbb{N}}$ and, in what follows, we say that a sequence of events A_n is *eventual* if $\lim_m \Pr(\cap_{n\geq m} A_n) = 1$.

Theorem 2 *Let $X_1, ..., X_n$ be i.i.d. on $\mathbb{R}$ with common law P. Let $F_n^K(\hat{h}_n)$ be defined as above, where K is a kernel of order $T+1$, $0 \leq T < \infty$, right-continuous and of bounded variation. Then*

$$\sqrt{n}\left(F_n^K(\hat{h}_n) - F\right) \rightsquigarrow_{\ell^\infty(\mathbb{R})} G_P, \tag{24}$$

the convergence being uniform over the set of all probability measures P on $\mathbb{R}$, in any distance that metrizes convergence in law. If furthermore P possesses a bounded and uniformly continuous density p_0 with respect to Lebesgue measure, then

$$\{\textit{the Lebesgue density } p_n^K(\hat{h}_n) \textit{ of } F_n^K(\hat{h}_n) \textit{ exists}\}$$

is eventual, and, if C is a precompact subset of $\mathsf{C}(\mathbb{R})$, then

$$\sup_{p_0 \in C} \|p_n^K(\hat{h}_n) - p_0\|_\infty = o_P(1). \tag{25}$$

If, in addition, $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some $0 < t \leq T$, then also

$$\sup_{p_0 : \|p_0\|_{t,\infty} \leq D} \|p_n^K(\hat{h}_n) - p_0\|_\infty = O_P\left(\left(\frac{\log n}{n}\right)^{t/(2t+1)}\right). \tag{26}$$

Proof Uniformity in p_0, which follows from control of the constants involved, will be left implicit in the derivations. First we have the following three observations.

(I) The class of functions $\{1_{(-\infty,x]} : x \in \mathbb{R}\}$ is uniform Donsker, and since $\{\sqrt{h}(\log(1/h))^2 : h \in \mathcal{H}\}$ is bounded in absolute value, by the constant D say, we have

$$\|F_n^K(\hat{h}_n) - F_n\|_\infty \leq \frac{D}{\sqrt{n \log n}}$$

by construction of the estimator, which proves (24).

(II) We need to obtain the bias and variance of the kernel density estimator in the sup-norm for given h under the assumptions of the second part of the theorem. For the variance, we use Corollary 3.4 in [14] to obtain

$$E\|p_n^K(h) - Ep_n^K(h)\|_\infty^2 \leq D^2 \frac{\log(1/h)}{nh} := D^2\sigma^2(h,n) \tag{27}$$

for $h \in \mathcal{H}$ and some $0 < D < \infty$ depending only on $\|p_0\|_\infty$ and K.

For the bias, we have the following: If p_0 is in $\mathsf{C}^t(\mathbb{R})$ for some $t > 0$, we have by standard Taylor-series arguments

$$|Ep_n^K(h,x) - p_0(x)| = \left| \int_{\mathbb{R}} K(u)[p_0(x-uh) - p_0(x)]du \right|$$
$$\leq h^t \frac{\|p_0\|_{t,\infty}}{[t]!} \int_{\mathbb{R}} |K(u)||u|^t du := B(h, p_0) \tag{28}$$

since the kernel is of order t. If p_0 is only bounded and uniformly continuous, then one still has

$$\sup_{x\in\mathbb{R}} |Ep_n^K(h,x) - p_0(x)| = \|K_h * p_0 - p_0\|_\infty = o(1), \tag{29}$$

see Theorem 8.14b in Folland [12], and in this case we define $B(h, p_0) := \|K_h * p_0 - p_0\|_\infty$.

(III) We need to control the probability that $\tilde{M} > 1.2C\|p_0\|_\infty$ or $\tilde{M} < 0.8C\|p_0\|_\infty$ if p_0 is bounded and uniformly continuous. For some $0 < L < \infty$ and n large enough we have

$$\begin{aligned}
&\Pr\left(|\tilde{M} - C\|p_0\|_\infty| > 0.2C\|p_0\|_\infty\right) \\
&= \Pr\left(|\|p_n^K(h_{\min})\|_\infty - \|p_0\|_\infty| > 0.2\|p_0\|_\infty\right) \\
&\leq \Pr\left(\|p_n^K(h_{\min}) - p_0\|_\infty > 0.2\|p_0\|_\infty\right) \\
&\leq \Pr\left(\|p_n^K(h_{\min}) - Ep_n^K(h_{\min})\|_\infty > 0.2\|p_0\|_\infty - B(h_{\min}, p_0)\right) \\
&\leq \Pr\left(n\|p_n^K(h_{\min}) - Ep_n^K(h_{\min})\|_\infty > 0.1n\|p_0\|_\infty\right) \\
&\leq L\exp\left\{-\frac{(\log n)^2}{L}\right\}
\end{aligned}$$

by (28), $h_{\min} \simeq (\log n)^2/n$, Lemma 3 and Talagrand's inequality as given in Corollary 2.2 in [14]. [One could also proceed as in Lemma 1.]

Proof of (25) *and* (26)*:* First we observe that if P has a bounded density, then the event $\{\hat{h}_n \geq h_{\min}\} = \{p_n^K(\hat{h}_n) \text{ exists}\}$ is eventual in view of Theorem 1 with $\lambda = \sqrt{h_{\min}}\,(\log(1/h_{\min}))^2\,(\log n)^{-1/2}$ and $h_{\min} \simeq n^{-1}(\log n)^2$ (and the Borel–Cantelli lemma). We will assume that $p_n^K(\hat{h}_n)$ exists throughout the rest of the proof. In particular, expectations in the following derivations are taken over the event $\{\hat{h}_n \geq h_{\min}\}$ so that $p_n^K(\hat{h}_n)$ exists.

Set

$$M = C\|p_0\|_\infty$$

and, if $t > 0$, define $h_p := h(p_0)$ by the balance equation

$$h_p = \max\left\{h \in \mathcal{H} : B(h, p_0) \le \frac{\sqrt{0.8M}}{4}\sigma(h, n)\right\}.$$

Using the results from (II), it is easily verified that

$$h_p \simeq \left(\frac{\log n}{n}\right)^{\frac{1}{2t+1}}$$

if $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some $0 < t \le T$. If p_0 is only bounded and uniformly continuous but not contained in $\mathsf{C}^t(\mathbb{R})$ for any $t > 0$, we set $h_p = h_{\min}$. Then we define $\tilde{\sigma}(h_p, n)$ as $\sigma(h_p, n)$ if $t > 0$ and set

$$\tilde{\sigma}(h_p, n) = \max\left(\sigma(h_p, n), \frac{4}{\sqrt{0.8M}}B(h_p, p_0)\right)$$

otherwise, so that

$$B(h_p, p_0) \le \frac{\sqrt{0.8M}}{4}\tilde{\sigma}(h_p, n)$$

always holds. Clearly $\sigma(h_p, n) = O(\tilde{\sigma}(h_p, n))$ and we note that for $t > 0$

$$\tilde{\sigma}(h_p, n) = \sigma(h_p, n) \simeq \left(\frac{\log n}{n}\right)^{\frac{t}{2t+1}} \tag{30}$$

is the rate of convergence required in (26), but $\tilde{\sigma}(h_p, n) \to 0$ as soon as P has a bounded and uniformly continuous density.

We will consider the cases $\{\hat{h}_n \ge h_p\}$ and $\{\hat{h}_n < h_p\}$ separately. In the first case we also distinguish $\tilde{M} \le 1.2C\|p_0\|_\infty = 1.2M$ and $\tilde{M} > 1.2C\|p_0\|_\infty = 1.2M$. First, by definition of $\hat{h}_n$, (27) and (28) we have

$$\begin{aligned}
&E\left\|p_n^K(\hat{h}_n) - p_0\right\|_\infty I_{\{\hat{h}_n \ge h_p\}\cap\{\tilde{M}\le 1.2M\}} \\
&\le E\Big(\|p_n^K(\hat{h}_n) - p_n^K(h_p)\|_\infty + \|p_n^K(h_p) - Ep_n^K(h_p)\|_\infty \\
&\quad + \|Ep_n^K(h_p) - p_0\|_\infty\Big) I_{\{\hat{h}_n \ge h_p\}\cap\{\tilde{M}\le 1.2M\}} \\
&\le \sqrt{1.2M}\sigma(h_p, n) + D\sigma(h_p, n) + \frac{\sqrt{0.8M}}{4}\tilde{\sigma}(h_p, n) = O(\tilde{\sigma}(h_p, n)).
\end{aligned}$$

Also, for some constant c,

$$
\begin{aligned}
&E\left\|p_n^K(\hat{h}_n)-p_0\right\|_\infty I_{\{\hat{h}_n\geq h_p\}\cap\{\tilde{M}>1.2M\}}\\
&\quad\leq\sum_{h\in\mathcal{H}:h\geq h_p}E\left(\left[\|p_n^K(h)-Ep_n^K(h)\|_\infty+B(h,p_0)\right]I_{\{\hat{h}_n=h\}}I_{\{\tilde{M}>1.2M\}}\right)\\
&\quad\leq c\log n\left[D\sigma(h_p,n)+B(1,p_0)\right]\cdot\sqrt{E1_{\{\tilde{M}>1.2M\}}}\\
&\quad=O\left((\log n)\sqrt{\exp\left\{-\frac{(\log n)^2}{L}\right\}}\right)=o(\tilde{\sigma}(h_p,n))
\end{aligned}
$$

by the results in (II), (III).

We now turn to $\{\hat{h}_n<h_p\}$. If P has a uniformly continuous density not contained in $\mathsf{C}^t(\mathbb{R})$ for any t—in which case we have $h_p=h_{\min}$—we have that $\{\hat{h}_n<h_p\}=\{\hat{h}_n<h_{\min}\}$ has empty intersection with $\{\hat{h}_n\geq h_{\min}\}$. So the last two bounds already prove (25) in this case, and (25) will follow from (26) in the case $t>0$, which we assume for the rest of the proof. [Note that then $\tilde{\sigma}(h_p,n)=\sigma(h_p,n)$]. Now we have, first,

$$
\begin{aligned}
&E\left\|p_n^K(\hat{h}_n)-p_0\right\|_\infty I_{\{\hat{h}_n<h_p\}\cap\{\tilde{M}<0.8M\}}\\
&\quad\leq\sum_{h\in\mathcal{H}:h<h_p}E\left(\left[\|p_n^K(h)-Ep_n^K(h)\|_\infty+B(h,p_0)\right]I_{\{\hat{h}_n=h\}}I_{\{\tilde{M}<0.8M\}}\right)\\
&\quad\leq c'\log n\left[D\sigma(h_{\min},n)+B(h_p,p_0)\right]\cdot\sqrt{EI_{\{\tilde{M}<0.8M\}}}\\
&\quad=O\left(\sqrt{\exp\left\{-\frac{(\log n)^2}{L}\right\}}\right)=o(\sigma(h_p,n)),
\end{aligned}
$$

again by the results in (II), (III); and second,

$$
\begin{aligned}
&E\left\|p_n^K(\hat{h}_n)-p_0\right\|_\infty I_{\{\hat{h}_n<h_p\}\cap\{0.8M\leq\tilde{M}\}}\\
&\quad\leq\sum_{h\in\mathcal{H}:h<h_p}E\left[\left(\|p_n^K(h)-Ep_n^K(h)\|_\infty+\|Ep_n^K(h)-p_0\|_\infty\right)I_{\{\hat{h}_n=h\}\cap\{0.8M\leq\tilde{M}\}}\right]\\
&\quad\leq\sum_{h\in\mathcal{H}:h<h_p}\left(E\left\|p_n^K(h)-Ep_n^K(h)\right\|_\infty^2\right)^{1/2}\left(EI_{\{\hat{h}_n=h\}\cap\{0.8M\leq\tilde{M}\}}\right)^{1/2}+B(h_p,p_0)\\
&\quad\leq\sum_{h\in\mathcal{H}:h<h_p}D\sigma(h,n)\cdot\sqrt{\Pr\left(\{\hat{h}_n=h\}\cap\{0.8M\leq\tilde{M}\}\right)}+O(\sigma(h_p,n)).
\end{aligned}
$$

It remains to show that

$$\sum_{h\in\mathcal{H}:h<h_p} \sigma(h,n)\cdot\sqrt{\Pr(\{\hat{h}_n=h\}\cap\{0.8M\le\tilde{M}\})} = O(\sigma(h_p,n)) \tag{31}$$

is satisfied. Pick any $h\in\mathcal{H}$ so that $h<h_p$, denote by h^+ the previous element in the grid (i.e., $h^+=\rho h$) and observe that

$$\begin{aligned}
&\sqrt{\Pr(\{\hat{h}_n=h\}\cap\{0.8M\le\tilde{M}\})}\\
&\le\left(\sum_{g\in\mathcal{H}:g\le h}\Pr\left(\left\|p_n^K(h^+)-p_n^K(g)\right\|_\infty>\sqrt{0.8M}\sigma(g,n)\right)\right)^{1/2}\\
&\quad+\left(\Pr\left(\sqrt{n}\|F_n^K(h^+)-F_n\|_\infty>\frac{\sqrt{h^+}\left(\log(1/h^+)\right)^2}{\log n}\right)\right)^{1/2}:=A+B
\end{aligned}\tag{32}$$

First we observe that by Theorem 1 with $\lambda=(\sqrt{h^+}\left(\log(1/h^+)\right)^2)/\log n$, the definition of the grid and (27),

$$\begin{aligned}
&\sum_{h\in\mathcal{H}:h<h_p}\sigma(h,n)\cdot B \qquad (33)\\
&\le c(\log n)\,\sigma(h_{\min},n)\sqrt{2\exp\left\{-L\min\left(\frac{(\log(1/h_p))^4}{(\log n)^2},\frac{\sqrt{nh_{\min}}(\log(1/h_p))^2}{\log n}\right)\right\}}\\
&\le c(\log n)\,\sigma(h_{\min},n)\sqrt{2\exp\left\{-L'(\log n)^2\right\}}\\
&=o\left(\sigma(h_p,n)\right) \qquad (34)
\end{aligned}$$

from some n onwards since $h_{\min}\le h^+\le h_p$ implies that $\max(\sqrt{h^+\log(1/h^+)},\sqrt{n}(h^+)^{t+1})=o((\sqrt{h^+}\left(\log(1/h^+)\right)^2)/\log n)$ by definition of $h_{\min},h_p$.

For the term including A we first observe that

$$\begin{aligned}
\left\|p_n^K(h^+)-p_n^K(g)\right\|_\infty\le{}&\left\|p_n^K(h^+)-Ep_n^K(h^+)\right\|_\infty+\left\|p_n^K(g)-Ep_n^K(g)\right\|_\infty\\
&+B(h^+,p_0)+B(g,p_0),
\end{aligned}$$

where

$$B(h^+,p_0)+B(g,p_0)\le 2B(h_p,p_0)\le(1/2)\sqrt{0.8M}\sigma(h_p,n)\le(1/2)\sqrt{0.8M}\sigma(g,n)$$

since $g < h^+ \le h_p$. Consequently,

$$\begin{aligned}
&\Pr\left(\left\|p_n^K(h^+) - p_n^K(g)\right\|_\infty > \sqrt{0.8M}\sigma(g,n)\right)\\
&\quad\le \Pr\left(\left\|p_n^K(h^+) - Ep_n^K(h^+)\right\|_\infty + \left\|p_n^K(g) - Ep_n^K(g)\right\|_\infty > (1/2)\sqrt{0.8M}\sigma(g,n)\right)\\
&\quad\le \Pr\left(\left\|p_n^K(h^+) - Ep_n^K(h^+)\right\|_\infty > (1/4)\sqrt{0.8M}\sigma(h^+,n)\right)\\
&\qquad+ \Pr\left(\left\|p_n^K(g) - Ep_n^K(g)\right\|_\infty > (1/4)\sqrt{0.8M}\sigma(g,n)\right).
\end{aligned}$$

We will now need the following inequality.

Lemma 1 *We have*

$$\Pr\left(\left\|p_n^K(g) - Ep_n^K(g)\right\|_\infty > (1/4)\sqrt{0.8M}\sigma(g,n)\right) \le 2g$$

for every $g \le h_p$, $g \in \mathcal{H}$ and n large enough (depending only on $\|p_0\|_\infty$ and K).

Proof For g fixed set

$$f_t(x) = (2\|K\|_\infty)^{-1}\left(K\left(\frac{t-x}{g}\right) - EK\left(\frac{t-X}{g}\right)\right)$$

so that

$$\left\|p_n^K(g) - Ep_n^K(g)\right\|_\infty = 2\frac{\|K\|_\infty}{gn}\sup_{t\in\mathbb{R}}\left|\sum_{j=1}^n f_t(X_j)\right| = 2\frac{\|K\|_\infty}{gn}\sup_{t\in\mathbb{Q}}\left|\sum_{j=1}^n f_t(X_j)\right|,$$

by right-continuity of K, and define $\mathcal{F} = \{f_t : t \in \mathbb{Q}\}$, which is P-centered and uniformly bounded by 1. Consequently

$$\begin{aligned}
&\Pr\left(\left\|p_n^K(g) - Ep_n^K(g)\right\|_\infty > (1/4)\sqrt{0.8M}\sigma(g,n)\right)\\
&\quad= \Pr\left(\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}} > \frac{\sqrt{0.8M}\sqrt{ng\log(1/g)}}{8\|K\|_\infty}\right).
\end{aligned}$$

To bound the latter probability, we apply Talagrand's inequality with constants as given in [3], see (37). We first compute the variance σ^2 for this class of functions. Clearly

$$\sup_t Ef_t^2(X) \le \frac{\|p_0\|_\infty}{4\|K\|_\infty^2}\|K\|_2^2 g =: \sigma^2.$$

Choosing $x = \log(1/g)$ in (37), the lemma will be proved if we show that

$$E\|\sum_{j=1}^{n} f(X_j)\|_{\mathcal{F}} + \sqrt{2Vx} + x/3 \leq \frac{\sqrt{0.8M}\sqrt{ng\log(1/g)}}{8\|K\|_\infty} \tag{35}$$

where $V = n\sigma^2 + 2E\|\sum_{j=1}^{n} f(X_j)\|_{\mathcal{F}}$. We first need to obtain a good bound (with constants) for the expectation term $E\|\sum_{j=1}^{n} f(X_j)\|_{\mathcal{F}}$. By Proposition 3 (and Lemma 3) we obtain for n large enough, since $(ng/\log(1/g)) \to \infty$ for $g \geq h_{\min}$,

$$E\|\sum_{j=1}^{n} f(X_j)\|_{\mathcal{F}} \leq 42.5\frac{\|p_0\|_\infty^{1/2}\|K\|_2}{\|K\|_\infty}\sqrt{ng\log(1/g)}. \tag{36}$$

This shows, in particular, that $V \leq 1.1n\sigma^2$ for n large enough since $g \geq h_{\min}$. So summarizing, since $g \geq h_{\min}$, we have

$$E\|\sum_{j=1}^{n} f(X_j)\|_{\mathcal{F}} + \sqrt{2V\log(1/g)} + \frac{\log(1/g)}{3} \leq 43\frac{\|p_0\|_\infty^{1/2}\|K\|_2}{\|K\|_\infty}\sqrt{ng\log(1/g)},$$

and (35) holds in view of $M = C\|p_0\|_\infty$ and $\sqrt{C} = 384\|K\|_2 \geq (43\cdot 8\|K\|_2)/\sqrt{0.8}$. □

This lemma and the above give

$$\sum_{g\in\mathcal{H}:g\leq h} \Pr\left(\left\|p_n^K(h^+) - p_n^K(g)\right\|_\infty > (1/4)\sqrt{0.8M}\sigma(g,n)\right)$$
$$\leq \sum_{g\in\mathcal{H}:g\leq h} 2\left[h^+ + g\right] \leq Lh\log n$$

for n large enough and then

$$\sum_{h\in\mathcal{H}:h<h_p} \sigma(h,n)\cdot A = \sqrt{L}\sum_{h\in\mathcal{H}:h<h_p}\sqrt{\frac{\log(1/h)}{nh}}\sqrt{h\log n} = O(n^{-1/2}(\log n)^2)$$
$$= o(\sigma(h_p,n)).$$

Now this, (32) and (34) verify (31), which completes the proof of the theorem. □

Remark 3 (Case when P is not absolutely continuous) Although the first step (22) in the definition of $\hat{h}_n$ can be viewed as a test of absolute continuity of P, we do not advocate its use out of context. [As pointed out by a referee—much simpler tests exist, for example, one can check whether any sample point occurred twice.] As soon as P has a discrete part, we just estimate P by P_n, without even attempting to estimate separately different (discrete and continuous) components of P.

Remark 4 (Modification of Lepski's method) We modified Lepski's method here by adding an additional test at each step. In principle, it would have been enough to apply Lepski's method restricted to density estimators whose distribution functions are in a sup-norm ball of radius $o(1/\sqrt{n})$ around F_n, that is, Theorem 2 can still be proved if we replace $\sqrt{h}(\log(1/h))^2/(\sqrt{n}\log n)$ in (23) (and, mutatis mutandis, in the relevant previous steps) by, e.g., $1/(\sqrt{n}\log n)$. However, for small bandwidths h, the results from Sect. 3 show that $\sqrt{n}\|F_n^K(h) - F_n\|_\infty$ is of the stochastic order $\sqrt{h\log(1/h)}$, whereas, if the bandwidth is chosen too large, the bias $\sqrt{n}h^{t+1}$ starts to dominate, depending on the unknown smoothness t. Hence restricting Lepski's method to estimators $p_n^K(h)$ whose distribution functions are in a sup-norm ball of radius $\sqrt{h}(\log(1/h))^2/(\sqrt{n}\log n)$ around F_n should improve the asymptotic precision of the estimator $\hat{h}_n$. In fact, it is for these sizes that one has to use the full strength of Theorem 1, cf. (34) in the proof of Theorem 2.

Remark 5 (Order of the kernel) The range of adaptation in the above theorem is restricted to the Hölder classes with smoothness smaller than T, the reason being that the order of the kernel has to be fixed at a certain degree to control the bias term. If one wants to adapt up to order $T = 1$, then a standard symmetric positive kernel (of order 2) can be used. If one wants to adapt to smoothness larger than $T = 1$, one has to use higher order kernels, and then it might be advisable to allow for kernels of varying order in the construction of the estimator. This can be achieved by using similar methods as in Sect. 2.8 of Lepski and Spokoiny [27], and might be important for improving the actual implementability of the procedure proposed above.

Acknowledgments We thank two referees and Jon Wellner for useful comments on the manuscript. A question from Jon Wellner led us to proving Proposition 1.

5 Appendix: Inequalities

5.1 Talagrand's inequality

Let $X_1, \ldots, X_n$ be i.i.d. with law P on $\mathbb{R}$, and let $\mathcal{F}$ be a P-centered (i.e., $\int f dP = 0$ for all $f \in \mathcal{F}$) countable class of real-valued functions on $\mathbb{R}$, uniformly bounded by the constant U. Let σ be any positive number such that $\sigma^2 \geq \sup_{f\in\mathcal{F}} E(f^2(X))$, and set $V := n\sigma^2 + 2UE\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}}$. Then, Bousquet's [3] version of Talagrand's inequality ([32]), with constants, is as follows (see Theorem 7.3 in [3]): For every $x \geq 0$

$$\Pr\left\{\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}} \geq E\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}} + \sqrt{2Vx} + Ux/3\right\} \leq 2e^{-x}. \quad (37)$$

We note that Rio [30] obtained this inequality with $Ux/2$ instead of $Ux/3$, and we could have used his inequality to obtain exactly the same results in the present article.

5.2 Moment inequalities for VC classes, with constants

To apply Talagrand's inequality in Theorem 1 and Lemma 1, we need bounds for $E\|\sum_{j=1}^n f(X_j)\|_{\mathcal{F}}$ with explicit constants, which we obtain in this subsection. Talagrand [31] showed that one can obtain simple bounds for the moments of the empirical process indexed by a VC class of sets just by means of the usual entropy bound and the contraction principle for Rademacher processes. This was extended to VC classes of functions by several authors ([11,13,15], among others). Here we sketch a proof of these extensions, with explicit constants, but only for the case we are using (constant envelope, VC classes).

Proposition 3 *Let $\mathcal{F}$ be a countable (or stochastically separable) P-centered class of real valued functions uniformly bounded by a constant U and such that, for all finitely supported probability measures Q, the $\mathcal{L}^2(Q)$-covering numbers satisfy*

$$N(\mathcal{F}, \mathcal{L}^2(Q), \tau) \leq \left(\frac{AU}{\tau}\right)^v, \quad 0 < \tau < U,$$

for some $A > e$ and $v \geq 2$. Let $\sigma \leq U$ be as before (37). Then, for all $n \in \mathbb{N}$,

$$E\left\|\sum_{j=1}^n f(X_j)\right\|_{\mathcal{F}} \leq 30\sqrt{2v}\sqrt{n\sigma^2 \log\frac{5AU}{\sigma}} + 15^2 2^5 vU \log\frac{5AU}{\sigma}. \tag{38}$$

Proof (Sketch) It suffices to prove this inequality for $U = 1$. By the entropy bound with constants in the lemma below, we have

$$\frac{1}{\sqrt{n}} E_\varepsilon \left\|\sum_{j=1}^n \varepsilon_j f(X_j)\right\|_{\mathcal{F}} \leq 30\sqrt{2v} \int_0^{\sqrt{\|P_n f^2\|_{\mathcal{F}}}/4} \sqrt{\log\frac{2^{1/(2v)}A}{\tau}}\, d\tau,$$

where ε_i are i.i.d. Rademacher variables independent of the variables X_j and E_ε indicates expectation with respect to the ε_i's. As in [17], if $(\log C/c) \geq 2$ then

$$\int_0^c \left(\log\frac{C}{x}\right)^{1/2} dx \leq 2c\left(\log\frac{C}{c}\right)^{1/2}.$$

Since $4 \cdot 2^{1/(2v)} A/\sqrt{\|P_n f^2\|_{\mathcal{F}}} \geq 4A \geq e^2$, we conclude

$$\frac{1}{\sqrt{n}} E_\varepsilon \left\|\sum_{j=1}^n \varepsilon_j f(X_j)\right\|_{\mathcal{F}} \leq 15\sqrt{2v}\sqrt{\|P_n f^2\|_{\mathcal{F}}}\sqrt{\log\frac{5A}{\sqrt{\|P_n f^2\|_{\mathcal{F}}}}}.$$

By concavity of $\sqrt{x}\sqrt{-\log x}$ on $(0, e^{-1})$, this yields

$$\frac{1}{\sqrt{n}} E \left\| \sum_{j=1}^{n} \varepsilon_j f(X_j) \right\|_{\mathcal{F}} \leq 15\sqrt{2v}\sqrt{E\|P_n f^2\|_{\mathcal{F}}}\sqrt{\log \frac{5A}{\sqrt{E\|P_n f^2\|_{\mathcal{F}}}}} =: 15\sqrt{2v}B.$$

Now we apply the contraction principle for Rademacher processes [31], to the effect that

$$nE\|P_n f^2\|_{\mathcal{F}} \leq n\sigma^2 + 8E \left\| \sum_{j=1}^{n} \varepsilon_j f(X_j) \right\|_{\mathcal{F}}.$$

The last two inequalities give

$$E\|P_n f^2\|_{\mathcal{F}} \leq \sigma^2 + \frac{120\sqrt{2v}B}{\sqrt{n}},$$

which, replaced into the definition of B, implies that B satisfies the inequality

$$B^2 \leq \left(\sigma^2 + \frac{120\sqrt{2v}B}{\sqrt{n}}\right) \log \frac{5A}{\sigma}.$$

The result follows by solving for B, and applying a simple symmetrization inequality. □

For lack of reference, here we sketch how to get the usual entropy bound with sensible (not necessarily best possible) constants:

Lemma 2 *Let $X(t) = \sum_{i=1}^{N} a_i(t)\varepsilon_i$, $t \in T$, (for any $N \in \mathbb{N}$) be a Rademacher process indexed by the set T such that $X(t_0) = 0$ a.s. for some $t_0 \in T$, and let $D = 2\sup_{t\in T}(EX^2(t))^{1/2} = 2\sup_{t\in T}\left(\sum_{i=1}^{N} a_i^2(t)\right)^{1/2}$. Then,*

$$E \sup_{t\in T} |X(t)| \leq 30 \int_0^{D/4} \sqrt{2\log(\sqrt{2}N(T, \delta, \tau))}d\tau,$$

where $\delta^2(s, t) = E|X(t) - X(s)|^2$.

Proof (Sketch) If $\xi = \sum a_i\varepsilon_i$, then the well known hypercontractivity inequality (e.g., de la Peña and Giné [6, p. 113]) states

$$E|\xi|^q \leq (q-1)^{q/2}(E\xi^2)^{q/2},$$

which, by development of the exponential gives

$$Ee^{\xi^2/6\tau^2} < 2$$

where $\tau^2 = \sum a_i^2$. This gives, by a convexity argument e.g., (4.3.1), p. 188, in de la Peña and Giné [6], that if ξ_i, $1 \le i \le N$, are Rademacher polynomials and $\sigma^2 = \max_{i \le N} E\xi_i^2$, then

$$E \max_{i \le N} |\xi_i| \le \sqrt{6}\sigma \sqrt{\log(2N)}. \tag{39}$$

It can be shown that (T, δ) is stochastically separable (as it is isometric to a subset of a cube in $\mathbb{R}^N$), so it suffices to show that

$$E \sup_{t \in S} |X(t)| \le 30 \int_0^{D/4} \sqrt{2 \log(\sqrt{2} N(T, \delta, \tau))} d\tau,$$

for all finite subsets $S \subseteq T$, S containing t_0 without loss of generality. Then, a computation (chaining) similar to the one leading to (5.1.14), p. 217 in de la Peña and Giné [6], applied with $d = \sqrt{6}\delta$, but for $E \max_{s \in S} |X(s)|$ instead of $\| \max_{s \in S} |X(s)| \|_\psi$, and using (39) and the properties of log, and assuming that the d-diameter of T is at most 1, gives

$$E \sup_{t \in S} |X(t)| \le 12 \int_0^{1/4} \sqrt{2 \log(\sqrt{2} N(T, d, \tau))} d\tau.$$

Applying this bound to $Y(t) = X(t)/(\sqrt{6}D)$ if the δ-diameter of T is dominated by D yields the result. □

We use the above bounds for classes of functions of the form

$$\mathcal{F} := \mathcal{F}_h = \left\{ H\left(\frac{t - \cdot}{h}\right) : t \in \mathbb{R} \right\},$$

where H is of bounded variation. We record here the entropy bound for these classes:

Lemma 3 *Let $H : \mathbb{R} \mapsto \mathbb{R}$ be a function of bounded variation. Then there exists $A < \infty$ independent of h and of H such that, for all probability measures Q on $\mathbb{R}$ and all $0 < \varepsilon < 1$,*

$$N(\mathcal{F}_h, \mathcal{L}^2(Q), \varepsilon) \le \left(\frac{2\|H\|_V A}{\varepsilon} \right)^4$$

where $\|H\|_V$ is the total variation norm of the function H.

Proof Let H^+ and H^- be, respectively, the positive and negative variations of H. Then $H(\cdot/h) = H^+(\cdot/h) - H^-(\cdot/h)$. Set $\mathcal{F}_h^+ = \{H^+((t - \cdot)/h) : t \in \mathbb{R}\}$ and likewise define $\mathcal{F}_h^-$. Then, a simple estimate of covering numbers gives

$$N(\mathcal{F}_h, \mathcal{L}_2(Q), 2\varepsilon) \le N(\mathcal{F}_h^+, \mathcal{L}_2(Q), \varepsilon) N(\mathcal{F}_h^-, \mathcal{L}_2(Q), \varepsilon).$$

Applying Theorem 2.6.7 and Lemma 2.6.16 in [34] to $\mathcal{F}_h^+$, we obtain that there is a universal constant A such that

$$N(\mathcal{F}_h^+, \mathcal{L}_2(Q), \varepsilon) \leq \left(AH^+(\infty-)/\varepsilon\right)^2,$$

since $\sup_x H^+((t-x)/h) = H^+(\infty-)$, and similarly for $\mathcal{F}_h^-$. The lemma follows. □

References

1. Balabdaoui, F., Wellner, J.A.: A Kiefer–Wolfowitz theorem for convex densities. IMS Lect. Notes Monogr. Ser. **55**, 1–31 (2007)
2. Bickel, J.P., Ritov, Y.: Nonparametric estimators which can be 'plugged-in'. Ann. Stat. **31**, 1033–1053 (2003)
3. Bousquet, O.: Concentration inequalities for sub-additive functions using the entropy method. In: Giné, E., Houdré, C., Nualart, D. (eds.) Stochastic inequalities and applications., Progr. Probab., vol 56, pp 213–247. Birkhäuser, Boston (2003)
4. Butucea, C.: Exact adaptive pointwise estimation on Sobolev classes of densities. ESAIM Probab. Stat. **5**, 1–31 (2001)
5. Dalalyan, A.S., Golubev, G.K., Tsybakov, A.B.: Penalized maximum likelihood and semiparametric second-order efficiency. Ann. Stat. **34**, 169–201 (2006)
6. de la Peña, V., Giné, E.: Decoupling From Dependence to Independence. Springer, New York (1999)
7. Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Density estimation by wavelet thresholding. Ann. Stat. **24**, 508–539 (1996)
8. Dümbgen, L., Rufibach, K.: Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. Preprint (2007)
9. Durot, C., Tocquet, A.S.: On the distance between the empirical process and its concave majorant in a monotone regression framework. Ann. Inst. H. Poincaré Probab. Stat. **39**, 217–240 (2003)
10. Dvoretzky, A., Kiefer, J., Wolfowitz, J.: Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Stat. **27**, 642–669 (1956)
11. Einmahl, U., Mason, D.M.: An empirical process approach to the uniform consistency of kernel-type function estimators. J. Theor. Probab. **13**, 1–37 (2000)
12. Folland, G.B.: Real Analysis. Wiley, New York (1999)
13. Giné, E., Guillou, A.: On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. Ann. Inst. H. Poincaré Probab. Stat. **37**, 503–522 (2001)
14. Giné, E., Guillou, A.: Rates of strong uniform consistency for multivariate kernel density estimators. Ann. Inst. H. Poincaré Probab. Stat. **38**, 907–921 (2002)
15. Giné, E., Koltchinskii, V.: Concentration inequalities and asymptotic results for ratio type empirical processes. Ann. Probab. **34**, 1143–1216 (2006)
16. Giné, E., Mason, D.M.: Laws of the iterated logarithm for the local U-statistic process. J. Theor. Probab. **20**, 457–486 (2007)
17. Giné, E., Nickl, R.: Uniform limit theorems for wavelet density estimators. preprint (2007)
18. Giné, E., Nickl, R.: Uniform central limit theorems for kernel density estimators. Probab. Theory Related Fields (forthcoming) (2008)
19. Hall, P., Kerkyacharian, G., Picard, D.: Block threshold rules for curve estimation using kernel and wavelet methods. Ann. Stat. **26**, 922–942 (1998)
20. Ibragimov, I.A., Khasminski, R.Z.: On estimation of distribution density. Zap. Nauchnyh Seminarov LOMI **98**, 61–86 (1980)
21. Juditsky, A., Lambert-Lacroix, S.: On minimax density estimation on $\mathbb{R}$. Bernoulli **10**, 187–220 (2004)
22. Kerkyacharian, G., Picard, D., Tribouley, K: L^p adaptive density estimation. Bernoulli **2**, 229–247 (1996)
23. Kiefer, J., Wolfowitz, J.: Asymptotically minimax estimation of concave and convex distribution functions. Z. Wahr. Verw. Gebiete. **34**, 73–85 (1976)
24. Ledoux, M., Talagrand, M.: Probability in Banach spaces. Springer, Berlin (1991)

25. Lepski, O.V.: Asymptotic minimax adaptive estimation. 1. Upper bounds. Theory Probab. Appl. **36**, 682–697 (1991)
26. Lepski, O.V., Mammen, E., Spokoiny, V.G.: Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. Ann. Stat. **25**, 929–947 (1997)
27. Lepski, O.V., Spokoiny, V.G.: Optimal pointwise adaptive methods in nonparametric estimation. Ann. Stat. **25**, 2512–2546 (1997)
28. Marron, J.S., Wand, M.P.: Exact mean integrated squared error. Ann. Stat. **20**, 712–736 (1992)
29. Nickl, R.: Donsker-type theorems for nonparametric maximum likelihood estimators. Probab. Theory Related Fields **138**, 411–449 (2007)
30. Rio, E.: Une inégalité de Bennett pour les maxima de processus empiriques. Ann. Inst. H. Poincaré Probab. Stat. **38**, 1053–1057 (2002)
31. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. **22**, 28–76 (1994)
32. Talagrand, M.: New concentration inequalities in product spaces. Invent. Math. **126**, 505–563 (1996)
33. Tsybakov, A.B.: Pointwise and sup-norm sharp adaptive estimation of the functions on the Sobolev classes. Ann. Stat. **26**, 2420–2469 (1998)
34. van der Vaart, A.W., Wellner, J.A.: Weak Cconvergence and Empirical Processes. Springer, New York (1996)