Bernoulli **16**(4), 2010, 1137–1163
DOI: 10.3150/09-BEJ239

Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections

EVARIST GINÉ[1] and RICHARD NICKL[2]

[1]*Department of Mathematics, University of Connecticut, Storrs, CT 06269-3009, USA. E-mail: gine@math.uconn.edu*
[2]*Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. E-mail: nickl@statslab.cam.ac.uk*

Given an i.i.d. sample from a distribution F on $\mathbb{R}$ with uniformly continuous density p_0, purely data-driven estimators are constructed that efficiently estimate F in sup-norm loss and simultaneously estimate p_0 at the best possible rate of convergence over Hölder balls, also in sup-norm loss. The estimators are obtained by applying a model selection procedure close to Lepski's method with random thresholds to projections of the empirical measure onto spaces spanned by wavelets or B-splines. The random thresholds are based on suprema of Rademacher processes indexed by wavelet or spline projection kernels. This requires Bernstein-type analogs of the inequalities in Koltchinskii [*Ann. Statist.* **34** (2006) 2593–2656] for the deviation of suprema of empirical processes from their Rademacher symmetrizations.

Keywords: adaptive estimation; Lepski's method; Rademacher processes; spline estimator; sup-norm loss; wavelet estimator

1. Introduction

If $X_1, \ldots, X_n$ are i.i.d. with unknown distribution function F on $\mathbb{R}$, then classical results of mathematical statistics establish optimality of the empirical distribution function F_n as an estimator of F. That is to say, if we assume no a priori knowledge whatsoever on F and equip the set of all probability distribution functions with some natural loss function such as sup-norm loss, then F_n is asymptotically sharp minimax for estimating F. (The same is true even if more is known about F, for instance, if F is known to have a uniformly continuous density.) However, this does not preclude the existence of other estimators that are also asymptotically minimax for estimating F in sup-norm loss, but which improve upon F_n in other respects. What we have in mind is a purely data-driven estimator that is efficient for F, but, at the same time, also estimates the density f of F at the best rate of convergence in some relevant loss function over some prescribed classes of densities. More precisely, our goal in the present article is to construct estimators that satisfy the functional central limit theorem (CLT) for the distribution function *and* which adapt to the unknown smoothness of the density in *sup-norm loss*. Whereas this article is concerned with the mathematical problem of the existence and construction of such estimators, it does not deal with the practical implementation of estimation procedures.

1350-7265 © 2010 ISI/BS

To achieve adaptation, one can opt for several approaches, all of which are related. Among them, we mention the penalization method of Barron, Birgé and Massart [1], wavelet thresholding [7] and Lepski's [26] method. Our choice for the goal at hand consists of using Lepski's method, with random thresholds, applied to wavelet and spline projection estimators of a density.

The linear estimators underlying our procedure are projections of the empirical measure onto spaces spanned by wavelets, and wavelet theory is central to some of the derivations of this article. The wavelets most commonly used in statistics are those that are compactly supported (for example, Daubechies wavelets), and our results readily apply to these. However, for computational and other purposes, projections onto spline spaces are also interesting candidates for the estimators. Density estimators obtained by projecting the empirical measure onto Schoenberg spaces spanned by B-splines were studied by Huang and Studden [19]. As is well known in wavelet theory, the Schoenberg spline spaces with equally spaced knots have an orthonormal basis consisting of the Battle–Lemarié wavelets so that the spline projection estimator is, in fact, exactly equal to the wavelet estimator based on Battle–Lemarié wavelets. These wavelets do not have compact support, but they are exponentially localized. Although we cannot, in general, handle exponentially decaying wavelets, we can still work with Battle–Lemarié wavelets because the B-spline expansion of the projections allows us to show that the relevant classes of functions are of Vapnik–Chervonenkis type so that empirical process techniques can be applied. In particular, the adaptive estimators we devise in Theorem 3 may be based either on spline projections or on compactly supported wavelets. In the process of proving the main theorem, we also provide new asymptotic results for spline projection density estimators similar to those for wavelet estimators in [14].

We need to use Talagrand's exponential inequality with sharp constants [3,21] in the proofs, but to do this, we have to estimate the expectation of suprema of certain empirical processes that appear in the centering of Talagrand's inequality. The use of entropy-based moment inequalities for empirical processes typically results in too conservative constants (for example, in [13]). In order to remedy this problem, we adapt recent ideas due to Koltchinskii [22,23] and Bartlett, Boucheron and Lugosi [2] to density estimation: the entropy-based moment bounds are replaced by the sup-norm of the associated Rademacher averages, which are, with high probability, better estimates of the expected value of the supremum of the empirical process. We derive a Bernstein-type analog of an exponential inequality in [23] that shows how the supremum of an empirical process deviates from the supremum of the associated Rademacher processes. This Bernstein-type version allows one to use partial knowledge of the variance of the empirical processes involved, which is crucial for applications in our context of adaptive density estimation. Moreover, we show that one can use, instead of the supremum of the Rademacher process, its conditional expectation given the data.

Adaptive estimation in sup-norm loss is a relatively recent subject. We should mention the results in Tsybakov [34], Golubev, Lepski and Levit [16] – who only considered Sobolev-type smoothness conditions – and [15]. All of these results were obtained in the Gaussian white noise model. If one is interested in adapting to a Hölder-continuous density in sup-norm loss in the i.i.d. density model on $\mathbb{R}$, this simplifying Gaussian structure is not available and novel techniques are needed. In the i.i.d. density model on $\mathbb{R}$, a direct 'competitor' to the estimators constructed in this article is the hard thresholding wavelet density estimator introduced in [7]: as proved in [14], its distribution function satisfies the functional CLT and it is adaptive in the

sup-norm over Hölder balls; however, the proofs there seem to require the additional assumption that $\mathrm{d}F$ integrates $|x|^\delta$ for some $\delta > 0$, and the constants appearing in the threshold and the risk become quite large for δ small. The results in the present article hold under no moment condition whatsoever.

2. Wavelet expansions and estimators

We start with some basic notation. If $(S, \mathcal{S})$ is a measurable space, then for Borel-measurable functions $h : S \to \mathbb{R}$ and Borel measures μ on S, we set $\mu h := \int_S h \,\mathrm{d}\mu$. We will denote by $L^p(Q) := L^p(S, Q)$, $1 \le p \le \infty$, the usual Lebesgue spaces on S with respect to a Borel measure Q, and if Q is Lebesgue measure on $S = \mathbb{R}$, then we simply denote this space by $L^p(\mathbb{R})$, and its norm by $\|\cdot\|_p$, if $p < \infty$. We will use $\|h\|_\infty$ to denote $\sup_{x\in\mathbb{R}} |h(x)|$ for $h : \mathbb{R} \to \mathbb{R}$. For $s \in \mathbb{N}$, denote by $\mathsf{C}^s(\mathbb{R})$ the spaces of functions $f : \mathbb{R} \to \mathbb{R}$ that are s-times differentiable with bounded uniformly continuous $D^r f$, $0 < r \le s$, equipped with the norm $\|f\|_{s,\infty} = \sum_{0\le\alpha\le s} \|D^\alpha f\|_\infty$, with the convention that $D^0 =: id$ and that $\mathsf{C}(\mathbb{R}) := \mathsf{C}^0(\mathbb{R})$ is then the space of bounded uniformly continuous functions. For non-integer $s > 0$ and $[s]$ the integer part of s, set

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

2.1. Multiresolution analysis and wavelet bases

We recall here a few well-known facts about wavelet expansions; see, for example, Sections 8 and 9 in [17]. Let $\phi \in L^2(\mathbb{R})$ be a scaling function, that is, ϕ is such that $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ is an orthonormal system in $L^2(\mathbb{R})$ and, moreover, the linear spaces $V_0 = \{f(x) = \sum_k c_k \phi(x - k) : \{c_k\}_{k\in\mathbb{Z}} \in \ell^2\}$, $V_1 = \{h(x) = f(2x) : f \in V_0\}, \ldots, V_j = \{h(x) = f(2^j x) : f \in V_0\}, \ldots$ are nested ($V_{j-1} \subseteq V_j$ for $j \in \mathbb{N}$) and their union is dense in $L^2(\mathbb{R})$. In the case where ϕ is a bounded function that decays exponentially at infinity (that is, $|\phi(x)| \le C\mathrm{e}^{-\gamma|x|}$ for some $C, \gamma > 0$) – which we assume for the rest of this subsection – the kernel of the projection onto the space V_j has certain properties. First, the series

$$K(y, x) := K(\phi, y, x) = \sum_{k\in\mathbb{Z}} \phi(y - k)\phi(x - k) \tag{1}$$

converges pointwise and we set $K_j(y, x) := 2^j K(2^j y, 2^j x)$, $j \in \mathbb{N} \cup \{0\}$. Furthermore, we have

$$|K(y, x)| \le \Phi(|y - x|) \quad \text{and} \quad \sup_{x\in\mathbb{R}} \sum_k \bigl|\phi(x - k)\bigr| < \infty, \tag{2}$$

where $\Phi : \mathbb{R} \to \mathbb{R}^+$ is bounded and has exponential decay (cf. Lemma 8.6 in [33]). For any j fixed, if $f \in L^p(\mathbb{R})$, $1 \le p \le \infty$, then the series

$$K_j(f)(y) := \int K_j(x, y) f(x) \,\mathrm{d}x = \sum_{k\in\mathbb{Z}} 2^j \phi(2^j y - k) \int \phi(2^j x - k) f(x) \,\mathrm{d}x, \qquad y \in \mathbb{R},$$

converges pointwise and, for $f \in L^2(\mathbb{R})$, $K_j(f)$ coincides with the orthogonal projection $\pi_j : L^2(\mathbb{R}) \to V_j$ of f onto V_j. For $f \in L^1(\mathbb{R})$, which is the main case in this article, the convergence of the series in fact takes place in $L^p(\mathbb{R})$, $1 \le p \le \infty$. This still holds true if $f(x)\,\mathrm{d}x$ is replaced by $\mathrm{d}\mu(x)$, where μ is any finite signed measure. If, now, ϕ is a scaling function and ψ the associated mother wavelet so that $\{\phi(\cdot - k), 2^{l/2}\psi(2^l(\cdot) - k) : k \in \mathbb{Z}, l \in \mathbb{N}\}$ is an orthonormal basis of $L^2(\mathbb{R})$, then any $f \in L^p(\mathbb{R})$ admits the formal expansion

$$f(y) = \sum_k \alpha_k(f)\phi(y-k) + \sum_{l=0}^{\infty}\sum_k \beta_{lk}(f)\psi_{lk}(y), \tag{3}$$

where $\psi_{lk}(y) = 2^{l/2}\psi(2^l y - k)$, $\alpha_k(f) = \int f(x)\phi(x-k)\,\mathrm{d}x$, $\beta_{lk}(f) = \int f(x)\psi_{lk}(x)\,\mathrm{d}x$. Since $(K_{l+1} - K_l)f = \sum_k \beta_{lk}(f)\psi_{lk}$, the partial sums of the series (3) are in fact given by

$$K_j(f)(y) = \sum_k \alpha_k(f)\phi(y-k) + \sum_{l=0}^{j-1}\sum_k \beta_{lk}(f)\psi_{lk}(y) \tag{4}$$

and if ϕ, ψ are bounded and have exponential decay, then convergence of the series (4) holds pointwise; it also holds in $L^p(\mathbb{R})$, $1 \le p \le \infty$, if $f \in L^1(\mathbb{R})$ or if f is replaced by a finite signed measure. Now, using these facts, one can furthermore show that the wavelet series (3) converges in $L^p(\mathbb{R})$, $p < \infty$, for $f \in L^p(\mathbb{R})$ and we also note that if p_0 is a uniformly continuous density, then its wavelet series converges uniformly.

2.2. Density estimation using wavelet and spline projection kernels

Let $X_1, \dots, X_n$ be i.i.d. random variables with common law P and density p_0 on $\mathbb{R}$, and denote by $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the associated empirical measure. A natural first step is to estimate the projection $K_j(p_0)$ of p_0 onto V_j by

$$p_n(y) := p_n(y, j) = \frac{1}{n}\sum_{i=1}^n K_j(y, X_i) = \sum_k \hat{\alpha}_k \phi(y-k) + \sum_{l=0}^{j-1}\sum_k \hat{\beta}_{lk}\psi_{lk}(y), \qquad y \in \mathbb{R}, \tag{5}$$

where K is as in (1), $j \in \mathbb{N}$, and where $\hat{\alpha}_k = \int \phi(x-k)\,\mathrm{d}P_n(x)$, $\hat{\beta}_{lk} = \int \psi_{lk}(x)\,\mathrm{d}P_n(x)$ are the empirical wavelet coefficients. We note that for ϕ, ψ compactly supported (for example, Daubechies wavelets), there are only finitely many k's for which these coefficients are non-zero. This estimator was first studied by Kerkyacharian and Picard [20] for compactly supported wavelets.

If the wavelets ϕ and ψ do not have compact support, it may be impossible to compute the estimator exactly since the sums over k consist of infinitely many summands. However, in the special case of the Battle–Lemarié family $\phi_r, r \ge 1$ (see, for example, Section 6.1 in [17]) – which is a class of non-compactly supported but exponentially decaying wavelets – the estimator has a simple form in terms of splines: the associated spaces $V_{j,r} = \{\sum_k c_k 2^{j/2}\phi_r(2^j(\cdot) - k) : \sum_k c_k^2 < \infty\}$

are, in fact, equal to the *Schoenberg spaces* generated by the Riesz basis of B-splines of order r so that the sum in (5) can be computed by

$$p_n(y, j) := \frac{1}{n}\sum_{i=1}^{n}\kappa_j(y, X_i) = \frac{2^j}{n}\sum_{i=1}^{n}\sum_{k}\sum_{l} b_{kl}N_{j,k,r}(X_i)N_{j,l,r}(y), \qquad y \in \mathbb{R}, \tag{6}$$

where the $N_{j,k,r}$ are (suitably translated and dilated) B-splines of order r, the kernel κ is as in (29) below and the b_{kl}'s are the entries of the inverse of the matrix defined in (28) below. An exact derivation of this spline projection, its wavelet representation and detailed definitions are given in Section 3.2. It turns out that for every sample point X_i and for every y, each of the last two sums extends over only r terms. We should note that this 'spline projection' estimator was first studied (outside the wavelet setting) by Huang and Studden [19], who derived pointwise rates of convergence; see also [18], where some comparison between Daubechies and spline wavelets can be found.

In the course of proving the main theorem of this article, we will derive some basic results for the linear spline projection estimator (6), which we now state. For classical kernel estimators, results similar to those that follow were obtained in [5,11,13], and for wavelet estimators based on compactly supported wavelets, this was done in [14].

Theorem 1. *Suppose that P has a bounded density p_0. Assume that $j_n \to \infty$, $n/(j_n 2^{j_n}) \to \infty$, $j_n/\log\log n \to \infty$ and $j_{2n} - j_n \le \tau$ for some τ positive. Let $p_n(y) = p_n(y, j_n)$ be the estimator from (6) for some $r \ge 1$. Then*

$$\limsup_n \sqrt{\frac{n}{2^{j_n} j_n}}\sup_{y\in\mathbb{R}}|p_n(y) - Ep_n(y)| = C \qquad a.s.$$

and, for $1 \le p < \infty$,

$$\sup_n \sqrt{\frac{n}{2^{j_n} j_n}}\Big(E\sup_{y\in\mathbb{R}}|p_n(y) - Ep_n(y)|^p\Big)^{1/p} \le C',$$

where C and C' depend only on $\|p_0\|_\infty$ and on r, p, τ ... and on r, p, τ. Moreover, if $p_0 \in C^t(\mathbb{R})$, then

$$\sup_{y\in\mathbb{R}}|p_n(y) - p_0(y)| = O\left(\sqrt{\frac{2^{j_n} j_n}{n}} + 2^{-tj_n}\right) \qquad a.s.\ and\ in\ L^p(P).$$

For rates of convergence in probability, the conditions on j_n can be weakened (see Proposition 3 below). The last bound in this theorem gives, for $p_0 \in \mathsf{C}^t(\mathbb{R})$ with $t \le r$ and $2^{j_n} \simeq (n/\log n)^{1/(2t+1)}$, that

$$\sup_{y\in\mathbb{R}}|p_n(y) - p_0(y)| = \mathrm{O}\left(\left(\frac{\log n}{n}\right)^{t/(2t+1)}\right), \qquad \text{both a.s. and in } L^p(P).$$

For the following central limit theorem, we denote by $\rightsquigarrow_{\ell^\infty(\mathbb{R})}$ convergence in law for sample-bounded processes in the Banach space of bounded functions on $\mathbb{R}$, and by G_P the usual P-Brownian bridge (for example, Chapter 3 in [8]). We should emphasize that the optimal bandwidth choice $2^{-j_n} \simeq n^{-1/(2t+1)}$ (if sup-norm loss is being considered, replace n by $n/\log n$) is admissible for every $t > 0$ in the theorem below.

Theorem 2. *Assume that the density p_0 of P is a bounded function ($t = 0$) or that $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some t, $0 < t \le r$. Let j_n satisfy $n/(2^{j_n} j_n) \to \infty$ and $\sqrt{n}2^{-j_n(t+1)} \to 0$ as $n \to \infty$. If F is the distribution function of P and we set $F_n^S(s) := \int_{-\infty}^s p(y, j_n)\,\mathrm{d}y$, then*

$$\sqrt{n}(F_n^S - F) \rightsquigarrow_{\ell^\infty(\mathbb{R})} G_P.$$

Proof. Given $\varepsilon > 0$, apply Proposition 4 below with $\lambda = \varepsilon$ so that $\|F_n^S - F_n\|_\infty = o_P(1/\sqrt{n})$ follows and use the fact that $\sqrt{n}(F_n - F)$ converges in law in $\ell^\infty(\mathbb{R})$ to G_P. □

3. The adaptive estimation procedures

In this section, we construct data-driven choices of the resolution level j and state the main adaptation results. As mentioned in the Introduction, we will use Rademacher symmetrization for this. Generate a Rademacher sequence ε_i, $i = 1, \dots, n$, independent of the sample (that is, ε_i takes values $1, -1$ with probability $1/2$) and set, for $j < l$,

$$\begin{aligned} R(n, j) &= 2\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i K_j(X_i, \cdot)\right\|_\infty \quad \text{and} \\ T(n, j, l) &= 2\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i (K_j - K_l)(X_i, \cdot)\right\|_\infty, \end{aligned} \tag{7}$$

where K_j is the kernel of the wavelet projection π_j onto V_j (both for Battle–Lemarié and compactly supported wavelets). In both cases, these are suprema of fixed random functions that depend only on known quantities that can be computed in a numerically effective way. For more details on Rademacher processes, see Section 3.1.1.

To construct the estimators, we first need a grid indexing the spaces V_j onto which we project P_n. For $r \ge 1$, $n > 1$, choose integers $j_{\min} := j_{\min,n}$ and $j_{\max} := j_{\max,n}$ such that $0 < j_{\min} < j_{\max}$,

$$2^{j_{\min}} \simeq \left(\frac{n}{\log n}\right)^{1/(2r+1)} \quad \text{and} \quad 2^{j_{\max}} \simeq \frac{n}{(\log n)^2}, \tag{8}$$

and set

$$\mathcal{J} := \mathcal{J}_n = [j_{\min}, j_{\max}] \cap \mathbb{N}.$$

Note that the number of elements in this grid is of order $\log n$. We will consider two preliminary estimators, $\bar{j}_n$ and $\tilde{j}_n$, of the resolution level (of course, only one is needed, but we offer a choice

between two, as discussed below). Let $p_n(j)$ be as in (5) or (6). First, we set

$$\bar{j}_n = \min\Bigg\{ j \in \mathcal{J} : \|p_n(j) - p_n(l)\|_\infty \leq T(n, j, l) + 7\|\Phi\|_2 \|p_n(j_{\max})\|_\infty^{1/2} \sqrt{\frac{2^l l}{n}}, \forall l > j, l \in \mathcal{J} \Bigg\}, \tag{9}$$

where the function Φ is as in (2), and we discuss an explicit way to construct Φ in Remark 2 below. If the minimum does not exist, then we set $\bar{j}_n$ equal to $j_{\max}$. An alternative estimator of the resolution level is

$$\tilde{j}_n = \min\Bigg\{ j \in \mathcal{J} : \|p_n(j) - p_n(l)\|_\infty \leq \big(B(\phi) + 1\big) R(n, l) + 7\|\Phi\|_2 \|p_n(j_{\max})\|_\infty^{1/2} \sqrt{\frac{2^l l}{n}}, \forall l > j, l \in \mathcal{J} \Bigg\}, \tag{10}$$

where $B(\phi)$ is a bound, uniform in j, for the operator norm in $L^\infty(\mathbb{R})$ of the projection π_j; see Remark 3 below. Again, if the minimum does not exist, we set $\tilde{j}_n$ equal to $j_{\max}$.

Before we state the main result, we briefly discuss these procedures. The data-driven resolution level $\tilde{j}_n$ in (10) is based on tests that use Rademacher-type analogs of the usual thresholds in Lepski's method: starting with $j_{\min}$, the main contribution to $\|p_n(j) - p_n(l)\|_\infty$ is the bias $\|Ep_n(j) - p_0\|_\infty$. The procedure should stop when the 'variance term' $\|p_n(l) - Ep_n(l)\|_\infty$ starts to dominate. Since this is an unknown quantity and since we know no good non-random upper bound for it, we estimate it by the supremum of the associated Rademacher process, that is, by $R(n, l)$. The constant $B(\phi)$ is necessary in order to correct for the lack of monotonicity of the $R(n, l)$'s in the resolution level l.

The estimator $\bar{j}_n$ in (9) is somewhat more refined: it attempts to take advantage of the fact that in the 'small bias' domain, and using the results from Section 3.1.1,

$$\|p_n(j) - p_n(l)\|_\infty = \left\| \frac{1}{n} \sum_{i=1}^n (K_j - K_l)(X_i, \cdot) \right\|_\infty$$

should not exceed its Rademacher symmetrization

$$T(n, j, l) = 2 \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (K_j - K_l)(X_i, \cdot) \right\|_\infty .$$

We now state the main result, whose proof is deferred to the next section. As usual, we say that a wavelet basis is *s-regular*, $s \in \mathbb{N} \cup \{0\}$, if either the scaling function ϕ has s weak derivatives contained in $L^p(\mathbb{R})$ for some $p \geq 1$ or if the mother wavelet ψ satisfies $\int x^\alpha \psi(x)\,\mathrm{d}x = 0$ for $\alpha = 0, \ldots, s$. Note that any compactly supported element of $\mathsf{C}^s(\mathbb{R})$, $0 < s \leq 1$, is of bounded $(1/s)$-variation so that the p-variation condition in the following theorem is satisfied, for example,

for all Daubechies wavelets. The estimators below achieve the optimal rate of convergence for estimating p_0 in sup-norm loss in the minimax sense (over Hölder balls); see, for example, [24] for optimality of these rates.

Theorem 3. *Let $X_1, \dots, X_n$ be i.i.d. on $\mathbb{R}$ with common law P that possesses a uniformly continuous density p_0. Let $p_n(j) := p_n(y, j)$ be as in (5), where ϕ is either compactly supported, of bounded p-variation ($p \geq 1$) and $(r-1)$-regular or $\phi = \phi_r$ equals a Battle–Lemarié wavelet. Let the sequence $\{\hat{j}_n\}_{n\in\mathbb{N}}$ be either $\{\bar{j}_n\}_{n\in\mathbb{N}}$ or $\{\tilde{j}_n\}_{n\in\mathbb{N}}$ and let $F_n(\hat{j}_n)(t) = \int_{-\infty}^t p_n(y, \hat{j}_n)\,\mathrm{d}y$. Then*

$$\sqrt{n}\big(F_n(\hat{j}_n) - F\big) \rightsquigarrow_{\ell^\infty(\mathbb{R})} G_P, \tag{11}$$

the convergence being uniform over the set of all probability measures P on $\mathbb{R}$ with densities p_0 bounded by a fixed constant, in any distance that metrizes convergence in law. Furthermore, if C is any precompact subset of $\mathsf{C}(\mathbb{R})$, then

$$\sup_{p_0\in C} E \sup_{y\in\mathbb{R}} |p_n(y, \hat{j}_n) - p_0(y)| = \mathrm{o}(1). \tag{12}$$

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

$$\sup_{p_0 : \|p_0\|_{t,\infty}\leq D} E \sup_{y\in\mathbb{R}} |p_n(y, \hat{j}_n) - p_0(y)| = \mathrm{O}\left(\left(\frac{\log n}{n}\right)^{t/(2t+1)}\right). \tag{13}$$

Remark 1 (Relaxing the uniform continuity assumption). The assumption of uniform continuity of the density of F can be relaxed by modifying the definition of $\bar{j}_n$ (or $\tilde{j}_n$) along the lines of [13]. The idea is to constrain all candidate estimators to lie in a ball of size $\mathrm{o}(1/\sqrt{n})$ around the empirical distribution function F_n so that (11) holds automatically. Formally, this can be done by adding the requirement

$$\sup_{t\in\mathbb{R}} \left| \int_{-\infty}^t p_n(y, j)\,\mathrm{d}y - F_n(t) \right| \leq \frac{1}{\sqrt{n}\log n}$$

to each test in (9) or (10). If this requirement does not even hold for $j_{\max}$, then it can be seen as evidence that F has no density and one just uses F_n as the estimator so as to obtain at least the functional CLT. If F has a bounded density, then one can use the exponential bound in Proposition 4 in the proof to control rejection probabilities of these test in the 'small bias' domain $\hat{j}_n > j^*$ and Theorem 3 can then still be proven for this procedure without *any* assumptions on F. See Theorem 2 in [13] for more details on this procedure and its proof.

Remark 2 (The constant $\|\Phi\|_2$). Once the wavelet ϕ have been chosen, $\hat{j}_n$ is purely data-driven since the function Φ depends only on ϕ. For the Haar basis ($\phi = I_{[0,1)}$), we can take $\Phi = \phi$ because, in this case, $K(x, y) \leq I_{[0,1)}(|x - y|)$ so that $\|\Phi\|_2 = 1$. A general way to obtain majorizing kernels Φ is described in Section 8.6 of [17]. For Battle–Lemarié wavelets, the spline representation of the projection kernel is again useful for estimating $\|\Phi\|_2$. See [19] for explicit computations.

Remark 3 (The constant $B(\phi)$). To construct $\tilde{j}_n$, one requires knowledge of the constant $B(\phi)$ that bounds the operator norm $\|\pi_j\|'_\infty$ of π_j, viewed as an operator $L^\infty(\mathbb{R})$. A simple way of obtaining a bound is as follows: for any $f \in L^\infty(\mathbb{R})$, we have, by (2),

$$|\pi_j(f)(x)| = \left| \int K_j(x,y) f(y)\,\mathrm{d}y \right| \le \|\Phi\|_1 \|f\|_\infty,$$

that is, $\|\pi_j\|'_\infty \le \|\Phi\|_1$. In combination with the previous remark, one readily obtains possible values for $B(\phi)$. For instance, for the Haar wavelet, $B(\phi) \le 1$. For spline wavelets, other methods are available. For example, for Battle–Lemarié wavelets arising from linear B-splines, $\|\pi_j\|'_\infty$ is bounded by 3, and [30], page 135, conjectures the bound $2r-1$ for general order r. See [6], Chapter 13.4, [30] and references therein for more information.

We also note that – as the results in Section 3.1.1, in particular Proposition 2, show – all of our proofs go through if one replaces $R(n,j)$, $T(n,j,l)$ by their respective Rademacher expectations $E^\varepsilon R(n,j)$, $E^\varepsilon T(n,j,l)$ in the definitions of $\tilde{j}_n$, $\bar{j}_n$.

3.1. Estimating suprema of empirical processes

Talagrand's [33] exponential inequality for empirical processes (see also [25]), which is a uniform Prohorov-type inequality, is not specific about constants. Constants in its Bernstein-type version have been specified by several authors [3,21,27]. Let X_i be the coordinates of the product probability space $(S,\mathcal{S},P)^{\mathbb{N}}$, where P is any probability measure on $(S,\mathcal{S})$ and let $\mathcal{F}$ be a countable class of measurable functions on S that take values in $[-1/2,1/2]$ or, if $\mathcal{F}$ is P-centered, in $[-1,1]$. Let $\sigma \le 1/2$ and V be any two numbers satisfying

$$\sigma^2 \ge \|Pf^2\|_{\mathcal{F}}, \qquad V \ge n\sigma^2 + 2E\left\| \sum_{i=1}^n \bigl(f(X_i) - Pf\bigr) \right\|_{\mathcal{F}}, \tag{14}$$

in which case V is also an upper bound for $E\|\sum (f(X_i)-Pf)^2\|_{\mathcal{F}}$ [21]. Then, noting that $\sup_{f\in\mathcal{F}\cup(-\mathcal{F})} \sum_{i=1}^n f(X_i) = \sup_{\mathcal{F}} |\sum_{i=1}^n f(X_i)|$, Bousquet's [3] version of Talagrand's inequality is as follows: for every $t>0$,

$$\Pr\left\{ \left\| \sum_{i=1}^n \bigl(f(X_i)-Pf\bigr) \right\|_{\mathcal{F}} \ge E\left\| \sum_{i=1}^n \bigl(f(X_i)-Pf\bigr) \right\|_{\mathcal{F}} + t \right\} \le \exp\left(-\frac{t^2}{2V + (2/3)t} \right). \tag{15}$$

In the other direction, the Klein and Rio [21] result is that for every $t>0$,

$$\Pr\left\{ \left\| \sum_{i=1}^n \bigl(f(X_i)-Pf\bigr) \right\|_{\mathcal{F}} \le E\left\| \sum_{i=1}^n \bigl(f(X_i)-Pf\bigr) \right\|_{\mathcal{F}} - t \right\} \le \exp\left(-\frac{t^2}{2V + 2t} \right). \tag{16}$$

These inequalities can be applied in conjunction with an estimate of the expected value obtained via empirical process methods. Here, we describe one such result for VC-type classes, that

is, for $\mathcal{F}$ satisfying the uniform metric entropy condition

$$\sup_Q N(\mathcal{F}, L^2(Q), \tau) \le \left(\frac{A}{\tau}\right)^v, \qquad 0 < \tau \le 1 \ (A \ge e, v \ge 2), \tag{17}$$

with the supremum extending over all Borel probability measures on $(S, \mathcal{S})$. We denote here by $N(\mathcal{G}, L^2(Q), \tau)$ the usual covering numbers of a class $\mathcal{G}$ of functions by balls of radius less than or equal to τ in $L^2(Q)$-distance. One then has, for every n,

$$E\left\|\sum_{i=1}^n \big(f(X_i) - Pf\big)\right\|_{\mathcal{F}} \le 2\left[15\sqrt{2vn\sigma^2 \log\frac{5A}{\sigma}} + 1350v \log\frac{5A}{\sigma}\right]; \tag{18}$$

see Proposition 3 in [13] with a change obtained by using V as in (14) instead of an earlier bound due to Talagrand for $E\|\sum(f(X_i) - Pf)^2\|_{\mathcal{F}}$. Inequalities of this type also have some historical precedents ([9,10,12,32] among others). The constants on the right-hand side of (18) may be far from the best possible, but we prefer them over unspecified 'universal' constants.

As is the case of Bernstein's inequality in $\mathbb{R}$, Talagrand's inequality is especially useful in the Gaussian tail range and, combining (15) and (18), one can obtain such a 'Gaussian tail' bound for the supremum of the empirical process that depends only on σ (similar to a bound in [10]).

Proposition 1. *Let $\mathcal{F}$ be a countable class of measurable functions that satisfies* (17) *and is uniformly bounded* (*in absolute value*) *by* $1/2$. *Assume, further, that for some* $\lambda > 0$,

$$n\sigma^2 \ge \frac{\lambda^2 v}{2} \log\frac{5A}{\sigma}. \tag{19}$$

Set $c_1(\lambda) = 2[15 + 1350\lambda^{-1}]$ *and let* $c_2(\lambda) \ge 1 + 120\lambda^{-1} + 10{,}800\lambda^{-2}$. *Then, if*

$$c_1(\lambda)\sqrt{2vn\sigma^2 \log\frac{5A}{\sigma}} \le t \le \frac{3}{2}c_2(\lambda)n\sigma^2, \tag{20}$$

we have

$$\Pr\left\{\left\|\sum_{i=1}^n \big(f(X_i) - Pf\big)\right\|_{\mathcal{F}} \ge 2t\right\} \le \exp\left(-\frac{t^2}{3c_2(\lambda)n\sigma^2}\right). \tag{21}$$

Proof. In the light of (19), inequality (18) gives

$$E\left\|\sum_{i=1}^n \big(f(X_i) - Pf\big)\right\|_{\mathcal{F}} \le c_1(\lambda)\sqrt{2vn\sigma^2 \log\frac{5A}{\sigma}}$$

and (14) implies that we can take $V = c_2(\lambda)n\sigma^2$. The result now follows from (15), taking into account that in the range of t's, $E\|\sum_{i=1}^n (f(X_i) - Pf)\|_{\mathcal{F}} \le t \le 3V/2$, (15) becomes

$$\Pr\left\{\left\|\sum_{i=1}^n \big(f(X_i) - Pf\big)\right\|_{\mathcal{F}} \ge 2t\right\} \le \exp\left(-\frac{t^2}{3V}\right).$$

□

The constants here may be too large for some applications, but they are not so in situations where λ can be taken very large, in particular, in asymptotic considerations. (Then $c_1(\lambda) \to 30$ and $c_2(\lambda) \to 1$ as $\lambda \to \infty$.)

3.1.1. *Estimating the size of empirical processes by Rademacher averages*

The constants one could obtain from Proposition 1 are not satisfactory for the applications to adaptive estimation which we have in mind. We now propose a remedy for this problem, inspired by a nice idea of Koltchinskii [22] and Bartlett, Boucheron and Lugosi [2] which they used in other contexts, namely in risk minimization and model selection. This consists of replacing the expectation of the supremum of an empirical process by the supremum of the associated Rademacher process. An inequality of this type (see [23], page 2602) is

$$\Pr\left\{\left\|\sum_{i=1}^{n}\bigl(f(X_i)-Pf\bigr)\right\|_{\mathcal{F}} \geq 2\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 3t\right\} \leq \exp\left(-\frac{2t^2}{3n}\right), \tag{22}$$

where ε_i, $i \in \mathbb{N}$, are i.i.d. Rademacher random variables, independent of the X_i's, all defined as coordinates on a large product probability space. Note that this bound does not take the variance V in (15) into account, but in the applications to density estimation that we have in mind, V is much smaller than n (it is of order $n2^{-j_n}$, $j_n \to \infty$). We need a similar inequality, with the quantity n in the bound replaced by V, valid over a large enough range of t's.

It will be convenient to use the following well-known symmetrization inequality (see, for example, [8], page 343):

$$\frac{1}{2}E\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}} - \frac{\sqrt{n}}{2}\|Pf\|_{\mathcal{F}} \leq E\left\|\sum_{i=1}^{n}\bigl(f(X_i)-Pf\bigr)\right\|_{\mathcal{F}} \leq 2E\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}. \tag{23}$$

The following exponential bound is the Bernstein-type analog of (22). Denote by E^{ε} expectation with respect to the Rademacher variables only.

Proposition 2. *Let $\mathcal{F}$ be a countable class of measurable functions, uniformly bounded (in absolute value) by $1/2$. Then, for every $t > 0$,*

$$\Pr\left\{\left\|\sum_{i=1}^{n}\bigl(f(X_i)-Ef(X)\bigr)\right\|_{\mathcal{F}} \geq 2\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 3t\right\} \leq 2\exp\left(-\frac{t^2}{2V'+2t}\right), \tag{24}$$

as well as

$$\Pr\left\{\left\|\sum_{i=1}^{n}\bigl(f(X_i)-Ef(X)\bigr)\right\|_{\mathcal{F}} \geq 2E^{\varepsilon}\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 3t\right\} \leq 2\exp\left(-\frac{t^2}{2V'+2t}\right), \tag{25}$$

where $V' = n\sigma^2 + 4E\|\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}}$.

Proof. We have

$$\Pr\left\{\left\|\sum_{i=1}^n \big(f(X_i) - Pf\big)\right\|_{\mathcal{F}} \geq 2\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 3t\right\}$$
$$\leq \Pr\left\{\left\|\sum_{i=1}^n \big(f(X_i) - Pf\big)\right\|_{\mathcal{F}} \geq 2E\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} + t\right\}$$
$$+ \Pr\left\{\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} \leq E\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} - t\right\}.$$

For the first term, combining (23) with (15) gives

$$\Pr\left\{\left\|\sum_{i=1}^n \big(f(X_i) - Pf\big)\right\|_{\mathcal{F}} \geq 2E\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} + t\right\} \leq \exp\left(-\frac{t^2}{2V' + (2/3)t}\right).$$

For the second term, note that (16) applies to the randomized sums $\sum_{i=1}^n \varepsilon_i f(X_i)$ as well, by just taking the class of functions

$$\mathcal{G} = \{g(\tau, x) = \tau f(x) : f \in \mathcal{F}\},$$

$\tau \in \{-1, 1\}$, instead of $\mathcal{F}$ and the probability measure $\bar{P} = 2^{-1}(\delta_{-1} + \delta_1) \times P$ instead of P. Hence,

$$\Pr\left\{\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} \leq E\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} - t\right\} \leq \exp\left(-\frac{t^2}{2V' + 2t}\right) \tag{26}$$

since $V' \geq n\sigma^2 + 2E\|\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}}$. Combining the bounds completes the proof of (24).

It remains to prove (25). Let $\mathcal{G}$, $\bar{P}$ be as above, let $Y_i = (\varepsilon_i, X_i)$ and note that $\bar{P}$ is the law of Y_i. By convexity,

$$E\mathrm{e}^{-tE^{\varepsilon}\|\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}}} \leq E\mathrm{e}^{-t\|\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}}} = E\mathrm{e}^{-t\|\sum_{i=1}^n g(Y_i)\|_{\mathcal{G}}}$$

for all t. The Klein and Rio [21] version (16) of Talagrand's inequality is, in fact, established by estimating the Laplace transform $E\mathrm{e}^{-t\|\sum_{i=1}^n g(Y_i)\|_{\mathcal{G}}}$ and Theorem 1.2a in [21] implies that

$$E\mathrm{e}^{-tE^{\varepsilon}\|\sum_{i=1}^n \varepsilon_i (f(X_i) - Pf)\|_{\mathcal{F}}} \leq -tE\left\|\sum_{i=1}^n g(Y_i)\right\|_{\mathcal{G}} + \frac{V}{9}(\mathrm{e}^{3t} - 3t + 1)$$

for $V \geq n\sigma^2 + 2E\|\sum_{i=1}^n g(Y_i)\|_{\mathcal{G}}$, which, by their proof of the implication $(a) \Rightarrow (c)$ in that theorem, gives

$$\Pr\left\{ E^{\varepsilon} \left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \leq E \left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} - t \right\} \leq \exp\left(-\frac{t^2}{2V' + 2t} \right).$$

The proof of (25) now follows as in the previous case. □

For $\mathcal{F}$ of VC-type, the moment bound (18) is usually proved as a consequence of a bound for the Rademacher process. In fact, the proof of Proposition 3 in [13] shows that

$$E \left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \leq 15\sqrt{2vn\sigma^2 \log \frac{5A}{\sigma}} + 1350v \log \frac{5A}{\sigma}, \tag{27}$$

where σ is as in (14), which we use in the following corollary, together with the previous proposition. The constant $c_2(\lambda)$ in the exponent below is still potentially large, but tends to one if $\lambda \to \infty$.

Corollary 1. *Let $\mathcal{F}$ be a countable class of measurable functions that satisfies* (17) *and assume it to be uniformly bounded* (*in absolute value*) *by* 1/2. *Assume, further,* (19) *for some $\lambda > 0$. Then, for $0 < t \leq \frac{1}{20} c_2(\lambda) n\sigma^2$ with $c_2(\lambda)$ as in Proposition* 1, *we have*

$$\Pr\left\{ \left\| \sum_{i=1}^n \big(f(X_i) - Ef(X) \big) \right\|_{\mathcal{F}} \geq 2 \left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} + 3t \right\} \leq 2\exp\left(-\frac{t^2}{2.1 c_2(\lambda) n\sigma^2} \right)$$

and the same inequality holds if $\|\sum_{i=1}^n \varepsilon_i f(X_i)\|_{\mathcal{F}}$ is replaced by its E^{ε} expectation.

Proof. By (19) and (27), we have $V' \leq c_2(\lambda) n\sigma^2$, and the condition on t together with (24) gives the result. □

3.2. Projections onto spline spaces and their wavelet representation

In this section, we briefly review how the wavelet estimator (5) for Battle–Lemarié wavelets can be represented as a spline projection estimator (6). We shall need the spline representation in some proofs, while the wavelet representation will be useful in others.

Let $T := T_j = \{t_i(j)\}_{-\infty}^{\infty} = 2^{-j}\mathbb{Z}$, $j \in \mathbb{Z}$, be a bi-infinite sequence of equally spaced knots, $t_i := t_i(j)$. A function S is a spline of order r, or of degree $m = r - 1$, if, on each interval (t_i, t_{i+1}), it is a polynomial of degree less than or equal to m (and of degree exactly m on at least one interval) and, at each breakpoint t_i, S is at least $(m-1)$-times differentiable. The Schoenberg space $\mathcal{S}_r(T) := \mathcal{S}_r(T, \mathbb{R})$ is defined as the set of all splines of order (less than or equal to) r and it coincides with the space $\mathcal{S}_r(T, 1, \mathbb{R})$ in [6], page 135. The space $\mathcal{S}_r(T_j)$ has a Riesz basis formed

by B-splines $\{N_{j,k,r}\}_{k\in\mathbb{Z}}$ that we now describe; see Section 4.4 in [31] and page 138f in [6] for more details. Define

$$N_{0,r}(x) = 1_{[0,1)} * \cdots * 1_{[0,1)}(x), \qquad r\text{-times} := \sum_{i=0}^{r} \frac{(-1)^i \binom{r}{i} (x-i)_+^{r-1}}{(r-1)!}.$$

For $r = 2$, this is the linear B-spline (the usual 'hat' function), for $r = 3$, it is the quadratic and for $r = 4$, it is the cubic B-spline. Set $N_{k,r}(x) := N_{0,r}(x-k)$. The elements of the Riesz basis are then given by

$$N_{j,k,r}(x) := N_{k,r}(2^j x) = N_{0,r}(2^j x - k).$$

By the Curry–Schoenberg theorem, any $S \in \mathcal{S}_r(T_j)$ can be uniquely represented as $S(x) = \sum_{k\in\mathbb{Z}} c_k N_{j,k,r}(x)$. The orthogonal projection $\pi_j(f)$ of $f \in L^2(\mathbb{R})$ onto $\mathcal{S}_r(T_j) \cap L^2(\mathbb{R})$ is derived, for example, in [6], page 401f, where it is shown that $\pi_j(f) = 2^{j/2} \sum_{k\in\mathbb{Z}} c_k N_{j,k,r}$, with the coefficients $c_k := c_k(f)$ satisfying $(Ac)_k = 2^{j/2} \int N_{j,k,r}(x) f(x)\,\mathrm{d}x$, the matrix A being given by

$$a_{kl} = \int 2^j N_{j,k,r}(x) N_{j,l,r}(x)\,\mathrm{d}x = \int N_{k,r}(x) N_{l,r}(x)\,\mathrm{d}x. \tag{28}$$

The inverse A^{-1} of A exists (see Corollary 4.2 on page 404 in [6]) and if we denote its entries by b_{kl} so that $c_k = 2^{j/2} \int \sum_l b_{kl} N_{j,l,r}(x) f(x)\,\mathrm{d}x$, then we have

$$\pi_j(f)(y) = 2^j \int \sum_k \sum_l b_{kl} N_{j,l,r}(x) N_{j,k,r}(y) f(x)\,\mathrm{d}x = \int \kappa_j(x,y) f(x)\,\mathrm{d}x,$$

where $\kappa_j(x,y) = 2^j \kappa(2^j x, 2^j y)$ with

$$\kappa(x,y) = \sum_k \sum_l b_{kl} N_{l,r}(x) N_{k,r}(y) \tag{29}$$

is the spline projection kernel. Note that κ is symmetric in its arguments.

In fact, diagonalization of the kernel κ of the projection operator π_j led to one of the first examples of wavelets; see, for example, page 21f and Section 2.3 in [28], Section 5.4 in [4] or Section 6.1 in [17]. There, it is shown that there exists an $(r-1)$-times differentiable scaling function ϕ_r with exponential decay, the Battle–Lemarié wavelet of order r, such that

$$\mathcal{S}_r(T_j) \cap L^2(\mathbb{R}) = V_{j,r} = \left\{ \sum_k c_k 2^{j/2} \phi_r\big(2^j(\cdot) - k\big) : \sum_k c_k^2 < \infty \right\}.$$

This necessarily implies that the kernels κ and $K = K(\phi_r)$ describe the same projections in $L^2(\mathbb{R})$ and the following simple lemma shows that these kernels are, in fact, pointwise the same.

Lemma 1. *Let $\{N_{k,r}\}_{k\in\mathbb{Z}}$ be the Riesz basis of B-splines of order $r \geq 1$ and let ϕ_r be the associated Battle–Lemarié scaling function. If K is as in (1) and κ is as in (29), then, for all $x, y \in \mathbb{R}$, we have*

$$K(x,y) = \kappa(x,y).$$

Proof. If $r = 1$, then $N_{0,1} = \phi_1$ since this is just the Haar basis. So, consider $r > 1$. Since $\{\phi_r(\cdot - k) : k \in \mathbb{Z}\}$ is an orthonormal basis of $\mathcal{S}_r(\mathbb{Z}) \cap L^2(\mathbb{R})$ (see, for example, Theorem 1 on page 26 in [28]), it follows that K and κ are the kernels of the same L^2-projection operator and, therefore, for all $f, g \in L^2(\mathbb{R})$,

$$\int\int \big(K(x, y) - \kappa(x, y)\big) f(x) g(y)\,\mathrm{d}x\,\mathrm{d}y = 0.$$

By density in $L^2(\mathbb{R} \times \mathbb{R})$ of linear combinations of products of elements of $L^2(\mathbb{R})$, this implies that κ and K are almost everywhere equal in $\mathbb{R}^2$. We complete the proof by showing that both functions are continuous on $\mathbb{R}^2$. For K, this follows from the decomposition

$$\begin{aligned}|K(x, y) - K(x', y')| \le{}& \sum_k |\phi_r(x - k) - \phi_r(x' - k)||\phi_r(y - k)| \\ &+ \sum_k |\phi_r(y - k) - \phi_r(y' - k)||\phi_r(x' - k)|,\end{aligned}$$

the uniform continuity of ϕ_r ($r > 1$) and relation (2). For κ, we use the relation (31) below,

$$\begin{aligned}|\kappa(x, y) - \kappa(x', y')| \le{}& \sum_i |N_{i,r}(x) - N_{i,r}(x')||H(y - i)| \\ &+ \sum_i |H(y - i) - H(y' - i)||N_{i,r}(x')|,\end{aligned}$$

which implies continuity of κ on $\mathbb{R}^2$ since $N_{0,r}$ and H are uniformly continuous (as $N_{0,r}$ is, and $\sum_i |g(|i|)| < \infty$) and since $N_{0,r}$ has compact support. □

3.3. An exponential inequality for the uniform deviations of the linear estimator

To control the uniform deviations of the linear estimators from their means, one can use inequalities for the empirical process indexed by classes of functions $\mathcal{F}$ contained in

$$\mathcal{K} = \big\{2^{-j} K_j(\cdot, y) : y \in \mathbb{R}, j \in \mathbb{N} \cup \{0\}\big\}, \tag{30}$$

together with suitable bounds on the 'weak' variance σ.

If ϕ has compact support (and is of finite p-variation), it is proved in Lemma 2 of [14] that the class $\mathcal{K}$ also satisfies the bound (17). However, the proof there does not apply to Battle–Lemarié wavelets. A different proof, using the Toeplitz and band-limited structure of the spline projection kernel, still enables us to prove that these classes of functions are of Vapnik–Chervonenkis type.

Lemma 2. *Let $\mathcal{K}$ be as in (30), where ϕ_r is a Battle–Lemarié wavelet for some $r \geq 1$. There then exist finite constants $A \geq 2$ and $v \geq 2$ such that*

$$\sup_Q N(\mathcal{K}, L^2(Q), \varepsilon) \leq \left(\frac{A}{\varepsilon}\right)^v$$

for $0 < \varepsilon < 1$ and where the supremum extends over all Borel probability measures on $\mathbb{R}$.

Proof. In the case $r = 1$, ϕ_1 is just the Haar wavelet, in which case the result follows from Lemma 2 of [14]. Hence, we assume that $r \geq 2$.

The matrix A is Toeplitz since, by a change of variables in (28), $a_{kl} = a_{k+1,l+1}$ for all $k, l \in \mathbb{Z}$, and it is band-limited because $N_{0,r}$ has compact support. It follows that A^{-1} is also Toeplitz and we denote its entries by $b_{kl} = g(|k-l|)$ for some function g. Furthermore, it is known (for example, Theorem 4.3 on page 404 of [6]) that the entries of the inverse of any positive definite band-limited matrix satisfy $|b_{kl}| \leq c\lambda^{|k-l|}$ for some $0 < \lambda < 1$ and c finite. Now, following [19], we write

$$\sum_k g(|l-k|)N_{k,r}(x) = \sum_k g(|l-k|)N_{k-l,r}(x-l) = \sum_k g(|k|)N_{k,r}(x-l),$$

so that

$$2^{-j}\kappa_j(\cdot, y) = \sum_{l\in\mathbb{Z}} N_{j,l,r}(y)H\big(2^j(\cdot) - l\big), \tag{31}$$

where $H(x) = \sum_{k\in\mathbb{Z}} g(|k|)N_{k,r}(x)$ is a function of bounded variation. To see the last claim, note that $N_{0,r}$ is of bounded variation and hence $\|N_{k,r}\|_{\mathrm{TV}} = \|N_{0,r}\|_{\mathrm{TV}}$ (where $\|\cdot\|_{\mathrm{TV}}$ denotes the usual total variation norm) so that $\|H\|_{\mathrm{TV}} \leq \|N_{0,r}\|_{\mathrm{TV}} \times \sum_{k\in\mathbb{Z}} |g(|k|)| < \infty$ because $\sum_k |b_{l,l-k}| \leq \sum_k c\lambda^{|k|} < \infty$. The last fact implies that

$$\mathcal{H} = \big\{H\big(2^j(\cdot) - l\big) : l \in \mathbb{Z}, j \in \mathbb{N} \cup \{0\}\big\}$$

satisfies, for finite constants $B > 1$ and $w \geq 1$,

$$\sup_Q N(\mathcal{H}, L^2(Q), \varepsilon) \leq \left(\frac{B\|H\|_{\mathrm{TV}}}{\varepsilon}\right)^w \qquad \text{for } 0 < \varepsilon < \|H\|_\infty,$$

as proved in [29]. Since $N_{j,0,r}$ is zero if y is not contained in $[0, 2^{-j}r]$, the sum in (31), for fixed y and j, extends over only those l's such that $2^j y - r \leq l < 2^j y$, hence it consists of at most r terms. This implies that $\mathcal{K}$ is contained in the set $\mathcal{H}_r$ of linear combinations of at most r functions from $\mathcal{H}$, with coefficients bounded in absolute value by $\|N_{j,l,r}\|_\infty = \|N_{0,r}\|_\infty < \infty$. Given ε, let $\varepsilon' = \varepsilon/(2r\max(\|H\|_\infty, \|N_{0,r}\|_\infty))$. Let $\alpha_1, \ldots, \alpha_{n_1}$ be an ε'-dense subset of $[-\|N_{0,r}\|_\infty, \|N_{0,r}\|_\infty]$ which, for $\varepsilon' < \|N_{0,r}\|_\infty$, has cardinality $n_1 \leq 3\|N_{0,r}\|_\infty/\varepsilon'$. Furthermore, let $h_1, \ldots, h_{n_2}$ be a subset of $\mathcal{H}$ of cardinality $n_2 = N(\mathcal{H}, L^2(Q), \varepsilon')$ which is ε'-dense in $\mathcal{H}$ in the $L^2(Q)$-metric. It follows that for $\varepsilon' < \min(\|H\|_\infty, \|N_{0,r}\|_\infty)$, every $\sum_{l\in\mathbb{Z}} N_{j,l,r}(y)H(2^j(\cdot) - l)$ is at $L^2(Q)$-distance at most ε from

$\sum_{l=1}^{r} \alpha_{i(l)} h_{i'(l)}$ for some $1 \le i(l) \le n_1$ and $1 \le i'(l) \le n_2$. The total number of such linear combinations is dominated by $(n_1 n_2)^r \le (B'/\varepsilon)^{(w+1)r}$. This shows that the lemma holds for $\varepsilon < 2r \min\{\|H\|_\infty, \|N_{0,r}\|_\infty\} \max\{\|H\|_\infty, \|N_{0,r}\|_\infty\} = 2r\|H\|_\infty \|N_{0,r}\|_\infty = U$, which completes the proof by taking $A = \max(B', U, e)$ (for $\varepsilon \in [U, A]$, one ball covers the whole set). □

Proposition 3. *Let K be as in (1) and assume either that ϕ has compact support and is of bounded p-variation ($p < \infty$) or that ϕ is a Battle–Lemarié scaling function for some $r \ge 1$. Suppose that P has a bounded density p_0. Given $C, T > 0$, there exist finite positive constants $C_1 = C_1(C, K, \|p_0\|_\infty)$ and $C_2 = C_2(C, T, K, \|p_0\|_\infty)$ such that, if*

$$\frac{n}{2^j j} \ge C \quad \text{and} \quad C_1 \sqrt{\frac{2^j j}{n}} \le t \le T,$$

then

$$\Pr\Big\{ \sup_{y \in \mathbb{R}} |p_n(y, j) - E p_n(y, j)| \ge t \Big\} \le \exp\Big(-C_2 \frac{n t^2}{2^j} \Big). \tag{32}$$

Proof. We first prove the Battle–Lemarié wavelet case. If $r > 1$, then the function K is continuous (see the proof of Lemma 1) and therefore the supremum in (32) is over a countable set. That this is also true for $r = 1$ follows from Remark 1 in [14]. We apply Proposition 1 and Lemma 2 to the supremum of the empirical process indexed by the classes of functions

$$\mathcal{K}_j := \{2^{-j} K_j(\cdot, y)/(2\|\Phi\|_\infty) : y \in \mathbb{R}\},$$

where Φ is a function majorizing K (as in (2)) so that $\mathcal{K}_j$ is uniformly bounded by $1/2$. We next bound the second moments $E(2^{-2j} K_j^2(X, y))$. We have, using (2), that

$$\begin{aligned} \int 2^{-2j} K_j^2(x, y) p_0(x)\, \mathrm{d}x &\le \int \Phi^2\big(|2^j(x - y)|\big) p_0(x)\, \mathrm{d}x \\ &\le 2^{-j} \int \Phi^2(|u|) p_0(y + 2^{-j}u)\, \mathrm{d}u \le 2^{-j} \|p_0\|_\infty \|\Phi\|_2^2. \end{aligned} \tag{33}$$

We may hence take $\sigma = \sqrt{2^{-j}\|\Phi\|_2^2 \|p_0\|_\infty}/(2\|\Phi\|_\infty)$ and the result is then a direct consequence of Proposition 1, which applies by Lemma 2. For compactly supported wavelets, the same proof applies, using Lemma 2 (and Remark 1) in [14]. □

Proof of Theorem 1. Using Lemma 2, the first two claims of the theorem follow by the same proof as in [14], Theorem 1 and Remark 4. For the bias term, we argue as in Theorem 8.1 in [17] – using the fact that ϕ_r is $(r-1)$-times differentiable – and obtain, for $p_0 \in \mathsf{C}^t(\mathbb{R})$,

$$|E p_n(x) - p_0(x)| \le 2^{-jt} \|p_0\|_{t,\infty} C, \tag{34}$$

where $C := C(\Phi) = \int \Phi(|u|)|u|^t\, \mathrm{d}u$. □

3.4. An exponential inequality for the distribution function of the linear estimator

The quantity of interest in this subsection is the distribution function F_n^S of the linear projection estimator p_n from (6). More precisely, we will study the stochastic process

$$\sqrt{n}\big(F_n^S(s) - F(s)\big) = \sqrt{n}\int_{-\infty}^{s} \big(p_n(y, j) - p_0(y)\big)\,\mathrm{d}y, \qquad s \in \mathbb{R}.$$

To prove a functional CLT for this process, it turns out that it is easier to compare F_n^S to F_n rather than to F. With $\mathcal{F} = \{1_{(-\infty,s]} : s \in \mathbb{R}\}$, the decomposition

$$(F_n^S - F_n)(s) = (P_n - P)\big(\pi_j(f) - f\big) + \int \big(\pi_j(p_0) - p_0\big)f, \qquad f \in \mathcal{F}, \tag{35}$$

will be useful, since it splits the quantity of interest into a deterministic 'bias' term and an empirical process.

Lemma 3. *Assume that p_0 is a bounded function ($t = 0$) or that $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some $0 < t \le r$. Let $\mathcal{F} = \{1_{(-\infty,s]} : s \in \mathbb{R}\}$. We then have*

$$\left|\int_{\mathbb{R}} \big(\pi_j(p_0) - p_0\big)f\right| \le C2^{-j(t+1)} \tag{36}$$

for some constant C depending only on r and $\|p_0\|_{t,\infty}$.

Proof. Let $\psi := \psi_r$ be the mother wavelet associated with ϕ_r. Since the wavelet series of $p_0 \in L^1(\mathbb{R})$ converges in $L^1(\mathbb{R})$, we have $\pi_j(p_0) - p_0 = -\sum_{l=j}^{\infty}\sum_k \beta_{lk}(p_0)\psi_{lk}$ in the $L^1(\mathbb{R})$-sense and then, since $f = 1_{(-\infty,s]} \in L^\infty(\mathbb{R})$,

$$-\int_{\mathbb{R}} \big(\pi_j(p_0) - p_0\big)f = \int_{\mathbb{R}} \left(\sum_{l=j}^{\infty}\sum_k \beta_{lk}(p_0)\psi_{lk}(x)\right) f(x)\,\mathrm{d}x = \sum_{l=j}^{\infty}\sum_k \beta_{lk}(p_0)\beta_{lk}(f).$$

The lemma now follows from an estimate for the decay of the wavelet coefficients of p_0 and f, namely, the bounds

$$\sup_{f\in\mathcal{F}} \sum_k |\beta_{lk}(f)| \le c2^{-l/2} \quad \text{and} \quad \sup_k |\beta_{lk}(p_0)| \le c'2^{-l(t+1/2)}. \tag{37}$$

The first bound is proved as in the proof of Lemma 3 in [14], noting that the identity before equation (37) in that proof also holds for spline wavelets by their exponential decay property. The second bound follows from

$$\begin{aligned}\sup_k |\beta_{lk}(p_0)| &\le c''2^{-l/2}\|K_{l+1}(p_0) - K_l(p_0)\|_\infty \\ &\le c''2^{-l/2}\big(\|K_l(p_0) - p_0\|_\infty + \|K_{l+1}(p_0) - p_0\|_\infty\big) \le c'2^{-l/2}2^{-lt},\end{aligned}$$

where we used (9.35) in [17] for the first inequality and (34) in the last. □

To control the fluctuations of the stochastic term, one applies Talagrand's inequality to the empirical process indexed by the 'shrinking' classes of functions $\{\pi_j(f) - f : f \in \mathcal{F}\}$. These classes consist of differences of elements in $\mathcal{F}$ and in

$$\mathcal{K}'_j := \left\{ \int_{-\infty}^{t} K_j(\cdot, y)\,\mathrm{d}y : t \in \mathbb{R} \right\},$$

and we have to show that for each j, this class satisfies the entropy condition (17). Again, for ϕ with compact support (and of finite p-variation), this result was proven in Lemma 2 of [14] and we now extend it to the Battle–Lemarié wavelets considered here.

Lemma 4. *Let $\mathcal{K}'_j$ be as above, where ϕ_r is a Battle–Lemarié wavelet for $r \geq 1$. There then exist finite constants $A \geq e$ and $v \geq 2$, independent of j and such that*

$$\sup_{Q} N(\mathcal{K}'_j, L^2(Q), \varepsilon) \leq \left(\frac{A}{\varepsilon}\right)^{v}, \qquad 0 < \varepsilon < 1,$$

where the supremum extends over all Borel probability measures on $\mathbb{R}$.

Proof. In analogy to the proof of Lemma 2, one can write

$$\int_{-\infty}^{t} K_j(\cdot, y)\,\mathrm{d}y = \sum_{l \in \mathbb{Z}} \int_{-\infty}^{t} 2^j N_{j,l,r}(y)\,\mathrm{d}y\, H\big(2^j(\cdot) - l\big)$$

since the series (31) converges absolutely (in view of

$$\sum_l |H(2^j x - l)| \leq \sum_k |g(|k|)| \sum_l N_{k,r}(2^j x - l) \leq r \|N_{0,r}\|_\infty \sum_k |g(|k|)| < \infty).$$

Recall that $N_{j,l,r}$ is supported in the interval $[2^{-j}l, 2^{-j}(r+l)]$. Hence, if $l > 2^j t$, then the last integral is zero. For $l \leq 2^j t - r$, the integral equals the constant $c = \int_{\mathbb{R}} N_{0,r}(y)\,\mathrm{d}y$ and for $l \in [2^j t - r, 2^j t]$, the integral $c_{j,l,r}$ is bounded by c, so this sum, in fact, equals

$$c \sum_{l \leq 2^j t - r} H\big(2^j(\cdot) - l\big) + \sum_{2^j t - r < l < 2^j t} c_{j,l,r} H\big(2^j(\cdot) - l\big).$$

The second sum is contained in the set $\mathcal{H}_r$ from the proof of Lemma 2, which satisfies the required entropy bound independent of j. For the first sum, decompose H into its positive and negative parts, so that the two resulting collections of functions are linearly ordered (in t) by inclusion and are hence a VC-subgraph of index 1; see Theorems 4.2.6 and 4.8.1 in [8]. Moreover, we can take the envelope $r\|N_{0,r}\|_\infty \sum_k |g(|k|)|$ independent of j. Combining entropy bounds, this proves the lemma. □

Combining these observations, one can prove the following inequality, which parallels Theorem 1 of [13] for the classical kernel density estimator, and Lemma 4 of [13] for the wavelet density estimator (with ϕ compactly supported).

Proposition 4. *Let $F_n(s) = \int_{-\infty}^{s} \mathrm{d}P_n$ and $F_n^S(s) := F_n^S(s, j) = \int_{-\infty}^{s} p_n(y, j)\,\mathrm{d}y$, where p_n is as in (6). Assume that the density p_0 of P is a bounded function ($t = 0$) or that $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some t, $0 < t \le r$. Let $j \in \mathbb{Z}$ satisfy $2^{-j} \ge d(\log n/n)$ for some $0 < d < \infty$. There then exist finite positive constants $L := L(\|p_0\|_\infty, K, d)$, $\Lambda_0 := \Lambda_0(\|p_0\|_{t,\infty}, K, d)$ such that for all $n \in \mathbb{N}$ and $\lambda \ge \Lambda_0 \max(\sqrt{j2^{-j}}, \sqrt{n}2^{-j(t+1)})$, we have*

$$\Pr\big(\sqrt{n}\|F_n^S - F_n\|_\infty > \lambda\big) \le L \exp\left\{-\frac{\min(2^j\lambda^2, \sqrt{n}\lambda)}{L}\right\}.$$

Proof. Given the preceding lemmas, the proposition follows from Talagrand's inequality applied to the class $\{\pi_j(1_{(-\infty,x]}) - 1_{(-\infty,x]}\}$ in the same way as in the proof of Lemma 4 in [14], so we omit it. □

3.5. Proof of Theorem 3

We can now prove the main result, Theorem 3. We will prove it only for Battle–Lemarié wavelets. For compactly supported wavelets, the proof is exactly the same, replacing the results from steps (I) and (II) below and from Sections 3.3 and 3.4 for spline wavelets by the corresponding ones for compactly supported wavelets obtained in [14]. Also, uniformity in p_0 – which is proved by controlling the respective constants – is left implicit in the derivations. We start with some preliminary observations.

(I) Since, uniformly in $j \in \mathcal{J}$, we have $n/(2^j j) > c \log n$ for some $c > 0$ independent of n, we have from Theorem 1 that

$$E\|p_n(j) - Ep_n(j)\|_\infty^p \le D^p \left(\frac{2^j j}{n}\right)^{p/2} := D^p \sigma^p(j, n) \tag{38}$$

for every $j \in \mathcal{J}$, $1 \le p < \infty$ and some $0 < D < \infty$ depending only on $\|p_0\|_\infty$ and Φ.

For the bias, we recall from (34) that for $0 < t \le r$,

$$|Ep_n(y, j) - p_0(y)| \le 2^{-jt}\|p_0\|_{t,\infty} C(\Phi) := B(j, p_0). \tag{39}$$

If the density p_0 is only uniformly continuous, then one still has from (2) and integrability of Φ that, uniformly in $y \in \mathbb{R}$,

$$|Ep_n(y, j) - p_0(y)| \le \left|\int |\Phi(|u|)||p_0(y - 2^{-j}u) - p_0(y)|\,\mathrm{d}u\right| := B(j, p_0) = \mathrm{o}(1). \tag{40}$$

(II) Define $\tilde{M} := \tilde{M}_n = C\|p_n(j_{\max})\|_\infty$ and set $C = 49\|\Phi\|_2^2$. Also, define $M = C\|p_0\|_\infty$ for the same C. We need to control the probability that $\tilde{M} > 1.01M$ or $\tilde{M} < 0.99M$ if p_0 is uniformly

continuous. For some $0 < L < \infty$ and n large enough, we have

$$\begin{aligned}
&\Pr(|\tilde{M} - M| > 0.01C\|p_0\|_\infty) \\
&\quad = \Pr\bigl(|\|p_n(j_{\max})\|_\infty - \|p_0\|_\infty| > 0.01\|p_0\|_\infty\bigr) \\
&\quad \le \Pr\bigl(\|p_n(j_{\max}) - p_0\|_\infty > 0.01\|p_0\|_\infty\bigr) \\
&\quad \le \Pr\bigl(\|p_n(j_{\max}) - Ep_n(j_{\max})\|_\infty > 0.01\|p_0\|_\infty - B(j_{\max}, p_0)\bigr) \\
&\quad \le \Pr\bigl(\|p_n(j_{\max}) - Ep_n(j_{\max})\|_\infty > 0.009\|p_0\|_\infty\bigr) \\
&\quad \le \exp\left\{-\frac{(\log n)^2}{L}\right\},
\end{aligned}$$

by Proposition 3 and step (I). Furthermore, there exists a constant L' such that $E\tilde{M} \le L'$ for every n, in view of

$$E\|p_n(j_{\max})\|_\infty \le E\|p_n(j_{\max}) - Ep_n(j_{\max})\|_\infty + \|Ep_n(j_{\max})\|_\infty \le c + \|\Phi\|_1\|p_0\|_\infty,$$

where we have used (2) and (38).

(III) We need some observations on the Rademacher processes used in the definition of $\hat{j}_n$. First, for the symmetrized empirical measure $\tilde{P}_n = 2n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}$, we have

$$R(n, j) = \|\pi_j(\tilde{P}_n)\|_\infty = \|\pi_j(\pi_l(\tilde{P}_n))\|_\infty \le \|\pi_j\|'_\infty R(n, l) \le B(\phi)R(n, l) \tag{41}$$

for every $l > j$. Here, $\|\pi_j\|'_\infty$ is the operator norm in $L^\infty(\mathbb{R})$ of the projection π_j, which admits bounds $B(\phi)$ independent of j. (Clearly, π_j acts on finite signed measures μ by duality, taking values in $L^\infty(\mathbb{R})$ since $|\pi_j(\mu)| = |\int K_j(\cdot, y)\,\mathrm{d}\mu(y)| \le 2^j\|\Phi\|_\infty|\mu|(\mathbb{R})$.) See Remark 3 for details on how to obtain $B(\phi)$. Furthermore, for $j < l$,

$$T(n, j, l) \le R(n, j) + R(n, l) \le \bigl(1 + B(\phi)\bigr)R(n, l) \tag{42}$$

and the same inequality holds for the Rademacher expectations of $T(n, j, l)$. We also record the following bound for the (full) expectation of $R(n, l)$, $l \in \mathcal{J}$: using inequality (27) and the variance computation (33), we have that there exists a constant L depending only on $\|p_0\|_\infty$ and Φ such that, for every $l \in \mathcal{J}$, $ER(n, l) \le L\sqrt{2^l l/n}$.

Proof of (11). Let $\mathcal{F} = \{1_{(-\infty,s]} : s \in \mathbb{R}\}$ and let $f \in \mathcal{F}$. We have

$$\sqrt{n}\int \bigl(p_n(\hat{j}_n) - p_0\bigr)f = \sqrt{n}\int \bigl(p_n(j_{\max}) - p_0\bigr)f + \sqrt{n}\int \bigl(p_n(\hat{j}_n) - p_n(j_{\max})\bigr)f.$$

The first term satisfies the CLT from Theorem 2 for the linear estimator with $j_n = j_{\max}$. We now show that the second term converges to zero in probability. First, observe that

$$p_n(\hat{j}_n)(y) - p_n(j_{\max})(y) = P_n\bigl(K_{\hat{j}_n}(\cdot, y) - K_{j_{\max}}(\cdot, y)\bigr) = -\sum_{l=\hat{j}_n}^{j_{\max}-1}\sum_k \hat{\beta}_{lk}\psi_{lk}(y),$$

with convergence in $L^1(\mathbb{R})$. Next, we have, by (9.35) in [17], for all $l \in [\hat{j}_n, j_{\max} - 1]$ and all k, by the definition of $\hat{j}_n$, that for some $0 < D' < \infty$,

$$\begin{aligned}(1/D')2^{l/2}|\hat{\beta}_{lk}| &\le \sup_{y\in\mathbb{R}} |P_n(K_{l+1}(\cdot, y)) - P_n(K_l(\cdot, y))| = \|p_n(l+1) - p_n(l)\|_\infty \\ &\le \|p_n(l+1) - p_n(\hat{j}_n)\|_\infty + \|p_n(l) - p_n(\hat{j}_n)\|_\infty \\ &\le \big(1 + B(\phi)\big)\big(R(n, l+1) + R(n, l)\big) + 3\sqrt{\tilde{M}2^l l/n},\end{aligned}$$

in the case $\hat{j}_n = \bar{j}_n$, also using the inequality $T(n, \bar{j}_n, l) \le (1 + B(\phi))R(n, l)$ for $l \ge \bar{j}_n$; see (42). Consequently, uniformly in $f \in \mathcal{F}$,

$$\begin{aligned}&E\left|\int \big(p_n(\hat{j}_n) - p_n(j_{\max})\big) f\right| \\ &= E\left|\sum_{l=\hat{j}_n}^{j_{\max}-1} \sum_k \hat{\beta}_{lk} \int \psi_{lk}(y) f(y)\,\mathrm{d}y\right| \\ &\le E \sum_{l=j_{\min}}^{j_{\max}-1} D' 2^{-l/2}\big(\big(B(\phi) + 1\big)\big(R(n, l+1) + R(n, l)\big) + 3\sqrt{\tilde{M}2^l l/n}\big) \sum_k |\beta_{lk}(f)| \\ &\le \left(\frac{D''}{\sqrt{n}}\right) \sum_{l=j_{\min}}^{j_{\max}-1} 2^{-l/2}\sqrt{l} = \mathrm{o}\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

using the moment bounds in (II) and (III), $\hat{j}_n \ge j_{\min} \to \infty$ as $n \to \infty$ (by definition of $\mathcal{J}$) and the fact that $\sup_{f\in\mathcal{F}} \sum_k |\beta_{lk}(f)| \le c2^{-l/2}$ by (37) for some constant c. □

Proof of (12) and (13). The proof of the case $t = 0$ follows from a simple modification of the arguments below as in Theorem 2 of [13], so we omit it. (In this case, one defines j^* as $j_{\max}$ if $t = 0$ so that only the case $\hat{j}_n \le j^*$ has to be considered.) For $t > 0$, define $j^* := j(p_0)$ by the balance equation

$$j^* = \min\big\{j \in \mathcal{J} : B(j, p_0) \le \sqrt{2\log 2}\|p_0\|_\infty^{1/2}\|\Phi\|_2\sigma(j, n)\big\}. \tag{43}$$

Using the results from (I), it is easily verified that $2^{j^*} \simeq (n/\log n))^{1/(2t+1)}$ if $p_0 \in \mathsf{C}^t(\mathbb{R})$ for some $0 < t \le r$ and that

$$\sigma(j^*, n) = \mathrm{O}\left(\left(\frac{\log n}{n}\right)^{t/(2t+1)}\right)$$

is the rate of convergence required in (13).

We will consider the cases $\{\hat{j}_n \le j^*\}$ and $\{\hat{j}_n > j^*\}$ separately. First, if $\hat{j}_n$ is $\bar{j}_n$, then we have, by the definition of $\bar{j}_n$, (42), the definitions of M and j^*, (38) and the moment bound in (III),

$$\begin{aligned} &E\|p_n(\bar{j}_n) - p_0\|_\infty I_{\{\bar{j}_n \le j^*\}\cap\{\tilde{M}\le 1.01M\}} \\ &\quad\le E\big(\|p_n(\bar{j}_n) - p_n(j^*)\|_\infty + E\|p_n(j^*) - p_0\|_\infty\big) I_{\{\bar{j}_n \le j^*\}\cap\{\tilde{M}\le 1.01M\}} \\ &\quad\le \big(B(\phi)+1\big) ER(n, j^*) + \sqrt{1.01M}\sigma(j^*, n) + \|p_n(j^*) - p_0\|_\infty \\ &\quad\le B'\sqrt{\frac{2^{j^*} j^*}{n}} + B''\sigma(j^*, n) = \mathrm{O}(\sigma(j^*, n)). \end{aligned} \tag{44}$$

If $\hat{j}_n$ is $\tilde{j}_n$, then one has the same bound (without even using (42)).

Also, by the results in (I) and (II), we have

$$\begin{aligned} &E\|p_n(\hat{j}_n) - p_0\|_\infty I_{\{\hat{j}_n \le j^*\}\cap\{\tilde{M} > 1.01M\}} \\ &\quad\le \sum_{j\in\mathcal{J}: j\le j^*} E\big([\|p_n(j) - Ep_n(j)\|_\infty + B(j, p_0)] I_{\{\hat{j}_n = j\}} I_{\{\tilde{M}>1.01M\}}\big) \\ &\quad\le c\log n[D\sigma(j^*, n) + B(j_{\min}, p_0)] \cdot \sqrt{E1_{\{\tilde{M}>1.01M\}}} \\ &\quad= \mathrm{o}\left((\log n)\sqrt{\exp\left\{-\frac{(\log n)^2}{L}\right\}}\right) = \mathrm{o}(\sigma(j^*, n)). \end{aligned}$$

We now turn to $\{\hat{j}_n > j^*\}$. First,

$$\begin{aligned} &E\|p_n(\hat{j}_n) - p_0\|_\infty I_{\{\hat{j}_n > j^*\}\cap\{\tilde{M}<0.99M\}} \\ &\quad\le \sum_{j\in\mathcal{J}:\, j> j^*} E\big([\|p_n(j) - Ep_n(j)\|_\infty + B(j, p_0)] I_{\{\hat{j}_n = j\}} I_{\{\tilde{M}<0.99M\}}\big) \\ &\quad\le c'\log n[D\sigma(j_{\max}, n) + B(j^*, p_0)] \cdot \sqrt{EI_{\{\tilde{M}<0.99M\}}} \\ &\quad= \mathrm{O}\left(\sqrt{(\log n)\exp\left\{-\frac{(\log n)^2}{L}\right\}}\right) = \mathrm{o}(\sigma(j^*, n)), \end{aligned}$$

again by the results in (I) and (II), and, second, for any $1 < p < \infty$, $1/p + 1/q = 1$, using (38) and the definition of j^*, we have

$$\begin{aligned} &E\|p_n(\hat{j}_n) - p_0\|_\infty I_{\{\hat{j}_n > j^*\}\cap\{0.99M\le\tilde{M}\}} \\ &\quad\le \sum_{j\in\mathcal{J}: j> j^*} \big(E\|p_n(j) - p_0\|_\infty^p\big)^{1/p} \big(EI_{\{\hat{j}_n = j\}\cap\{0.99M\le\tilde{M}\}}\big)^{1/q} \\ &\quad\le \sum_{j\in\mathcal{J}: j> j^*} D'\sigma(j, n) \cdot \Pr(\{\hat{j}_n = j\}\cap\{0.99M\le\tilde{M}\})^{1/q}. \end{aligned}$$

We show below that for n large enough, some constant c, some $\delta > 0$ and some $q > 1$,

$$\Pr(\{\hat{j}_n = j\} \cap \{0.99M \le \tilde{M}\}) \le c2^{-j(q/2+\delta)}, \tag{45}$$

which gives the bound

$$\sum_{j \in \mathcal{J} : j > j^*} D''\sigma(j,n) \cdot 2^{-j/2 - j\delta/q} = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right) = \mathrm{o}(\sigma(j^*,n)),$$

completing the proof, modulo verification of (45).

To verify (45), we split the proof into two cases. Pick any $j \in \mathcal{J}$ such that $j > j^*$ and denote by j^- the previous element in the grid (that is, $j^- = j - 1$).

Case I: $\hat{j}_n = \bar{j}_n$. We have

$$\begin{aligned}&\Pr(\{\bar{j}_n = j\} \cap \{0.99M \le \tilde{M}\})\\ &\quad\le \sum_{l \in \mathcal{J} : l \ge j} \Pr\big(\|p_n(j^-) - p_n(l)\|_\infty > T(n, j^-, l) + \sqrt{0.99M}\sigma(l,n)\big).\end{aligned}$$

We first observe that

$$\begin{aligned}&\|p_n(j^-) - p_n(l)\|_\infty\\ &\quad\le \|p_n(j^-) - p_n(l) - Ep_n(j^-) + Ep_n(l)\|_\infty + B(j^-, p_0) + B(l, p_0),\end{aligned} \tag{46}$$

where, setting $\sqrt{2\log 2}\|p_0\|_\infty^{1/2}\|\Phi\|_2 =: U(p_0, \Phi)$,

$$B(j^-, p_0) + B(l, p_0) \le 2B(j^*, p_0) \le 2U(p_0, \Phi)\sigma(j^*, n) \le 2U(p_0, \Phi)\sigma(l, n),$$

by definition of j^* and since $l > j^- \ge j^*$. Consequently, the lth probability in the last sum is bounded by

$$\begin{aligned}&\Pr\big(\|p_n(j^-) - p_n(l) - Ep_n(j^-) + Ep_n(l)\|_\infty\\ &\qquad> T(n, j^-, l) + \big(\sqrt{0.99M} - 2U(p_0, \Phi)\big)\sigma(l,n)\big)\end{aligned} \tag{47}$$

and we now apply Corollary 1 to this bound. Define the class of functions

$$\mathcal{F} := \mathcal{F}_{j^-,l} = \big\{2^{-l}\big(K_{j^-}(\cdot, y) - K_l(\cdot, y)\big)/(4\|\Phi\|_\infty)\big\},$$

which is uniformly bounded by $1/2$ and satisfies (17) for some A and v independent of l and j^-, by Lemma 2 (and a computation on covering numbers). We compute σ, using (33) and $l > j^-$:

$$\begin{aligned}\big(2^{-l}E(K_{j^-} - K_l)(X, y)\big)^2 &\le 2^{-2l+1}\big(EK_{j^-}^2(X, y) + EK_l^2(X, y)\big)\\ &\le 2^{-2l+1}\|\Phi\|_2^2\|p_0\|_\infty(2^{j^-} + 2^l) \le 3 \cdot 2^{-l}\|\Phi\|_2^2\|p_0\|_\infty,\end{aligned}$$

so that we can take $\sigma^2 = 3 \cdot 2^{-l}\|\Phi\|_2^2\|p_0\|_\infty/(16\|\Phi\|_\infty^2)$. The probability in (47) is then equal to

$$\begin{aligned}
&\Pr\Bigg(\frac{2^l 4\|\Phi\|_\infty}{n}\left\|\sum_{i=1}^n f(X_i) - Pf\right\|_{\mathcal{F}} \\
&\qquad > \frac{2^l 4\|\Phi\|_\infty}{n} 2\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} + \big(\sqrt{0.99M} - 2U(p_0,\Phi)\big)\sigma(l,n)\Bigg) \\
&\quad = \Pr\Bigg(\left\|\sum_{i=1}^n f(X_i) - Pf\right\|_{\mathcal{F}} > 2\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} + 3\frac{n(\sqrt{0.99M} - 2U(p_0,\Phi))\sigma(l,n)}{3\cdot 2^l\cdot 4\|\Phi\|_\infty}\Bigg).
\end{aligned}$$

Since $n\sigma^2/\log(1/\sigma) \simeq n/(2^l l) \to \infty$ uniformly in $l \in \mathcal{J}$, there exists $\lambda_n \to \infty$ independent of l such that (19) is satisfied and the choice

$$t = \frac{n(\sqrt{0.99M} - 2U(p_0,\Phi))\sigma(l,n)}{3\cdot 2^l\cdot 4\|\Phi\|_\infty}$$

is admissible in Corollary 1 for $c_2(\lambda_n) = 1 + 120\lambda_n^{-1} + 10{,}800\lambda_n^{-2}$. Hence, using Corollary 1, the last probability is bounded by

$$\leq 2\exp\left(-\frac{n^2(\sqrt{0.99M} - 2U(p_0,\Phi))^2(2^l l/n)16\|\Phi\|_\infty^2}{9\cdot 6.3\cdot c_2(\lambda_n)2^{2l}n2^{-l}\|\Phi\|_2^2\|p_0\|_\infty 16\|\Phi\|_\infty^2}\right) \leq 2^{-l((q/2)+\delta)} \tag{48}$$

for some $\delta > 0$ and $q > 1$, by the definition of M. Since $\sum_{l\in\mathcal{J}:l\geq j} 2^{-l(q/2)+\delta)} \leq c2^{-j((q/2)+\delta)}$, we have proven (45).

Case II: $\hat{j}_n = \tilde{j}_n$. The proof reduces to the previous case since, by inequality (42), one has

$$\begin{aligned}
&\Pr(\{\tilde{j}_n^\varepsilon = j\} \cap \{0.99M \leq \tilde{M}\}) \\
&\quad \leq \sum_{l\in\mathcal{J}:l\geq j} \Pr\big(\|p_n(j^-) - p_n(l)\|_\infty > \big(B(\phi)+1\big)R(n,l) + \sqrt{0.99M}\sigma(l,n)\big) \\
&\quad \leq \sum_{l\in\mathcal{J}:l\geq j} \Pr\big(\|p_n(j^-) - p_n(l)\|_\infty > T(n,j^-,l) + \sqrt{0.99M}\sigma(l,n)\big).
\end{aligned}$$

□

Acknowledgements

We would like to thank Patricia Reynaud-Bouret and Benedikt Pötscher for helpful comments. The idea of using Rademacher thresholds in Lepski's method arose from a conversation with Patricia Reynaud-Bouret.

References

[1] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[2] Bartlett, P., Boucheron, S. and Lugosi, G. (2002). Model selection and error estimation. *Mach. Learn.* **48** 85–113.

[3] Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications* (E. Giné, C. Houdré and D. Nualart, eds.). *Progr. Probab.* **56** 213–247. Boston: Birkhäuser. MR2073435

[4] Daubechies, I. (1992). *Ten Lectures on Wavelets*. *CBMS-NSF Reg. Conf. Ser. in Appl. Math.* **61**. Philadelphia, PA: SIAM. MR1162107

[5] Deheuvels, P. (2000). Uniform limit laws for kernel density estimators on possibly unbounded intervals. In *Recent Advances in Reliability Theory* (N. Limnios and M. Nikulin, eds.) 477–492. Boston: Birkhäuser. MR1783500

[6] DeVore, R.A. and Lorentz, G.G. (1993). *Constructive Approximation*. Berlin: Springer. MR1261635

[7] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974

[8] Dudley, R.M. (1999). *Uniform Central Limit Theorems.* Cambridge: Cambridge Univ. Press. MR1720712

[9] Einmahl, U. and Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13** 1–37. MR1744994

[10] Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: Rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist.* **37** 503–522. MR1876841

[11] Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **38** 907–921. MR1955344

[12] Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143–1216. MR2243881

[13] Giné, E. and Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* **143** 569–596. MR2475673

[14] Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.* **37** 1605–1646. MR2546757

[15] Goldenshluger, A. and Lepski, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** 41–71. MR2449122

[16] Golubev, Y., Lepski, O. and Levit, B. (2001). On adaptive estimation for the sup-norm losses. *Math. Methods Statist.* **10** 23–37. MR1841807

[17] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics* **129**. New York: Springer. MR1618204

[18] Huang, S.-Y. (1999). Density estimation by wavelet-based reproducing kernels. *Statist. Sinica* **9** 137–151. MR1678885

[19] Huang, S.-Y. and Studden, W.J. (1993). Density estimation using spline projection kernels. *Comm. Statist. Theory Methods* **22** 3263–3285. MR1245544

[20] Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statist. Probab. Lett.* **13** 15–24. MR1147634

[21] Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33** 1060–1077. MR2135312

[22] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory* **47** 1902–1914. MR1842526
[23] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442
[24] Korostelev, A. and Nussbaum, M. (1999). The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli* **5** 1099–1118. MR1735786
[25] Ledoux, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs* **89**. Providence, RI: Amer. Math. Soc. MR1849347
[26] Lepski, O.V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697. MR1147167
[27] Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–884. MR1782276
[28] Meyer, Y. (1992). *Wavelets and Operators I.* Cambridge: Cambridge Univ. Press. MR1228209
[29] Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Ann. Statist.* **15** 780–799. MR0888439
[30] Shadrin, A.Y. (2001). The L_∞-norm of the L_2-spline projector is bounded independently of the knot sequence: A proof of de Boor's conjecture. *Acta Math.* **187** 59–137. MR1864631
[31] Schumaker, L.L. (1993). *Spline Functions: Basic Theory.* Malabar: Krieger. MR1226234
[32] Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22** 28–76. MR1258865
[33] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006
[34] Tsybakov, A.B. (1998). Pointwise and sup-norm sharp adaptive estimation of the functions on the Sobolev classes. *Ann. Statist.* **26** 2420–2469. MR1700239

Received May 2008 and revised August 2009