The Annals of Statistics
2011, Vol. 39, No. 5, 2383–2409
DOI: 10.1214/11-AOS903

ON ADAPTIVE INFERENCE AND CONFIDENCE BANDS

BY MARC HOFFMANN AND RICHARD NICKL

ENSAE-CREST and University of Cambridge

The problem of existence of adaptive confidence bands for an unknown density f that belongs to a nested scale of Hölder classes over $\mathbb{R}$ or $[0, 1]$ is considered. Whereas honest adaptive inference in this problem is impossible already for a pair of Hölder balls $\Sigma(r), \Sigma(s), r \neq s$, of fixed radius, a nonparametric distinguishability condition is introduced under which adaptive confidence bands can be shown to exist. It is further shown that this condition is necessary and sufficient for the existence of honest asymptotic confidence bands, and that it is strictly weaker than similar analytic conditions recently employed in Giné and Nickl [*Ann. Statist.* **38** (2010) 1122–1170]. The exceptional sets for which honest inference is not possible have vanishingly small probability under natural priors on Hölder balls $\Sigma(s)$. If no upper bound for the radius of the Hölder balls is known, a price for adaptation has to be paid, and near-optimal adaptation is possible for standard procedures. The implications of these findings for a general theory of adaptive inference are discussed.

1. Introduction.

One of the intriguing problems in the paradigm of adaptive nonparametric function estimation as developed in the last two decades is what one could call the "*hiatus*" between estimation and inference, or, to be more precise, between the existence of adaptive risk bounds and the nonexistence of adaptive confidence statements. In a nutshell the typical situation in nonparametric statistics could be described as follows: one is interested in a functional parameter f that could belong either to Σ or to Σ', two sets that can be distinguished by a certain "structural property," such as smoothness, with the possibility that $\Sigma \subset \Sigma'$. Based on a sample whose distribution depends on f, one aims to find a statistical procedure that adapts to the unknown structural property, that is, that performs optimally without having to know whether $f \in \Sigma$ or $f \in \Sigma'$. Now while such procedures can often be proved to exist, the statistician cannot take advantage of this optimality for inference: To cite Robins and van der Vaart [29], "An adaptive estimator can adapt to an underlying model, but does not reveal which model it adapts to, with the consequence that nonparametric confidence sets are necessarily much larger than the actual discrepancy between an adaptive estimator and the true parameter."

We argue in this article that adaptive inference is possible if the structural property that defines Σ and Σ' is statistically identifiable, by which we shall mean here

Received September 2010; revised March 2011.
MSC2010 subject classifications. Primary 62G15; secondary 62G10, 62G05.
Key words and phrases. Adaptive confidence sets, nonparametric hypothesis testing.

that the nonparametric hypotheses $H_0: f \in \Sigma$ and $H_1: f \in \Sigma' \setminus \Sigma$ are asymptotically consistently distinguishable (in the sense of Ingster [16–18]). In common adaptation problems this will necessitate that certain unidentified parts of the parameter space be removed, in other words, that the alternative hypothesis H_1 be restricted to a subset $\tilde{\Sigma}$ of $\Sigma' \setminus \Sigma$. One is in turn interested in choosing $\tilde{\Sigma}$ as large as possible, which amounts to imposing minimal identifiability conditions on the parameter space. We shall make these ideas rigorous in one key example of adaptive inference: confidence bands for nonparametric density functions f that adapt to the unknown smoothness of f. The general approach, however, is not specific to this example as we shall argue at the end of this introduction, and the heuristic mentioned above is valid more generally.

The interest in the example of confidence bands comes partly from the fact that the discrepancy between estimation and inference in this case is particularly pronounced. Let us highlight the basic problem in a simple "toy adaptation" problem. Consider $X_1, \dots, X_n$ independent and identically distributed random variables taking values in $[0, 1]$ with common probability density function f and joint law $\Pr_f$. We are interested in the existence of confidence bands for f that are adaptive over two nested balls in the classical Hölder spaces $\mathcal{C}^s([0, 1]) \subset \mathcal{C}^r([0, 1])$, $s > r$, of smooth functions with norm given by $\|\cdot\|_{s,\infty}$; see Definition 1 below. Define the class of densities

$$\Sigma(s) := \Sigma(s, B) = \left\{ f : [0, 1] \to [0, \infty), \int_0^1 f(x)\,dx = 1, \|f\|_{s,\infty} \le B \right\} \tag{1.1}$$

and note that $\Sigma(s) \subset \Sigma(r)$ for $s > r$. We shall assume throughout that $B \ge 1$ to ensure that $\Sigma(s)$ is nonempty.

A confidence band $C_n = C_n(X_1, \dots, X_n)$ is a family of random intervals

$$\{C_n(y) = [c_n(y), c_n'(y)]\}_{y \in [0,1]}$$

that contains graphs of densities $f : [0, 1] \to [0, \infty)$. We denote by $|C_n| = \sup_{y \in [0,1]} |c_n'(y) - c_n(y)|$ the maximal diameter of C_n. Following Li [24] the band C_n is called asymptotically *honest* with level α for a family of probability densities $\mathcal{P}$ if it satisfies the asymptotic coverage inequality

$$\liminf_n \inf_{f \in \mathcal{P}} \Pr_f\bigl(f(y) \in C_n(y)\ \forall y \in [0, 1]\bigr) \ge 1 - \alpha. \tag{1.2}$$

We shall usually only write $\Pr_f(f \in C_n)$ for the coverage probability if no confusion may arise. Note that $\mathcal{P}$ may (and later typically will have to) depend on the sample size n. Suppose the goal is to find a confidence band that is honest for the class

$$\mathcal{P}^{\mathrm{all}} := \Sigma(s) \cup \Sigma(r) = \Sigma(r)$$

and that is simultaneously *adaptive* in the sense that the expected diameter $E_f|C_n|$ of C_n satisfies, for every n (large enough),

$$\sup_{f \in \Sigma(s)} E_f|C_n| \le L r_n(s), \qquad \sup_{f \in \Sigma(r)} E_f|C_n| \le L r_n(r), \tag{1.3}$$

where L is a finite constant independent of n and where

$$r_n(s) = \left(\frac{\log n}{n}\right)^{s/(2s+1)}.$$

Indeed even if s were known no band could have expected diameter of smaller order than $r_n(s)$ uniformly over $\Sigma(s)$ (e.g., Proposition 1 below), so that we are looking for a band that is asymptotically honest for $\mathcal{P}^{\text{all}}$ and that shrinks at the fastest possible rate over $\Sigma(s)$ and $\Sigma(r)$ simultaneously. It follows from Theorem 2 in Low [26] (see also [4, 8]) that such bands do not exist.

THEOREM 1 (Low). *Any confidence band C_n that is honest over $\mathcal{P}^{\text{all}}$ with level $\alpha < 1$ necessarily satisfies*

$$\lim_n \sup_{f \in \Sigma(s)} \frac{E_f|C_n|}{r_n(s)} = \infty.$$

The puzzling fact is that this is in stark contrast to the situation in estimation: adaptive estimators $\hat{f}_n$ such as those based on Lepski's method [23] or wavelet thresholding [7] can be shown to satisfy simultaneously

$$\sup_{f \in \Sigma(s)} E_f \|\hat{f}_n - f\|_\infty = O(r_n(s)), \qquad \sup_{f \in \Sigma(r)} E_f \|\hat{f}_n - f\|_\infty = O(r_n(r));$$

see [10, 11, 13] and Theorem 5 below. So while $\hat{f}_n$ adapts to the unknown smoothness s, Theorem 1 reflects the fact that knowledge of the smoothness is still not accessible for the statistician.

Should we therefore abstain from using adaptive estimators such as $\hat{f}_n$ for inference? Giné and Nickl [12] recently suggested a new approach to this problem, partly inspired by Picard and Tribouley [28]. In [12] it was shown that one can construct confidence bands C_n and subsets $\bar{\Sigma}(\varepsilon, r) \subset \Sigma(r)$, defined by a concrete analytical condition that involves the constant $\varepsilon > 0$, such that C_n is asymptotically honest for

$$\mathcal{P}_\varepsilon = \Sigma(s) \cup \bar{\Sigma}(\varepsilon, r)$$

for every fixed $\varepsilon > 0$, and such that C_n is adaptive in the sense of (1.3). Moreover, these subsets were shown to be topologically generic in the sense that the set

$$\{f \in \Sigma(r) \text{ but } f \notin \bar{\Sigma}(\varepsilon, r) \text{ for any } \varepsilon > 0\}$$

that was removed is *nowhere dense* in the Hölder norm topology of $\mathcal{C}^r$ (in fact in the relevant trace topology on densities). This says that the functions $f \in \mathcal{P}^{\text{all}}$ that prevent adaptation in Theorem 1 are in a certain sense negligible.

In this article we shall give a more statistical interpretation of when, and if, why, adaptive inference is possible over certain subsets of Hölder classes. Our approach will also shed new light on why adaptation is possible over the sets $\bar{\Sigma}(\varepsilon, r)$. Define, for $s > r$, the following class:

$$\tilde{\Sigma}(r, \rho_n) := \tilde{\Sigma}(r, s, \rho_n, B) = \Big\{ f \in \Sigma(r, B) : \inf_{g \in \Sigma(s)} \|g - f\|_\infty \geq \rho_n \Big\}, \tag{1.4}$$

where ρ_n is a sequence of nonnegative real numbers. Clearly $\tilde{\Sigma}(r, 0) = \Sigma(r)$, but if $\rho_n > 0$, then we are removing those elements from $\Sigma(r)$ that are not separated away from $\Sigma(s)$ in sup-norm distance by at least ρ_n. Inspection of the proof of Theorem 2 shows that the set removed from $\Sigma(r) \setminus \Sigma(s)$ is nonempty as soon as $\rho_n > 0$.

Similar to above we are interested in finding a confidence band that is honest over the class

$$\mathcal{P}(\rho_n) := \Sigma(s) \cup \tilde{\Sigma}(r, \rho_n),$$

and that is *adaptive* in the sense of (1.3), in fact only in the sense that

$$\sup_{f \in \Sigma(s)} E_f |C_n| \leq L r_n(s), \qquad \sup_{f \in \tilde{\Sigma}(r, \rho_n)} E_f |C_n| \leq L r_n(r) \tag{1.5}$$

for every n (large enough). We know from Low's results that this is impossible if $\rho_n = 0$, but the question arises as to whether this changes if $\rho_n > 0$, and if so, what the smallest admissible choice for ρ_n is.

It was already noted or implicitly used in [1, 5, 15, 19, 29] that there is a generic connection between adaptive confidence sets and minimax distinguishability of certain nonparametric hypotheses. In our setting consider, for instance, testing the hypothesis

$$H_0 : f_0 = 1 \quad \text{against} \quad H_1 : f_0 \in \mathcal{M}, \qquad \mathcal{M} \text{ finite}, \mathcal{M} \subset \tilde{\Sigma}(r, \rho_n).$$

As we shall see in the proof of Theorem 2 below, an adaptive confidence band over $\mathcal{P}(\rho_n)$ can be used to test any such hypothesis consistently, and intuitively speaking an adaptive confidence band should thus only exist if ρ_n is of larger order than the minimax rate of testing between H_0 and H_1 in the sense of Ingster [16, 17]; see also the monograph [18]. For confidence bands a natural separation metric is the supremum-norm (see, however, also the discussion in the last paragraph of the Introduction), and an exploration of the corresponding testing problems gives our main result, which confirms this intuition and shows moreover that this lower bound is sharp up to constants at least in the case where B is known.

THEOREM 2. *Let $s > r > 0$. An adaptive and honest confidence band over*

$$\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$$

exists if and only if ρ_n is greater than or equal to the minimax rate of testing between $H_0 : f_0 \in \Sigma(s)$ and $H_1 : f_0 \in \tilde{\Sigma}(r, \rho_n)$, and this rate equals $r_n(r)$. More precisely:

(a) *Suppose that* C_n *is a confidence band that is asymptotically honest with level* $\alpha < 0.5$, *over* $\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$ *and that is adaptive in the sense of* (1.5). *Then necessarily*

$$\liminf_n \frac{\rho_n}{r_n(r)} > 0.$$

(b) *Suppose* B, r, s *and* $0 < \alpha < 1$ *are given. Then there exists a sequence* ρ_n *satisfying*

$$\limsup_n \frac{\rho_n}{r_n(r)} < \infty$$

and a confidence band $C_n = C_n(B, r, s, \alpha; X_1, \dots, X_n)$ *that is asymptotically honest with level* α *and adaptive over* $\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$ *in the sense of* (1.5).

(c) *Claims* (a) *and* (b) *still hold true if* $\Sigma(s)$ *is replaced by the set*

$$\Big\{ f \in \Sigma(s), \inf_{g \in \Sigma(t)} \|g - f\|_\infty \geq B r_n(s)/2 \Big\}$$

for any $t > s$.

The last claim shows that the situation does not change if one removes similar subsets from the smaller Hölder ball $\Sigma(s)$, in particular removing the standard null-hypothesis $f_0 = 1$ used in the nonparametric testing literature, or other very smooth densities, cannot improve the lower bound for ρ_n.

Part (b) of Theorem 2 implies the following somewhat curious corollary: since any $f \in \Sigma(r) \setminus \Sigma(s)$ satisfies $\inf_{g \in \Sigma(s)} \|g - f\|_\infty > 0$ (note that $\Sigma(s)$ is $\| \cdot \|_\infty$-compact), we conclude that $f \in \tilde{\Sigma}(r, L r_n(r))$ for every $L > 0$, $n \geq n_0(f, r, L)$ large enough. We thus have:

COROLLARY 1. *There exists a "dishonest" adaptive confidence band* $C_n := C_n(B, r, s, \alpha; X_1, \dots, X_n)$ *that has asymptotic coverage for every fixed* $f \in \mathcal{P}^{\text{all}}$; *that is*, C_n *satisfies*

$$\liminf_n \Pr_f(f \in C_n) \geq 1 - \alpha \qquad \forall f \in \mathcal{P}^{\text{all}}$$

and

$$f \in \Sigma(s) \quad \Rightarrow \quad E_f|C_n| = O(r_n(s)),$$
$$f \in \Sigma(r) \quad \Rightarrow \quad E_f|C_n| = O(r_n(r)).$$

A comparison to Theorem 1 highlights the subtle difference between the minimax paradigm and asymptotic results that hold pointwise in f: if one relaxes "honesty," that is, if one removes the infimum in (1.2), then Low's impossibility result completely disappears. Note, however, that the index n from which onwards coverage holds in Corollary 1 depends on f, so that the asymptotic result cannot be confidently used for inference at a fixed sample size. This is a reflection of the often neglected fact that asymptotic results that are pointwise in f have to be used with care for statistical inference; see [3, 22] for related situations of this kind.

In contrast to the possibly misleading conclusion of Corollary 1, Theorem 2 characterizes the boundaries of "honest" adaptive inference, and several questions arise.

(i) What is the relationship between the sets $\tilde{\Sigma}(r, \rho_n)$ from Theorem 2 and the classes $\bar{\Sigma}(\varepsilon, r)$ considered in [12]? Moreover, is there a "Bayesian" interpretation of the exceptional sets that complements the topological one?

(ii) The typical adaptation problem is not one over two classes, but over a scale of classes indexed by a possibly continuous smoothness parameter. Can one extend Theorem 2 to such a setting and formulate natural, necessary and sufficient conditions for the existence of confidence bands that adapt over a continuous scale of Hölder classes?

(iii) Can one construct "practical" adaptive nonparametric confidence bands? For instance, can one use bands that are centered at wavelet or kernel estimators with data-driven bandwidths? In particular can one circumvent having to know the radius B of the Hölder balls in the construction of the bands?

We shall give some answers to these questions in the remainder of the article, and summarize our main findings here.

About question (i): we show in Proposition 3 that the "statistical" separation of $\Sigma(r)$ and $\Sigma(s)$ using the sup-norm distance as in (1.4) enforces a weaker condition on $f \in \Sigma(r)$ than the analytic approach in [12], so that the present results are strictly more general for fixed smoothness parameters s. We then move on to give a Bayesian interpretation of the classes $\tilde{\Sigma}(r, \rho_n)$ and $\bar{\Sigma}(\varepsilon, r)$: we show in Proposition 4 that a natural Bayesian prior arising from "uniformly" distributing suitably scaled wavelets on $\Sigma(r)$ concentrates on the classes $\tilde{\Sigma}(r, \rho_n)$ and $\bar{\Sigma}(\varepsilon, r)$ with overwhelming probability.

About question (ii): if the radius B of the Hölder balls involved is known, then one can combine a natural testing approach with recent results in [10, 11, 13] to prove the existence of adaptive nonparametric confidence bands over a scale of Hölder classes indexed by a grid of smoothness parameters that grows dense in any fixed interval $[r, R] \subset (0, \infty)$ as $n \to \infty$; see Theorems 3, 4.

A full answer to question (iii) lies beyond the scope of this paper. Some partial findings that seem of interest are the following: note first that our results imply that the logarithmic penalties that occurred in the diameters of the adaptive confidence bands in [12] are not necessary if one knows the radius B. On the other hand we show in Proposition 1 that if the radius B is unknown, then a certain price in the rate of convergence of the confidence band cannot be circumvented, as B cannot reliably be estimated without additional assumptions on the model. This partly justifies the practice of undersmoothing in the construction of confidence bands, dating back to Bickel and Rosenblatt [2]. It leads us to argue that near-adaptive confidence bands that can be used in practice, and that do not require the knowledge of B, are more likely to follow from the classical adaptive techniques, like Lepski's method applied to classical kernel or wavelet estimators, rather than from the "testing approach" that we employ here to prove existence of optimal procedures.

To conclude: the question as to whether adaptive methods should be used for inference clearly remains a "philosophical" one, but we believe that our results shed new light on the problem. That full adaptive inference is not possible is a consequence of the fact that the typical smoothness classes over which one wants to adapt, such as Hölder balls, contain elements that are indistinguishable from a testing point of view. On the other hand Hölder spaces are used by statisticians to model regularity properties of unknown functions f, and it may seem sensible to exclude functions whose regularity is not statistically identifiable. Our main results give minimal identifiability conditions of a certain kind that apply in this particular case.

Our findings apply also more generally to the adaptation problem discussed at the beginning of this introduction with two abstract classes Σ, Σ'. We are primarily interested in confidence statements that Cai and Low [4] coin *strongly adaptive* (see Section 2.2 in their paper) and in our case this corresponds precisely to requiring (1.2) and (1.3). If Σ, Σ' are convex, and if one is interested in a confidence interval for a linear functional of the unknown parameter, Cai and Low show that whether strong adaptation is possible or not is related to the so-called "inter-class modulus" between Σ, Σ', and their results imply that in several relevant adaptation problems strongly adaptive confidence statements are impossible. The "separation-approach" put forward in the present article (following [12]) shows how strong adaptation can be rendered possible at the expense of imposing statistical identifiability conditions on Σ, Σ', as follows: one first proves existence of a risk-adaptive estimator $\hat{f}_n$ over Σ, Σ' in some relevant loss function. Subsequently one chooses a functional $\mathbb{F}: \Sigma \times \Sigma' \to [0, \infty)$, defines the nonparametric model

$$\mathcal{P}_n := \Sigma \cup \Big\{ f \in \Sigma' \setminus \Sigma : \inf_{g \in \Sigma} \mathbb{F}(g, f) \geq \rho_n \Big\}$$

and derives the minimax rate ρ_n of testing $H_0: f \in \Sigma$ against the generally non-convex alternative $\{f \in \Sigma' \setminus \Sigma : \inf_{g \in \Sigma} \mathbb{F}(g, f) \geq \rho_n\}$. Combining consistent tests for these hypotheses with $\hat{f}_n$ allows for the construction of confidence statements under sharp conditions on ρ_n. A merit of this approach is that the resulting confidence statements are naturally compatible with the statistical accuracy of the adaptive estimator used in the first place. An important question in this context, which is beyond the scope of the present paper, is the optimal choice of the functional $\mathbb{F}$: for confidence bands it seems natural to take $\mathbb{F}(f, g) = \|f - g\|_\infty$, but formalizing this heuristic appears not to be straightforward. In more general settings it may be less obvious to choose $\mathbb{F}$. These remain interesting directions for future research.

2. Proof of Theorem 2 and further results. Let $X_1, \dots, X_n$ be i.i.d. with probability density f on T which we shall take to equal either $T = [0, 1]$ or $T = \mathbb{R}$. We shall use basic wavelet theory [6, 14, 27] freely throughout this article, and we shall say that the wavelet basis is S-*regular* if the corresponding scaling functions ϕ_k and wavelets ψ_k are compactly supported and S-times continuously differentiable on T. For instance, we can take Daubechies wavelets of sufficiently large order $N = N(S)$ on $T = \mathbb{R}$ (see [27]) or on $T = [0, 1]$ (Section 4 in [6]).

We define Hölder spaces in terms of the moduli of the wavelet coefficients of continuous functions. The wavelet basis consists of the translated scaling functions ϕ_k and wavelets $\psi_{lk} = 2^{l/2}\psi_k(2^l(\cdot))$, where we add the boundary corrected scaling functions and wavelets in case $T = [0, 1]$. If $T = \mathbb{R}$ the indices k, l satisfy $l \in \mathbb{N} \cup \{0\}$, $k \in \mathbb{Z}$, but if $T = [0, 1]$ we require $l \geq J_0$ for some fixed integer $J_0 = J_0(N)$ and then $k = 1, \dots, 2^l$ for the ψ_{lk}'s, $k = 1, \dots, N < \infty$ for the ϕ_k's. Note that $\psi_{lk} = 2^{l/2}\psi(2^l(\cdot) - k)$ for a fixed wavelet ψ if either $T = \mathbb{R}$ or if ψ_{lk} is supported in the interior of $[0, 1]$. Write shorthand $\alpha_k(h) = \int h\phi_k$, $\beta_{lk}(h) = \int h\psi_{lk}$.

DEFINITION 1. Denote by $C(T)$ the space of bounded continuous real-valued functions on T, and let ϕ_k and ψ_k be S-regular Daubechies scaling and wavelet functions, respectively. For $s < S$, the Hölder space $\mathcal{C}^s(T)$ (=$\mathcal{C}^s$ when no confusion may arise) is defined as the set of functions

$$\Big\{f \in C(T) : \|f\|_{s,\infty} \equiv \max\Big(\sup_k |\alpha_k(f)|, \sup_{k,l} 2^{l(s+1/2)}|\beta_{lk}(f)|\Big) < \infty\Big\}.$$

Define, moreover, for $s > 0$, $B \geq 1$, the class of densities

$$\Sigma(s) := \Sigma(s, B, T) = \Big\{f : T \to [0, \infty), \int_T f(x)\,dx = 1, \|f\|_{s,\infty} \leq B\Big\}. \tag{2.1}$$

It is a standard result in wavelet theory (Chapter 6.4 in [27] for $T = \mathbb{R}$ and Theorem 4.4 in [6] for $T = [0, 1]$) that $\mathcal{C}^s$ is equal, with equivalent norms, to the classical Hölder–Zygmund spaces C^s. For $T = \mathbb{R}$, $0 < s < 1$, these spaces consist of all functions $f \in C(\mathbb{R})$ for which $\|f\|_\infty + \sup_{x \neq y, x,y \in \mathbb{R}}(|f(x) - f(y)|/|x - y|^s)$ is finite. For noninteger $s > 1$ the space C^s is defined by requiring $D^{[s]}f$ of $f \in C(\mathbb{R})$ to exist and to be contained in $C^{s-[s]}$. The Zygmund class C^1 is defined by requiring $|f(x + y) + f(x - y) - 2f(x)| \leq C|y|$ for all $x, y \in \mathbb{R}$, some $0 < C < \infty$ and $f \in C(\mathbb{R})$, and the case $m < s \leq m + 1$ follows by requiring the same condition on the mth derivative of f. The definitions for $T = [0, 1]$ are similar; we refer to [6].

Define the projection kernel $K(x, y) = \sum_k \phi_k(x)\phi_k(y)$ and write

$$\begin{aligned} K_j(f)(x) &= 2^j \int_T K(2^j x, 2^j y) f(y)\,dy \\ &= \sum_k \alpha_k(f)\phi_k + \sum_{l=J_0}^{j-1} \sum_k \beta_{lk}(f)\psi_{lk} \end{aligned}$$

for the partial sum of the wavelet series of a function f at resolution level $j \geq J_0 + 1$, with the convention that $J_0 = 0$ if $T = \mathbb{R}$.

If $X_1, \dots, X_n$ are i.i.d. $\sim f$ then an unbiased estimate of $K_j(f)$ is, for $\hat{\alpha}_k = (1/n)\sum_{i=1}^n \phi_k(X_i)$, $\hat{\beta}_{lk} = (1/n)\sum_{i=1}^n \psi_{lk}(X_i)$ the empirical wavelet coefficients,

$$f_n(x, j) = \frac{2^j}{n}\sum_{i=1}^n K(2^j x, 2^j X_i) = \sum_k \hat{\alpha}_k \phi_k + \sum_{l=J_0}^{j-1}\sum_k \hat{\beta}_{lk}\psi_{lk}. \tag{2.2}$$

2.1. *Proof of Theorem* 2. We shall first prove Theorem 2 to lay out the main ideas. We shall prove claims (a) and (b), that this also solves the testing problem $H_0 : f_0 \in \Sigma(s)$ against $H_1 : f_0 \in \tilde{\Sigma}(r, \rho_n)$ follows from the proofs. The proof of claim (c) is postponed to Section 3. Let us assume $B \geq 2$ to simplify some notation. Take $j_n^* \in \mathbb{N}$ such that

$$2^{j_n^*} \simeq \left(\frac{n}{\log n}\right)^{1/(2r+1)}$$

is satisfied, where $\simeq$ denotes two-sided inequalities up to universal constants.

($\Leftarrow$): Let us show that $\liminf_n(\rho_n/r_n(r)) = 0$ leads to a contradiction. In this case $\rho_n/r_n(r) \to 0$ along a subsequence of n, and we shall still index this subsequence by n. Let $f_0 = 1$ on $[0, 1]$ and define, for $\varepsilon > 0$, the functions

$$f_m := f_0 + \varepsilon 2^{-j(r+1/2)}\psi_{jm},$$

where $m = 1, \dots, M$, $c_0 2^j \leq M < 2^j$, $j \geq 0$, $c_0 > 0$, and where ψ is a Daubechies wavelet of regularity greater than s, chosen in such a way that ψ_{jm} is supported in the interior of $[0, 1]$ for every m and j large enough. (This is possible using the construction in Theorem 4.4 in [6].) Since $\int_0^1 \psi = 0$ we have $\int_0^1 f_m = 1$ for every m and also $f_m \geq 0$ $\forall m$ if $\varepsilon > 0$ is chosen small enough depending only on $\|\psi\|_\infty$. Moreover, for any $t > 0$, using the definition of $\|\cdot\|_{t,\infty}$ and since $c(\phi) \equiv \sup_k |\int_0^1 \phi_k| \leq \sup_k \|\phi_k\|_2 = 1$,

$$\|f_m\|_{t,\infty} = \max\bigl(c(\phi), \varepsilon 2^{j(t-r)}\bigr), \qquad m = 1, \dots, M, \tag{2.3}$$

so $f_m \in \Sigma(r)$ for $\varepsilon \leq 2$ (recall $B \geq 2$) and every j but $f_m \notin \Sigma(s)$ for j large enough depending only on s, r, B, ε.

Note next that

$$|\beta_{lk}(h)| = \left|2^{l/2}\int \psi_k(2^l x)h(x)\,dx\right| \leq 2^{-l/2}\|\psi_k\|_1\|h\|_\infty \leq 2^{-l/2}\|h\|_\infty$$

for every l, k, and any bounded function h implies

$$\|h\|_\infty \geq \sup_{l \geq 0, k} 2^{l/2}|\beta_{lk}(h)| \tag{2.4}$$

so that, for $g \in \Sigma(s)$ arbitrary,

$$\begin{aligned}\|f_m - g\|_\infty &\geq \sup_{l \geq 0, k} 2^{l/2} |\beta_{lk}(f_m) - \beta_{lk}(g)| \\ &\geq \varepsilon 2^{-jr} - 2^{j/2}|\beta_{jk}(g)| \geq \varepsilon 2^{-jr} - B 2^{-js} \\ &\geq \frac{\varepsilon}{2} 2^{-jr}\end{aligned} \tag{2.5}$$

for every m and for $j \geq j_0$, $j_0 = j_0(s, r, B, \varepsilon)$. Summarizing we see that

$$f_m \in \tilde{\Sigma}\left(r, \frac{\varepsilon}{2} 2^{-jr}\right) \qquad \forall m = 1, \dots, M$$

for every $j \geq j_0$. Since $\rho_n = o(r_n(r))$, $r_n(r) \simeq 2^{-j_n^* r}$, we can find $j_n > j_n^*$ such that

$$\rho_n' := \max(\rho_n, r_n(s) \log n) \leq \frac{\varepsilon}{2} 2^{-j_n r} = o(2^{-j_n^* r}) \tag{2.6}$$

in particular $f_m \in \tilde{\Sigma}(r, \rho_n')$ for every $m = 1, \dots, M$ and every $n \geq n_0, n_0 = n_0(s, r, B, \varepsilon)$.

Suppose now C_n is a confidence band that is adaptive and honest over $\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$, and consider testing

$$H_0 : f = f_0 \quad \text{against} \quad H_1 : f \in \{f_1, \dots, f_M\} =: \mathcal{M}.$$

Define a test Ψ_n as follows: if no $f_m \in C_n$, then $\Psi_n = 0$, but as soon as one of the f_m's is contained in C_n, then $\Psi_n = 1$. We control the error probabilities of this test. Using (2.5), Markov's inequality, adaptivity of the band, (2.6) and noting $r_n(s) = o(\rho_n')$, we deduce

$$\begin{aligned}\Pr_{f_0}(\Psi_n \neq 0) &= \Pr_{f_0}(f_m \in C_n \text{ for some } m) \\ &= \Pr_{f_0}(f_m, f_0 \in C_n \text{ for some } m) \\ &\quad + \Pr_{f_0}(f_m \in C_n \text{ for some } m, f_0 \notin C_n) \\ &\leq \Pr_{f_0}(\|f_m - f_0\|_\infty \leq |C_n| \text{ for some } m) + \alpha + o(1) \\ &\leq \Pr_{f_0}(|C_n| \geq \rho_n') + \alpha + o(1) \\ &\leq E_{f_0}|C_n|/\rho_n' + \alpha + o(1) = \alpha + o(1).\end{aligned}$$

Under any alternative $f_m \in \tilde{\Sigma}(r, \rho_n')$, invoking honesty of the band we have

$$P_{f_m}(\Psi_n = 0) = \Pr_{f_m}(\text{no } f_k \in C_n) \leq \Pr_{f_m}(f_m \notin C_n) \leq \alpha + o(1)$$

so that summarizing we have

$$\limsup_n \Big(E_{f_0}\Psi_n + \sup_{f \in \mathcal{M}} E_f(1 - \Psi_n)\Big) \leq 2\alpha < 1. \tag{2.7}$$

On the other hand, if $\tilde\Psi$ is *any* test (any measurable function of the sample taking values 0 or 1), we shall now prove

$$\liminf_n \inf_{\tilde\Psi}\Big(E_{f_0}\tilde\Psi + \sup_{f\in\mathcal{M}} E_f(1-\tilde\Psi)\Big) \ge 1, \tag{2.8}$$

which contradicts (2.7) and completes this direction of the proof. The proof follows ideas in [16]. We have, for every $\eta > 0$,

$$\begin{aligned} E_{f_0}\tilde\Psi + \sup_{f\in\mathcal{M}} E_f(1-\tilde\Psi) &\ge E_{f_0}(1\{\tilde\Psi = 1\}) + \frac{1}{M}\sum_{m=1}^M E_{f_m}(1-\tilde\Psi) \\ &\ge E_{f_0}(1\{\tilde\Psi=1\} + 1\{\tilde\Psi = 0\}Z) \\ &\ge (1-\eta)\Pr_{f_0}(Z \ge 1-\eta), \end{aligned}$$

where $Z = M^{-1}\sum_{m=1}^M (dP_m^n/dP_0^n)$ with P_m^n the product probability measures induced by a sample of size n from the density f_m. By Markov's inequality,

$$\Pr_{f_0}(Z \ge 1-\eta) \ge 1 - \frac{E_{f_0}|Z-1|}{\eta} \ge 1 - \frac{\sqrt{E_{f_0}(Z-1)^2}}{\eta}$$

for every $\eta > 0$, and we show that the last term converges to zero. Writing (in abuse of notation) $\gamma_j = \varepsilon 2^{-j_n(r+1/2)}$, using independence, orthonormality of ψ_{jm} and $\int \psi_{jm} = 0$ repeatedly as well as $(1+x) \le e^x$, we see

$$\begin{aligned} E_{f_0}(Z-1)^2 &= \frac{1}{M^2}\int_{[0,1]^n}\left(\sum_{m=1}^M\left(\prod_{i=1}^n f_m(x_i) - 1\right)\right)^2 dx \\ &= \frac{1}{M^2}\int_{[0,1]^n}\left(\sum_{m=1}^M\left(\prod_{i=1}^n (1+\gamma_j\psi_{jm}(x_i)) - 1\right)\right)^2 dx \\ &= \frac{1}{M^2}\sum_{m=1}^M\int_{[0,1]^n}\left(\prod_{i=1}^n (1+\gamma_j\psi_{jm}(x_i)) - 1\right)^2 dx \\ &= \frac{1}{M^2}\sum_{m=1}^M\left(\int_{[0,1]^n}\prod_{i=1}^n (1+\gamma_j\psi_{jm}(x_i))^2\, dx - 1\right) \\ &= \frac{1}{M^2}\sum_{m=1}^M\left(\left(\int_{[0,1]} (1+\gamma_j\psi_{jm}(x))^2\, dx\right)^n - 1\right) \\ &= \frac{1}{M}\big((1+\gamma_j^2)^n - 1\big) \le \frac{e^{n\gamma_j^2}-1}{M}. \end{aligned}$$

Now using (2.6) we see $n\gamma_j^2 = \varepsilon^2 n 2^{-j_n(2r+1)} = o(\log n)$ so that $e^{n\gamma_j^2} = o(n^\kappa)$ for every $\kappa > 0$, whereas $M \simeq 2^{j_n} \geq 2^{j_n^*} \simeq r_n(r)^{-1/r}$ still diverges at a fixed polynomial rate in n, so that the last quantity converges to zero, which proves (2.8) since η was arbitrary.

($\Rightarrow$): Let us now show that an adaptive band C_n can be constructed if ρ_n equals $r_n(r)$ times a large enough constant, and if the radius B is known. The remarks after Definition 1 imply that $\|f\|_\infty \leq k\|f\|_{s,\infty} \leq kB$ for some $k > 0$. Set

$$\sigma(j) := \sigma(n, j) := \sqrt{kB\frac{2^j j}{n}}, \qquad \rho_n := L'\sigma(j_n^*) \simeq r_n(r) \tag{2.9}$$

for L' a constant to be chosen later. Using Definition 1 and $\sup_x \sum_k |\psi_k(x)| < \infty$, we have for f_n from (2.2) based on wavelets of regularity $S > s$

$$\|E_f f_n(j_n^*) - f\|_\infty = \|K_{j_n^*}(f) - f\|_\infty \leq b_0 2^{-j_n^* r} \leq b\sigma(j_n^*) \tag{2.10}$$

for some constants b_0, b that depend only on B, ψ.

Define the test statistic $\hat{d}_n := \inf_{g \in \Sigma(s)} \|f_n(j_n^*) - g\|_\infty$. Let now $\hat{f}_n(y)$ be any estimator for f that is exact rate adaptive over $\Sigma(s) \cup \Sigma(r)$ in sup-norm risk; that is, $\hat{f}_n$ satisfies simultaneously, for some fixed constant D depending only on B, s, r

$$\sup_{f \in \Sigma(r)} E_f\|\hat{f}_n - f\|_\infty \leq Dr_n(r), \qquad \sup_{f \in \Sigma(s)} E_f\|\hat{f}_n - f\|_\infty \leq Dr_n(s). \tag{2.11}$$

Such estimators exist; see Theorem 5 below. Define the confidence band $C_n \equiv \{C_n(y), y \in [0, 1]\}$ to equal

$$\hat{f}_n(y) \pm Lr_n(r) \quad \text{if } \hat{d}_n > \tau \quad \text{and} \quad \hat{f}_n(y) \pm Lr_n(s) \quad \text{if } \hat{d}_n \leq \tau,\ y \in [0, 1],$$

where $\tau = \kappa\sigma(j_n^*)$, and where κ and L are constants to be chosen below.

We first prove that C_n is an honest confidence band for $f \in \Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$ when ρ_n is as above with L' large enough depending only on κ, B. If $f \in \Sigma(s)$ we have coverage since adaptivity of $\hat{f}_n$ implies, by Markov's inequality,

$$\begin{aligned}\inf_{f \in \Sigma(s)} \Pr_f(f \in C_n) &\geq 1 - \sup_{f \in \Sigma(s)} \Pr_f\bigl(\|\hat{f}_n - f\|_\infty > Lr_n(s)\bigr) \\ &\geq 1 - \frac{1}{Lr_n(s)} \sup_{f \in \Sigma(s)} E_f\|\hat{f}_n - f\|_\infty \\ &\geq 1 - \frac{D}{L},\end{aligned}$$

which can be made greater than $1 - \alpha$ for any $\alpha > 0$ by choosing L large enough depending only on K, B, α, r, s. When $f \in \tilde{\Sigma}(r, \rho_n)$ there is the danger of $\hat{d}_n \leq \tau$ in which case the size of the band is too small. In this case, however, we have, using again Markov's inequality,

$$\inf_{f \in \tilde{\Sigma}(r,\rho_n)} \Pr_f(f \in C_n) \geq 1 - \frac{\sup_{f \in \tilde{\Sigma}(r,\rho_n)} E_f\|\hat{f}_n - f\|_\infty}{Lr_n(r)} - \sup_{f \in \tilde{\Sigma}(r,\rho_n)} \Pr_f(\hat{d}_n \leq \tau)$$

and the first term subtracted can be made smaller than α for L large enough in view of (2.11). For the second note that $\Pr_f(\hat{d}_n \le \tau)$ equals, for every $f \in \tilde{\Sigma}(r, \rho_n)$,

$$\begin{aligned}
&\Pr_f\Big(\inf_{g\in\Sigma(s)} \|f_n(j_n^*) - g\|_\infty \le \kappa\sigma(j_n^*)\Big) \\
&\quad \le \Pr_f\Big(\inf_g \|f - g\|_\infty - \|f_n(j_n^*) - E_f f_n(j_n^*)\|_\infty \\
&\qquad\qquad - \|K_{j_n^*}(f) - f\|_\infty \le \kappa\sigma(j_n^*)\Big) \\
&\quad \le \Pr_f\big(\rho_n - \|K_{j_n^*}(f) - f\|_\infty - \kappa\sigma(j_n^*) \le \|f_n(j_n^*) - E_f f_n(j_n^*)\|_\infty\big) \\
&\quad \le \Pr_f\big(\|f_n(j_n^*) - E_f f_n(j_n^*)\|_\infty \ge (L' - \kappa - b)\sigma(j_n^*)\big) \\
&\quad \le ce^{-cj_n^*} = o(1)
\end{aligned}$$

for some $c > 0$, by choosing $L' = L'(\kappa, B, K)$ large enough independent of $f \in \tilde{\Sigma}(r, \rho_n)$, in view of Proposition 5 below. This completes the proof of coverage of the band.

We now turn to adaptivity of the band and verify (1.5). By definition of C_n we have almost surely

$$|C_n| \le Lr_n(r),$$

so the case $f \in \tilde{\Sigma}(r, \rho_n)$ is proved. If $f \in \Sigma(s)$ then, using (2.10) and Proposition 5,

$$\begin{aligned}
E_f|C_n| &\le Lr_n(r)\Pr_f(\hat{d}_n > \tau) + Lr_n(s) \\
&\le Lr_n(r)\Pr_f\Big(\inf_{g\in\Sigma(s)} \|f_n(j_n^*) - g\|_\infty > \kappa\sigma(j_n^*)\Big) + Lr_n(s) \\
&\le Lr_n(r)\Pr_f\big(\|f_n(j_n^*) - f\|_\infty > \kappa\sigma(j_n^*)\big) + Lr_n(s) \\
&\le Lr_n(r)\Pr_f\big(\|f_n(j_n^*) - E_f f_n(j_n^*)\|_\infty > (\kappa - b)\sigma(j_n^*)\big) + Lr_n(s) \\
&\le Lr_n(r)ce^{-cj_n^*} + Lr_n(s) = O(r_n(s))
\end{aligned}$$

since c can be taken sufficiently large by choosing $\kappa = \kappa(K, B)$ large enough. This completes the proof of the second claim of Theorem 2.

2.2. *Unknown radius B*. The existence results in the previous section are not entirely satisfactory in that the bands constructed to prove existence of adaptive procedures cannot be easily implemented. Particularly the requirement that the radius B of the Hölder ball be known is restrictive. A first question is whether exact rate-adaptive bands exist if B is unknown, and the answer turns out to be no. This in fact is not specific to the adaptive situation, and occurs already for a fixed Hölder ball, as the optimal size of a confidence band depends on the radius B. The following proposition is a simple consequence of the formula for the exact asymptotic minimax constant for density estimation in sup-norm loss as derived in [21].

PROPOSITION 1. *Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in* $[0, 1]$ *with density* $f \in \Sigma(r, B, [0, 1])$ *where* $0 < r < 1$. *Let* C_n *be a confidence band that is asymptotically honest with level* α *for* $\Sigma(r, B, [0, 1])$. *Then*

$$\liminf_n \sup_{f\in\Sigma(r,B,[0,1])} \frac{E_f|C_n|}{r_n(r)} \geq cB^p(1-\alpha)$$

for some fixed constants $c, p > 0$ *that depend only on* r.

In particular if C_n does not depend on B, then $E_f|C_n|$ cannot be of order $r_n(r)$ uniformly over $\Sigma(r, B, [0, 1])$ for every $B > 0$, unless B can be reliably estimated, which for the full Hölder ball is impossible without additional assumptions. It can be viewed as one explanation for why undersmoothing is necessary to construct "practical" asymptotic confidence bands.

2.3. *Confidence bands for adaptive estimators.* The usual risk-adaptive estimators such as those based on Lepski's [23] method or wavelet thresholding [7] do not require the knowledge of the Hölder radius B. As shown in [12] (see also [20]) such estimators can be used in the construction of (near-)adaptive confidence bands under certain analytic conditions on the elements of $\Sigma(s)$. Let us briefly describe the results in [12, 20]. Let ℓ_n be a sequence of positive integers (typically $\ell_n \to \infty$ as $n \to \infty$) and define, for K the wavelet projection kernel associated to some S-regular wavelet basis, $S > s$

$$\bar{\Sigma}(\varepsilon, s, \ell_n) := \{f \in \Sigma(s) : \varepsilon 2^{-ls} \leq \|K_l(f) - f\|_\infty \leq B2^{-ls}\ \forall l \geq \ell_n\}. \tag{2.12}$$

The conditions in [12, 20] are slightly weaker in that they have to hold only for $l \in [\ell_n, \ell_n']$ where $\ell_n' - \ell_n \to \infty$. This turns out to be immaterial in what follows, however, so we work with these sets to simplify the exposition.

Whereas the upper bound in (2.12) is automatic for functions in $\Sigma(s)$, the lower bound is not. However one can show that a lower bound on $\|K_l(f) - f\|_\infty$ of order 2^{-ls} is "topologically" generic in the Hölder space $\mathcal{C}^s(T)$. The following is Proposition 4 in [12].

PROPOSITION 2. *Let K be S-regular with $S > s$. The set*

$$\{f : \text{there exists no } \varepsilon > 0, l_0 \geq 0 \text{ s.t. } \|K_l(f) - f\|_\infty \geq \varepsilon 2^{-l(s+1/2)}\ \forall l \geq l_0\}$$

is nowhere dense in the norm topology of $\mathcal{C}^s(\mathbb{R})$.

Using this condition, [12] constructed an estimator $\hat{f}_n$ based on Lepski's method applied to a kernel or wavelet density estimator such that

$$\hat{A}_n\left(\sup_{y\in[0,1]}\left|\frac{\hat{f}_n(y) - f(y)}{\hat{\sigma}_n\sqrt{\hat{f}_n(y)}}\right| - \hat{B}_n\right) \to^d Z \tag{2.13}$$

as $n \to \infty$, where Z is a standard Gumbel random variable and where $\hat{A}_n, \hat{B}_n, \hat{\sigma}_n$ are some random constants. If ℓ_n is chosen such that

$$2^{\ell_n} \simeq \left(\frac{n}{\log n}\right)^{1/(2R+1)}, \tag{2.14}$$

then the limit theorem (2.13) is uniform in relevant unions over $s \in [r, R], r > 0$, of Hölder classes $\bar{\Sigma}(\varepsilon, s, \ell_n)$. Since the constants $\hat{A}_n, \hat{B}_n, \hat{\sigma}_n$ in (2.13) are known, confidence bands can be retrieved directly from the limit distribution, and [12] further showed that so-constructed bands are near-adaptive: they shrink at rate $O_P(r_n(s)u_n)$ whenever $f \in \bar{\Sigma}(\varepsilon, s, \ell_n)$, where u_n can be taken of the size $\log n$. See Theorem 1 in [12] for detailed statements. As shown in Theorem 4 in [20], the restriction $u_n \simeq \log n$ can be relaxed to $u_n \to \infty$ as $n \to \infty$, at least if one is not after exact limiting distributions but only after asymptotic coverage inequalities, and this matches Proposition 1, so that these bands shrink at the optimal rate in the case where B is unknown.

Obviously it is interesting to ask how the sets in (2.12) constructed from analytic conditions compare to the classes considered in Theorems 2, 3 and 4 constructed from statistical separation conditions. The following result shows that the conditions in the present paper are strictly weaker than those in [12, 20] for the case of two fixed Hölder classes, and also gives a more statistical explanation of why adaptation is possible over the classes from (2.12).

PROPOSITION 3. *Let $t > s$.*

(a) *Suppose $f \in \bar{\Sigma}(\varepsilon, s, \ell_n)$ for some fixed $\varepsilon > 0$. Then $\inf_{g \in \Sigma(t)} \|f - g\|_\infty \geq c2^{-\ell_n s}$ for some constant $c \equiv c(\varepsilon, B, s, t, K)$. Moreover, if $2^{-\ell_n s}/r_n(s) \to \infty$ as $n \to \infty$, so in particular in the adaptive case as in (2.14), then, for every $L_0 > 0$,*

$$\bar{\Sigma}(\varepsilon, s, \ell_n) \subset \tilde{\Sigma}(s, L_0 r_n(s))$$

for $n \geq n_0(\varepsilon, B, s, t, L_0, K)$ large enough.

(b) *If ℓ_n is s.t. $2^{-\ell_n s}/r_n(s) \to \infty$ as $n \to \infty$, so in particular in the adaptive case (2.14), then $\forall L_0' > 0, \varepsilon > 0$ the set*

$$\tilde{\Sigma}(s, L_0' r_n(s)) \setminus \bar{\Sigma}(\varepsilon, s, \ell_n)$$

is nonempty for $n \geq n_0(s, t, K, B, L_0')$ large enough.

2.4. *A Bayesian perspective.* Instead of analyzing the topological capacity of the set removed, one can try to quantify its size by some measure on the Hölder space $\mathcal{C}^s$. As there is no translation-invariant measure available we consider certain probability measures on $\mathcal{C}^s$ that have a natural interpretation as nonparametric Bayes priors.

Take any S-regular wavelet basis $\{\phi_k, \psi_{lk} : k \in \mathbb{Z}, l \in \mathbb{N}\}$ of $L^2([0,1])$, $S > s$. The wavelet characterization of $\mathcal{C}^s([0,1])$ motivates to distribute the basis functions ψ_{lk}'s randomly on $\Sigma(s, B)$ as follows: take u_{lk} i.i.d. uniform random variables on $[-B, B]$ and define the random wavelet series

$$U_s(x) = 1 + \sum_{l=J}^{\infty} \sum_k 2^{-l(s+1/2)} u_{lk} \psi_{lk}(x),$$

which converges uniformly almost surely. It would be possible to set $J = 0$ and replace 1 by $\sum_k u_{0k}\phi_k$ below, but to stay within the density framework we work with this minor simplification, for which $\int_0^1 U_s(x)\,dx = 1$ as well as $U_s \geq 0$ almost surely if $J \equiv J(\|\psi\|_\infty, B, s)$ is chosen large enough. Conclude that U_s is a random density that satisfies

$$\|U_s\|_{s,\infty} \leq \max\Bigl(1, \sup_{k,l \geq J} |u_{lk}|\Bigr) \leq B \qquad \text{a.s.},$$

so its law is a natural prior on $\Sigma(s, B)$ that uniformly distributes suitably scaled wavelets on $\Sigma(s)$ around its expectation $EU_s = 1$.

PROPOSITION 4. *Let K be the wavelet projection kernel associated to a S-regular wavelet basis ϕ, ψ of $L^2([0,1])$, $S > s$, and let $\varepsilon > 0$, $j \geq 0$. Then*

$$\Pr\{\|K_j(U_s) - U_s\|_\infty < \varepsilon B 2^{-js}\} \leq e^{-\log(1/\varepsilon) 2^j}.$$

By virtue of part (a) of Proposition 3 the same bound can be established, up to constants, for the probability of the sets $\Sigma(s) \setminus \tilde{\Sigma}(s, \rho_n)$ under the law of U_s.

Similar results (with minor modifications) could be proved if one replaces the u_{lk}'s by i.i.d. Gaussians, which leads to measures that have a structure similar to Gaussian priors used in Bayesian nonparametrics; see, for example, [30]. If we choose j at the natural frequentist rate $2^j \simeq n^{1/(2s+1)}$, then the bound in Proposition 4 becomes $e^{-Cn\delta_n^2(s)}$, $\delta_n(s) = n^{-s/(2s+1)}$, where $C > 0$ can be made as large as desired by choosing ε small enough. In view of (2.3) in Theorem 2.1 in [9] one could therefore heuristically conclude that the exceptional sets are "effective null-sets" from the point of view of Bayesian nonparametrics.

2.5. *Adaptive confidence bands for collections of Hölder classes.* The question arises of how Theorem 2 can be extended to adaptation problems over collections of Hölder classes whose smoothness degree varies in a fixed interval $[r, R] \subset (0, \infty)$. A fixed finite number of Hölder classes can be handled by a straightforward extension of the proof of Theorem 2. Of more interest is to consider a continuum of smoothness parameters—adaptive estimators that attain the minimax sup-norm risk over each element of the collection $\bigcup_{0<s\leq R} \Sigma(s)$ exist; see

Theorem 5 below. Following Theorem 2 a first approach might seem to introduce analogues of the sets $\tilde{\Sigma}(s, \rho_n)$ as

$$\Big\{f \in \Sigma(s): \inf_{g \in \Sigma(t)} \|g - f\|_\infty \geq \rho_n(s) \ \forall t > s\Big\}.$$

However this does not make sense as the sets $\{\Sigma(t)\}_{t>s}$ are $\|\cdot\|_\infty$-dense in $\Sigma(s)$, so that so-defined $\tilde{\Sigma}(s, \rho_n(s))$ would be empty [unless $\rho_n(s) = 0$]. Rather one should note that any adaptation problem with a continuous smoothness parameter s and convergence rates that are polynomial in n can be recast as an adaptation problem with a discrete parameter set whose cardinality grows logarithmically in n. Indeed let us dissect $[r, R]$ into $|\mathcal{S}_n| \simeq \log n$ points

$$\mathcal{S}_n := \mathcal{S}_n(\zeta) = \{s_i, i = 1, \dots, |\mathcal{S}_n|\}$$

that include $r \equiv s_1$, $R \equiv s_{|\mathcal{S}_n|}$, $s_i < s_{i+1}\ \forall i$, and each of which has at most $2\zeta/\log n$ and at least $\zeta/\log n$ distance to the next point, where $\zeta > 0$ is a fixed constant. A simple calculation shows

$$r_n(s_i) \leq C r_n(s) \tag{2.15}$$

for some constant $C = C(\zeta, R)$ and every $s_i \leq s < s_{i+1}$, so that any estimator that is adaptive over $\Sigma(s)$, $s \in \mathcal{S}_n$, is also adaptive over $\Sigma(s)$, $s \in [r, R]$.

After this discretization we can define

$$\tilde{\Sigma}(s, \rho_n(s), \mathcal{S}_n) = \Big\{f \in \Sigma(s): \inf_{g \in \Sigma(t)} \|g - f\|_\infty \geq \rho_n(s) \ \forall t > s, t \in \mathcal{S}_n\Big\},$$

where $\rho_n(s)$ is a sequence of nonnegative integers. We are interested in the existence of adaptive confidence bands over

$$\Sigma(R) \cup \Big(\bigcup_{s \in \mathcal{S}_n \setminus \{R\}} \tilde{\Sigma}(s, \rho_n(s), \mathcal{S}_n)\Big)$$

under sharp conditions on $\rho_n(s)$.

Let us first address lower bounds, where we consider $T = [0, 1]$ for simplicity. Theorem 2 cannot be applied directly since the smoothness index s depends on n in the present setting, and any two $s, s' \in \mathcal{S}_n$ could be as close as $\zeta/\log n$ possibly. If the constant ζ is taken large enough (but finite) one can prove the following result.

THEOREM 3 (Lower bound). *Let $T = [0, 1]$, $L \geq 1$ and $0 < \alpha < 1/3$ be given, and let $\mathcal{S}_n(\zeta)$ be a grid as above. Let $s < s'$ be any two points in $\mathcal{S}_n(\zeta)$ and suppose that C_n is a confidence band that is asymptotically honest with level α over*

$$\Sigma(s') \cup \tilde{\Sigma}(s, \rho_n(s), \mathcal{S}_n),$$

and that is adaptive in the sense that

$$\sup_{f \in \Sigma(s')} E_f|C_n| \leq L r_n(s'), \qquad \sup_{f \in \tilde{\Sigma}(s, \rho_n(s), \mathcal{S}_n)} E_f|C_n| \leq L r_n(s)$$

for every n large enough. Then if $\zeta := \zeta(R, B, L, \alpha)$ is a large enough but finite constant, we necessarily have

$$\liminf_n \frac{\rho_n(s)}{r_n(s)} > 0.$$

A version of Theorem 3 for $T = \mathbb{R}$ can be proved as well, by natural modifications of its proof.

To show that adaptive procedures exist if B is known define

$$\tilde{\Sigma}_n(s) := \Big\{ f \in \Sigma(s) : \inf_{g \in \Sigma(t)} \|g - f\|_\infty \geq L_0 r_n(s)\ \forall t \in \mathcal{S}_n, t > s \Big\},$$

where s varies in $[r, R)$, and where $L_0 > 0$. Setting $\tilde{\Sigma}_n(R) \equiv \Sigma(R)$ for notational convenience, we now prove that an adaptive and honest confidence band exists, for L_0 large enough, over the class

$$\mathcal{P}_n(L_0) := \mathcal{P}(\mathcal{S}_n, B, L_0, n) := \bigcup_{s \in \mathcal{S}_n} \tilde{\Sigma}_n(s).$$

Analyzing the limit set (as $n \to \infty$) of $\mathcal{P}_n(L_0)$, or a direct comparison to the continuous scale of classes in (2.12), seems difficult, as $\mathcal{S}_n$ depends on n now. Note, however, that one can always choose $\{\mathcal{S}_n\}_{n \geq 1}$ in a nested way, and ζ large enough, such that $\mathcal{P}_n(L_0)$ contains, for every n, any fixed finite union (over s) of sets of the form $\bar{\Sigma}(\varepsilon, s, \ell_n)$ (using Proposition 3).

THEOREM 4 (Existence of adaptive bands). *Let $X_1, \dots, X_n$ be i.i.d. random variables on $T = [0, 1]$ or $T = \mathbb{R}$ with density $f \in \mathcal{P}_n(L_0)$ and suppose $B, r, R, 0 < \alpha < 1$ are given. Then, if L_0 is large enough depending only on B, a confidence band $C_n = C_n(B, r, R, \alpha; X_1, \dots, X_n)$ can be constructed such that*

$$\liminf_n \inf_{f \in \mathcal{P}_n(L_0)} \Pr_f(f \in C_n) \geq 1 - \alpha$$

and, for every $s \in \mathcal{S}_n, n \in \mathbb{N}$ and some constant L' independent of n,

$$\sup_{f \in \tilde{\Sigma}_n(s)} E_f|C_n| \leq L' r_n(s). \tag{2.16}$$

3. Proofs of remaining results.

PROOF OF PROPOSITION 1. On the events $\{f \in C_n\}$ we can find a random density $T_n \in C_n$ depending only on C_n such that $\{|C_n| \leq D, f \in C_n\} \subseteq \{\|T_n - f\|_\infty \leq D\}$ for any $D > 0$, and negating this inclusion we have

$$\{|C_n| > D\} \cup \{f \notin C_n\} \supseteq \{\|T_n - f\|_\infty > D\}$$

so that $\Pr_f(|C_n| > D) \ge \Pr_f(\|T_n - f\|_\infty > D) - \Pr_f(f \notin C_n)$. Thus, using coverage of the band

$$\liminf_n \sup_{f\in\Sigma(r,B)} \Pr_f\big(|C_n| > cB^p r_n(r)\big)$$
$$\ge \liminf_n \sup_{f\in\Sigma(r,B)} \Pr_f\big(\|T_n - f\|_\infty > cB^p r_n(r)\big) - \alpha.$$

The limit inferior in the last line equals 1 as soon as $c > 0$ is chosen small enough depending only on r, p in view of Theorem 1 in [21]; see also page 1114 as well as Lemma A.2 in that paper. Taking lim inf's in the inequality

$$\sup_{f\in\Sigma(r,B)} \frac{E_f|C_n|}{r_n(r)} \ge cB^p \sup_{f\in\Sigma(r,B)} \Pr_f\big(|C_n| > cB^p r_n(r)\big)$$

gives the result. □

PROOF OF PROPOSITION 3. (a) Observe first that for every $l_0 \ge \ell_n$,

$$\|\psi\|_\infty \sum_{l\ge l_0} 2^{l/2} \sup_k |\beta_{lk}(f)| \ge \|K_{l_0}(f) - f\|_\infty \ge \varepsilon 2^{-l_0 s}.$$

Let N be a fixed integer, and let $\ell_n' \ge \ell_n$ be a sequence of integers to be chosen later. Then for some $\bar l \in [\ell_n', \ell_n' + N - 1]$

$$\begin{aligned}
\sup_k |\beta_{\bar l k}(f)| &\ge \frac{1}{N} \sum_{l=\ell_n'}^{\ell_n'+N-1} \sup_k |\beta_{lk}(f)| \\
&\ge \frac{2^{-(\ell_n'+N)/2}}{N} \left(\sum_{l=\ell_n'}^{\infty} 2^{l/2} \sup_k |\beta_{lk}(f)| - \sum_{l=\ell_n'+N}^{\infty} 2^{l/2} \sup_k |\beta_{lk}(f)| \right) \\
&\ge \frac{2^{-(\ell_n'+N)/2}}{N} \left(\frac{\varepsilon}{\|\psi\|_\infty} 2^{-\ell_n' s} - c(B,s) 2^{-(\ell_n'+N)s} \right) \\
&\ge \frac{2^{-(\ell_n'+N)/2}}{2\|\psi\|_\infty N} \varepsilon 2^{-\ell_n' s} \ge d(\varepsilon, B, \psi, s) 2^{-\ell_n'(s+1/2)}
\end{aligned}$$

for some $d(\varepsilon, B, \psi, s) > 0$ if N is chosen large enough but finite depending only on ε, B, ψ, s. From (2.4) we thus have, for any $t > s$,

$$\begin{aligned}
\inf_{g\in\Sigma(t)} \|f - g\|_\infty &\ge \inf_{g\in\Sigma(t)} \sup_{l\ge \ell_n', k} 2^{l/2} |\beta_{lk}(f - g)| \\
&\ge d(\varepsilon, B, \psi, s) 2^{-\ell_n' s} - \sup_{g\in\Sigma(t)} \sup_{l\ge\ell_n', k} 2^{l/2} |\beta_{lk}(g)| \\
&\ge d(\varepsilon, B, \psi, s) 2^{-\ell_n' s} - B 2^{-\ell_n' t} \\
&\ge c(\varepsilon, B, s, t, \psi) 2^{-\ell_n s},
\end{aligned}$$

where we have chosen ℓ_n' large enough depending only on $B, s, t, d(\varepsilon, B, \psi, s)$ but still of order $O(\ell_n)$. This completes the proof of the first claim. The second claim is immediate in view of the definitions.

(b) Take $f = f_0 + 2^{-\ell_n(s+1/2)}\psi_{\ell_n m}$ for some m. Then $\|f\|_{s,\infty} \le 1$ so $f \in \Sigma(s, B)$ and the estimate in the last display of the proof of part (a) implies

$$\inf_{g\in\Sigma(t)} \|f - g\|_\infty \ge c2^{-\ell_n s} \ge L_0' r_n(s)$$

for n large enough depending only on B, s, t, L_0', ψ. On the other hand $\|K_{\ell_n+1}(f) - f\|_\infty = 0$ so $f \notin \bar{\Sigma}(\varepsilon, s, \ell_n)$ for any $\varepsilon > 0$. □

PROOF OF PROPOSITION 4. Using (2.4) we have

$$\|K_j(U_s) - U_s\|_\infty \ge \|\psi\|_1^{-1} \sup_{l\ge j,k} 2^{l/2}|\beta_{lk}(U_s)| \ge \|\psi\|_1^{-1}2^{-js} \max_{k=1,\dots,2^j} |u_{jk}|.$$

The variables u_{jk}/B are i.i.d. $U(-1, 1)$ and so the U_k's, $U_k := |u_{jk}/B|$, are i.i.d. $U(0, 1)$ with maximum equal to the largest order statistic $U_{(2^j)}$. Deduce

$$\Pr\big(\|K_j(U_s) - U_s\|_\infty < \varepsilon B2^{-js}\big) \le \Pr\big(U_{(2^j)} < \varepsilon\big) = \varepsilon^{2^j}$$

to complete the proof. □

PROOF OF THEOREM 3. The proof is a modification of the "necessity part" of Theorem 2. Let us assume w.l.o.g. $B \ge 2$, $R \ge 1$, let us write, in slight abuse of notation, s_n, s_n' for s, s' throughout this proof to highlight the dependence on n and choose $j_n(s_n) \in \mathbb{N}$ such that

$$(n/\log n)^{1/(2R+1)} \le c_0(n/\log n)^{1/(2s_n+1)} \le 2^{j_n(s_n)} \le (n/\log n)^{1/(2s_n+1)}$$

holds for some $c_0 > 1/(2R + 1)^{1/(2R+1)}$ and every n large enough. We shall assume that ζ is any fixed number satisfying

$$\zeta > (4R + 2) \max\bigg(\log_2\big((4R + 2)B\big), (2R + 1) \log \frac{(4R + 2)L}{\alpha}\bigg)$$

in the rest of the proof, and we shall establish $\liminf_n(\rho_n(s_n)/Lr_n(s_n^+)) > 0$, where $s_n^+ > s_n$ is the larger "neighbor" of s_n in $\mathcal{S}_n$. This completes the proof since $\liminf_n r_n(s_n^+)/r_n(s_n) \ge c(\zeta) > 0$ by definition of the grid.

Assume thus by way of contradiction that $\liminf_n(\rho_n(s_n)/Lr_n(s_n^+)) = 0$ so that, by passing to a subsequence of n if necessary, $\rho_n(s_n) \le Lr_n(s_n^+) + \delta$ for every $\delta > 0$ and every $n = n(\delta)$ large enough. Let $\varepsilon := 1/(2R + 1)$ and define

$$f_0 = 1, \qquad f_m = f_0 + \varepsilon 2^{-j(s_n+1/2)}\psi_{jm}, \qquad m = 1, \dots, M,$$

as in the proof of Theorem 2, $c_0'2^j \le M \le 2^j$, $c_0' > 0$. Then $f_m \in \Sigma(s_n)$ for every $j \ge j_0$ where j_0 can be taken to depend only on r, R, B, ψ. Moreover for

$j \geq (\log n)/(4R+2)$ we have, using (2.4) and the assumption on ζ, for any $g \in \Sigma(t), t \in \mathcal{S}_n, t > s_n$, and every m

$$\begin{aligned} \|f_m - g\|_\infty &\geq \sup_{l \geq 0, k} 2^{l/2} |\beta_{lk}(f_m) - \beta_{lk}(g)| \\ &\geq \varepsilon 2^{-js_n} - 2^{j/2}|\beta_{jk}(g)| \geq \varepsilon 2^{-js_n} - B2^{-jt} \\ &\geq 2^{-js_n}(\varepsilon - B2^{-j\zeta/\log n}) \geq \frac{\varepsilon}{2} 2^{-js_n}. \end{aligned} \tag{3.1}$$

We thus see that

$$f_m \in \tilde{\Sigma}\left(s_n, \frac{\varepsilon}{2} 2^{-js_n}, \mathcal{S}_n\right) \qquad \forall m = 1, \dots, M,$$

for every $j \geq J_0 := \max(j_0, (\log n)/(4R+2))$. Take now $j \equiv j_n(s_n)$ which exceeds J_0 for n large enough, and conclude

$$\begin{aligned} \frac{\varepsilon}{2} 2^{-j_n(s_n)s_n} &\geq \frac{\varepsilon}{2} r_n(s_n) \geq \frac{\varepsilon}{2} \frac{\alpha}{L} \frac{L}{\alpha} (e^{\zeta/2})^{1/(2R+1)^2} r_n(s_n^+) \\ &\geq \frac{L}{\alpha} r_n(s_n^+) \geq \rho_n(s_n) \end{aligned} \tag{3.2}$$

for n large enough, where we have used the definition of the grid $\mathcal{S}_n$, of ε, the assumption on ζ and the hypothesis on ρ_n. Summarizing $f_m \in \tilde{\Sigma}(s_n, \rho_n(s_n), \mathcal{S}_n)$ for every $m = 1, \dots, M$ and every $n \geq n_0, n_0 = n_0(r, R, B, \psi)$.

Suppose now C_n is a confidence band that is adaptive and asymptotically honest over $\Sigma(s'_n) \cup \tilde{\Sigma}(s_n, \rho_n(s_n), \mathcal{S}_n)$, and consider testing $H_0 : f = f_0$ against $H_1 : f \in \{f_1, \dots, f_M\} =: \mathcal{M}$. Define a test Ψ_n as follows: if no $f_m \in C_n$ then $\Psi_n = 0$, but as soon as one of the f_m's is contained in C_n then $\Psi_n = 1$. Now since $r_n(s'_n) \leq r_n(s_n^+)$ and using (3.1), (3.2) we have

$$\begin{aligned} \Pr_{f_0}(\Psi_n \neq 0) &= \Pr_{f_0}(f_m \in C_n \text{ for some } m) \\ &\leq \Pr_{f_0}(\|f_m - f_0\|_\infty \leq |C_n| \text{ for some } m) + \alpha + o(1) \\ &\leq \Pr_{f_0}\big(|C_n| \geq (L/\alpha) r_n(s_n^+)\big) + \alpha + o(1) \\ &\leq \alpha r_n(s'_n)/r_n(s_n^+) + \alpha + o(1) \leq 2\alpha + o(1). \end{aligned}$$

Under any alternative $f_m \in \tilde{\Sigma}(s_n)$, invoking honesty of the band we have

$$P_{f_m}(\Psi_n = 0) = \Pr_{f_m}(\text{no } f_k \in C_n) \leq \Pr_{f_m}(f_m \notin C_n) \leq \alpha + o(1)$$

so that summarizing we have

$$\limsup_n \Big(E_{f_0} \Psi_n + \sup_{f \in \mathcal{M}} E_f(1 - \Psi_n) \Big) \leq 3\alpha < 1.$$

But this has led to a contradiction by the same arguments as in the proof of Theorem 2, noting in the last step that $n\gamma_j^2 = \varepsilon^2 n 2^{-j_n(s_n)(2s_n+1)} \le (\varepsilon^2/(c_0)^{2R+1}) \log n$ and thus

$$\frac{e^{n\gamma_j^2}-1}{M} \le \frac{1}{c_0' c_0} e^{(\varepsilon^2/(c_0)^{2R+1})\log n}\left(\frac{\log n}{n}\right)^{1/(2R+1)} = o(1)$$

since $1/(2R+1) = \varepsilon < c_0^{2R+1}$. □

PROOF OF THEOREM 4. We shall only prove the more difficult case $T = \mathbb{R}$. Let j_i be such that $2^{j_i} \simeq (n/\log n)^{1/(2s_i+1)}$, let $f_n(j)$ be as in (2.2) based on wavelets of regularity $S > R$ and define test statistics

$$\hat{d}_n(i) := \inf_{g\in\Sigma(s_{i+1})} \|f_n(j_i) - g\|_\infty, \qquad i = 1, \dots, |\mathcal{S}_n| - 1.$$

Recall further $\sigma(j)$ from (2.9) and, for a constant L to be chosen below, define tests

$$\Psi(i) = \begin{cases} 0, & \text{if } \hat{d}_n(i) \le L\sigma(j_i), \\ 1, & \text{otherwise,} \end{cases}$$

to accept $H_0 : f \in \Sigma(s_{i+1})$ against the alternative $H_1 : f \in \tilde{\Sigma}_n(s_i)$. Starting from the largest model we first test $H_0 : f \in \Sigma(s_2)$ against $H_1 : f \in \tilde{\Sigma}_n(r)$. If H_0 is rejected we set $\hat{s}_n = r$, otherwise we proceed to test $H_0 : f \in \Sigma(s_3)$ against $H_1 : f \in \tilde{\Sigma}_n(s_2)$ and iterating this procedure downwards we define $\hat{s}_n$ to be the first element s_i in $\mathcal{S}$ for which $\Psi(i) = 1$ rejects. If no rejection occurs set $\hat{s}_n = R$.

For $f \in \mathcal{P}_n(L_0)$ define $s_{i_0} := s_{i_0}(f) = \max\{s \in \mathcal{S}_n : f \in \tilde{\Sigma}_n(s)\}$.

LEMMA 1. *We can choose the constants L and then L_0 depending only on B, ϕ, ψ such that*

$$\sup_{f\in\mathcal{P}_n(L_0)} \Pr_f\big(\hat{s}_n \ne s_{i_0}(f)\big) \le Cn^{-2}$$

for some constant C and every n large enough.

PROOF. If $\hat{s}_n < s_{i_0}$, then the test $\Psi(i)$ has rejected for some $i < i_0$. In this case $f \in \tilde{\Sigma}_n(s_{i_0}) \subset \Sigma(s_{i_0}) \subseteq \Sigma(s_{i+1})$ for every $i < i_0$, and thus, proceeding as in (2.10) and using Proposition 5 below, we have for L and then d large enough depending only on B, K

$$\begin{aligned}\Pr_f(\hat{s}_n < s_{i_0}) &= \Pr_f\Big(\bigcup_{i<i_0}\Big\{\inf_{g\in\Sigma(s_{i+1})} \|f_n(j_i) - g\|_\infty > L\sigma(j_i)\Big\}\Big) \\ &\le \sum_{i<i_0} \Pr_f\big(\|f_n(j_i) - E_f f_n(j_i)\|_\infty > (L-b)\sigma(j_i)\big) \\ &\le C'|\mathcal{S}_n| e^{-d\log n} \le Cn^{-2}.\end{aligned}$$

On the other hand if $\hat{s}_n > s_{i_0}$ (ignoring the trivial case $s_{i_0} = R$), then $\Psi(i_0)$ has accepted despite $f \in \tilde{\Sigma}_n(s_{i_0})$. Thus, using $r_n(s_{i_0}) \geq c\sigma(j_{i_0})$ for some $c = c(B)$ and proceeding as in (2.10) we can bound $\Pr_f(\hat{s}_n > s_{i_0})$ by

$$
\begin{aligned}
&\Pr_f\Big(\inf_{g\in\Sigma(s_{i_0+1})} \|f_n(j_{i_0}) - g\|_\infty \leq L\sigma(j_{i_0})\Big)\\
&\quad\leq \Pr_f\Big(\inf_{g\in\Sigma(s_{i_0+1})} \|f - g\|_\infty - \|f_n(j_{i_0}) - E_f f_n(j_{i_0})\|_\infty\\
&\qquad\qquad\qquad - \|E_f f_n(j_{i_0}) - f\|_\infty \leq L\sigma(j_{i_0})\Big)\\
&\quad\leq \Pr_f\big(L_0 r_n(s_{i_0}) - \|K_{j_{i_0}}(f) - f\|_\infty - L\sigma(j_{i_0}) \leq \|f_n(j_{i_0}) - E_f f_n(j_{i_0})\|_\infty\big)\\
&\quad\leq \Pr_f\big(\|f_n(j_{i_0}) - E_f f_n(j_{i_0})\|_\infty \geq (cL_0 - L - b)\sigma(j_{i_0})\big)\\
&\quad\leq c' e^{-c' j_{i_0}} \leq C/n^2
\end{aligned}
$$

for L_0 and then also $c' > 0$ large enough, using Proposition 5 below. □

Take now $\hat{f}_n$ to be an estimator of f that is adaptive in sup-norm loss over $\bigcup_{s\in[r,R]} \Sigma(s)$ as in Theorem 5 below and define the confidence band

$$
C_n = \hat{f}_n \pm M\left(\frac{\log n}{n}\right)^{\hat{s}_n/(2\hat{s}_n+1)},
$$

where M is chosen below. For $f \in \tilde{\Sigma}_n(s_{i_0})$ the lemma implies

$$
\begin{aligned}
E_f|C_n| &\leq 2M\left(\frac{\log n}{n}\right)^{s_{i_0}/(2s_{i_0}+1)} + 2M\left(\frac{\log n}{n}\right)^{r/(2r+1)} \times \Pr_f(\hat{s}_n < s_{i_0})\\
&\leq C(M)\left(\frac{\log n}{n}\right)^{s_{i_0}/(2s_{i_0}+1)},
\end{aligned}
$$

so this band is adaptive.

For coverage, we have, again from the lemma and Markov's inequality

$$
\begin{aligned}
\Pr_f(f \in C_n) &= \Pr_f\big(\|\hat{f}_n - f\|_\infty \leq M r_n(\hat{s}_n)\big)\\
&\geq 1 - \Pr_f\big(\|\hat{f}_n - f\|_\infty > M r_n(s_{i_0})\big) - \Pr(\hat{s}_n > s_{i_0})\\
&\geq 1 - \frac{E_f\|\hat{f}_n - f\|_\infty}{M r_n(s_{i_0})} - \frac{C}{n^2}\\
&\geq 1 - \frac{D(B,R,r)}{M} - \frac{C}{n^2},
\end{aligned}
$$

which is greater than or equal to $1 - \alpha$ for M and n large enough depending only on B, R, r. □

PROOF OF PART (c) OF THEOREM 2. The analog of case (b) is immediate. The analog of part (a) requires the following modifications: set again $f_0 = 1$ on $[0, 1]$, $0 \le j_n' < j_n$ to be chosen below, and define

$$f_m := f_0 + B2^{-j_n'(s+1/2)}\psi_{j_n' m_0} + \varepsilon 2^{-j_n(r+1/2)}\psi_{j_n m},$$

where $m = 1, \dots, M \simeq 2^j$, all ψ_{lk}'s are Daubechies wavelets supported in the interior of $[0, 1]$ and where $m_0 \neq m$ is chosen such that $\psi_{j_n' m_0}$ and $\psi_{j_n m}$ have disjoint support for every m (which is possible for j_n, j_n' large enough since Daubechies wavelets have localized support). Recalling j_n^* from the proof of part (a), we can choose j_n', j_n in such a way that $j_n' < j_n$, $2^{-j_n r} = o(2^{-j_n^* r})$,

$$f_m \in \tilde{\Sigma}(r, \rho_n) \qquad \forall m, \qquad f_0' := f_0 + B2^{-j_n'(s+1/2)}\psi_{j_n' m_0} \in \tilde{\Sigma}\big(s, (B/2)r_n(s)\big)$$

for every $n \ge n_0, n_0 = n_0(s, r, B, \varepsilon, \psi)$. Now if C_n is a confidence band that is adaptive and honest over $\tilde{\Sigma}(s, r_n(s)) \cup \tilde{\Sigma}(r, \rho_n)$ consider testing $H_0 : f = f_0'$ against $H_1 : f \in \{f_1, \dots, f_M\} =: \mathcal{M}$. The same arguments as before (2.7) show that there exists a test Ψ_n such that $\limsup_n (E_{f_0}\Psi_n + \sup_{f\in\mathcal{M}} E_f(1 - \Psi_n)) \le 2\alpha < 1$ along a subsequence of n, a claim that leads to a contradiction since we can lower bound the error probabilities of any test as in the original proof above, the only modification arising in the bound for the likelihood ratio. Let P_0' be the n-fold product probability measure induced by the density f_0' and set $Z = (1/M)\sum_{m=1}^{M}(dP_m/dP_0')$. We suppress now the dependence of j_n on n for notational simplicity, and define shorthand $\gamma_j = \varepsilon 2^{-j(r+1/2)}$, $\kappa_j = B2^{-j'(s+1/2)}$. To bound $E_{f_0'}(Z-1)^2$ we note that, using orthonormality of the ψ_{jm}'s, that $\int \psi_{jm} = 0$ and that $\psi_{j'm_0}$ has disjoint support with ψ_{jm}, $m = 1, \dots, M$, we have $(m \neq m')$

$$\int \frac{\psi_{lm}\psi_{lm'}}{(1 + \kappa_j\psi_{j'm_0})^2} f_0' = \int \psi_{jm}\psi_{jm'} = 0,$$

$$\int \frac{\psi_{jm}}{1 + \kappa_j\psi_{j'm_0}} f_0' = \int \psi_{jm} = 0,$$

$$\int \frac{\psi_{jm}^2}{(1 + \kappa_j\psi_{j'm_0})^2} f_0' = \int \psi_{jm}^2 = 1.$$

The identities in the last display can be used to bound $E_{f_0'}(Z-1)^2$ by

$$\begin{aligned}
&\frac{1}{M^2}\int_{[0,1]^n}\left(\sum_{m=1}^{M}\left(\prod_{i=1}^{n}\frac{f_m(x_i)}{f_0'(x_i)} - 1\right)\right)^2 \prod_{i=1}^{n} f_0'(x_i)\,dx \\
&\quad = \frac{1}{M^2}\sum_{m=1}^{M}\int_{[0,1]^n}\left(\prod_{i=1}^{n}\left(1 + \frac{\gamma_j\psi_{jm}(x_i)}{1 + \kappa_j\psi_{j'm_0}(x_i)}\right) - 1\right)^2 \prod_{i=1}^{n} f_0'(x_i)\,dx
\end{aligned}$$

$$= \frac{1}{M^2} \sum_{m=1}^{M} \left(\left(\int_{[0,1]} \left(1 + \frac{\gamma_j \psi_{jm}(x_i)}{1 + \kappa_j \psi_{j'm_0}(x_i)} \right)^2 f_0'(x)\,dx \right)^n - 1 \right)$$

$$= \frac{1}{M}\big((1+\gamma_j^2)^n - 1\big) \le \frac{e^{n\gamma_j^2} - 1}{M}.$$

The rest of the proof is as in part (a) of Theorem 2. □

3.1. *Auxiliary results.* The following theorem is due to [10, 11, 13]. We state a version that follows from Theorem 4 in [25] for $T = \mathbb{R}$. In case $T = [0, 1]$ it follows from the same proofs. The restriction that B be known is not necessary but suffices for our present purposes.

THEOREM 5. *Let $X_1, \dots, X_n$ be i.i.d. with uniformly continuous density f on $T = [0, 1]$ or $T = \mathbb{R}$. Then for every $r, R, 0 < r \le R$ there exists an estimator $\hat{f}_n(x) := \hat{f}_n(x, X_1, \dots, X_n, B, R)$ such that, for every s, $r \le s \le R$, some constant $D(B, r, R)$ and every $n \ge 2$ we have $\sup_{f\in\Sigma(s,B,T)} E\|\hat{f}_n - f\|_\infty \le D(B, r, R) r_n(s)$.*

The following inequality was proved in [11] (see also page 1167 in [12]) for $T = \mathbb{R}$ (the case $T = [0, 1]$ is similar, in fact simpler).

PROPOSITION 5. *Let ϕ, ψ be a compactly supported scaling and wavelet function, respectively, both S-Hölder for some $S > 0$. Suppose P has a bounded density f and let $f_n(x, j)$ be the estimator from (2.2). Given $C, C' > 0$, there exist finite positive constants $C_1 = C_1(C, K)$ and $C_2 = C_2(C, C', K)$ such that, if $(n/2^j j) \ge C$ and $C_1\sqrt{(\|f\|_\infty \vee 1)(2^j j/n)} \le t \le C'$, then, for every $n \in \mathbb{N}$,*

$$\Pr_f\Big\{ \sup_{x\in\mathbb{R}} |f_n(x, j) - Ef_n(x, j)| \ge t \Big\} \le C_2 \exp\left(-\frac{nt^2}{C_2(\|f\|_\infty \vee 1)2^j} \right).$$

Acknowledgments. We would like to thank Tony Cai, the Associate Editor, as well as two anonymous referees for valuable remarks and criticism. We are further grateful to Adam Bull and Jakob Söhl for pointing out mistakes in the first version of this article. We are also indebted to Yannick Baraud, Lutz Dümbgen, Vladimir Koltchinskii, Oleg Lepski, Axel Munk, Dominique Picard, Benedikt Pötscher, Markus Reiss, Vladimir Spokoiny and Aad van der Vaart for stimulating discussions on the subject over the past years. The second author would like to thank the hospitality of the Cafés Bräunerhof and Florianihof in Vienna for their hospitality.

REFERENCES

[1] BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* **32** 528–551. MR2060168

[2] BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. MR0348906

[3] BROWN, L. D., LOW, M. G. and ZHAO, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25** 2607–2625. MR1604424

[4] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494

[5] CAI, T. T. and LOW, M. G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. MR2275240

[6] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. MR1256527

[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] GENOVESE, C. and WASSERMAN, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36** 875–905. MR2396818

[9] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[10] 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

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

[12] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707

[13] GINÉ, E. and NICKL, R. (2010). Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli* **16** 1137–1163. MR2759172

[14] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics* **129**. Springer, New York. MR1618204

[15] HOFFMANN, M. and LEPSKI, O. (2002). Random rates in anisotropic regression. *Ann. Statist.* **30** 325–396. MR1902892

[16] INGSTER, Y. I. (1986). A minimax test of nonparametric hypotheses on the density of a distribution in L_p metrics. *Teor. Veroyatnost. i Primenen.* **31** 384–389. MR0851000

[17] INGSTER, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.* **2** 85–114. MR1257978

[18] INGSTER, Y. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446

[19] JUDITSKY, A. and LAMBERT-LACROIX, S. (2003). Nonparametric confidence set estimation. *Math. Methods Statist.* **12** 410–428. MR2054156

[20] KERKYACHARIAN, G., NICKL, R. and PICARD, D. (2011). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probab. Theory Related Fields*. To appear.

[21] KOROSTELEV, A. and NUSSBAUM, M. (1999). The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli* **5** 1099–1118. MR1735786

[22] LEEB, H. and PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34** 2554–2591. MR2291510

[23] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.* **35** 459–470. MR1091202

[24] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135

[25] LOUNICI, K. and NICKL, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Statist.* **39** 201–231. MR2797844

[26] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412

[27] MEYER, Y. (1992). *Wavelets and Operators*. *Cambridge Studies in Advanced Mathematics* **37**. Cambridge Univ. Press, Cambridge. MR1228209

[28] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913

[29] ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241

[30] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

ENSAE-CREST AND CNRS UMR 8050
LABORATOIRE D'ANALYSE
ET DE MATHÉMATIQUES APPLIQUÉES
3, AVENUE PIERRE LAROUSSE
92245 MALAKOFF CEDEX
FRANCE
E-MAIL: marc.hoffmann@ensae.fr

STATISTICAL LABORATORY
DEPARTMENT OF PURE MATHEMATICS
AND MATHEMATICAL STATISTICS
UNIVERSITY OF CAMBRIDGE
CB3 0WB CAMBRIDGE
UNITED KINGDOM
E-MAIL: r.nickl@statslab.cam.ac.uk