Electronic Journal of Statistics
Vol. 6 (2012) 1490–1516
ISSN: 1935-7524
DOI: 10.1214/12-EJS720

Honest adaptive confidence bands and self-similar functions

Adam D. Bull

Statistical Laboratory
University of Cambridge
Cambridge CB3 0WB, UK
e-mail: a.bull@statslab.cam.ac.uk

Abstract: Confidence bands are confidence sets for an unknown function f, containing all functions within some sup-norm distance of an estimator. In the density estimation, regression, and white noise models, we consider the problem of constructing adaptive confidence bands, whose width contracts at an optimal rate over a range of Hölder classes.

While adaptive estimators exist, in general adaptive confidence bands do not, and to proceed we must place further conditions on f. We discuss previous approaches to this issue, and show it is necessary to restrict f to fundamentally smaller classes of functions.

We then consider the self-similar functions, whose Hölder norm is similar at large and small scales. We show that such functions may be considered typical functions of a given Hölder class, and that the assumption of self-similarity is both necessary and sufficient for the construction of adaptive bands.

AMS 2000 subject classifications: Primary 62G15; secondary 62G07, 62G08, 62G20.
Keywords and phrases: Nonparametric statistics, adaptation, confidence sets, supremum norm, self-similar functions.

Received April 2012.

1. Introduction

Suppose we have an unknown function $f : [0, 1] \to \mathbb{R}$ we wish to estimate. Our data may come from:

(i) density estimation, where f is a density on $[0, 1]$, and we observe

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f;$$

(ii) fixed design regression, where we observe

$$Y_i \coloneqq f(x_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

for $x_i \coloneqq i/n$, $i = 1, \dots, n$; or

(iii) white noise, where we observe the process

$$Y_t \coloneqq \int_0^t f(s)\, ds + n^{-1/2} B_t,$$

for a standard Brownian motion B.

The performance of an estimator $\hat{f}_n$ depends on the smoothness of the function f. In the following, we will measure performance by the L^∞ loss, $\|\hat{f}_n - f\|_\infty$, where $\|f\|_\infty := \sup_{x\in[0,1]}|f(x)|$. L^∞ loss is the hardest of the L^p loss functions to estimate under, but provides intuitive risk bounds, simultaneously describing local and global performance. If the function f is known to lie in the smoothness class $C^s(M)$ of functions with s-Hölder norm at most M,

$$C^s(M) := \Bigg\{ f \in C([0,1]) : f \text{ has } k := \lceil s \rceil - 1 \text{ derivatives,} \\ \|f\|_\infty, \dots, \|f^{(k)}\|_\infty \le M, \sup_{x,y\in[0,1]} \frac{|f^{(k)}(x) - f^{(k)}(y)|}{|x-y|^{s-k}} \le M \Bigg\},$$

then the L^∞ minimax rate of estimation,

$$\inf_{\hat{f}_n} \sup_{f \in C^s(M)} \mathbb{E}_f \|\hat{f}_n - f\|_\infty,$$

decays like $(n/\log n)^{-s/(2s+1)}$ [24].

The simplest estimators attaining this rate depend on the quantities s and M, which in practice we will not know in advance. However, it is possible to estimate f *adaptively*: to choose an estimator $\hat{f}_n$, not depending on s or M, which nevertheless obtains the minimax rate over a range of classes $C^s(M)$,

$$\sup_{f\in C^s(M)} \mathbb{E}_f \|\hat{f}_n - f\|_\infty = O\left((n/\log n)^{-s/(2s+1)} \right).$$

Such estimators can be constructed, for example, using Lepski's method [19], wavelet thresholding [9], or model selection [1].

Of course, to make full use of an adaptive estimator $\hat{f}_n$, we must also quantify the uncertainty in our estimate. We would like to have a risk bound R_n, depending only on the data, which satisfies $\|f - \hat{f}_n\|_\infty \le R_n$ with high probability. Equivalently, we would like a *confidence band*,

$$S_n := \{ f \in C([0,1]) : \|f - \hat{f}_n\|_\infty \le R_n \}, \tag{1.1}$$

containing f with high probability. To benefit from the adaptive nature of $\hat{f}_n$, we would also like the radius R_n to be adaptive, decaying like $(n/\log n)^{-s/(2s+1)}$ over any class $C^s(M)$.

Unfortunately, this is impossible in general [21, 5]. The size of an adaptive confidence band must depend on the parameters s and M, which we cannot estimate from the data: the function f may be *deceptive*, superficially appearing to belong to one smoothness class $C^s(M)$, while instead belonging to a different, rougher class. If we wish to proceed, we must place further conditions on f.

Different conditions have been considered by several authors [22, 11, 13, 16]. Of note, Giné and Nickl place a self-similarity condition on f, requiring its regularity to be similar at large and small scales; they then obtain confidence bands which contract adaptively over classes $C^s(M)$, where $M > 0$ is fixed. Hoffmann and Nickl consider a weaker separation condition, which allows adaptation to finitely many classes $C^{s_1}(M), \dots, C^{s_k}(M)$.

The conditions in these two papers are qualitatively different. Hoffmann and Nickl consider a family of functions f which asymptotically contains the full model,

$$\mathcal{F} \coloneqq \bigcup_{i=1}^{k} C^{s_i}(M), \qquad 0 < s_1 < \dots < s_k,\ M > 0. \tag{1.2}$$

The confidence bands constructed are thus eventually valid for all functions $f \in \mathcal{F}$, although the time n after which a band is valid depends on the unknown f. The penalty for this generality comes in the nature of the adaptive result: the bands contract at rates $n^{-s_i/(2s_i+1)}$ for any $f \in C^{s_i}(M)$, but they do not attain the minimax rate $n^{-s/(2s+1)}$ for $f \in C^s(M)$, $s \notin \{s_1, \dots, s_k\}$.

Conversely, Giné and Nickl provide bands attaining $n^{-s/(2s+1)}$ for any $f \in C^s(M)$, $s \in [s_{\min}, s_{\max}]$. However, the family of functions considered does not, even in the limit, contain the full model,

$$\mathcal{F} \coloneqq \bigcup_{s=s_{\min}}^{s_{\max}} C^s(M), \qquad 0 < s_{\min} < s_{\max},\ M > 0. \tag{1.3}$$

Instead, some functions f must be permanently excluded from consideration.

We can describe this difference in terms of dishonest confidence sets. We say a confidence set S_n for f is *honest*, at level $1 - \gamma$, if it satisfies

$$\limsup_{n} \sup_{f \in \mathcal{F}} \mathbb{P}_f(f \notin S_n) \le \gamma, \tag{1.4}$$

where $\mathcal{F}$ is the entire family of functions f we wish to adapt to [20]. Honesty is necessary to produce practical confidence sets; it ensures that there is a known time n, not depending on f, after which the level of the confidence set is not much smaller than $1 - \gamma$. In contrast, a *dishonest* set satisfies the weaker condition

$$\sup_{f \in \mathcal{F}} \limsup_{n} \mathbb{P}_f(f \notin S_n) \le \gamma.$$

While dishonest confidence sets are not useful for inference, they can provide a useful benchmark of nonparametric procedures. Hoffman and Nickl's bands are dishonest confidence sets for the full model (1.2); Giné and Nickl's are not, for the model (1.3).

In the following, we will show that this distinction is intrinsic: that the problem of adapting to finitely many s_i is fundamentally different from adapting to continuous s. We will construct confidence bands which are adaptive in the model (1.3), under a weaker self-similarity condition than Giné and Nickl's; functions satisfying this condition may be considered typical members of any class $C^s(M)$. We will then show that our condition is as weak as possible for adaptation over (1.3), and that no adaptive confidence band can be valid, even dishonestly, for all of (1.3).

We also provide further improvements on past results. Firstly, past constructions of adaptive confidence sets under self-similarity have required *sample splitting*: splitting the data into two groups, one for estimating the function f, and the other for estimating its smoothness. In the construction of our bands, we will show that this procedure can be avoided, leading to smaller constants in the rate of contraction.

We also show that our bands, at no further cost, adapt honestly to the unknown norm M; they are valid even for the model

$$\mathcal{F} := \bigcup_{M=0}^{\infty} \bigcup_{s=s_{\min}}^{s_{\max}} C^s(M), \qquad 0 < s_{\min} < s_{\max}.$$

This is in contrast to previous results, where M is either assumed known, or adapted to only dishonestly.

As our bands make fundamental use of the self-similarity condition, their construction differs significantly from those given previously in the literature. We likewise describe new approaches to undersmoothing, and to linking the white noise model with density estimation and regression, which in this context are valid even for functions of unbounded norm M.

Our bands thus depend on self-similarity parameters ε and ρ, which determine the functions f to be excluded. In a practical setting, suitable values of these parameters might be found via preliminary experiments on a suite of example functions; any choice of parameters will give a confidence band which is maximally adaptive for this problem.

Alternatively, if maximal adaptation is not required, we might view our results as a vindication of the self-similarity approach described by previous authors. We could then, for example, use the simpler method of Giné and Nickl [13], assured that the assumptions demanded are not unreasonable.

In either case, using such methods in practice requires us to interpret the meaning of a self-similarity assumption. In some settings, we may believe such assumptions to be true; for example, in finance, turbulence, or other fractal systems, where some notion of self-similarity is often assumed.

In other settings, we may not believe such assumptions explicitly, but may still be willing to use them as a working model. We will show that functions not satisfying a self-similarity condition form a negligible subset of any Hölder class; our assumptions are therefore not too onerous. This is, of course, no guarantee: it is possible that some structure of the problem will cause our unknown function to lie in such a negligible set. Nevertheless, any method of nonparametric inference must make some assumptions; our results show that, in the context of adaptation, self-similarity is a natural assumption to make.

In Section 2, we describe our self-similarity condition, and in Section 3, we state our main results. We describe the construction of our confidence bands in Section 4, and provide proofs in Section 5.

2. Self-similar functions

To state our results, we must first define our self-similarity condition, using a wavelet basis of $L^2([0,1])$ [15]. We begin with φ and ψ, the scaling function and wavelet of an orthonormal multiresolution analysis on $L^2(\mathbb{R})$. We make the following assumptions on φ and ψ, which are satisfied, for example, by Daubechies wavelets and symlets, with $N \geq 6$ vanishing moments [8, 23].

Assumption 2.1.

(i) For $K \in \mathbb{N}$, φ and ψ are supported on the interval $[1-K, K]$.
(ii) For $N \in \mathbb{N}$, ψ has N vanishing moments:

$$\int_{\mathbb{R}} x^i \psi(x)\, dx = 0, \qquad i = 0, \dots, N-1.$$

(iii) φ is twice continuously differentiable.

Using the construction of Cohen, Daubechies and Vial [7, 6], we can then generate an orthonormal wavelet basis of $L^2([0,1])$, with basis functions

$$\varphi_{j_0,k}, \quad k \in [0, 2^{j_0}), \qquad \text{and} \qquad \psi_{j,k}, \quad j \geq j_0,\, k \in [0, 2^j),$$

for some suitable lower resolution level $j_0 > 0$. For $k \in [N, 2^j - N)$, the basis functions are given by scalings of φ and ψ,

$$\varphi_{j,k}(x) \coloneqq 2^{j/2}\varphi(2^j x - k), \qquad \psi_{j,k} \coloneqq 2^{j/2}\psi(2^j x - k).$$

For other values of k, the basis functions are specially constructed, so as to form an orthonormal basis of $L^2([0,1])$ with desired smoothness properties.

Using this wavelet basis, we may proceed to define the spaces C^s over which we wish to adapt. Given a function $f \in L^2([0,1])$,

$$f = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k},$$

for $s \in (0, N)$, define the C^s norm of f by

$$\|f\|_{C^s} \coloneqq \max\left(\sup_k |\alpha_{j_0,k}|, \sup_{j,\,k} 2^{j(s+1/2)}|\beta_{j,k}|\right).$$

Define the spaces

$$C^s \coloneqq \{f \in L^2([0,1]) : \|f\|_{C^s} < \infty\},$$

and for $M > 0$,

$$C^s(M) \coloneqq \{f \in L^2([0,1]) : \|f\|_{C^s} \leq M\}.$$

For $s \notin \mathbb{N}$, these spaces are equivalent to the classical Hölder spaces; for $s \in \mathbb{N}$, they are equivalent to the Zygmund spaces, which continuously extend the Hölder spaces [7]. In either case, we may therefore take this to be our definition of C^s in the following.

We are now ready to state our self-similarity condition. Denote the wavelet series of f, for resolution levels i to j, $j_0 \leq i \leq j$, by

$$f_{i,j} := \sum_{l=i}^{j-1} \sum_{k=0}^{2^l-1} \beta_{l,k}\psi_{l,k}.$$

Fix some $s_{\max} \in (0, N)$; for any $s \in (0, s_{\max})$, $M > 0$, $\varepsilon \in (0,1)$, and $\rho > 1$, we will say a function $f \in C^s(M)$ is *self-similar*, if

$$\|f_{j,\lceil \rho j \rceil}\|_{C^s} \geq \varepsilon M \ \forall\ j \geq j_0. \tag{2.1}$$

If $s = s_{\max}$, we will instead require (2.1) only for $j = j_0$. Denote the set of self-similar $f \in C^s(M)$ by $C_0^s(M, \varepsilon, \rho)$; for fixed ε, ρ, we will denote this set simply as $C_0^s(M)$.

The above condition ensures that the regularity of f is similar at small and large scales, and will be shown to be necessary to perform adaptive inference. To bound the bias of an adaptive estimator $\hat{f}_n$, we need to know the regularity of f at small scales, which we cannot observe. If f is self-similar, however, we can infer this regularity from the behaviour of f at large scales, which we can observe.

Similar conditions have been considered by previous authors, in the context of turbulence [10, 18], and more recently in statistical applications [22, 13]. We note that Picard and Tribouley's condition can be thought of as a stronger, pointwise version of Giné and Nickl's: if Picard and Tribouley's condition holds at some $x_0 \in [0,1]$, with their ρ_n a constant and $l_j = 1$, this implies the condition of Giné and Nickl.

We now show that our condition (2.1) is weaker than the condition of Giné and Nickl; we will see in Section 3 that it is, in a sense, as weak as possible.

Proposition 2.2. *Given $s_{\min} \in (0, s_{\max}]$, $b > 0$, $0 < b_1 \leq b_2$, and $J_1 \geq j_0$, there exist $M > 0$, $\varepsilon \in (0,1)$, and $\rho > 1$ such that, for any $s \in [s_{\min}, s_{\max}]$, the condition*

$$f \in C^s \cap C^{s_{\min}}(b), \qquad b_1 2^{-js} \leq \|f_{j,\infty}\|_\infty \leq b_2 2^{-js} \ \forall\ j \geq J_1, \tag{2.2}$$

implies $f \in C_0^s(M, \varepsilon, \rho)$. Conversely, given $s \in (0, s_{\max}]$, $M > 0$, $\varepsilon \in (0,1)$, and $\rho > 1$, there exist $f \in C_0^s(M, \varepsilon, \rho)$ which do not satisfy the above condition, for any $s_{\min} \in (0, s]$, $b > 0$, $0 < b_1 \leq b_2$, and $J_1 \geq j_0$.

In fact, we can show that self-similarity is a generic property: that the set $\mathcal{D}$ of self-dissimilar functions, which for some s never satisfy (2.1), is in more than one sense negligible. Firstly, we can show that $\mathcal{D}$ is nowhere dense: the self-dissimilar functions cannot approximate any open set in $C^s(M)$. In particular, this means that $\mathcal{D}$ is meagre. Secondly, we can show that $\mathcal{D}$ is a null set, for a natural probability measure π on $C^s(M)$. We thus have that π-almost-every function in $C^s(M)$ is self-similar.

Proposition 2.3. *For $s \in (0, s_{\max}]$ and $M > 0$, define*

$$\mathcal{D} \coloneqq C^s(M) \setminus \bigcup_{\varepsilon \in (0,1),\, \rho > 1} C_0^s(M, \varepsilon, \rho).$$

Further define a probability measure π on $f \in C^s(M)$, with f having independently distributed wavelet coefficients,

$$\alpha_{j_0,k} \sim M 2^{-j_0(s+1/2)} U([-1,1]), \qquad \beta_{j,k} \sim M 2^{-j(s+1/2)} U([-1,1]).$$

Then:

(i) $\mathcal{D}$ is nowhere dense in the norm topology of $C^s(M)$; and
(ii) $\pi(\mathcal{D}) = 0$.

These results are already known for Giné and Nickl's condition (2.2) [13, 16]; as a consequence of Proposition 2.2, they also hold for our condition (2.1). We conclude that the self-similar functions may be considered typical members of any class $C^s(M)$.

For smoother functions, we note that when $s = s_{\max}$, our condition (2.1) is weaker, and no longer requires a specific smoothness of f. Indeed, in this case, the smoother f is, the easier it becomes to satisfy our condition. We may therefore expect the condition to likewise often hold for smoother functions.

We further note that Proposition 2.3 shows, for a particular Bayesian prior on functions f, that self-similarity is implicitly assumed. In fact, this result is more general, applying to many priors which can be written as series expansions, including Gaussian processes with a Karhunen-Loève expansion. As these results will often involve different bases or scaling laws, we do not pursue this further, except to comment that self-similarity is thus also a common modelling assumption in Bayesian nonparametrics.

3. Self-similarity and adaptation

We are now ready to state our main results. First, however, we will require an additional assumption on our wavelet basis, allowing us to precisely control the variance of our estimators. This assumption has been verified analytically for Battle-Lemarié wavelets [12]; for compactly supported wavelets, it can be tested with provably good numerical approximations. The assumption is known to hold for Daubechies wavelets and symlets, with $N = 6, \ldots, 20$ vanishing moments. Larger values of N, and other wavelet bases, can be easily checked, and the assumption is conjectured to hold also in those cases [3].

Assumption 3.1. *The 1-periodic function*

$$\sigma_\varphi^2(t) \coloneqq \sum_{k \in \mathbb{Z}} \varphi(t-k)^2$$

attains its maximum $\overline{\sigma}_\varphi^2$ at a unique point $t_0 \in [0,1)$, and $(\sigma_\varphi^2)''(t_0) < 0$.

We may now begin with the construction of an exact confidence band, which has exact asymptotic level $1 - \gamma$. Exact confidence bands are often preferred in the literature, being simpler to compute, and offering more reliable control over coverage [14].

To obtain exact coverage, our bands are centred at an undersmoothed estimate of f: an estimate slightly rougher than optimal, chosen so that the known variance dominates the unknown bias. The larger variance does mean, however, that our bands adapt to s and M only up to a logarithmic rate penalty. We state our results for the white noise model, which serves as an idealisation of density estimation and regression; we will return later to consequences for the other models.

Theorem 3.2. *In the white noise model, fix $0 < \gamma < 1$, $s_{\min} \in (0, s_{\max}]$, and set*

$$r_n(s) \coloneqq (n/\log n)^{-s/(2s+1)} \log n, \qquad \mathcal{F} \coloneqq \bigcup_{s\in[s_{\min},s_{\max}],\, M>0} C_0^s(M).$$

There exists a confidence band $C_n^{ex} \coloneqq C_n^{ex}(\gamma, s_{\min}, s_{\max}, \varepsilon, \rho)$ as in (1.1), with radius R_n^{ex}, satisfying:

(i) $\sup_{f\in\mathcal{F}} |\mathbb{P}(f \notin C_n^{ex}) - \gamma| \to 0$; *and*
(ii) for a fixed constant $L > 0$, and any $s \in [s_{\min}, s_{\max}]$, $M > 0$,

$$\sup_{f\in C_0^s(M)} \mathbb{P}_f \left(R_n^{ex} > LM^{1/(2s+1)} r_n(s) \right) \to 0.$$

In asymptotic terms, we can do better by dropping the requirement of exactness. Intuitively, we may feel that an exact band should always be preferable: given an inexact band, surely we can modify it to produce something more accurate? In fact, this is not necessarily the case. Consider a simplified statistical model, where we wish to identify a parameter $\theta \in \mathbb{R}$, and have the luxury of observing data $X = \theta$. The optimal confidence set for θ is thus $\{X\}$, but this set is not exact at the 95% level. We can produce an exact set by adding noise: if $Z \sim N(0, 1)$, the confidence set

$$\{x \in \mathbb{R} : |X + Z - x| \le \Phi^{-1}(0.975)\}$$

is exact at the 95% level. However, the perfect, inexact set is more accurate than the imperfect, exact one.

The situation is similar in nonparametrics: we can obtain better asymptotic results using an inexact band, whose asymptotic level is unknown, but is guaranteed to be at least $1 - \gamma$. We now construct inexact bands, centred at an adaptive Lepski-type estimator, which are exact rate-adaptive with respect to s and M.

Theorem 3.3. *In the white noise model, fix $0 < \gamma < 1$, and set*

$$r_n(s) \coloneqq (n/\log n)^{-s/(2s+1)}, \qquad \mathcal{F} \coloneqq \bigcup_{s\in(0,s_{\max}],M>0} C_0^s(M).$$

There exists a confidence band $C_n^{ad} \coloneqq C_n^{ad}(\gamma, s_{\max}, \varepsilon, \rho)$ as in (1.1), with radius R_n^{ad}, satisfying:

(i) $\limsup_n \sup_{f \in \mathcal{F}} \mathbb{P}(f \notin C_n^{ad}) \leq \gamma$*; and*
(ii) for a fixed constant $L > 0$, and any $s \in (0, s_{\max}]$, $M > 0$,

$$\sup_{f \in C_0^s(M)} \mathbb{P}_f \left(R_n^{ad} > \frac{LM^{1/(2s+1)}}{2^s - 1} r_n(s) \right) \to 0.$$

The constant in the above rate contains an extra $1/(2^s - 1)$ term, which is present to allow for s tending to 0. Note that if, as before, we restrict to $s \geq s_{\min} > 0$, we may then fold this term into the constant L, producing a rate of the same form as in Theorem 3.2.

As is standard, the rates adapt only to smoothnesses $s \leq s_{\max}$; if f is smoother than our wavelet basis, we cannot reliably detect this from the wavelet coefficients. As noted in Section 2, however, our self-similarity condition (2.1) is weaker when $s = s_{\max}$, and the class $C_0^{s_{\max}}(M)$ contains many smoother functions f; in this case we obtain the rate of contraction optimal for $C^{s_{\max}}(M)$.

Theorem 3.3 is, in more than one sense, maximal. Firstly, we can verify that the minimax rate of estimation over $C_0^s(M)$ is the same as over $C^s(M)$. Since any adaptive confidence band must be centred at an adaptive estimator, we may conclude that the above results are indeed optimal.

Theorem 3.4. *In the white noise model, fix $0 < \gamma < \frac{1}{2}$, $s \in (0, s_{\max}]$, $M > 0$. An estimator $\hat{f}_n$ cannot satisfy*

$$\limsup_n \sup_{f \in C_0^s(M)} \mathbb{P}_f \left(\|\hat{f}_n - f\|_\infty \geq r_n \right) \leq \gamma,$$

for any rate $r_n = o((n/\log n)^{-s/(2s+1)})$.

Secondly, we can show that the self-similarity condition (2.1) is, in a sense, as weak as possible. In (2.1), the function f is required to have significant wavelet coefficients on resolution levels j growing at most geometrically. If we relax this assumption even slightly, allowing the significant coefficients to occur less often, then adaptive inference is impossible.

For $s \in (0, s_{\max})$, $M > 0$, denote by $C_1^s(M)$ the set of $f \in C^s(M)$ satisfying the slightly weaker self-similarity condition,

$$\|f_{j, \lceil \rho_j j \rceil}\|_{C^s} \geq \varepsilon M \;\forall\; j \geq j_0,$$

for fixed $\varepsilon > 0$, and $\rho_j > 1$, $\rho_j \to \infty$. Even allowing dishonesty, and with known bound M on the Hölder norm, we cannot construct a confidence band which adapts to classes $C_1^s(M)$.

Theorem 3.5. *In the white noise model, fix $0 < \gamma < \frac{1}{2}$, $0 < s_{\min} < s_{\max}$, and $M > 0$. Set*

$$r_n(s) \coloneqq (n/\log n)^{-s/(2s+1)}, \qquad \mathcal{F} \coloneqq \bigcup_{s \in (s_{\min}, s_{\max})} C_1^s(M).$$

A confidence band S_n, with radius R_n, cannot satisfy:

(i) $\limsup_n \mathbb{P}_f(f \notin S_n) \leq \gamma$, *for all* $f \in \mathcal{F}$; *and*
(ii) $R_n = O_p(r_n(s))$ *under* $\mathbb{P}_f$, *for all* $f \in C_1^s(M)$, $s \in (s_{\min}, s_{\max})$.

As a consequence, we firstly cannot adapt to the full classes $C^s(M)$. More importantly, we cannot obtain adaptation merely by removing elements of the classes $C^s(M)$ which are asymptotically negligible, as Hoffmann and Nickl do for the model (1.2). In order to construct adaptive bands, we must fully exclude some functions f from consideration, as Giné and Nickl do for the model (1.3).

The difference between these problems lies in the accuracy to which we must estimate s. To distinguish between finitely many classes, we need to know s only up to a constant; to adapt to a continuum of smoothness, we must know it with error shrinking like $1/\log n$. The finite-class problem is in this sense more like the L^2 adaptation problem [4]; the distinctive nature of the L^∞ adaptation problem is revealed only when requiring adaptation to continuous s.

While the above theorems are stated for the white noise model, we can prove similar results for density estimation and regression. The following theorem gives a construction of adaptive bands in these models; other results can be proved similarly to previous results in the literature [13, 4].

Theorem 3.6. *In the density estimation model, let* $s_{\min} \in (0, s_{\max}]$, *or in the regression model,* $s_{\min} \in [\frac{1}{2}, s_{\max}]$. *In either model, the statement of Theorem 3.3 remains true, for the family*

$$\mathcal{F} := \bigcup_{s \in [s_{\min}, s_{\max}],\, M > 0} C_0^s(M),$$

and with constants L, L' *depending on* s *and* M.

4. Constructing adaptive bands

To construct confidence bands satisfying the conditions in Section 3, we will use estimators $\hat{f}_n$ given by truncated empirical wavelet expansions,

$$\hat{f}(\hat{j}_n) := \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k} \varphi_{j_0,k} + \sum_{j=j_0}^{\hat{j}_n - 1} \sum_{k=0}^{2^j - 1} \hat{\beta}_{j,k} \psi_{j,k},$$

for the empirical wavelet coefficients

$$\hat{\alpha}_{j_0,k} := \int \varphi_{j_0,k}(t)\, dY_t, \qquad \hat{\beta}_{j,k} := \int \psi_{j,k}(t)\, dY_t.$$

The resolution levels $\hat{j}_n$ will also depend on the data Y, and will be chosen to produce adaptive estimators $\hat{f}(\hat{j}_n)$.

We will consider several different choices of resolution level, corresponding to different properties of the function f, and the class $C^s(M)$ to which it belongs.

We first consider the adaptive resolution choice j_n^{ad}, chosen in terms of the function f. Pick sequences $j_n^{\min}, j_n^{\max} \in \mathbb{N}$, $j_0 \le j_n^{\min} \le j_n^{\max}$, so that $2^{j_n^{\min}} \sim (n/\log n)^{1/(2N+1)}$, and $2^{j_n^{\max}} \sim n/\log n$. Further define

$$c_{n,\mu} \coloneqq (n/(\log n)^{\mu})^{-1/2},$$

and for $\kappa > 0$, $\mu \ge 1$, let

$$j_n^{\mathrm{ad}}(\kappa,\mu) \coloneqq \inf\left(\{j_n^{\max}\} \cup \{j \in [j_n^{\min}, j_n^{\max}) : \sup_k |\beta_{j,k}| \le \kappa c_{n,\mu}\}\right).$$

While j_n^{ad} is unknown, we can estimate it by a Lepski-type resolution choice,

$$\hat{j}_n^{\mathrm{ad}}(\kappa,\mu) \coloneqq \inf\left(\{j_n^{\max}\} \cup \{j \in [j_n^{\min}, j_n^{\max}) : \sup_k |\hat{\beta}_{j,k}| \le \kappa c_{n,\mu}\}\right),$$

which depends only on the data. Fix $\lambda > \sqrt{2}$, $\nu \ge 1$, and for convenience set $\hat{j}_n^{\mathrm{ad}} \coloneqq \hat{j}_n^{\mathrm{ad}}(\lambda,\nu)$. If $\nu = 1$, we will see $\hat{f}(\hat{j}_n^{\mathrm{ad}})$ is then an adaptive estimator of f; if $\nu > 1$, it is near-adaptive.

While the above statements are true for general f, they do not provide us with an estimate of the error in $\hat{f}_n$. To produce confidence bands, we must estimate the smoothness of f, and this is where self-similarity is required. We will consider values of the truncated Hölder norm,

$$M_{i,j}^s \coloneqq \|f_{i,j}\|_{C^s},$$

which measures the smoothness of f at resolution levels i to j. We may bound $M_{i,j}^s$ by the quantities

$$\begin{aligned}
\underline{M}_{i,j}^s &\coloneqq \sup_{l=i}^{j-1} \sup_k 2^{l(s+1/2)}(|\hat{\beta}_{l,k}| - \sqrt{2}c_{n,1})^+,\\
\overline{M}_{i,j}^s &\coloneqq \sup_{l=i}^{j-1} \sup_k 2^{l(s+1/2)}(|\hat{\beta}_{l,k}| + \sqrt{2}c_{n,1}),
\end{aligned}$$

where $x^+ \coloneqq \max(x,0)$, and we will show in Section 5.2 that for $j \le j_n^{\max}$, $M_{i,j}^s \in [\underline{M}_{i,j}^s, \overline{M}_{i,j}^s]$ with high probability.

Set $J_0 = j_0$, $J_1 = \lceil \rho J_0 \rceil$, $J_2 = \lfloor \hat{j}_n^{\mathrm{ad}}/\rho \rfloor$, $J_3 = \hat{j}_n^{\mathrm{ad}}$, and suppose n is large enough that $j_n^{\min} \ge \rho J_1$, so $J_0 < J_1 \le J_2 < J_3$. If $f \in C_0^s(M)$ for $s < s_{\max}$, then with high probability,

$$R(s) \coloneqq \frac{\overline{M}_{J_2,J_3}^s}{\underline{M}_{J_0,J_1}^s} \ge \frac{M_{J_2,J_3}^s}{M_{J_0,J_1}^s} \ge \varepsilon.$$

Assuming further $s \ge s_{\min}$, for some $s_{\min} \ge 0$, we can lower bound s by

$$\hat{s}_n \coloneqq \inf(\{s_{\max}\} \cup \{s \in [s_{\min}, s_{\max}) : R(s) \ge \varepsilon\}).$$

Since

$$R(s) = \frac{\overline{M}_{J_2,J_3}^s 2^{-J_1(s+1/2)}}{\underline{M}_{J_0,J_1}^s 2^{-J_1(s+1/2)}}$$

is increasing in s, $\hat{s}_n$ can be found efficiently using binary search.

Likewise, set

$$M(s) \coloneqq \varepsilon^{-1}\overline{M}^{s}_{J_0,J_1},$$

and $\hat{M}_n \coloneqq M(\hat{s}_n)$. With high probability,

$$M(s)2^{-J_1(s+1/2)} \geq \varepsilon^{-1}M^s_{J_0,J_1}2^{-J_1(s+1/2)} \geq M2^{-J_1(s+1/2)},$$

and as the LHS is decreasing in s, also

$$\hat{M}_n 2^{-J_1(\hat{s}_n+1/2)} \geq M2^{-J_1(s+1/2)}.$$

Using these bounds, we can control the error in $\hat{f}$, producing adaptive confidence bands for f.

To construct the bands, we will introduce some more resolution choices $\hat{j}_n$. Firstly, we consider the class resolution choice j_n^{cl}, chosen in terms of the class $C^s(M)$. For $\kappa > 0$, $\mu \geq 1$, define

$$\begin{aligned} j_n^{\mathrm{cl}}(\kappa,\mu) &\coloneqq \inf\{j \geq j_n^{\min} : M2^{-j(s+1/2)} \leq \kappa c_{n,\mu}\} \\ &= \max\left(j_n^{\min},\, \lceil \log_2(M/\kappa c_{n,\mu})/(s+\tfrac{1}{2})\rceil\right), \end{aligned} \tag{4.1}$$

which we can estimate by

$$\hat{j}_n^{\mathrm{cl}}(\kappa,\mu) \coloneqq \max\left(j_n^{\min},\, \lceil \log_2(\hat{M}_n/\kappa c_{n,\mu})/(\hat{s}_n+\tfrac{1}{2})\rceil\right). \tag{4.2}$$

Secondly, to produce exact confidence bands, we will need the undersmoothed resolution choice j_n^{ex}. Fix $u_n \in \mathbb{N}$, $2^{u_n} \sim \log n$, and set

$$j_n^{\mathrm{ex}}(\kappa,\mu) \coloneqq j_n^{\mathrm{cl}}(\kappa,\mu) + \lceil \log_2 j_n^{\mathrm{cl}}(\kappa,\mu)\rceil + u_n,$$

defining $\hat{j}_n^{\mathrm{ex}}$ similarly, in terms of $\hat{j}_n^{cl}$. Let $\overline{\lambda} \coloneqq \lambda + \sqrt{2}$, and $\underline{\lambda} \coloneqq \lambda - \sqrt{2}$; for convenience, write $j_n^{cl} \coloneqq j_n^{cl}(\overline{\lambda},1)$, $j_n^{\mathrm{ex}} \coloneqq j_n^{\mathrm{ex}}(\underline{\lambda},1)$, and likewise $\hat{j}_n^{cl}$, $\hat{j}_n^{\mathrm{ex}}$.

We may now proceed to define our bands. Let

$$\begin{aligned} a(j) &\coloneqq \sqrt{2\log(2)j}, \\ b(j) &\coloneqq a(j) - \frac{\log(\pi\log 2) + \log j - \frac{1}{2}\log(1+\upsilon_\varphi)}{2a(j)}, \\ c(j) &\coloneqq \overline{\sigma}_\varphi n^{-1/2}2^{j/2}, \\ x(\gamma) &\coloneqq -\log\left(-\log(1-\gamma)\right), \\ R_1(j,\gamma) &\coloneqq c(j)\left(\frac{x(\gamma)}{a(j)} + b(j)\right), \\ l(j) &\coloneqq \max(j, \min(\hat{j}_n^{cl}, j_n^{\max})), \\ R_2(j) &\coloneqq \tau_\varphi\overline{\lambda}(2^{l(j)/2} - 2^{j/2})c_{n,\nu}/(\sqrt{2}-1), \\ R_3(j) &\coloneqq \begin{cases} \tau_\varphi \hat{M}_n 2^{-l(j)\hat{s}_n}/(1-2^{-\hat{s}_n}) & \hat{s}_n > 0, \\ \infty, & \hat{s}_n = 0, \end{cases} \end{aligned}$$

where $\overline{\sigma}_\varphi$ is given by Assumption 3.1,

$$\tau_\varphi := \sup_{t\in[0,1]} 2^{-j_0/2}\sum_{k\in\mathbb{Z}}|\psi_{j_0,k}(t)|,$$

and

$$\upsilon_\varphi := -\frac{\sum_{k\in\mathbb{Z}}\varphi'(t_0-k)^2}{\overline{\sigma}_\varphi\sigma''_\varphi(t_0)}.$$

If we set $s_{\min} > 0$, $\nu > 1$, the undersmoothed resolution choice $\hat{j}_n^{\mathrm{ex}}$, with confidence radius

$$R_n^{\mathrm{ex}} := R_1(\hat{j}_n^{\mathrm{ex}},\gamma),$$

will be shown to give a band C_n^{ex} satisfying Theorem 3.2. If instead we set $s_{\min} = 0$, $\nu = 1$, and define

$$\gamma_n := \gamma/(j_n^{\max} - j_n^{\min} + 1),$$

then the adaptive resolution choice $\hat{j}_n^{\mathrm{ad}}$, with confidence radius

$$R_n^{\mathrm{ad}} := R_1(\hat{j}_n^{\mathrm{ad}},\gamma_n) + R_2(\hat{j}_n^{\mathrm{ad}}) + R_3(\hat{j}_n^{\mathrm{ad}}),$$

will be shown to give a band C_n^{ad} satisfying Theorem 3.3.

5. Proofs

5.1. Results on self-similarity

We begin by establishing that our self-similarity condition (2.1) is weaker than Giné and Nickl's condition (2.2).

Proof of Proposition 2.2. We first consider the case $s < s_{\max}$. Given (2.2), for $j \ge J_1$, $k \in [N, 2^j - N)$, we obtain

$$|\beta_{j,k}| = |\langle f_{j,\infty},\psi_{j,k}\rangle| \le \|f_{j,\infty}\|_\infty\|\psi_{j,k}\|_1 \le b_2\|\psi\|_1 2^{-j(s+1/2)},$$

and similar bounds for $k \in [0,N)\cup[2^j - N, 2^j)$. We thus conclude $f \in C^s(M)$, for a constant $M > 0$.

We will choose $\varepsilon \in (0,1)$ small, $\rho > 1$ large, so that $\rho j_0 \ge J_1$, and

$$C := M(\varepsilon + 2^{-(\rho j_0 - J_1)s})$$

is small. If $f \notin C_0^s(M)$, we have $J_2 \ge j_0$ such that

$$|\beta_{j,k}| < \varepsilon M 2^{-j(s+1/2)},$$

for all $j \in [J_2, \lceil\rho J_2\rceil)$, $k \in [0, 2^j)$. Let $J_3 := \max(J_1, J_2)$. Then

$$\begin{aligned}\|f_{J_3,\infty}\|_\infty &\lesssim M\left(\sum_{j=J_3}^{\lceil\rho J_2\rceil-1}\varepsilon 2^{-js} + \sum_{j=\lceil\rho J_2\rceil}^{\infty} 2^{-js}\right)\\ &\lesssim M\left(\varepsilon 2^{-J_3 s} + 2^{-\rho J_2 s}\right) \lesssim C 2^{-J_3 s},\end{aligned}$$

contradicting (2.2) for C small. Thus, given (2.2), we have M, ε, and ρ for which $f \in C_0^s(M)$.

Conversely, given $s \in (0, s_{\max}]$, $M > 0$, $\varepsilon \in (0,1)$, and $\rho > 1$, for $i \in \mathbb{N}$ set $j_i \coloneqq \lceil \rho j_{i-1} \rceil$, and consider the function

$$f \coloneqq \sum_{i=0}^{\infty} M 2^{-j_i(s+1/2)} \psi_{j_i, 2^{j_i - 1}}$$

in $C_0^s(M)$. We have

$$\|f_{j_n+1,\infty}\|_\infty \lesssim M \sum_{i=n+1}^{\infty} 2^{-j_i s} \lesssim 2^{-j_{n+1} s} = o(2^{-j_n s})$$

as $n \to \infty$, so f does not satisfy (2.2) for any $s_{\min}$, b, b_1, b_2, and J_1. As our self-similarity condition is weaker for $s = s_{\max}$, the same is true also in that case. □

5.2. Constructive results

We now prove our results on the existence of adaptive confidence bands. To proceed, we will decompose the error in estimates $\hat{f}(j)$ into variance and bias terms,

$$\|\hat{f}(j) - f\|_\infty \le \|\hat{f}(j) - \bar{f}(j)\|_\infty + \|\bar{f}(j) - f\|_\infty,$$

where

$$\bar{f}(j) \coloneqq \mathbb{E}_f[\hat{f}(j)] = f_{j_0,j}.$$

To control the variance, we will need the following result, which is a rephrasing of a Smirnov-Bickel-Rosenblatt theorem for compactly-supported wavelets [3].

Lemma 5.1. *Let $0 < \gamma_n \le \gamma_0 < 1$, and $\gamma_n^{-1} = o(n^{-\alpha})$, for all $\alpha > 0$. Then as $n \to \infty$, uniformly in $f \in L^2([0,1])$,*

$$\sup_{j \ge j_n^{\min}} \left| \gamma_n^{-1} \mathbb{P}\left(a(j) \left(\frac{\|\hat{f}(j) - \bar{f}(j)\|_\infty}{c(j)} - b(j) \right) > x(\gamma_n) \right) - 1 \right| \to 0.$$

To bound the bias, we must control the estimators $\hat{j}_n$, $\hat{s}_n$ and $\hat{M}_n$. We will show that, on events E_n with probability tending to 1, these estimators are close to the quantities they bound.

Lemma 5.2. *Set $\underline{j}_n^{ad} \coloneqq j_n^{ad}(\overline{\lambda}, \nu)$, $\overline{j}_n^{ad} \coloneqq j_n^{ad}(\underline{\lambda}, \nu)$. For $s \in [s_{\min}, s_{\max}]$, $M > 0$, and $f \in C_0^s(M)$, we have events E_n, with $\mathbb{P}(E_n) \to 1$ uniformly, on which:*

- *(i)* $\underline{j}_n^{ad} \le \hat{j}_n^{ad} \le \overline{j}_n^{ad}$;
- *(ii)* $\hat{s}_n \le s$, *and* $\hat{M}_n 2^{-J_1(\hat{s}_n+1/2)} \ge M 2^{-J_1(s+1/2)}$; *and*
- *(iii)* $\hat{s}_n \ge s_n$, *and* $\hat{M}_n \le M_n$;

for sequences M_n, s_n satisfying

$$M_n/M \to \varepsilon^{-1}, \qquad \log_2(n)(s - s_n) \to S,$$

uniformly over $f \in C_0^s(M)$, with constant $S > 0$ depending on N, ε, ρ, and λ. Also on E_n, for any $0 < \kappa \le \lambda + \sqrt{2}$, $1 \le \mu \le \nu$:

(iv) $\hat{j}_n^{cl}(\kappa,\mu) \ge \hat{j}_n^{ad}$;
(v) $j_n^{cl}(\kappa,\mu) \le \hat{j}_n^{cl}(\kappa,\mu) \le j_n^{cl}(\kappa,\mu) + J_n^{cl}(\kappa,\mu)$; *and*
(vi) $j_n^{ex}(\kappa,\mu) \le \hat{j}_n^{ex}(\kappa,\mu) \le j_n^{ex}(\kappa,\mu) + J_n^{ex}(\kappa,\mu)$;

for sequences $J_n^{cl}(\kappa,\mu), J_n^{ex}(\kappa,\mu) \to 2(1 + \log_2(\varepsilon^{-1}) + S)$, uniformly over $f \in C_0^s(M)$.

Proof. For n such that $j_n^{\min} < \rho\lceil \rho j_0 \rceil$, set $E_n \coloneqq \emptyset$. Otherwise, let E_n be the event that

$$\max\left(\sup_k |\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}|, \sup_{j=j_0}^{j_n^{\max}-1} \sup_k |\hat{\beta}_{j,k} - \beta_{j,k}|\right) \le \sqrt{2} c_{n,1}. \tag{5.1}$$

Now, for n large enough that $E_n \ne \emptyset$, we have

$$\begin{aligned}\mathbb{P}(E_n^c) &\le 2^{j_n^{\max}} \Phi(-\sqrt{2n} c_{n,1}) \\ &\le (\pi \log n)^{-1/2} 2^{j_n^{\max}} n^{-1} \\ &= O\left((\log n)^{-3/2}\right) = o(1),\end{aligned}$$

using the bound that $\Phi(-x) \le \phi(x)/x$ for $x > 0$.

(i) If $\underline{j}_n^{\mathrm{ad}} = j_n^{\min}$, then trivially $\hat{j}_n^{\mathrm{ad}} \ge \underline{j}_n^{\mathrm{ad}}$. Otherwise, for $j = \underline{j}_n^{\mathrm{ad}} - 1$, we have some k such that $|\beta_{j,k}| > \overline{\lambda} c_{n,\nu}$. Thus, on E_n,

$$|\hat{\beta}_{j,k}| \ge |\beta_{j,k}| - \sqrt{2} c_{n,1} > \lambda c_{n,\nu},$$

and again $\hat{j}_n^{\mathrm{ad}} \ge \underline{j}_n^{\mathrm{ad}}$. Similarly, for all $\overline{j}_n^{\mathrm{ad}} \le j < j_n^{\max}$, k,

$$|\hat{\beta}_{j,k}| \le |\beta_{j,k}| + \sqrt{2} c_{n,1} < \lambda c_{n,\nu},$$

so $\hat{j}_n^{\mathrm{ad}} \le \overline{j}_n^{\mathrm{ad}}$.

(ii) On E_n, we have

$$M_{i,j}^s \in [\underline{M}_{i,j}^s, \overline{M}_{i,j}^s],$$

for any $i \le j \le j_n^{\max}$. If $s < s_{\max}$, by the argument given in Section 4, we then obtain

$$\hat{s}_n \le s, \qquad \hat{M}_n 2^{-j_1(\hat{s}_n + 1/2)} \ge M 2^{-j_1(s+1/2)}.$$

If $s = s_{\max}$, the results follow similarly, noting that $\hat{s}_n \le s_{\max}$ by definition.

(iii) On E_n, $J_3 = \hat{j}_n^{\mathrm{ad}} \le \overline{j}_n^{\mathrm{ad}} \le j_n^{cl}(\underline{\lambda}, \nu)$, and for n large $j_n^{\mathrm{cl}}(\underline{\lambda}, \nu) > j_n^{\min}$, so

$$d_n \coloneqq c_{n,1} 2^{J_3(s+1/2)} \le c_{n,\nu} 2^{J_3(s+1/2)} \le M \underline{\lambda}^{-1},$$

and also

$$e_n \coloneqq c_{n,1} 2^{J_1(s+1/2)} \to 0.$$

We then obtain

$$R(s) \le \frac{\underline{M}^s_{J_2,J_3} + 2\sqrt{2} d_n}{\overline{M}^s_{J_0,J_1} - 2\sqrt{2} e_n} \le \frac{M + 2\sqrt{2} d_n}{\varepsilon M - 2\sqrt{2} e_n} \le R_n,$$

for a sequence

$$R_n \to \varepsilon^{-1}(1 + 2\sqrt{2}\underline{\lambda}^{-1}) =: R.$$

On E_n, $\hat{s}_n \le s \le s_{\max}$ by (ii), so if $\hat{s}_n = s_{\max}$, we are done. If not, then $R(\hat{s}_n) \ge \varepsilon$, and

$$2^{(J_2 - J_1)(s - \hat{s}_n)} \le \frac{\overline{M}^s_{J_2,J_3} / \overline{M}^{\hat{s}_n}_{J_2,J_3}}{\underline{M}^s_{J_0,J_1} / \underline{M}^{\hat{s}_n}_{J_0,J_1}} = \frac{R(s)}{R(\hat{s}_n)} \le \frac{R_n}{\varepsilon}.$$

Since

$$J_2 - J_1 \ge \lfloor j_n^{\min} / \rho \rfloor - J_1 =: \delta_n,$$

we have

$$\hat{s}_n \ge s - \log_2(\varepsilon^{-1} R_n)/\delta_n =: s_n,$$

and since $\delta_n \sim \log_2(n)/\rho(2N+1)$,

$$\log_2(n)(s - s_n) \to \rho(2N+1)\log_2(\varepsilon^{-1} R) =: S.$$

Likewise,

$$\hat{M}_n \le M(s) \le \varepsilon^{-1}(\underline{M}^s_{J_0,J_1} + 2\sqrt{2} e_n) \le \varepsilon^{-1}(M + 2\sqrt{2} e_n) \le M_n,$$

for a sequence $M_n > 0$, with $M_n / M \to \varepsilon^{-1}$.

(iv) If $\hat{j}_n^{\mathrm{ad}} = j_n^{\min}$, then trivially $\hat{j}_n^{\mathrm{cl}}(\kappa, \mu) \ge \hat{j}_n^{\mathrm{ad}}$. If not, for $j = \hat{j}_n^{\mathrm{ad}} - 1$, we have some k such that $|\hat{\beta}_{j,k}| > \lambda c_{n,\nu}$. Hence, on E_n,

$$\hat{M}_n 2^{-j(\hat{s}_n + 1/2)} > \varepsilon^{-1}(\lambda + \sqrt{2}) c_{n,\nu} \ge \kappa c_{n,\mu},$$

and again $\hat{j}_n^{\mathrm{cl}}(\kappa, \mu) \ge \hat{j}_n^{\mathrm{ad}}$.

(v) On E_n, by the above we have

$$M 2^{-\hat{j}_n^{\mathrm{cl}}(\kappa,\mu)(s+1/2)} \le \hat{M}_n 2^{-\hat{j}_n^{\mathrm{cl}}(\kappa,\mu)(\hat{s}_n + 1/2)} \le \kappa c_{n,\mu},$$

and so $\hat{j}_n^{\mathrm{cl}}(\kappa, \mu) \ge j_n^{\mathrm{cl}}(\kappa, \mu)$. Equally, from (4.1), (4.2) and the above, we obtain

$$\begin{aligned} \hat{j}_n^{cl}(\kappa, \mu) - j_n^{cl}(\kappa, \mu) &\le 2 + 2\log_2(\hat{M}_n / M) + 4\lceil \log_2(\sqrt{n} M / \kappa) \rceil (s - \hat{s}_n) \\ &\le J_n^{\mathrm{cl}}(\kappa, \mu), \end{aligned}$$

for a sequence $J_n^{\mathrm{cl}}(\kappa, \mu) \to 2(1 + \log_2(\varepsilon^{-1}) + S)$.

(vi) From (v), we also have

$$\hat{j}_n^{\mathrm{ex}}(\kappa,\mu) - j_n^{\mathrm{ex}}(\kappa,\mu) \le J_n^{\mathrm{ex}}(\kappa,\mu),$$

for a sequence $J_n^{\mathrm{ex}}(\kappa,\mu) \to 2(1+\log_2(\varepsilon^{-1})+S)$. □

We may now bound the bias of $\hat{f}$ with the estimators $\hat{j}_n$, $\hat{s}_n$ and $\hat{M}_n$, which bound the true parameters by the above lemma.

Lemma 5.3. *Let $\hat{j}_n \ge \hat{j}_n^{ad}$. On events E_n as in Lemma 5.2, for any $s \in [s_{\min}, s_{\max}]$, $M > 0$, and $f \in C_0^s(M)$,*

$$\|\bar{f}(\hat{j}_n) - f\|_\infty \le R_2(\hat{j}_n) + R_3(\hat{j}_n).$$

Proof. If $\hat{s}_n = 0$, this is trivial. If not, then by the construction of the wavelet basis,

$$\tau = \sup_{j=j_0}^{\infty} \sup_{t\in[0,1]} 2^{-j/2}\sum_{k\in\mathbb{Z}}|\psi_{j,k}(t)|.$$

Further, by Lemma 5.2, on E_n we have $\hat{j}_n \ge \hat{j}_n^{\mathrm{ad}} \ge \underline{j}_n^{ad}$, and for $j \ge \hat{j}_n$, $M2^{-j(s+1/2)} \le \hat{M}_n 2^{-j(\hat{s}_n+1/2)}$. Thus

$$\begin{aligned}
\|\bar{f}(\hat{j}_n) - f\|_\infty = \|f_{\hat{j}_n,\infty}\|_\infty &\le \tau_\varphi \sum_{j=\hat{j}_n}^{\infty} \sup_k 2^{j/2}|\beta_{j,k}| \\
&\le \tau_\varphi \left(\sum_{j=\hat{j}_n}^{l(\hat{j}_n)-1} 2^{j/2}\overline{\lambda} c_{n,\nu} + \sum_{j=l(\hat{j}_n)}^{\infty} \hat{M}_n 2^{-j\hat{s}_n} \right) \\
&\le R_2(\hat{j}_n) + R_3(\hat{j}_n).
\end{aligned}$$

□

We are now ready to prove our theorems. First, we consider the exact band C_n^{ex}.

Proof of Theorem 3.2.

(i) Let $d(j,x) \coloneqq a(j)\left(c(j)^{-1}x - b(j)\right)$, and define the terms

$$\begin{aligned}
F(j) &\coloneqq d(j, \|\hat{f}(j) - f\|_\infty), \\
G(j) &\coloneqq d(j, \|\hat{f}(j) - \bar{f}(j)\|_\infty), \\
H(j) &\coloneqq d\left(j, \left\|\hat{f}(j)_{\overline{j}_n^{\mathrm{ad}},\infty} - \bar{f}(j)_{\overline{j}_n^{\mathrm{ad}},\infty}\right\|_\infty\right).
\end{aligned} \tag{5.2}$$

We will show that uniformly in j, F, G and H are close, and H is independent of $\hat{j}_n^{\mathrm{ex}}$, so we may bound $F(\hat{j}_n^{\mathrm{ex}})$ by Lemma 5.1.
By definition, $\hat{s}_n \ge s_{\min} > 0$, and $\hat{j}_n^{\mathrm{ex}} \ge \hat{j}_n^{cl}(\underline{\lambda}, 1) \ge \hat{j}_n^{cl}$, so on the events E_n, by Lemma 5.3,

$$\begin{aligned}
|F(\hat{j}_n^{\mathrm{ex}}) - G(\hat{j}_n^{\mathrm{ex}})| \le \frac{a(\hat{j}_n^{\mathrm{ex}})}{c(\hat{j}_n^{\mathrm{ex}})} R_3(\hat{j}_n^{\mathrm{ex}}) &\lesssim \sqrt{\frac{n\hat{j}_n^{\mathrm{ex}}}{2^{\hat{j}_n^{\mathrm{ex}}}}} \frac{\hat{M}_n 2^{-\hat{j}_n^{\mathrm{ex}}\hat{s}_n}}{2^{\hat{s}_n}-1} \\
&\lesssim \sqrt{\frac{\hat{j}_n^{\mathrm{ex}}}{\hat{j}_n^{cl}(\underline{\lambda},1)}} \left(\hat{j}_n^{cl}(\underline{\lambda},1)\log(n)\right)^{-s_{\min}} = o(1),
\end{aligned}$$

since $\hat{j}_n^{cl}(\underline{\lambda}, 1) \geq j_n^{\min}$, and

$$\frac{\hat{j}_n^{\text{ex}}}{\hat{j}_n^{\text{cl}}(\underline{\lambda}, 1)} - 1 = \frac{\log_2 \hat{j}_n^{\text{cl}}(\underline{\lambda}, 1) + u_n}{\hat{j}_n^{\text{cl}}(\underline{\lambda}, 1)} \leq \frac{\log_2 j_n^{\min} + u_n}{j_n^{\min}} \to 0.$$

Similarly, for $j_n \geq j_n^{\text{ex}}$, on E_n,

$$\begin{aligned}
&|G(j_n) - H(j_n)| \\
&\quad \lesssim \frac{a(j_n)}{c(j_n)} \left(\sup_{k=0}^{2^{j_0}-1} |\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}| + \sum_{j=j_0}^{\overline{j}_n^{\text{ad}}-1} \sup_{k=0}^{2^j-1} 2^{j/2} |\hat{\beta}_{j,k} - \beta_{j,k}| \right) \\
&\quad \lesssim (j_n^{\text{ex}} / j_n^{\text{cl}}(\underline{\lambda}, 1))^{1/2} 2^{-(j_n^{\text{cl}}(\underline{\lambda},1) - \overline{j}_n^{\text{ad}})/2} \\
&\quad \lesssim 2^{-(j_n^{\text{cl}}(\underline{\lambda},1) - j_n^{cl}(\underline{\lambda},\nu))/2} = o(1),
\end{aligned}$$

since

$$j_n^{\text{cl}}(\underline{\lambda}, 1) - j_n^{\text{cl}}(\underline{\lambda}, \nu) \geq \frac{\nu - 1}{2s_{\max} + 1} \log_2(\log(n)) \to \infty.$$

On E_n, $\hat{j}_n^{\text{ex}}$ depends only on $\hat{\beta}_{j,k}$ for $j \leq \hat{j}_n^{\text{ad}} < \overline{j}_n^{\text{ad}}$, and $H(j)$ depends only on $\hat{\beta}_{j,k}$ for $j \geq \overline{j}_n^{\text{ad}}$, so $H(j)$ is independent of $\hat{j}_n^{\text{ex}}$. Hence, given $x, \varepsilon > 0$, for n large, and any $j \geq j_n^{\text{ex}}$,

$$\begin{aligned}
\mathbb{P}(F(j) \leq x \mid E_n, \hat{j}_n^{\text{ex}} = j) &\geq \mathbb{P}(G(j) \leq x - \varepsilon \mid E_n, \hat{j}_n^{\text{ex}} = j) \\
&\geq \mathbb{P}(H(j) \leq x - 2\varepsilon \mid E_n, \hat{j}_n^{\text{ex}} = j) \\
&= \mathbb{P}(H(j) \leq x - 2\varepsilon \mid E_n) \\
&\geq \mathbb{P}(G(j) \leq x - 3\varepsilon \mid E_n) \\
&\geq \mathbb{P}(G(j) \leq x - 3\varepsilon) - \mathbb{P}(E_n^c) \\
&\geq \exp\left(-e^{-(x-3\varepsilon)}\right) - o(1).
\end{aligned}$$

Likewise,

$$\mathbb{P}(F(j) \geq x \mid E_n, \hat{j}_n^{\text{ex}} = j) \leq \exp\left(-e^{-(x+3\varepsilon)}\right) + o(1).$$

As these results are uniform in $j \geq j_n^{\min}$, and true for any $\varepsilon > 0$, we have

$$\sup_{j \geq j_n^{ex}} \left| \mathbb{P}\left(F(j) \geq x \mid E_n, \hat{j}_n^{\text{ex}} = j\right) - \exp\left(-e^{-x}\right) \right| \to 0.$$

On E_n, we have $\hat{j}_n^{\text{ex}} \geq j_n^{\text{ex}}$, so

$$\begin{aligned}
\mathbb{P}(F(\hat{j}_n^{\text{ex}}) \leq x \mid E_n) &= \sum_{j=j_n^{\text{ex}}}^{\infty} \mathbb{P}(F(j) \leq x \mid E_n, \hat{j}_n^{\text{ex}} = j) \mathbb{P}(\hat{j}_n^{\text{ex}} = j \mid E_n) \\
&= \left(\exp\left(-e^{-x}\right) + o(1)\right) \sum_{j=j_n^{\text{ex}}}^{\infty} \mathbb{P}(\hat{j}_n^{\text{ex}} = j \mid E_n) \\
&= \exp\left(-e^{-x}\right) + o(1).
\end{aligned}$$

Since $\mathbb{P}(E_n) \to 1$, we obtain $\mathbb{P}(F(\hat{j}_n^{\text{ex}}) \leq x) \to \exp\left(-e^{-x}\right)$, and rearranging,

$$\mathbb{P}(f \notin C_n^{\text{ex}}) \to \gamma.$$

As the limits are all uniform in f, the result follows.

(ii) Let $J_n^{\text{ex}} := J_n^{\text{ex}}(\underline{\lambda}, 1)$, so on E_n, $\hat{j}_n^{\text{ex}} \leq j_n^{\text{ex}} + J_n^{\text{ex}}$ by Lemma 5.2. For n large, $j_n^{\text{cl}} > j_n^{\min}$, so

$$2^{j_n^{\text{cl}}/2} \approx \left(\frac{M}{c_{n,1}}\right)^{1/(2s+1)}, \qquad 2^{j_n^{\text{ex}}/2} \approx \log(n) 2^{j_n^{\text{cl}}/2}, \tag{5.3}$$

and

$$R_n^{\text{ex}} \lesssim \sqrt{j_n^{\text{ex}} + J_n^{\text{ex}}}\, 2^{(j_n^{\text{ex}} + J_n^{\text{ex}})/2} n^{-1/2} \lesssim M^{1/(2s+1)} r_n(s).$$

As $\mathbb{P}(E_n) \to 1$ uniformly, and the limits are uniform over $f \in C_0^s(M)$, the result follows. □

We now move on to the adaptive band C_n^{ad}. As the variance term is no longer independent of $\hat{j}_n$, we must use a different method to establish the validity of our band. We will instead consider $j_n^{\max} - j_n^{\min} + 1$ confidence bands, one for each possible choice of $\hat{j}_n$, and show that the effect of this change is asymptotically negligible.

Proof of Theorem 3.3.

(i) Let $G(j)$ be given by (5.2). From Lemma 5.1, we have

$$\begin{aligned}
\mathbb{P}(G(\hat{j}_n^{\text{ad}}) > x(\gamma_n)) &\leq \mathbb{P}\left(\exists\, j \in [j_n^{\min}, j_n^{\max}] : G(j) > x(\gamma_n)\right) \\
&\leq \sum_{j=j_n^{\min}}^{j_n^{\max}} \mathbb{P}\left(G(j) > x(\gamma_n)\right) \\
&= (j_n^{\max} - j_n^{\min} + 1)(1 + o(1))\gamma_n \\
&= \gamma + o(1).
\end{aligned}$$

Rearranging, we get

$$\mathbb{P}\left(\|\hat{f}(\hat{j}_n^{\text{ad}}) - \bar{f}(\hat{j}_n^{\text{ad}})\|_\infty > R_1(\hat{j}_n^{\text{ad}}, \gamma_n)\right) \leq \gamma + o(1).$$

By Lemma 5.3, on the events E_n,

$$\|\bar{f}(\hat{j}_n^{\text{ad}}) - f\|_\infty \leq R_2(\hat{j}_n^{\text{ad}}) + R_3(\hat{j}_n^{\text{ad}})$$

and by Lemma 5.2, $\mathbb{P}(E_n) \to 1$. Since

$$\|f - \hat{f}(\hat{j}_n^{\text{ad}})\|_\infty \leq \|\hat{f}(\hat{j}_n^{\text{ad}}) - \bar{f}(\hat{j}_n^{\text{ad}})\|_\infty + \|\bar{f}(\hat{j}_n^{\text{ad}}) - f\|_\infty,$$

we obtain

$$\mathbb{P}(f \notin C_n^{\text{ad}}) \leq \gamma + o(1).$$

As the limits are uniform in f, the result follows.

(ii) Since $\hat{j}_n^{\mathrm{ad}} \geq j_n^{\min}$, and $x(\gamma_n) = O(\log\log n)$, we have that $R_1(\hat{j}_n^{\mathrm{ad}}, \gamma_n)$ is dominated by $b(\hat{j}_n^{\mathrm{ad}})c(\hat{j}_n^{\mathrm{ad}})$. Let $J_n^{\mathrm{cl}} := J_n^{\mathrm{cl}}(\overline{\lambda}, 1)$, so on E_n, $\hat{j}_n^{\mathrm{ad}} \leq \hat{j}_n^{\mathrm{cl}} \leq j_n^{\mathrm{cl}} + J_n^{\mathrm{cl}}$ by Lemma 5.2. For n large, $j_n^{\mathrm{cl}} > j_n^{\min}$, so by (5.3), we obtain

$$R_1(\hat{j}_n^{\mathrm{ad}}, \gamma_n) \lesssim \sqrt{j_n^{\mathrm{cl}} + J_n^{\mathrm{cl}}}\, 2^{(j_n^{\mathrm{cl}} + J_n^{\mathrm{cl}})/2} n^{-1/2} \lesssim M^{1/(2s+1)} r_n(s).$$

Likewise on E_n, for n large $j_n^{\mathrm{cl}} + J_n^{\mathrm{cl}} \leq j_n^{\max}$, so $l(\hat{j}_n^{\mathrm{ad}}) = \hat{j}_n^{\mathrm{cl}}$, and

$$R_2(\hat{j}_n^{\mathrm{ad}}) \lesssim 2^{(j_n^{\mathrm{cl}} + J_n^{\mathrm{cl}})/2} c_{n,1} \lesssim M^{1/(2s+1)} r_n(s).$$

Also for n large, $\hat{s}_n \geq s_n > 0$, so

$$R_3(\hat{j}_n^{\mathrm{ad}}) \lesssim \frac{M_n}{2^{s_n} - 1} 2^{-j_n^{\mathrm{cl}} s_n} \lesssim \frac{M^{1/(2s+1)}}{2^s - 1} r_n(s).$$

As $\mathbb{P}(E_n) \to 1$ uniformly, and the limits are uniform over $f \in C_0^s(M)$, the result follows. □

Finally, we prove our result on confidence bands in density estimation and regression.

Proof of Theorem 3.6. We can prove the result analogously to Theorem 3.3. To bound the bias term, we will sketch a version of Lemma 5.2 for the density estimation and regression models. It is possible to also adapt the variance bound in Lemma 5.1 [3]; however, we will provide a weaker bound, as a consequence of our lemma.

Consider the empirical wavelet coefficents

$$\hat{\alpha}_{j_0,k} := \frac{1}{n}\sum_{i=1}^{n} \varphi_{j_0,k}(X_i), \qquad \hat{\beta}_{j,k} := \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(X_i),$$

in density estimation, or

$$\hat{\alpha}_{j_0,k} := \frac{1}{n}\sum_{i=1}^{n} \varphi_{j_0,k}(x_i)Y_i, \qquad \hat{\beta}_{j,k} := \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(x_i)Y_i,$$

in regression. To prove the lemma, we must find an event E_n on which, with high probability, these estimates are close to the true wavelet coefficients $\alpha_{j_0,k}$, $\beta_{j,k}$.

In density estimation, we note that, for $j \geq j_0$, $k \in [N, 2^j - N)$, the empirical wavelet coefficients satisfy

$$\mathbb{E}[\hat{\beta}_{j,k}] = \beta_{j,k}, \qquad \mathbb{V}\mathrm{ar}[\hat{\beta}_{j,k}] \leq \frac{\|f\|_\infty}{n}, \qquad |\hat{\beta}_{j,k}| \leq 2^{j/2}\|\psi\|_\infty.$$

Using Bernstein's inequality, we then obtain that, for a constant $A = A(\|f\|_\infty)$, uniformly for n large,

$$\mathbb{P}(|\hat{\beta}_{j,k} - \beta_{j,k}| > Ac_{n,1}) \leq n^{-1},$$

with similar bounds for the other coefficients. Thus, on an event E_n, with probability tending uniformly to 1 as $n \to \infty$,

$$\max\left(\sup_k |\hat{\alpha}_{j_0,k} - \alpha_{j_0,k}|, \sup_{j=j_0}^{j_n^{\max}-1} \sup_k |\hat{\beta}_{j,k} - \beta_{j,k}|\right) \leq A c_{n,1}. \tag{5.4}$$

The regression model is often identified with the white noise model, for f in classes $C^s(M)$, $s \geq \frac{1}{2}$ [2]. In this case, however, we wish to consider functions with unbounded Hölder norm, so we must discuss regression explicitly. To control the empirical wavelet coefficients, we use a Gaussian tail bound, noting that for j, k as before,

$$\hat{\beta}_{j,k} \sim N\left(\frac{1}{n}\sum_{i=1}^n \psi_{j,k}(x_i) f(x_i), \frac{\sigma^2}{n^2}\sum_{i=1}^n \psi_{j,k}(x_i)^2\right).$$

For $j \leq j_n^{\max}$, as $n \to \infty$, the mean and variance are thus

$$\beta_{j,k} + O(n^{-1/2}\|f\|_{C^{1/2}}) \qquad \text{and} \qquad \sigma^2 n^{-1}(1 + o(1)),$$

uniformly. We obtain that, for a constant $A = A(\|f\|_{C^{1/2}})$, uniformly for n large,

$$\mathbb{P}(|\hat{\beta}_{j,k} - \beta_{j,k}| > A c_{n,1}) \leq n^{-1},$$

again with similar bounds for the other coefficients, leading to an event E_n as above.

In both cases, we therefore have events E_n comparable to those in Lemma 5.2, but with constant A now depending on $\|f\|_\infty$ or $\|f\|_{C^{1/2}}$. To proceed, we require an estimator $\hat{A}$ of A, which satisfies

$$\sup_{f \in \mathcal{F}} \mathbb{P}_f(\hat{A} < A) \to 0,$$

and for any $s \in [s_{\min}, s_{\max}]$, $M > 0$, and constants $B = B(s, M) > 0$,

$$\sup_{f \in C_0^s(M)} \mathbb{P}_f(\hat{A} > B) \to 0.$$

We will describe such an estimator $\hat{A}$, and plug it into our bounds (5.4). We may then obtain a bound on the bias term, as in Theorem 3.3. To bound the variance term, we note that on the event E_n,

$$\|\hat{f}(j_n) - \bar{f}(j_n)\|_\infty \lesssim A 2^{j_n/2} c_{n,1},$$

uniformly in all $j_n \leq j_n^{\max}$; as this bound is of the same order as the one arising from Lemma 5.1, we may then proceed as before.

It remains to construct the estimator $\hat{A}$. We note that, in density estimation and regression respectively, the quantities $\|f\|_\infty$ and $\|f\|_{C^{1/2}}$ are bounded by a constant times $T \coloneqq \|f\|_{C^{s_{\min}}}$. We may therefore consider A as a function of T; it can be checked that in each case $A \leq CT + D$, for constants $C, D > 0$. We deduce that the estimator $\hat{A} \coloneqq C'\|\hat{f}(j_1)\|_{C^{s_{\min}}} + D'$ satisfies our conditions, for large enough constants $C', D' > 0$. $\square$

5.3. Negative results

We now prove our negative results. First, we will need a testing inequality for normal means experiments, proved using standard arguments [17]. We will prove a modified result, which controls the performance of tests also under small perturbations of the means.

Lemma 5.4. *Suppose we have observations* $X_1, \dots, X_n, Y_1, Y_2, \dots,$ *and we wish to test the hypothesis*

$$H_0 : X_i, Y_i \overset{i.i.d.}{\sim} N(0, 1),$$

against alternatives

$$H_k(\nu) : X_i \sim N(\mu\delta_{ik}, 1),\ Y_i \sim N(\nu_i, 1),\ \textit{independently,}$$

for $k = 1, \dots, n$, *and* $\mu, \nu_i \in \mathbb{R}$, $\|\nu\|^2 \leq \xi^2$. *Let* $T = 0$ *if we accept* H_0, *or* $T = 1$ *if we reject. There is a choice of* k, *not depending on* ν, *for which the sum of the Type I and Type II errors satisfies*

$$\mathbb{P}_{H_0}(T = 1) + \inf_{\|\nu\|^2 \leq \xi^2} \mathbb{P}_{H_k(\nu)}(T = 0) \geq 1 - n^{-1/2}(e^{\mu^2} - 1)^{1/2} - (e^{\xi^2} - 1)^{1/2}.$$

Proof. Consider first the case $\nu = 0$. The density of $\mathbb{P}_{H_k(0)}$ w.r.t. $\mathbb{P}_{H_0}$ is

$$Z_k := e^{\mu X_k - \mu^2/2}.$$

Let $Z := n^{-1} \sum_{k=1}^n Z_k$. Then $\mathbb{E}_{H_0} Z = 1$, and $\mathbb{E}_{H_0} Z^2 = 1 + n^{-1}(e^{\mu^2} - 1)$, so

$$\mathbb{E}_{H_0}(Z - 1)^2 = \mathbb{V}\mathrm{ar}_{H_0} Z = n^{-1}(e^{\mu^2} - 1).$$

We thus have

$$\begin{aligned}
\mathbb{P}_{H_0}(T = 1) + \max_{k=1}^{n} \mathbb{P}_{H_k(0)}(T = 0) &\geq \mathbb{P}_{H_0}(T = 1) + n^{-1} \sum_{k=1}^{n} \mathbb{P}_{H_k(0)}(T = 0) \\
&= 1 + \mathbb{E}_{H_0}[(Z - 1)1(T = 0)] \\
&\geq 1 - \mathbb{V}\mathrm{ar}_{H_0}(Z)^{1/2} \\
&= 1 - n^{-1/2}(e^{\mu^2} - 1)^{1/2}.
\end{aligned}$$

Fix k maximizing the above expression, and consider a hypothesis $H_k(\nu)$ with $\|\nu\|^2 \leq \xi^2$. The density of $\mathbb{P}_{H_k(\nu)}$ w.r.t. $\mathbb{P}_{H_k(0)}$ is

$$Z' := e^{\sum_i \nu_i Y_i - \|\nu\|^2/2},$$

and similarly we have

$$\mathbb{E}_{H_k(0)}(Z' - 1)^2 = \mathbb{V}\mathrm{ar}_{H_k(0)} Z' = e^{\|\nu\|^2} - 1.$$

Thus

$$\begin{aligned}
&\mathbb{P}_{H_0}(T=1)+\mathbb{P}_{H_k(\nu)}(T=0)\\
&\qquad\qquad = \mathbb{P}_{H_0}(T=1)+\mathbb{P}_{H_k(0)}(T=0)+\mathbb{E}_{H_k(0)}[(Z'-1)1(T=0)]\\
&\qquad\qquad \geq \mathbb{P}_{H_0}(T=1)+\mathbb{P}_{H_k(0)}(T=0)-\mathbb{V}\mathrm{ar}_{H_k(0)}[Z']^{1/2}\\
&\qquad\qquad \geq 1-n^{-1/2}(e^{\mu^2}-1)^{1/2}-(e^{\xi^2}-1)^{1/2}.
\end{aligned}$$

As this is true for all $\|\nu\|^2 \leq \xi^2$, the result follows. □

We may now prove our result on minimax rates in $C_0^s(M)$. For $f \in C^s(M)$, the argument is standard [24], but we must check that we can construct suitable alternative hypotheses lying within the restricted class $C_0^s(M)$.

Proof of Theorem 3.4. Suppose such an estimator $\hat{f}_n$ exists. For $i \in \mathbb{N}$, set $j_i := \lceil \rho j_{i-1} \rceil$, and consider functions

$$f_0 \coloneqq \sum_{i=0}^{\infty} \beta_{j_i}\psi_{j_i,0}, \qquad f_k \coloneqq f_0 + \beta_j \psi_{j,k},$$

where $\beta_j \coloneqq M2^{-j(s+1/2)}$, $j \geq j_0$ is to be determined, and $k \in [N, 2^j - N)$. By definition, these functions are in $C_0^s(M)$. By standard arguments, $\hat{f}_n$ must be able to distinguish the hypothesis $H_0 : f = f_0$ from alternatives $H_k : f = f_k$, contradicting Lemma 5.4. □

Finally, we will show that the self-similarity condition (2.1) is as weak as possible.

Proof of Theorem 3.5. We argue in a similar fashion to Theorem 3.4, taking care to account for the dishonesty of S_n. Suppose such a band S_n exists. For $m = 1, 2, \dots, \infty$, we will construct functions f_m which serve as hypotheses for the function f. We will choose these functions so that $f_m \in C_1^{s_m}(M)$, for a sequence $s_m \in (s_{\min}, s_{\max})$ with limit $s_\infty \in (s_{\min}, s_{\max})$. We will then find a subsequence n_m such that, for $\delta \coloneqq \frac{1}{4}(1-2\gamma)$,

$$\inf_{m=2}^{\infty} \mathbb{P}_{f_\infty}(f_\infty \notin S_{n_m}) \geq \gamma + \delta,$$

contradicting our assumptions on S_n.

Taking infimums if necessary, we may assume ρ_j increasing; for $i \in \mathbb{N}$, set $j_i \coloneqq \lceil \rho_{j_{i-1}} j_{i-1} \rceil$. Then for $m = 1, 2, \dots, \infty$, set

$$f_m \coloneqq \sum_{i=0}^{\infty} b_{i,m}\psi_i + \sum_{l=1}^{m} b'_l \psi'_l,$$

where

$$\psi_i \coloneqq \psi_{j_i, 2^{j_i - 1}}, \qquad \psi'_l \coloneqq \psi_{j_{i_l}, k_l},$$

and $b_{i,m}, b'_l \in \mathbb{R}$, $i_l \in \mathbb{N}$, and $k_l \in [N, 2^{j_i} - N) \setminus \{2^{j_i - 1}\}$ are to be determined. We will set $-1 = i_0 < i_1 < \dots$,

$$b_{i,m} := \begin{cases} M2^{-j_i(s_l+1/2)}, & i_l < i \le i_{l+1} \text{ for some } l < m, \\ M2^{-j_i(s_m+1/2)}, & i > i_m, \end{cases}$$

and

$$b'_l := M2^{-j_{i_l}(s_l+1/2)}.$$

Set

$$\begin{aligned} s_0 &:= s_{\max}, & s_m &:= s_{m-1} - (j_{i_m}^{-1} - j_{i_m+1}^{-1}) \log_2(\varepsilon^{-1}), \quad m > 0, \\ t_0 &:= s_{\min}, & t_m &:= s_m - j_{i_m+1}^{-1} \log_2(\varepsilon^{-1}), \quad m > 0, \end{aligned}$$

and choose i_1 large enough that:

(i) $t_1 > t_0$;
(ii) for $i \ge i_1$, the ψ_i are interior wavelets, supported inside $(0, 1)$; and
(iii) the set of choices for k_1 is non-empty.

By definition, s_m is decreasing, t_m increasing, and $s_m - t_m \searrow 0$. For $m \ge 1$, both sequences thus lie in $(s_{\min}, s_{\max})$, and tend to a limit $s_\infty \in (s_{\min}, s_{\max})$. For all $m = 1, 2, \dots, \infty$, $l \in \mathbb{N}$, and $i_l \le i \le i_{l+1}$,

$$M2^{-j_i(s_l+1/2)} \ge \varepsilon M 2^{-j_i(t_{l+1}+1/2)} \ge \varepsilon M 2^{-j_i(s_m+1/2)},$$

so indeed $f_m \in C_1^{s_m}(M)$.

We have thus defined f_1, making an arbitrary choice of k_1; for convenience, set $n_1 = 1$. Inductively, suppose we have defined f_{m-1} and n_{m-1}, and set $r_n := r_n(s_{m-1})$. For $n_m > n_{m-1}$ and $D > 0$ both large, we have:

(i) $\mathbb{P}_{f_{m-1}}(f_{m-1} \notin S_{n_m}) \le \gamma + \delta$; and
(ii) $\mathbb{P}_{f_{m-1}}(|S_{n_m}| \ge D r_{n_m}) \le \delta$.

Setting $T_n = 1\left(\exists\, f \in S_n : \|f - f_{m-1}\|_\infty \ge 2 D r_n\right)$, we then have

$$\begin{aligned} \mathbb{P}_{f_{m-1}}(T_{n_m} = 1) &\le \mathbb{P}_{f_{m-1}}(f_{m-1} \notin S_{n_m}) + \mathbb{P}_{f_{m-1}}(|S_{n_m}| \ge D r_{n_m}) \\ &\le \gamma + 2\delta. \end{aligned} \tag{5.5}$$

We claim it is possible to choose f_m and n_m so that also, for any further choice of functions f_l,

$$\|f_\infty - f_{m-1}\|_\infty \ge 2 D r_{n_m}, \tag{5.6}$$

and

$$\mathbb{P}_{f_\infty}(T_{n_m} = 0) \ge 1 - \gamma - 3\delta = \gamma + \delta. \tag{5.7}$$

We may then conclude that

$$\mathbb{P}_{f_\infty}(f_\infty \notin S_{n_m}) \ge \mathbb{P}_{f_\infty}(T_{n_m} = 0) \ge \gamma + \delta,$$

as required.

It remains to verify the claim. Letting $i_m \to \infty$, choose n_m so that

$$r_{n_m} \sim D' 2^{-j_{i_m} s_m}, \tag{5.8}$$

for $D' > 0$ to be determined. Now,

$$\begin{aligned}
D''(i_m) &:= \sum_{l=m}^{\infty} \left(2^{-j_{i_{l+1}} s_{l+1}} + \sum_{i=i_l+1}^{i_{l+1}} 2^{-j_i s_l} \right) \\
&\le \sum_{l=m}^{\infty} \left(2^{-j_{i_{l+1}} s_{\min}} + \sum_{i=i_l+1}^{i_{l+1}} 2^{-j_i s_{\min}} \right) \\
&\le 2 \sum_{j=j_{i_m+1}}^{\infty} 2^{-j s_{\min}} \\
&= \frac{2^{1-j_{i_m+1} s_{\min}}}{1 - 2^{-s_{\min}}},
\end{aligned}$$

so, for i_m large,

$$\begin{aligned}
\|f_{m-1} - f_\infty\|_\infty &\ge \|b'_m \psi'_m\|_\infty - \left\| \sum_{l=m+1}^{\infty} b'_l \psi'_l + \sum_{i=i_m+1}^{\infty} (b_{i,\infty} - b_{i,m-1}) \psi_i \right\|_\infty \\
&\ge M \|\psi\|_\infty \left(2^{-j_{i_m} s_m} - D''(i_m) \right) \\
&\ge M \|\psi\|_\infty \left(2^{-j_{i_m} s_m} - \frac{2^{1-j_{i_m+1} s_{\min}}}{1 - 2^{-s_{\min}}} \right) \\
&\ge \tfrac{1}{2} M \|\psi\|_\infty 2^{-j_{i_m} s_m}.
\end{aligned}$$

We have thus satisfied (5.6), for a suitable choice of D'.

To satisfy (5.7), we will apply Lemma 5.4, testing $H_0 : f = f_{m-1}$ against $H_1 : f = f_\infty$. The observations X_i will correspond to $\int \psi'_m(t)\, dY_t$, for all possible choices of k_m, and the Y_i to the other empirical wavelet coefficients. From (5.8),

$$n_m = O\left(j_{i_m} 2^{j_{i_m} (2 + s_{m-1}^{-1}) s_m} \right),$$

so the quantity

$$\begin{aligned}
\mu^2 = n_m (b'_m)^2 &= n_m M^2 2^{-j_{i_m} (2 s_m + 1)} \\
&= O\left(j_{i_m} 2^{j_{i_m} (s_m / s_{m-1} - 1)} \right) \\
&= O\left(j_{i_m} \varepsilon^{(j_{i_m} / j_{i_m - 1} - 1)/s_{m-1}} \right) \\
&= o(j_{i_m}),
\end{aligned}$$

and likewise

$$\begin{aligned}
\xi^2 &= n_m \sup_{f_\infty} \left(\sum_{l=m}^{\infty} (b'_{l+1})^2 + \sum_{i=i_m+1}^{\infty} (b_{i,m-1} - b_{i,\infty})^2 \right) \\
&\le n_m M^2 \sum_{l=m}^{\infty} \left(2^{-j_{i_{l+1}}(2s_{l+1}+1)} + \sum_{i=i_l+1}^{i_{l+1}} 2^{-j_i(2s_l+1)} \right) \\
&= O\left(n_m 2^{-j_{i_m+1}(2s_m+1)} \right) \\
&= O\left(j_{i_m} 2^{j_{i_m} s_m/s_{m-1} - j_{i_m+1}} \right) \\
&= o(1).
\end{aligned}$$

Thus, for i_m large,

$$(2^{j_{i_m}} - (2N+1))^{-1/2}(e^{\mu^2} - 1)^{1/2} + (e^{\xi^2} - 1)^{1/2} \le \delta.$$

Hence by Lemma 5.4, if we take i_m large enough also that (5.5) holds, then (5.7) holds for a suitable choice of k_m, and our claim is proved. □

Acknowledgements

We would like to thank Richard Nickl, Adam McCloskey, David Ruppert and an anonymous referee for their valuable comments and suggestions.

References

[1] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301–413. MR1679028
[2] Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics* **24** 2384–2398. MR1425958
[3] Bull, A. D. (2011). A Smirnov-Bickel-Rosenblatt theorem for compactly-supported wavelets. arXiv:1110.4961.
[4] Bull, A. D. and Nickl, R. (2011). Adaptive confidence sets in L^2. *Probability Theory and Related Fields* to appear.
[5] Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics* **32** 1805–1840. MR2102494
[6] Chyzak, F., Paule, P., Scherzer, O., Schoisswohl, A. and Zimmermann, B. (2001). The construction of orthonormal wavelets using symbolic methods and a matrix analytical approach for wavelets on the interval. *Experimental Mathematics* **10** 67–86. MR1822852

[7] Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis* **1** 54–81. MR1256527

[8] Daubechies, I. (1992). *Ten lectures on wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR1162107

[9] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B. Methodological* **57** 301–369. MR1323344

[10] Frisch, U. and Parisi, G. (1980). On the singularity structure of fully developed turbluence. *Annals of the New York Academy of Sciences* **357** 84–88.

[11] Genovese, C. and Wasserman, L. (2008). Adaptive confidence bands. *The Annals of Statistics* **36** 875–905. MR2396818

[12] Giné, E., Güntürk, C. S. and Madych, W. R. (2011). On the periodized square of L^2 cardinal splines. *Experimental Mathematics* **20** 177–188. MR2821389

[13] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics* **38** 1122–1170. MR2604707

[14] Hall, P. (1992). Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. *The Annals of Statistics* **20** 675–694. MR1165587

[15] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications. Lecture Notes in Statistics* **129**. Springer-Verlag, New York. MR1618204

[16] Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *The Annals of Statistics* **39** 2383–2409. MR2906872

[17] Ingster, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the L^p metrics. *Theory of Probability and its Applications* **31** 333.

[18] Jaffard, S. (2000). On the Frisch-Parisi conjecture. *Journal de Mathématiques Pures et Appliquées. Neuvième Série* **79** 525–552. MR1770660

[19] Lepski, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teoriya Veroyatnostei i ee Primeneniya* **35** 459–470. MR1091202

[20] Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics* **17** 1001–1008. MR1015135

[21] Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics* **25** 2547–2554. MR1604412

[22] Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics* **28** 298–335. MR1762913

[23] Rioul, O. (1992). Simple regularity criteria for subdivision schemes. *SIAM Journal on Mathematical Analysis* **23** 1544–1576. MR1185641

[24] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. MR2724359