

## ANTI-CONCENTRATION AND HONEST, ADAPTIVE CONFIDENCE BANDS

BY VICTOR CHERNOZHUKOV<sup>1</sup>, DENIS CHETVERIKOV<sup>1</sup> AND KENGO KATO<sup>2</sup>

*Massachusetts Institute of Technology, University of California, Los Angeles  
and University of Tokyo*

Modern construction of uniform confidence bands for nonparametric densities (and other functions) often relies on the classical Smirnov–Bickel–Rosenblatt (SBR) condition; see, for example, Giné and Nickl [*Probab. Theory Related Fields* **143** (2009) 569–596]. This condition requires the existence of a limit distribution of an extreme value type for the supremum of a studentized empirical process (equivalently, for the supremum of a Gaussian process with the same covariance function as that of the studentized empirical process). The principal contribution of this paper is to remove the need for this classical condition. We show that a considerably weaker sufficient condition is derived from an anti-concentration property of the supremum of the approximating Gaussian process, and we derive an inequality leading to such a property for separable Gaussian processes. We refer to the new condition as a generalized SBR condition. Our new result shows that the supremum does not concentrate too fast around any value.

We then apply this result to derive a Gaussian multiplier bootstrap procedure for constructing honest confidence bands for nonparametric density estimators (this result can be applied in other nonparametric problems as well). An essential advantage of our approach is that it applies generically even in those cases where the limit distribution of the supremum of the studentized empirical process does not exist (or is unknown). This is of particular importance in problems where resolution levels or other tuning parameters have been chosen in a data-driven fashion, which is needed for adaptive constructions of the confidence bands. Finally, of independent interest is our introduction of a new, practical version of Lepski’s method, which computes the optimal, nonconservative resolution levels via a Gaussian multiplier bootstrap method.

**1. Introduction.** Let  $X_1, \dots, X_n$  be i.i.d. random vectors with common unknown density  $f$  on  $\mathbb{R}^d$ . We are interested in constructing confidence bands for  $f$  on a subset  $\mathcal{X} \subset \mathbb{R}^d$  that are *honest* to a given class  $\mathcal{F}$  of densities on  $\mathbb{R}^d$ . Typically,  $\mathcal{X}$  is a compact set on which  $f$  is bounded away from zero, and  $\mathcal{F}$  is a

---

Received December 2013; revised April 2014.

<sup>1</sup>Supported by a National Science Foundation grant.

<sup>2</sup>Supported by the Grant-in-Aid for Young Scientists (B) (25780152), the Japan Society for the Promotion of Science.

*MSC2010 subject classifications.* 62G07, 62G15.

*Key words and phrases.* Anti-concentration of separable Gaussian processes, honest confidence bands, Lepski’s method, multiplier method, non-Donsker empirical processes.

class of smooth densities such as a subset of a Hölder ball. A confidence band  $\mathcal{C}_n = \mathcal{C}_n(X_1, \dots, X_n)$  is a family of random intervals

$$\mathcal{C}_n := \{C_n(x) = [c_L(x), c_U(x)] : x \in \mathcal{X}\}$$

that contains the graph of  $f$  on  $\mathcal{X}$  with a guaranteed probability. Following [31], a band  $\mathcal{C}_n$  is said to be *asymptotically honest with level  $\alpha \in (0, 1)$  for the class  $\mathcal{F}$*  if

$$\liminf_{n \rightarrow \infty} \inf_{f \in \mathcal{F}} \mathbb{P}_f(f(x) \in C_n(x), \forall x \in \mathcal{X}) \geq 1 - \alpha.$$

Also, we say that a band  $\mathcal{C}_n$  is *asymptotically honest at a polynomial rate with level  $\alpha \in (0, 1)$  for the class  $\mathcal{F}$*  if

$$(1) \quad \inf_{f \in \mathcal{F}} \mathbb{P}_f(f(x) \in C_n(x), \forall x \in \mathcal{X}) \geq 1 - \alpha - Cn^{-c}$$

for some constants  $c, C > 0$ .

Let  $\hat{f}_n(\cdot, l)$  be a generic estimator of  $f$  with a smoothing parameter  $l$ , say bandwidth or resolution level, where  $l$  is chosen from a candidate set  $\mathcal{L}_n$ ; see [26, 42, 44] for a textbook level introduction to the theory of density estimation. Let  $\hat{l}_n = \hat{l}_n(X_1, \dots, X_n)$  be a possibly data-dependent choice of  $l$  in  $\mathcal{L}_n$ . Denote by  $\sigma_{n,f}(x, l)$  the standard deviation of  $\sqrt{n} \hat{f}_n(x, l)$ , that is,  $\sigma_{n,f}(x, l) := (n \text{Var}_f(\hat{f}_n(x, l)))^{1/2}$ . Then we consider a confidence band of the form

$$(2) \quad C_n(x) = \left[ \hat{f}_n(x, \hat{l}_n) - \frac{c(\alpha)\sigma_{n,f}(x, \hat{l}_n)}{\sqrt{n}}, \hat{f}_n(x, \hat{l}_n) + \frac{c(\alpha)\sigma_{n,f}(x, \hat{l}_n)}{\sqrt{n}} \right],$$

where  $c(\alpha)$  is a (possibly data-dependent) critical value determined to make the confidence band to have level  $\alpha$ . Generally,  $\sigma_{n,f}(x, l)$  is unknown and has to be replaced by an estimator.

A crucial point in construction of confidence bands is the computation of the critical value  $c(\alpha)$ . Assuming that  $\sigma_{n,f}(x, l)$  is positive on  $\mathcal{X} \times \mathcal{L}_n$ , define the stochastic process

$$(3) \quad Z_{n,f}(v) := Z_{n,f}(x, l) := \frac{\sqrt{n}(\hat{f}_n(x, l) - \mathbb{E}_f[\hat{f}_n(x, l)])}{\sigma_{n,f}(x, l)},$$

where  $v = (x, l) \in \mathcal{X} \times \mathcal{L}_n =: \mathcal{V}_n$ . We refer to  $Z_{n,f}$  as a “studentized process.” If, for the sake of simplicity, the bias  $|f(x) - \mathbb{E}_f[\hat{f}_n(x, l)]_{l=\hat{l}_n}|$  is sufficiently small compared to  $\sigma_{n,f}(x, \hat{l}_n)$ , then

$$\begin{aligned} \mathbb{P}_f(f(x) \in C_n(x), \forall x \in \mathcal{X}) &\approx \mathbb{P}_f\left(\sup_{x \in \mathcal{X}} |Z_{n,f}(x, \hat{l}_n)| \leq c(\alpha)\right) \\ &\geq \mathbb{P}_f\left(\sup_{v \in \mathcal{V}_n} |Z_{n,f}(v)| \leq c(\alpha)\right), \end{aligned}$$

so that band (2) will be of level  $\alpha \in (0, 1)$  by taking

$$(4) \quad c(\alpha) = (1 - \alpha)\text{-quantile of } \|Z_{n,f}\|_{\mathcal{V}_n} := \sup_{v \in \mathcal{V}_n} |Z_{n,f}(v)|.$$

The critical value  $c(\alpha)$ , however, is infeasible since the finite sample distribution of the process  $Z_{n,f}$  is unknown. Instead, we estimate the  $(1 - \alpha)$ -quantile of  $\|Z_{n,f}\|_{\mathcal{V}_n}$ .

Suppose that one can find an appropriate centered Gaussian process  $G_{n,f}$  indexed by  $\mathcal{V}_n$  with known or estimable covariance structure such that  $\|Z_{n,f}\|_{\mathcal{V}_n}$  is close to  $\|G_{n,f}\|_{\mathcal{V}_n}$ . Then we may approximate the  $(1 - \alpha)$ -quantile of  $\|Z_{n,f}\|_{\mathcal{V}_n}$  by

$$c_{n,f}(\alpha) := (1 - \alpha)\text{-quantile of } \|G_{n,f}\|_{\mathcal{V}_n}.$$

Typically, one computes or approximates  $c_{n,f}(\alpha)$  by one of the following two methods:

- (1) Analytical method: derive analytically an approximated value of  $c_{n,f}(\alpha)$ , by using an explicit limit distribution or large deviation inequalities.
- (2) Simulation method: simulate the Gaussian process  $G_{n,f}$  to compute  $c_{n,f}(\alpha)$  numerically, by using, for example, a multiplier method.

The main purpose of this paper is to introduce a general approach to establishing the validity of the so-constructed confidence band. Importantly, our analysis does not rely on the existence of an explicit (continuous) limit distribution of any kind, which is a major difference from the previous literature. For the density estimation problem, if  $\mathcal{L}_n$  is a singleton, that is, the smoothing parameter is chosen deterministically, the existence of such a continuous limit distribution, which is typically a Gumbel distribution, has been established for convolution kernel density estimators and *some* wavelet projection kernel density estimators; see [1, 4, 5, 17, 18, 20, 40]. We refer to the existence of the limit distribution as the Smirnov–Bickel–Rosenblatt (SBR) condition. However, the SBR condition has not been obtained for other density estimators such as nonwavelet projection kernel estimators based, for example, on Legendre polynomials or Fourier series. In addition, to guarantee the existence of a continuous limit distribution often requires more stringent regularity conditions than a Gaussian approximation itself. More importantly, if  $\mathcal{L}_n$  is not a singleton, which is typically the case when  $\hat{l}_n$  is data-dependent, and so the randomness of  $\hat{l}_n$  has to be taken into account, it is often hard to determine an exact limit behavior of  $\|G_{n,f}\|_{\mathcal{V}_n}$ .

We thus take a different route and significantly generalize the SBR condition. Our key ingredient is the *anti-concentration* property of suprema of Gaussian processes that shows that suprema of Gaussian processes do not concentrate too fast. To some extent, this is a reverse of numerous concentration inequalities for Gaussian processes. In studying the effect of approximation and estimation errors on the coverage probability, it is required to know how the random variable

$\|G_{n,f}\|_{\mathcal{V}_n} := \sup_{v \in \mathcal{V}_n} |G_{n,f}(v)|$  concentrates or “anti-concentrates” around, say, its  $(1 - \alpha)$ -quantile. It is not difficult to see that  $\|G_{n,f}\|_{\mathcal{V}_n}$  itself has a continuous distribution, so that *with keeping  $n$  fixed*, the probability that  $\|G_{n,f}\|_{\mathcal{V}_n}$  falls into the interval with center  $c_{n,f}(\alpha)$  and radius  $\varepsilon$  goes to 0 as  $\varepsilon \rightarrow 0$ . However, what we need to know is the behavior of those probabilities when  $\varepsilon$  depends on  $n$  and  $\varepsilon = \varepsilon_n \rightarrow 0$ . In other words, bounding explicitly “anti-concentration” probabilities for suprema of Gaussian processes is desirable. We will first establish bounds on the Lévy concentration function (see Definition 2.1) for suprema of Gaussian processes and then use these bounds to quantify the effect of approximation and estimation errors on the finite sample coverage probability. We say that a *generalized SBR condition* or simply an anti-concentration condition holds if  $\|G_{n,f}\|_{\mathcal{V}_n}$  concentrates sufficiently slowly, so that this effect is sufficiently small to yield asymptotically honest confidence bands.

As a substantive application of our results, we consider the problem of constructing honest *adaptive* confidence bands based on either convolution or wavelet projection kernel density estimators in Hölder classes  $\mathcal{F} \subset \bigcup_{t \in [\underline{t}, \bar{t}]} \Sigma(t, L)$  for some  $0 < \underline{t} < \bar{t} < \infty$  where  $\Sigma(t, L)$  is the Hölder ball of densities with radius  $L$  and smoothness level  $t$ . Following [6], we say the confidence band  $\mathcal{C}_n$  is *adaptive* if for every  $t, \varepsilon > 0$  there exists  $C > 0$  such that for all  $n \geq 1$ ,

$$\sup_{f \in \mathcal{F} \cap \Sigma(t, L)} \mathbb{P}_f \left( \sup_{x \in \mathcal{X}} \lambda(\mathcal{C}_n(x)) > Cr_n(t) \right) \leq \varepsilon,$$

where  $\lambda$  denotes the Lebesgue measure on  $\mathbb{R}$  and  $r_n(t) := (\log n/n)^{t/(2t+d)}$ , the minimax optimal rate of convergence for estimating a density  $f$  in the function class  $\Sigma(t, L)$  in the sup-metric  $d_\infty(\hat{f}, f) = \sup_{x \in \mathcal{X}} |\hat{f}(x) - f(x)|$ . We use Lepski’s method [2, 30] to find an adaptive value of the smoothing parameter. Here our contribution is to introduce a *Gaussian multiplier bootstrap implementation* of Lepski’s method. This is a practical proposal since previous implementations relied on conservative (one-sided) maximal inequalities and are not necessarily recommended for practice; see, for example, [19] for a discussion.

We should also emphasize that our techniques can also be used for constructing honest and/or adaptive confidence bands in many other nonparametric problems, but in this paper we focus on the density problem for the sake of clarity. Our techniques [anti-concentration of separable Gaussian processes (Theorem 2.1), and coupling inequalities (Theorems A.1 and A.2)] are of particular importance in non-Donsker settings since they allow us to prove validity of the Gaussian multiplier bootstrap for approximating distributions of suprema of sequences of empirical processes of VC type function classes where the metric entropy of the process may increase with  $n$ . Thus these techniques may be important in many nonparametric problems. For example, applications of our anti-concentration bounds can be found in [10] and [11], which consider the problems of nonparametric inference on a minimum of a function and nonparametric testing of qualitative hypotheses about functions, respectively.

1.1. *Related references.* Confidence bands in nonparametric estimation have been extensively studied in the literature. A classical approach, which goes back to [40] and [1], is to use explicit limit distributions of normalized suprema of studentized processes. A “Smirnov–Bickel–Rosenblatt type limit theorem” combines Gaussian approximation techniques and extreme value theory for Gaussian processes. It was argued that the convergence to normal extremes is rather slow despite that the Gaussian approximation is relatively fast [24]. To improve the finite sample coverage, bootstrap is often used in construction of confidence bands; see [3, 12]. However, to establish the validity of bootstrap confidence bands, researchers relied on the existence of continuous limit distributions of normalized suprema of original studentized processes. In the deconvolution density estimation problem, Lounici and Nickl [32] considered confidence bands without using Gaussian approximation. In the current density estimation problem, their idea reads as bounding the deviation probability of  $\|\hat{f}_n - E[\hat{f}_n(\cdot)]\|_\infty$  by using Talagrand’s [41] inequality and replacing the expected supremum by the Rademacher average. Such a construction is indeed general and applicable to many other problems, but is likely to be more conservative than our construction.

1.2. *Organization of the paper.* In the next section, we give a new anti-concentration inequality for suprema of Gaussian processes. Section 3 contains a theory of generic confidence band construction under high-level conditions. These conditions are easily satisfied both for convolution and projection kernel techniques under mild primitive assumptions, which are also presented in Section 3. Section 4 is devoted to constructing honest adaptive confidence bands in Hölder classes. Finally, most proofs are contained in the [Appendix](#), and some proofs and discussions are put into the supplemental material [9].

1.3. *Notation.* In what follows, constants  $c, C, c_1, C_1, c_2, C_2, \dots$  are understood to be positive and independent of  $n$ . The values of  $c$  and  $C$  may change at each appearance but constants  $c_1, C_1, c_2, C_2, \dots$  are fixed. Throughout the paper,  $\mathbb{E}_n[\cdot]$  denotes the average over index  $1 \leq i \leq n$ , that is, it simply abbreviates the notation  $n^{-1} \sum_{i=1}^n [\cdot]$ . For example,  $\mathbb{E}_n[g(X_i)] = n^{-1} \sum_{i=1}^n g(X_i)$ . For a set  $T$ , denote by  $\ell^\infty(T)$  the set of all bounded functions, that is, all functions  $z: T \rightarrow \mathbb{R}$  such that

$$\|z\|_T := \sup_{t \in T} |z(t)| < \infty.$$

Moreover, for a generic function  $g$ , we also use the notation  $\|g\|_\infty := \sup_x |g(x)|$  where the supremum is taken over the domain of  $g$ . For two random variables  $\xi$  and  $\eta$ , we write  $\xi \stackrel{d}{=} \eta$  if they share the same distribution. The standard Euclidean norm is denoted by  $|\cdot|$ .

**2. Anti-concentration of suprema of Gaussian processes.** The main purpose of this section is to derive an upper bound on the *Lévy concentration function* for suprema of separable Gaussian processes, where the terminology is adapted from [39]. Let  $(\Omega, \mathcal{A}, \mathbb{P})$  be the underlying (complete) probability space.

DEFINITION 2.1 (Lévy concentration function). Let  $Y = (Y_t)_{t \in T}$  be a separable stochastic process indexed by a semimetric space  $T$ . For all  $x \in \mathbb{R}$  and  $\varepsilon \geq 0$ , let

$$(5) \quad p_{x,\varepsilon}(Y) := \mathbb{P}\left(\left|\sup_{t \in T} Y_t - x\right| \leq \varepsilon\right).$$

Then the *Lévy concentration function* of  $\sup_{t \in T} Y_t$  is defined for all  $\varepsilon \geq 0$  as

$$(6) \quad p_\varepsilon(Y) := \sup_{x \in \mathbb{R}} p_{x,\varepsilon}(Y).$$

Likewise, define  $p_{x,\varepsilon}(|Y|)$  by (5) with  $\sup_{t \in T} Y_t$  replaced by  $\sup_{t \in T} |Y_t|$ , and define  $p_\varepsilon(|Y|)$  by (6) with  $p_{x,\varepsilon}(Y)$  replaced by  $p_{x,\varepsilon}(|Y|)$ .

Let  $X = (X_t)_{t \in T}$  be a separable Gaussian process indexed by a semimetric space  $T$  such that  $\mathbb{E}[X_t] = 0$  and  $\mathbb{E}[X_t^2] = 1$  for all  $t \in T$ . Assume that  $\sup_{t \in T} X_t < \infty$  a.s. Our aim here is to obtain a qualitative bound on the concentration function  $p_\varepsilon(X)$ . In a trivial example where  $T$  is a singleton, that is,  $X$  is a real standard normal random variable, it is immediate to see that  $p_\varepsilon(X) \asymp \varepsilon$  as  $\varepsilon \rightarrow 0$ . A nontrivial case is that when  $T$  is not a singleton, and both  $T$  and  $X$  are indexed by  $n = 1, 2, \dots$ , that is,  $T = T_n$  and  $X = X^n = (X_{n,t})_{t \in T_n}$ , and the complexity of the set  $\{X_{n,t} : t \in T_n\}$  [in  $L^2(\Omega, \mathcal{A}, \mathbb{P})$ ] is increasing in  $n$ . In such a case, it is typically not known whether  $\sup_{t \in T_n} X_{n,t}$  has a limiting distribution as  $n \rightarrow \infty$ , and therefore it is not trivial at all whether, for any sequence  $\varepsilon_n \rightarrow 0$ ,  $p_{\varepsilon_n}(X^n) \rightarrow 0$  as  $n \rightarrow \infty$ .

The following is the first main result of this paper.

THEOREM 2.1 (Anti-concentration for suprema of separable Gaussian processes). Let  $X = (X_t)_{t \in T}$  be a separable Gaussian process indexed by a semimetric space  $T$  such that  $\mathbb{E}[X_t] = 0$  and  $\mathbb{E}[X_t^2] = 1$  for all  $t \in T$ . Assume that  $\sup_{t \in T} X_t < \infty$  a.s. Then  $a(X) := \mathbb{E}[\sup_{t \in T} X_t] \in [0, \infty)$  and

$$(7) \quad p_\varepsilon(X) \leq 4\varepsilon(a(X) + 1),$$

for all  $\varepsilon \geq 0$ .

The similar conclusion holds for the concentration function of  $\sup_{t \in T} |X_t|$ .

COROLLARY 2.1. Let  $X = (X_t)_{t \in T}$  be a separable Gaussian process indexed by a semimetric space  $T$  such that  $\mathbb{E}[X_t] = 0$  and  $\mathbb{E}[X_t^2] = 1$  for all  $t \in T$ . Assume that  $\sup_{t \in T} X_t < \infty$  a.s. Then  $a(|X|) := \mathbb{E}[\sup_{t \in T} |X_t|] \in [\sqrt{2/\pi}, \infty)$  and

$$(8) \quad p_\varepsilon(|X|) \leq 4\varepsilon(a(|X|) + 1),$$

for all  $\varepsilon \geq 0$ .

We refer to (7) and (8) as anti-concentration inequalities because they show that suprema of separable Gaussian processes can not concentrate too fast. The proof of Theorem 2.1 and Corollary 2.1 follows by extending the results in [8] where we derived anti-concentration inequalities for maxima of Gaussian random vectors. See the [Appendix](#) for a detailed exposition.

**3. Generic construction of honest confidence bands.** We go back to the analysis of confidence bands. Recall that we consider the following setting. We observe i.i.d. random vectors  $X_1, \dots, X_n$  with common unknown density  $f \in \mathcal{F}$  on  $\mathbb{R}^d$ , where  $\mathcal{F}$  is a nonempty subset of densities on  $\mathbb{R}^d$ . We denote by  $P_f$  the probability distribution corresponding to the density  $f$ . We first state the result on the construction of honest confidence bands under certain high-level conditions and then show that these conditions hold for most commonly used kernel density estimators.

3.1. *Main result.* Let  $\mathcal{X} \subset \mathbb{R}^d$  be a set of interest. Let  $\hat{f}_n(\cdot, l)$  be a generic estimator of  $f$  with a smoothing parameter  $l \in \mathcal{L}_n$  where  $\mathcal{L}_n$  is the candidate set. Denote by  $\sigma_{n,f}(x, l)$  the standard deviation of  $\sqrt{n}\hat{f}_n(x, l)$ . We assume that  $\sigma_{n,f}(x, l)$  is positive on  $\mathcal{V}_n := \mathcal{X} \times \mathcal{L}_n$  for all  $f \in \mathcal{F}$ . Define the studentized process  $Z_{n,f} = \{Z_{n,f}(v) : v = (x, l) \in \mathcal{V}_n\}$  by (3). Let

$$W_{n,f} := \|Z_{n,f}\|_{\mathcal{V}_n}$$

denote the supremum of the studentized process. We assume that  $W_{n,f}$  is a well-defined random variable. Let  $c_1, C_1$  be some positive constants. We will assume the following high-level conditions.

CONDITION H1 (Gaussian approximation). For every  $f \in \mathcal{F}$ , there exists (on a possibly enriched probability space) a sequence of random variables  $W_{n,f}^0$  such that (i)  $W_{n,f}^0 \stackrel{d}{=} \|G_{n,f}\|_{\mathcal{V}_n}$  where  $G_{n,f} = \{G_{n,f}(v) : v \in \mathcal{V}_n\}$  is a tight Gaussian random element in  $\ell^\infty(\mathcal{V}_n)$  with  $E[G_{n,f}(v)] = 0, E[G_{n,f}(v)^2] = 1$  for all  $v \in \mathcal{V}_n$ , and  $E[\|G_{n,f}\|_{\mathcal{V}_n}] \leq C_1\sqrt{\log n}$ ; and moreover (ii)

$$(9) \quad \sup_{f \in \mathcal{F}} P_f(|W_{n,f} - W_{n,f}^0| > \varepsilon_{1n}) \leq \delta_{1n},$$

where  $\varepsilon_{1n}$  and  $\delta_{1n}$  are some sequences of positive numbers bounded from above by  $C_1n^{-c_1}$ .

Analysis of uniform confidence bands often relies on the classical Smirnov–Bickel–Rosenblatt (SBR) condition that states that for some sequences  $A_n$  and  $B_n$ ,

$$(10) \quad A_n(\|G_{n,f}\|_{\mathcal{V}_n} - B_n) \xrightarrow{d} Z, \quad \text{as } n \rightarrow \infty,$$

where  $Z$  is a Gumbel random variable; see, for example, [20]. Here both  $A_n$  and  $B_n$  are typically of order  $\sqrt{\log n}$ . However, this condition is often difficult to verify. Therefore, we propose to use a weaker condition (recall the definition of the Lévy concentration function given in Definition 2.1):

**CONDITION H2** (Anti-concentration or generalized SBR condition). For any sequence  $\varepsilon_n$  of positive numbers, we have

- (a)  $\sup_{f \in \mathcal{F}} p_{\varepsilon_n}(|G_{n,f}|) \rightarrow 0$  if  $\varepsilon_n \sqrt{\log n} \rightarrow 0$  or
- (b)  $\sup_{f \in \mathcal{F}} p_{\varepsilon_n}(|G_{n,f}|) \leq C_1 \varepsilon_n \sqrt{\log n}$ .

Note that Condition H2(a) follows trivially from Condition H2(b). In turn, under Condition H1, Condition H2(b) is a simple consequence of Corollary 2.1. Condition H2(a) (along with Conditions H1 and H3–H6 below) is sufficient to show that the confidence bands are asymptotically honest, but we will use Condition H2(b) to show that the confidence bands are asymptotically honest at a polynomial rate. We refer to Condition H2 as a generalized SBR condition because Condition H2(a) holds if (10) holds with  $A_n$  of order  $\sqrt{\log n}$ . An advantage of Condition H2 in comparison with the classical condition (10) is that Condition H2 follows easily from Corollary 2.1.

Let  $\alpha \in (0, 1)$  be a fixed constant (confidence level). Recall that  $c_{n,f}(\alpha)$  is the  $(1 - \alpha)$ -quantile of the random variable  $\|G_{n,f}\|_{\mathcal{V}_n}$ . If  $G_{n,f}$  is pivotal, that is, independent of  $f$ ,  $c_{n,f}(\alpha) = c_n(\alpha)$  can be directly computed, at least numerically. Otherwise, we have to approximate or estimate  $c_{n,f}(\alpha)$ . Let  $\hat{c}_n(\alpha)$  be an estimator or approximated value of  $c_{n,f}(\alpha)$ , where we assume that  $\hat{c}_n(\alpha)$  is nonnegative [which is reasonable since  $c_{n,f}(\alpha)$  is nonnegative]. The following is concerned with a generic regularity condition on the accuracy of the estimator  $\hat{c}_n(\alpha)$ .

**CONDITION H3** [Estimation error of  $\hat{c}_n(\alpha)$ ]. For some sequences  $\tau_n, \varepsilon_{2n}$ , and  $\delta_{2n}$  of positive numbers bounded from above by  $C_1 n^{-c_1}$ , we have

- (a)  $\sup_{f \in \mathcal{F}} P_f(\hat{c}_n(\alpha) < c_{n,f}(\alpha + \tau_n) - \varepsilon_{2n}) \leq \delta_{2n}$  and
- (b)  $\sup_{f \in \mathcal{F}} P_f(\hat{c}_n(\alpha) > c_{n,f}(\alpha - \tau_n) + \varepsilon_{2n}) \leq \delta_{2n}$ .

In the next subsection, we shall verify this condition for the estimator  $\hat{c}_n(\alpha)$  based upon the Gaussian multiplier bootstrap method. Importantly, in this condition, we introduce the sequence  $\tau_n$  and compare  $\hat{c}_n(\alpha)$  with  $c_{n,f}(\alpha + \tau_n)$  and  $c_{n,f}(\alpha - \tau_n)$  instead of directly comparing it with  $c_{n,f}(\alpha)$ , which considerably simplifies verification of this condition. With  $\tau_n = 0$  for all  $n$ , we would need to



have an upper bound on  $c_{n,f}(\alpha) - c_{n,f}(\alpha + \tau_n)$  and  $c_{n,f}(\alpha - \tau_n) - c_{n,f}(\alpha)$ , which might be difficult to obtain in general.

The discussion in the [Introduction](#) presumes that  $\sigma_{n,f}(x, l)$  were known, but of course it has to be replaced by a suitable estimator in practice. Let  $\hat{\sigma}_n(x, l)$  be a generic estimator of  $\sigma_{n,f}(x, l)$ . Without loss of generality, we may assume that  $\hat{\sigma}_n(x, l)$  is nonnegative. Condition [H4](#) below states a high-level assumption on the estimation error of  $\hat{\sigma}_n(x, l)$ . Verifying Condition [H4](#) is rather standard for specific examples.

CONDITION [H4](#) [Estimation error of  $\hat{\sigma}_n(\cdot)$ ]. For some sequences  $\varepsilon_{3n}$  and  $\delta_{3n}$  of positive numbers bounded from above by  $C_1 n^{-c_1}$ ,

$$\sup_{f \in \mathcal{F}} \mathbb{P}_f \left( \sup_{v \in \mathcal{V}_n} \left| \frac{\hat{\sigma}_n(v)}{\sigma_{n,f}(v)} - 1 \right| > \varepsilon_{3n} \right) \leq \delta_{3n}.$$

We now consider strategies to deal with the bias term. We consider two possibilities. The first possibility is to control the bias explicitly, so that the confidence band contains the bias controlling term. This construction is inspired by [\[4\]](#). The advantage of this construction is that it yields the confidence band the length of which shrinks at the minimax optimal rate with no additional inflating terms; see [Theorem 4.1](#) below. The disadvantage, however, is that this construction yields a conservative confidence band in terms of coverage probability. We consider this strategy in [Conditions H5](#) and [H6](#) and [Theorem 3.1](#). The other possibility is to undersmooth, so that the bias is asymptotically negligible, and hence the resulting confidence band contains no bias controlling terms. This is an often used strategy; see, for example, [\[20\]](#). The advantage of this construction is that it sometimes yields an exact (nonconservative) confidence band, so that the confidence band covers the true function with probability  $1 - \alpha$  asymptotically exactly; see [Corollary 3.1](#) below. The disadvantages, however, are that this method yields the confidence band that shrinks at the rate slightly slower than the minimax optimal rate, and that is centered around a nonoptimal estimator. We consider the possibility of undersmoothing in [Corollary 3.1](#) below. Note that [Conditions H5](#) and [H6](#) below are not assumed in [Corollary 3.1](#).

We now consider the first possibility, that is, we assume that the smoothing parameter  $\hat{l}_n := \hat{l}_n(X_1, \dots, X_n)$ , which is allowed to depend on the data, is chosen so that the bias can be controlled sufficiently well. Specifically, for all  $l \in \mathcal{L}_n$ , define

$$\Delta_{n,f}(l) := \sup_{x \in \mathcal{X}} \frac{\sqrt{n} |f(x) - \mathbb{E}_f[\hat{f}_n(x, l)]|}{\sigma_n(x, l)}.$$

We assume that there exists a sequence of random variables  $c'_n$ , which are known or can be calculated via simulations, that control  $\Delta_{n,f}(\hat{l}_n)$ . In particular, the theory in the next subsection assumes that  $c'_n$  is chosen as a multiple of the estimated high quantile of the supremum of certain Gaussian process.

CONDITION H5 [Bound on  $\Delta_{n,f}(\hat{l}_n)$ ]. For some sequence  $\delta_{4n}$  of positive numbers bounded from above by  $C_1 n^{-c_1}$ ,

$$\sup_{f \in \mathcal{F}} P_f(\Delta_{n,f}(\hat{l}_n) > c'_n) \leq \delta_{4n}.$$

In turn, we assume that  $c'_n$  can be controlled by  $u_n \sqrt{\log n}$  where  $u_n$  is a sequence of nonnegative positive numbers. Typically,  $u_n$  is either a bounded or slowly growing sequence; see, for example, our construction under primitive conditions in the next section.

CONDITION H6 (Bound on  $c'_n$ ). For some sequences  $\delta_{5n}$  and  $u_n$  of positive numbers where  $\delta_{5n}$  is bounded from above by  $C_1 n^{-c_1}$ ,

$$\sup_{f \in \mathcal{F}} P_f(c'_n > u_n \sqrt{\log n}) \leq \delta_{5n}.$$

When  $\mathcal{L}_n$  is a singleton, conditions like Conditions H5 and H6 have to be assumed. When  $\mathcal{L}_n$  contains more than one element, that is, we seek for an adaptive procedure, verification of Conditions H5 and H6 is nontrivial. In Section 4, we provide an example of such analysis.

We consider the confidence band  $\mathcal{C}_n = \{\mathcal{C}_n(x) : x \in \mathcal{X}\}$  defined by

$$(11) \quad \mathcal{C}_n(x) := [\hat{f}_n(x, \hat{l}_n) - s_n(x, \hat{l}_n), \hat{f}_n(x, \hat{l}_n) + s_n(x, \hat{l}_n)],$$

where

$$(12) \quad s_n(x, \hat{l}_n) := (\hat{c}_n(\alpha) + c'_n) \hat{\sigma}_n(x, \hat{l}_n) / \sqrt{n}.$$

Define

$$\begin{aligned} \bar{\varepsilon}_{n,f} &:= \varepsilon_{1n} + \varepsilon_{2n} + \varepsilon_{3n}(c_{n,f}(\alpha) + u_n \sqrt{\log n}), \\ \delta_n &:= \delta_{1n} + \delta_{2n} + \delta_{3n} + \delta_{4n} + \delta_{5n}. \end{aligned}$$

We are now in position to state the main result of this section. Recall the definition of Lévy concentration function (Definition 2.1).

THEOREM 3.1 (Honest generic confidence bands). *Suppose that Conditions H1 and H3–H6 are satisfied. Then*

$$(13) \quad \inf_{f \in \mathcal{F}} P_f(f \in \mathcal{C}_n) \geq (1 - \alpha) - \delta_n - \tau_n - p_{\bar{\varepsilon}_{n,f}}(|G_{n,f}|).$$

*If, in addition, Condition H2(a) is satisfied and  $\varepsilon_{3n} u_n \sqrt{\log n} \leq C_1 n^{-c_1}$ , then*

$$(14) \quad \liminf_{n \rightarrow \infty} \inf_{f \in \mathcal{F}} P_f(f \in \mathcal{C}_n) \geq 1 - \alpha,$$

*and if, in addition, Condition H2(b) is satisfied, then*

$$(15) \quad \inf_{f \in \mathcal{F}} P_f(f \in \mathcal{C}_n) \geq 1 - \alpha - C n^{-c},$$

*where  $c$  and  $C$  are constants depending only on  $\alpha$ ,  $c_1$  and  $C_1$ .*

COMMENT 3.1 (Honest confidence bands). Theorem 3.1 shows that the confidence band defined in (11) and (12) is asymptotically honest with level  $\alpha$  for the class  $\mathcal{F}$ . Moreover, under Condition H2(b), the coverage probability can be smaller than  $1 - \alpha$  only by a polynomially small term  $Cn^{-c}$  uniformly over the class  $\mathcal{F}$ . That is, in this case the confidence band is *asymptotically honest at a polynomial rate* as defined in (1).

COMMENT 3.2 (Advantages of Theorem 3.1). An advantage of Theorem 3.1 is that it does not require the classical SBR condition that is often difficult to obtain. Instead, it only requires a weaker generalized SBR Condition H2, which allows us to control the effect of estimation and approximation errors on the coverage probabilities. In the next subsection, we will show that as long as the bias  $\Delta_{n,f}(\hat{l}_n)$  can be controlled, our theorem applies when  $\hat{f}_n(\cdot)$  is defined using either convolution or projection kernels under mild conditions, and, as far as projection kernels are concerned, it covers estimators based on compactly supported wavelets, Battle–Lemarié wavelets of *any* order as well as other nonwavelet projection kernels such as those based on Legendre polynomials and Fourier series. When  $\mathcal{L}_n$  is a singleton, the SBR condition for compactly supported wavelets was obtained in [5] under certain assumptions that can be verified *numerically* for any given wavelet, for Battle–Lemarié wavelets of degree up-to 4 in [20], and for Battle–Lemarié wavelets of degree higher than 4 in [17]. To the best of our knowledge, the SBR condition for nonwavelet projection kernel functions (such as those based on Legendre polynomials and Fourier series) has not been obtained in the literature. In addition, and perhaps most importantly, there are no results in the literature on the SBR condition when  $\mathcal{L}_n$  is not a singleton. Finally, the SBR condition, being based on extreme value theory, yields only a logarithmic (in  $n$ ) rate of approximation of coverage probability; that is, this approach is *asymptotically honest at a logarithmic rate*. In contrast, our approach can lead to confidence bands that are asymptotically honest at a polynomial rate; see (15). Note also that one can obtain confidence bands that would be asymptotically honest at a polynomial rate with level  $\alpha' < \alpha$ , but such confidence bands would in general be wider than those provided by our approach.

COMMENT 3.3 [On dependence of constants  $c, C$  on  $\alpha$  in (15)]. We note that (15) is a *nonasymptotic* bound. In addition, it immediately follows from the proof of Theorem 3.1 that the constants  $c$  and  $C$  in (15) can be chosen to be independent of  $\alpha$  (thus, they depend only on  $c_1$  and  $C_1$ ) as long as

$$(16) \quad |\log \alpha| \leq C_1 \log n.$$

Therefore, (15) can be applied with  $\alpha = \alpha_n$  depending on  $n$  as long as (16) holds (and Condition H3 is satisfied for the given sequence  $\alpha = \alpha_n$ ).

COMMENT 3.4 (On the condition  $\varepsilon_{3n}u_n\sqrt{\log n} \leq C_1n^{-c_1}$ ). The second part of Theorem 3.1 requires the condition that  $\varepsilon_{3n}u_n\sqrt{\log n} \leq C_1n^{-c_1}$ . This is a very mild assumption. Indeed, under Condition H4,  $\varepsilon_{3n} \leq C_1n^{-c_1}$ , so that the assumption that  $\varepsilon_{3n}u_n\sqrt{\log n} \leq C_1n^{-c_1}$  is met (with possibly different constants  $c_1$  and  $C_1$ ) as long as  $u_n$  is bounded from above by a slowly growing sequence, for example,  $u_n \leq C_1 \log n$ , which is typically the case; see, for example, our construction in Section 4.

The confidence band defined in (11) and (12) is constructed so that the bias  $\Delta_{n,f}(\hat{l}_n)$  is controlled explicitly via the random variable  $c'_n$ . Alternatively, one can choose to undersmooth so that the bias is negligible asymptotically. To cover this possibility, we note that it follows from the proof of Theorem 3.1 that if  $u_n \log n \rightarrow 0$  or  $u_n \log n \leq C_1n^{-c_1}$ , then conclusions (14) or (15) of Theorem 3.1 continue to hold, respectively, with  $s_n(x, \hat{l}_n)$  in (12) replaced by  $\hat{c}_n(\alpha)\hat{\sigma}_n(x, \hat{l}_n)/\sqrt{n}$ . Thus, obtaining the asymptotically honest at a polynomial rate confidence band requires polynomial undersmoothing ( $u_n \log n \leq C_1n^{-c_1}$ ), but on the other hand, logarithmic undersmoothing ( $u_n \log n \rightarrow 0$ ) suffices if polynomial rate is not required. Moreover, if  $\mathcal{L}_n$  is a singleton, it is possible to show that the confidence band is asymptotically exact, with a polynomial convergence rate (21) under the condition  $u_n \log n \leq C_1n^{-c_1}$ . We collect these observations into the following corollary, the detailed proof of which can be found in the supplemental material [9].

COROLLARY 3.1 (Honest generic confidence bands with undersmoothing). Consider the confidence band  $\tilde{\mathcal{C}}_n = \{\tilde{\mathcal{C}}_n(x) : x \in \mathcal{X}\}$  defined by

$$\tilde{\mathcal{C}}_n(x) := [\hat{f}_n(x, \hat{l}_n) - \tilde{s}_n(x, \hat{l}_n), \hat{f}_n(x, \hat{l}_n) + \tilde{s}_n(x, \hat{l}_n)],$$

where

$$\tilde{s}_n(x, \hat{l}_n) := \hat{c}_n(\alpha)\hat{\sigma}_n(x, \hat{l}_n)/\sqrt{n}.$$

Suppose that Conditions H1, H3 and H4 are satisfied. In addition, assume that for some sequences  $\delta_{6n}$  and  $u_n$  of positive numbers,

$$(17) \quad \sup_{f \in \mathcal{F}} \mathbb{P}_f(\Delta_{n,f}(\hat{l}_n) > u_n\sqrt{\log n}) \leq \delta_{6n},$$

where  $\delta_{6n}$  is bounded from above by  $C_1n^{-c_1}$ . If Condition H2(a) holds and  $u_n \log n \rightarrow 0$ , then

$$(18) \quad \liminf_{n \rightarrow \infty} \inf_{f \in \mathcal{F}} \mathbb{P}_f(f \in \tilde{\mathcal{C}}_n) \geq 1 - \alpha.$$

If Condition H2(b) holds and  $u_n \log n \leq C_1n^{-c_1}$ , then

$$(19) \quad \inf_{f \in \mathcal{F}} \mathbb{P}_f(f \in \tilde{\mathcal{C}}_n) \geq 1 - \alpha - Cn^{-c}.$$

Moreover, assume in addition that  $\mathcal{L}_n$  is a singleton. If Condition H2(a) holds and  $u_n \log n \rightarrow 0$ , then

$$(20) \quad \lim_{n \rightarrow \infty} \sup_{f \in \mathcal{F}} |\mathbb{P}_f(f \in \tilde{\mathcal{C}}_n) - (1 - \alpha)| = 0.$$

If Condition H2(b) and  $u_n \log n \leq C_1 n^{-c_1}$ , then

$$(21) \quad \sup_{f \in \mathcal{F}} |\mathbb{P}_f(f \in \tilde{\mathcal{C}}_n) - (1 - \alpha)| \leq C n^{-c}.$$

Here  $c$  and  $C$  are constants depending only on  $\alpha$ ,  $c_1$  and  $C_1$ .

COMMENT 3.5 (Other methods for controlling bias term). In practice, there can be other methods for controlling the bias term. For example, an alternative approach is to estimate the bias function in a pointwise manner and construct bias corrected confidence bands; see, for example, [45] in the nonparametric regression case. A yet alternative approach to controlling the bias based upon bootstrap in construction of confidence bands is proposed and studied by the recent paper of [25].

COMMENT 3.6 [On dependence of constants  $c$ ,  $C$  on  $\alpha$  in (19) and (21)]. Similar to Comment 3.3, we note that (19) and (21) are *nonasymptotic* bounds, and it immediately follows from the proof of Corollary 3.1 that these bounds apply with  $\alpha = \alpha_n$  depending on  $n$  and constants  $c$  and  $C$  depending only on  $c_1$  and  $C_1$  as long as  $|\log \alpha| \leq C_1 \log n$  [in case of (19)] and  $|\log(\alpha - \tau_n)| \leq C_1 \log n$  [in case of (21)].

3.2. *Verifying Conditions H1–H4 for confidence bands constructed using common density estimators via Gaussian multiplier bootstrap.* We now argue that when  $\hat{c}_n(\alpha)$  is constructed via Gaussian multiplier bootstrap, Conditions H1–H4 hold for common density estimators—specifically, both for convolution and for projection kernel density estimators under mild assumptions on the kernel function.

Let  $\{K_l\}_{l \in \mathcal{L}_n}$  be a family of kernel functions where  $K_l : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  and  $l$  is a smoothing parameter. We consider kernel density estimators of the form

$$(22) \quad \hat{f}_n(x, l) := \mathbb{E}_n[K_l(X_i, x)] = \frac{1}{n} \sum_{i=1}^n K_l(X_i, x),$$

where  $x \in \mathcal{X}$  and  $l \in \mathcal{L}_n$ . The variance of  $\sqrt{n} \hat{f}_n(x, l)$  is given by

$$\sigma_{n,f}^2(x, l) := \mathbb{E}_f[K_l(X_1, x)^2] - (\mathbb{E}_f[K_l(X_1, x)])^2.$$

We estimate  $\sigma_{n,f}^2(x, l)$  by

$$(23) \quad \hat{\sigma}_n^2(x, l) := \frac{1}{n} \sum_{i=1}^n K_l(X_i, x)^2 - \hat{f}_n(x, l)^2.$$

This is a sample analogue estimator.

*Examples.* Our general theory covers a wide class of kernel functions, such as convolution, wavelet projection and nonwavelet projection kernels.

(i) *Convolution kernel.* Consider a function  $K : \mathbb{R} \rightarrow \mathbb{R}$ . Let  $\mathcal{L}_n \subset (0, \infty)$ . Then for  $x = (x_1, \dots, x_d)' \in \mathbb{R}^d$ ,  $y = (y_1, \dots, y_d)' \in \mathbb{R}^d$  and  $l \in \mathcal{L}_n$ , the convolution kernel function is defined by

$$(24) \quad K_l(y, x) := 2^{ld} \prod_{1 \leq m \leq d} K(2^l(y_m - x_m)).$$

Here  $2^{-l}$  is the bandwidth parameter.

(ii) *Wavelet projection kernel.* Consider a father wavelet  $\phi$ , that is, a function  $\phi$  such that: (a)  $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$  is an orthonormal system in  $L_2(\mathbb{R})$ , (b) the spaces  $V_j = \{\sum_k c_k \phi(2^j x - k) : \sum_k c_k^2 < \infty\}$ ,  $j = 0, 1, 2, \dots$ , are nested in the sense that  $V_j \subset V_{j'}$  whenever  $j \leq j'$  and (c)  $\bigcup_{j \geq 0} V_j$  is dense in  $L_2(\mathbb{R})$ . Let  $\mathcal{L}_n \subset \mathbb{N}$ . Then for  $x = (x_1, \dots, x_d)' \in \mathbb{R}^d$ ,  $y = (y_1, \dots, y_d)' \in \mathbb{R}^d$ , and  $l \in \mathcal{L}_n$ , the wavelet projection kernel function is defined by

$$(25) \quad K_l(y, x) := 2^{ld} \sum_{k_1, \dots, k_d \in \mathbb{Z}} \prod_{1 \leq m \leq d} \phi(2^l y_m - k_m) \prod_{1 \leq m \leq d} \phi(2^l x_m - k_m).$$

Here  $l$  is the resolution level. We refer to [13] and [26] as basic references on wavelet theory.

(iii) *Nonwavelet projection kernel.* Let  $\{\varphi_j : j = 1, \dots, \infty\}$  be an orthonormal basis of  $L_2(\mathcal{X})$ , the space of square integrable (with respect to Lebesgue measure) functions on  $\mathcal{X}$ . Let  $\mathcal{L}_n \subset (0, \infty)$ . Then for  $x = (x_1, \dots, x_d)' \in \mathbb{R}^d$ ,  $y = (y_1, \dots, y_d)' \in \mathbb{R}^d$  and  $l \in \mathcal{L}_n$ , the nonwavelet projection kernel function is defined by

$$(26) \quad K_l(y, x) := \sum_{j=1}^{\lfloor 2^{ld} \rfloor} \varphi_j(y) \varphi_j(x),$$

where  $\lfloor a \rfloor$  is the largest integer that is smaller than or equal to  $a$ . Here  $\lfloor 2^{ld} \rfloor$  is the number of series (basis) terms used in the estimation. When  $d = 1$  and  $\mathcal{X} = [-1, 1]$ , examples of orthonormal bases are Fourier basis

$$(27) \quad \{1, \cos(\pi x), \cos(2\pi x), \dots\}$$

and Legendre polynomial basis

$$(28) \quad \{1, (3/2)^{1/2}x, (5/8)^{1/2}(3x^2 - 1), \dots\}.$$

When  $d > 1$  and  $\mathcal{X} = [-1, 1]^d$ , one can take tensor products of bases for  $d = 1$ .

We assume that the critical value  $\hat{c}_n(\alpha)$  is obtained via the multiplier bootstrap method:

ALGORITHM 1 (Gaussian multiplier bootstrap). *Let  $\xi_1, \dots, \xi_n$  be independent  $N(0, 1)$  random variables that are independent of the data  $X_1^n := \{X_1, \dots, X_n\}$ . Let  $\xi_1^n := \{\xi_1, \dots, \xi_n\}$ . For all  $x \in \mathcal{X}$  and  $l \in \mathcal{L}_n$ , define a Gaussian multiplier process*

$$(29) \quad \hat{\mathbb{G}}_n(x, l) := \hat{\mathbb{G}}_n(X_1^n, \xi_1^n)(x, l) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \frac{K_l(X_i, x) - \hat{f}_n(x, l)}{\hat{\sigma}_n(x, l)}.$$

Then the estimated critical value  $\hat{c}_n(\alpha)$  is defined as

$$\hat{c}_n(\alpha) = \text{conditional } (1 - \alpha)\text{-quantile of } \|\hat{\mathbb{G}}_n\|_{\mathcal{V}_n} \text{ given } X_1^n.$$

Gaussian multiplier bootstrap is a special case of a more general exchangeable bootstrap; see, for example, [37]. We refer the reader to [22] for the first systematic use of the Gaussian multipliers and to [29] and [23] for conditional multiplier central limit theorems in the Donsker setting.

Let

$$\mathcal{K}_{n,f} := \left\{ \frac{K_l(\cdot, x)}{\sigma_{n,f}(x, l)} : (x, l) \in \mathcal{X} \times \mathcal{L}_n \right\}$$

denote the class of studentized kernel functions, and define

$$\sigma_n = \sup_{f \in \mathcal{F}} \sup_{g \in \mathcal{K}_{n,f}} (\mathbb{E}_f[g(X_1)^2])^{1/2}.$$

Note that  $\sigma_n \geq 1$ .

For a given class  $\mathcal{G}$  of measurable functions on a probability space  $(S, \mathcal{S}, Q)$  and  $\varepsilon > 0$ , the  $\varepsilon$ -covering number of  $\mathcal{G}$  with respect to the  $L_2(Q)$ -semimetric is denoted by  $N(\mathcal{G}, L_2(Q), \varepsilon)$ ; see Chapter 2 of [43] on details of covering numbers. We will use the following definition of VC type classes:

DEFINITION 3.1 (VC type class). Let  $\mathcal{G}$  be a class of measurable functions on a measurable space  $(S, \mathcal{S})$ , and let  $b > 0$ ,  $a \geq e$  and  $v \geq 1$  be some constants. Then the class  $\mathcal{G}$  is called VC( $b, a, v$ ) type class if it is uniformly bounded in absolute value by  $b$  (i.e.,  $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq b$ ), and the covering numbers of  $\mathcal{G}$  satisfy

$$\sup_Q N(\mathcal{G}, L_2(Q), b\tau) \leq (a/\tau)^v, \quad 0 < \tau < 1,$$

where the supremum is taken over all finitely discrete probability measures  $Q$  on  $(S, \mathcal{S})$ .

Then we will assume the following condition.

CONDITION VC. There exist sequences  $b_n > 0$ ,  $a_n \geq e$  and  $v_n \geq 1$  such that for every  $f \in \mathcal{F}$ , the class  $\mathcal{K}_{n,f}$  is VC( $b_n, a_n, v_n$ ) type and pointwise measurable.

We refer to Chapter 2.3 of [43] for the definition of pointwise measurable classes of functions. We note that Condition VC is a mild assumption, which we verify for common constructions in Appendix F (as a part of proving results for the next section; see Comment 3.5 below); see also Appendix I (supplemental material [9]).

For some sufficiently large absolute constant  $A$ , take

$$K_n := Av_n(\log n \vee \log(a_nb_n/\sigma_n)).$$

We will assume without loss of generality that  $K_n \geq 1$  for all  $n$ . The following theorem verifies Conditions H1–H4 with so defined  $\hat{\sigma}_n^2(x, l)$  and  $\hat{c}_n(\alpha)$  under Condition VC, using the critical values constructed via Algorithm 1.

**THEOREM 3.2** (Conditions H1–H4 hold for our construction). *Suppose that Condition VC is satisfied and there exist constants  $c_2, C_2 > 0$  such that  $b_n^2\sigma_n^4K_n^4/n \leq C_2n^{-c_2}$ . Then Conditions H1–H4, including both Conditions H2(a) and H2(b), hold with some constants  $c_1, C_1 > 0$  that depend only on  $c_2, C_2$ .*

**COMMENT 3.7** (Convolution and wavelet projection kernels). The assumption of Theorem 3.2 holds for convolution and wavelet projection kernels under mild conditions on the resolution level  $l$ . It follows from Lemma F.2 in Appendix F (supplemental material [9]) that, under mild regularity conditions, for convolution and wavelet projection kernel functions,  $\sigma_n \leq C$  and Condition VC holds with  $b_n \leq C2^{l_{\max,n}d/2}$ ,  $a_n \leq C$ , and  $v_n \leq C$  for some  $C > 0$  where  $l_{\max,n} = \sup\{\mathcal{L}_n\}$ . Hence, for these kernel functions, the assumption that  $b_n^2\sigma_n^4K_n^4/n \leq C_2n^{-c_2}$  reduces to

$$2^{l_{\max,n}d}(\log^4 n)/n \leq C_2n^{-c_2}$$

(with possibly different constants  $c_2, C_2$ ), which is a mild requirement on the bandwidth value or resolution level. This is a very mild assumption on the possible resolution levels. Similar comments apply to nonwavelet projection kernels with Fourier and Legendre polynomial bases. See Appendix I in the supplemental material [9].

**COMMENT 3.8** (On Condition H3). We note that under conditions of Theorem 3.2, Condition H3 remains true with the same constants  $c_1$  and  $C_1$  even if  $\alpha = \alpha_n$  depends on  $n$  [if we define  $c_{n,f}(\beta) = 0$  for  $\beta \geq 1$  and  $c_{n,f}(\beta) = \infty$  for  $\beta \leq 0$ ]. To see this, note that according to Theorem 3.2, constants  $c_1$  and  $C_1$  in Condition H3 depend only on constants  $c_2$  and  $C_2$ , and do not depend on  $\alpha$ .

**4. Honest and adaptive confidence bands in Hölder classes.** In this section, we study the problem of constructing honest adaptive confidence bands in



Hölder smoothness classes. Recall that for  $t, L > 0$ , the Hölder ball of densities with radius  $L$  and smoothness level  $t$  is defined by

$$\Sigma(t, L) := \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} : f \text{ is a } \lfloor t \rfloor\text{-times continuously differentiable density,} \right. \\ \left. \|D^\alpha f\|_\infty \leq L, \forall |\alpha| \leq \lfloor t \rfloor, \sup_{x \neq y} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x - y|^{t - \lfloor t \rfloor}} \leq L, \right. \\ \left. \forall |\alpha| = \lfloor t \rfloor \right\},$$

where  $\lfloor t \rfloor$  denotes the largest integer smaller than  $t$ , and for a multi-index  $\alpha = (\alpha_1, \dots, \alpha_d)$  with  $|\alpha| = \alpha_1 + \dots + \alpha_d$ ,  $D^\alpha f(x) := \partial^{|\alpha|} f(x) / \partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}$ ; see, for example, [42]. We assume that for some  $0 < \underline{t} \leq \bar{t} < \infty$  and  $L \geq 1$ ,

$$(30) \quad \mathcal{F} \subset \bigcup_{t \in [\underline{t}, \bar{t}]} \Sigma(t, L),$$

and consider the confidence band  $\mathcal{C}_n = \{\mathcal{C}_n(x) : x \in \mathcal{X}\}$  of the form (11) and (12), where  $\mathcal{X}$  is a (suitable) compact set in  $\mathbb{R}^d$ .

We begin by stating our assumptions. First, we restrict attention to kernel density estimators  $\hat{f}_n$  based on either convolution or wavelet projection kernel functions. Let  $r$  be an integer such that  $r \geq 2$  and  $r > \bar{t}$ .

**CONDITION L1 (Density estimator).** The density estimator  $\hat{f}_n$  is either a convolution or wavelet projection kernel density estimator defined in (22), (24) and (25). For convolution kernels, the function  $K : \mathbb{R} \rightarrow \mathbb{R}$  has compact support and is of bounded variation, and moreover is such that  $\int K(s) ds = 1$  and  $\int s^j K(s) dx = 0$  for  $j = 1, \dots, r - 1$ . For wavelet projection kernels, the function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  is either a compactly supported father wavelet of regularity  $r - 1$  [i.e.,  $\phi$  is  $(r - 1)$ -times continuously differentiable], or a Battle–Lemarié wavelet of regularity  $r - 1$ .

The assumptions stated in Condition L1 are commonly used in the literature. See [16] for a more general class of convolution kernel functions that would suffice for our results. Details on compactly supported and Battle–Lemarié wavelets can be found in Chapters 6 and 5.4 of [13], respectively.

It is known that if the function class  $\mathcal{F}$  is sufficiently large [e.g., if  $\mathcal{F} = \Sigma(t, L) \cup \Sigma(t', L)$  for  $t' > t$ ], the construction of honest adaptive confidence bands is not possible; see [33]. Therefore, following [20], we will restrict the function class  $\mathcal{F} \subset \bigcup_{t \in [\underline{t}, \bar{t}]} \Sigma(t, L)$  in a suitable way, as follows:

**CONDITION L2 (Bias bounds).** There exist constants  $l_0, c_3, C_3 > 0$  such that for every  $f \in \mathcal{F} \subset \bigcup_{t \in [\underline{t}, \bar{t}]} \Sigma(t, L)$ , there exists  $t \in [\underline{t}, \bar{t}]$  with

$$(31) \quad c_3 2^{-lt} \leq \sup_{x \in \mathcal{X}} |E_f[\hat{f}_n(x, l)] - f(x)| \leq C_3 2^{-lt},$$

for all  $l \geq l_0$ .

This condition is inspired by the path-breaking work of [20]; see also [36]. It can be interpreted as the requirement that the functions  $f$  in the class  $\mathcal{F}$  are “self-similar” in the sense that their regularity remains the same at large and small scales; see also [4]. To put it differently, “self-similarity” could be understood as the requirement that the bias of the kernel approximation to  $f$  with bandwidth  $2^{-l}$  remains approximately proportional to  $(2^{-l})^t$ —that is, not much smaller or not much bigger—for all small values of the bandwidth  $2^{-l}$ .

It is useful to note that the upper bound in (31) holds for all  $f \in \Sigma(t, L)$  (for sufficiently large  $C_3$ ) under Condition L1; see, for example, Theorem 9.3 in [26]. In addition, Giné and Nickl [20] showed that under Condition L1, the restriction due to the lower bound in (31) is weak in the sense that the set of elements of  $\Sigma(t, L)$  for which the lower bound in (31) does not hold is “topologically small.” Moreover, they showed that the minimax optimal rate of convergence in the sup-norm over  $\Sigma(t, L)$  coincides with that over the set of elements of  $\Sigma(t, L)$  for which Condition L2 holds. We refer to [20] for a detailed and deep discussion of these conditions and results.

We also note that, depending on the problem, construction of honest adaptive confidence bands is often possible under somewhat weaker conditions than that in Condition L2. For example, if we are interested in the function class  $\Sigma(t, L) \cup \Sigma(t', L)$  for some  $t' > t$ , Hoffman and Nickl [27] showed that it is necessary and sufficient to exclude functions  $\Sigma(t, L) \setminus \Sigma(t, L, \rho_n)$  where  $\Sigma(t, L, \rho_n) = \{f \in \Sigma(t, L) : \inf_{g \in \Sigma(t', L)} \|g - f\|_\infty \geq \rho_n\}$  and where  $\rho_n > 0$  is allowed to converge to zero as  $n$  increases but sufficiently slowly. If we are interested in the function class  $\bigcup_{t \in [\underline{t}, \bar{t}]} \Sigma(t, L)$ , Bull [4] showed that (essentially) necessary and sufficient condition can be written in the form of the bound from below on the rate with which wavelet coefficients of the density  $f$  are allowed to decrease. Here we prefer to work with Condition L2 directly because it is directly related to the properties of the estimator  $\hat{f}_n$  and does not require any further specifications of the function class  $\mathcal{F}$ .

In order to introduce the next condition, we need to observe that under Condition L2, for every  $f \in \mathcal{F}$ , there exists a *unique*  $t \in [\underline{t}, \bar{t}]$  satisfying (31); indeed, if  $t_1 < t_2$ , then for any  $c, C > 0$ , there exists  $\bar{l}$  such that  $C2^{-lt_2} < c2^{-lt_1}$  for all  $l \geq \bar{l}$ , so that for each  $f \in \mathcal{F}$  condition (31) can hold for all  $l \geq l_0$  for at most one value of  $t$ . This defines the map

$$(32) \quad t : \mathcal{F} \rightarrow [\underline{t}, \bar{t}], \quad f \mapsto t(f).$$

The next condition states our assumptions on the candidate set  $\mathcal{L}_n$  of the values of the smoothing parameter:

CONDITION L3 (Candidate set). There exist constants  $c_4, C_4 > 0$  such that for every  $f \in \mathcal{F}$ , there exists  $l \in \mathcal{L}_n$  with

$$(33) \quad \left(\frac{c_4 \log n}{n}\right)^{1/(2t(f)+d)} \leq 2^{-l} \leq \left(\frac{C_4 \log n}{n}\right)^{1/(2t(f)+d)},$$

for the map  $t : f \mapsto t(f)$  defined in (32). In addition, the candidate set is  $\mathcal{L}_n = [l_{\min,n}, l_{\max,n}] \cap \mathbb{N}$ .

This condition thus ensures via (33) that the candidate set  $\mathcal{L}_n$  contains an appropriate value of the smoothing parameter that leads to the optimal rate of convergence for every density  $f \in \mathcal{F}$ .

Finally, we will make the following mild condition:

CONDITION L4 (Density bounds). There exist constants  $\delta, \underline{f}, \bar{f} > 0$  such that for all  $f \in \mathcal{F}$ ,

$$(34) \quad f(x) \geq \underline{f} \quad \text{for all } x \in \mathcal{X}^\delta \quad \text{and} \quad f(x) \leq \bar{f} \quad \text{for all } x \in \mathbb{R}^d,$$

where  $\mathcal{X}^\delta$  is the  $\delta$ -enlargement of  $\mathcal{X}$ , that is,  $\mathcal{X}^\delta = \{x \in \mathbb{R}^d : \inf_{y \in \mathcal{X}} |x - y| \leq \delta\}$ .

We now discuss how we choose various parameters in the confidence band  $\mathcal{C}_n$ . In the previous section, we have shown how to obtain honest confidence bands as long as we can control the bias  $\Delta_{n,f}(\hat{l}_n)$  appropriately. So to construct honest adaptive confidence bands, we seek a method to choose the smoothing parameter  $\hat{l}_n \in \mathcal{L}_n$  so that the bias  $\Delta_{n,f}(\hat{l}_n)$  can be controlled, and at the same time, the confidence band  $\mathcal{C}_n$  is adaptive.

Let  $\bar{\mathcal{V}}_n := \{(x, l, l') : x \in \mathcal{X}, l, l' \in \mathcal{L}_n, l < l'\}$ , and for  $(x, l, l') \in \bar{\mathcal{V}}_n$ , denote

$$\tilde{\sigma}_n(x, l, l') := \left( \frac{1}{n} \sum_{i=1}^n (K_l(X_i, x) - K_{l'}(X_i, x))^2 - (\hat{f}_n(x, l) - \hat{f}_n(x, l'))^2 \right)^{1/2}.$$

Also, for some small  $c_\sigma > 0$ , let

$$\hat{\sigma}_n(x, l, l') := (c_\sigma \hat{\sigma}_n(x, l')) \vee \tilde{\sigma}_n(x, l, l')$$

denote the truncated version of  $\tilde{\sigma}_n(x, l, l')$ . In practice, we suggest setting  $c_\sigma = 0.5(1 - 2^{-d/2})$  (the constant  $c_\sigma$  is chosen so that with probability approaching one,  $\hat{\sigma}_n(x, l, l') = \tilde{\sigma}_n(x, l, l')$  for all  $(x, l, l') \in \bar{\mathcal{V}}_n$  for convolution kernel estimators, and for all  $(x, l, l') \in \bar{\mathcal{V}}_n$  with  $l \leq l' - s$  for wavelet projection kernel estimators where  $s$  is some constant; see Lemmas F.2 and F.4 in the supplemental material [9]).

There exist several techniques in the literature to construct  $\hat{l}_n$  so that  $\Delta_{n,f}(\hat{l}_n)$  can be controlled and the confidence band  $\mathcal{C}_n$  is adaptive; see, for example, [35] for a thorough introduction. One of the most important such techniques is the Lepski method; see [30] for a detailed explanation of the method. In this paper, we introduce a new implementation of the Lepski method, which we refer to as a multiplier bootstrap implementation of the Lepski method.

ALGORITHM 2 (Multiplier bootstrap implementation of the Lepski method). Let  $\gamma_n$  be a sequence of positive numbers converging to zero. Let  $\xi_1, \dots, \xi_n$  be independent  $N(0, 1)$  random variables that are independent of the data  $X_1^n :=$

$\{X_1, \dots, X_n\}$ . Let  $\xi_1^n := (\xi_1, \dots, \xi_n)$ . For all  $(x, l, l') \in \bar{\mathcal{V}}_n$ , define a Gaussian multiplier process

$$\begin{aligned} \tilde{\mathbb{G}}_n(x, l, l') &:= \tilde{\mathbb{G}}_n(X_1^n, \xi_1^n)(x, l, l') \\ &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \frac{(K_l(X_i, x) - K_{l'}(X_i, x)) - (\hat{f}_n(x, l) - \hat{f}_n(x, l'))}{\hat{\sigma}_n(x, l, l')}. \end{aligned}$$

Also, define

$$\tilde{c}_n(\gamma_n) = \text{conditional } (1 - \gamma_n)\text{-quantile of } \|\tilde{\mathbb{G}}_n\|_{\bar{\mathcal{V}}_n} \text{ given } X_1^n.$$

Moreover, for all  $l \in \mathcal{L}_n$ , let

$$\mathcal{L}_{n,l} := \{l' \in \mathcal{L}_n : l' > l\}.$$

Finally, for some constant  $q > 1$ , which is independent of  $n$ , define a Lepski-type estimator

$$(35) \quad \hat{l}_n := \inf \left\{ l \in \mathcal{L}_n : \sup_{l' \in \mathcal{L}_{n,l}} \sup_{x \in \mathcal{X}} \frac{\sqrt{n} |\hat{f}_n(x, l) - \hat{f}_n(x, l')|}{\hat{\sigma}_n(x, l, l')} \leq q \tilde{c}_n(\gamma_n) \right\}.$$

COMMENT 4.1 (On our implementation of Lepski’s method). We refer to (35) as a (Gaussian) multiplier bootstrap implementation of the Lepski method because  $\tilde{c}_n(\gamma_n)$  is obtained as the conditional  $(1 - \gamma_n)$ -quantile of  $\|\tilde{\mathbb{G}}\|_{\bar{\mathcal{V}}_n}$  given  $X_1^n$ . Previous literature on the Lepski method used Talagrand’s inequality combined with some bounds on expected suprema of certain empirical processes (obtained via symmetrization and entropy methods) to choose the threshold level for the estimator [the right-hand side of the inequality in (35)]; see [19] and [21]. Because of the one-sided nature of the aforementioned inequalities, however, it was argued that the resulting threshold turned out to be too high leading to limited applicability of the estimator in small and moderate samples. In contrast, an advantage of our construction is that we use  $q\tilde{c}_n(\gamma_n)$  as a threshold level, which is essentially the minimal possible value of the threshold that suffices for good properties of the estimator.

Once we have  $\hat{l}_n$ , to define the confidence band  $\mathcal{C}_n$ , we need to specify  $\hat{\sigma}_n(x, l)$ ,  $\hat{c}_n(\alpha)$  and  $c'_n$ . We assume that  $\hat{\sigma}_n(x, l)$  is obtained via (23) and  $\hat{c}_n(\alpha)$  via Algorithm 1. To specify  $c'_n$ , let  $u'_n$  be a sequence of positive numbers such that  $u'_n$  is sufficiently large for large  $n$ . Specifically, for large  $n$ ,  $u'_n$  is assumed to be larger than some constant  $C(\mathcal{F})$  depending only on the function class  $\mathcal{F}$ . Set

$$c'_n := u'_n \tilde{c}_n(\gamma_n).$$

COMMENT 4.2 (On the choice of  $\gamma_n$ ,  $q$  and  $u'_n$ ). As follows from Lemmas F.7 and F.8 (supplemental material [9]), the parameter  $\gamma_n$  appearing in (35)

determines the probability that the estimator  $\hat{l}_n$  fails to select an appropriate value of the smoothing parameter. Thus, in practice  $\gamma_n$  should be chosen small relative to the nominal coverage level  $\alpha$ . Also, for fixed  $n$  and  $\gamma_n$ , the choice of the parameters  $q$  and  $u'_n$  depends on the trade-off between the error in the coverage probability and length of the confidence bands: smaller values of  $q$  yield higher values of  $\hat{l}_n$  leading to undersmoothing and good control of the coverage probability; larger values of  $q$  yield lower values of  $\hat{l}_n$  leading to oversmoothing and narrow confidence bands; similarly, larger values of  $u'_n$  yield wider confidence bands but better control of the coverage probability. Finding the optimal value of  $q$  is a difficult theoretical problem and is beyond the scope of the paper. Also, in principle, it is possible to trace out the value  $C(\mathcal{F})$  from the proof of the theorem below and set  $u'_n = C(\mathcal{F})$ . However, since the function class  $\mathcal{F}$  is typically unknown in practice,  $u'_n$  can be set as a slowly growing sequence of positive numbers. In our small-scale simulation study presented in Section J of the supplemental material [9], we find that the values  $q = 1.1$  and  $u'_n = 0.5$  strike a good balance between coverage probability control and the length of the confidence bands in one-dimensional examples. We should note, however, that the empirical researchers should always test out different values of  $q$  and  $u'_n$  in Monte Carlo examples that mimic the data at hand.

The following theorem shows that the confidence band  $\mathcal{C}_n$  defined in this way is honest and adaptive for  $\mathcal{F}$ :

**THEOREM 4.1** (Honest and adaptive confidence bands via our method). *Suppose that Conditions L1–L4 are satisfied. In addition, suppose that there exist constants  $c_5, C_5 > 0$  such that: (i)  $2^{l_{\max, n}^d}(\log^4 n)/n \leq C_5 n^{-c_5}$ , (ii)  $l_{\min, n} \geq c_5 \log n$ , (iii)  $\gamma_n \leq C_5 n^{-c_5}$ , (iv)  $|\log \gamma_n| \leq C_5 \log n$ , (v)  $u'_n \geq C(\mathcal{F})$  and (vi)  $u'_n \leq C_5 \log n$ . Then Conditions H1–H6 in Section 3 and (15) in Theorem 3.1 hold and*

$$(36) \quad \sup_{f \in \mathcal{F}} \mathbb{P}_f \left( \sup_{x \in \mathcal{X}} \lambda(\mathcal{C}_n(x)) > C(1 + u'_n) r_n(t(f)) \right) \leq C n^{-c},$$

where  $\lambda(\cdot)$  denotes the Lebesgue measure on  $\mathbb{R}$  and  $r_n(t) := (\log n/n)^{t/(2t+d)}$ . Here the constants  $c, C > 0$  depend only on  $c_5, C_5$ , the constants that appear in Conditions L1–L4,  $c_\sigma, \alpha$  and the function  $K$  (when convolution kernels are used) or the father wavelet  $\phi$  (when wavelet projection kernels are used). Moreover,

$$(37) \quad \sup_{f \in \mathcal{F} \cap \Sigma(t, L)} \mathbb{P}_f \left( \sup_{x \in \mathcal{X}} \lambda(\mathcal{C}_n(x)) > C(1 + u'_n) r_n(t) \right) \leq C n^{-c},$$

with the same constants  $c, C$  as those in (36).

**COMMENT 4.3** (Honest and adaptive confidence bands). Equation (15) implies that the confidence band  $\mathcal{C}_n$  constructed above is asymptotically honest at a

polynomial rate for the class  $\mathcal{F}$ . In addition, recall that  $r_n(t)$  is the minimax optimal rate of convergence in the sup-metric for the class  $\mathcal{F} \cap \Sigma(t, L)$ ; see [20]. Therefore, (37) implies that the confidence band  $\mathcal{C}_n$  is adaptive whenever  $u'_n$  is bounded or almost adaptive if  $u'_n$  is slowly growing; see the discussion in front of Theorem 4.1 on selecting  $u'_n$ .

COMMENT 4.4 (On inflating terms). When  $u'_n$  is bounded, the rate of convergence of the length of the confidence band to zero  $(1 + u'_n)r_n(t)$  coincides with the minimax optimal rate of estimation of over  $\Sigma(t, L)$  with *no additional inflating terms*. This shows an advantage of the method of constructing confidence bands based on the explicit control of the bias term in comparison with the method based on undersmoothing where inflating terms seem to be necessary. This type of construction is inspired by the interesting ideas in [4].

COMMENT 4.5 (Extensions). Finally, we note that the proof of (15) and (36) in Theorem 4.1 did not use (30) directly. The proof only relies on Conditions L1–L4 whereas (30) served to motivate these conditions. Therefore, results (15) and (36) of Theorem 4.1 apply more generally as long as Conditions L1–L4 hold, not just for Hölder smoothness classes.

#### APPENDIX A: COUPLING INEQUALITIES FOR SUPREMA OF EMPIRICAL AND RELATED PROCESSES

The purpose of this section is to provide two coupling inequalities based on Slepian–Stein methods that are useful for the analysis of uniform confidence bands. The first inequality is concerned with suprema of empirical processes and is proven in Corollary 2.2 in [7]. The second inequality is new, is concerned with suprema of Gaussian multiplier processes, and will be obtained from a Gaussian comparison theorem derived in [8].

Let  $X_1, \dots, X_n$  be i.i.d. random variables taking values in a measurable space  $(S, \mathcal{S})$ . Let  $\mathcal{G}$  be a pointwise-measurable VC( $b, a, v$ ) type function class for some  $b > 0$ ,  $a \geq e$ , and  $v \geq 1$  (the definition of VC type classes is given in Section 3). Let  $\sigma^2 > 0$  be any constant such that  $\sup_{g \in \mathcal{G}} \mathbb{E}[g(X_1)^2] \leq \sigma^2 \leq b^2$ . Define the empirical process

$$\mathbb{G}_n(g) := \frac{1}{\sqrt{n}} \sum_{i=1}^n (g(X_i) - \mathbb{E}[g(X_1)]), \quad g \in \mathcal{G},$$

and let

$$W_n := \|\mathbb{G}_n\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} |\mathbb{G}_n(g)|$$

denote the supremum of the empirical process. Note that  $W_n$  is a well-defined random variable since  $\mathcal{G}$  is assumed to be pointwise-measurable. Let  $B = \{B(g) : g \in$

$\mathcal{G}$  be a tight Gaussian random element in  $\ell^\infty(\mathcal{F})$  with mean zero and covariance function

$$E[B(g_1)B(g_2)] = E[g_1(X_1)g_2(X_1)] - E[g_1(X_1)]E[g_2(X_1)],$$

for all  $g_1, g_2 \in \mathcal{G}$ . It is well known that such a process exists under the VC type assumption; see [43], pages 100–101. Finally, for some sufficiently large absolute constant  $A$ , let

$$K_n := Av(\log n \vee \log(ab/\sigma)).$$

In particular, we will assume that  $K_n \geq 1$ . The following theorem shows that  $W_n$  can be well approximated by the supremum of the corresponding Gaussian process  $B$  under mild conditions on  $b, \sigma$  and  $K_n$ . The proof of this theorem can be found in Corollary 2.2 in [7].

**THEOREM A.1** (Slepian–Stein type coupling for suprema of empirical processes). *Consider the setting specified above. Then for every  $\gamma \in (0, 1)$  one can construct on an enriched probability space a random variable  $W^0$  such that:*

- (i)  $W^0 \stackrel{d}{=} \|B\|_{\mathcal{G}}$  and (ii)

$$\begin{aligned} P\left(|W_n - W^0| > \frac{bK_n}{(\gamma n)^{1/2}} + \frac{(b\sigma)^{1/2}K_n^{3/4}}{\gamma^{1/2}n^{1/4}} + \frac{b^{1/3}\sigma^{2/3}K_n^{2/3}}{\gamma^{1/3}n^{1/6}}\right) \\ \leq A'\left(\gamma + \frac{\log n}{n}\right), \end{aligned}$$

where  $A'$  is an absolute constant.

**COMMENT A.1** (Comparison with the Hungarian couplings). The main advantage of the coupling provided in this theorem in comparison with, say, Hungarian coupling [28], which can be used to derive a similar result, is that our coupling does not depend on total variation norm of functions  $g \in \mathcal{G}$  leading to sharper inequalities than those obtained via Hungarian coupling when the function class  $\mathcal{G}$  consists, for example, of Fourier series or Legendre polynomials; see [7]. In addition, our coupling does not impose any side restrictions. In particular, it does not require bounded support of  $X$  and allows for point masses on the support. In addition, if the density of  $X$  exists, our coupling does not assume that this density is bounded away from zero on the support. See, for example, [38] for the construction of the Hungarian coupling and the use of aforementioned conditions.

Let  $\xi_1, \dots, \xi_n$  be independent  $N(0, 1)$  random variables independent of  $X_1^n := \{X_1, \dots, X_n\}$ , and let  $\xi_1^n := \{\xi_1, \dots, \xi_n\}$ . We assume that random variables  $X_1, \dots, X_n, \xi_1, \dots, \xi_n$  are defined as coordinate projections from the product probability space. Define the Gaussian multiplier process

$$\tilde{\mathbb{G}}_n(g) := \tilde{\mathbb{G}}_n(X_1^n, \xi_1^n)(g) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i(g(X_i) - \mathbb{E}_n[g(X_i)]), \quad g \in \mathcal{G},$$

and for  $x_1^n \in S^n$ , let  $\widetilde{W}_n(x_1^n) := \|\widetilde{\mathbb{G}}_n(x_1^n, \xi_1^n)\|_{\mathcal{G}}$  denote the supremum of this process calculated for fixed  $X_1^n = x_1^n$ . Note that  $\widetilde{W}_n(x_1^n)$  is a well-defined random variable. In addition, let

$$\psi_n := \sqrt{\frac{\sigma^2 K_n}{n}} + \left(\frac{b^2 \sigma^2 K_n^3}{n}\right)^{1/4} \quad \text{and} \quad \gamma_n(\delta) := \frac{1}{\delta} \left(\frac{b^2 \sigma^2 K_n^3}{n}\right)^{1/4} + \frac{1}{n}.$$

The following theorem shows that  $\widetilde{W}_n(X_1^n)$  can be well approximated with high probability by the supremum of the Gaussian process  $B$  under mild conditions on  $b, \sigma$  and  $K_n$ . The proof of this theorem can be found in the supplemental material [9].

**THEOREM A.2** (Slepian–Stein type coupling for suprema of conditional multiplier processes). *Consider the setting specified above. Suppose that  $b^2 K_n \leq n\sigma^2$ . Then for every  $\delta > 0$ , there exists a set  $S_{n,0} \in S^n$  such that  $\mathbb{P}(X_1^n \in S_{n,0}) \geq 1 - 3/n$  and for every  $x_1^n \in S_{n,0}$  one can construct on an enriched probability space a random variable  $W^0$  such that: (i)  $W^0 \stackrel{d}{=} \|B\|_{\mathcal{G}}$  and (ii)*

$$\mathbb{P}(|\widetilde{W}_n(x_1^n) - W^0| > (\psi_n + \delta)) \leq A'' \gamma_n(\delta),$$

where  $A''$  is an absolute constant.

**COMMENT A.2** (On the use of Slepian–Stein couplings). Theorems A.1 and A.2 combined with anti-concentration inequalities (Theorem 2.1 and Corollary 2.1) can be used to prove validity of Gaussian multiplier bootstrap for approximating distributions of suprema of empirical processes of VC type function classes *without weak convergence arguments*. This allows us to cover cases where complexity of the function class  $\mathcal{G}$  is increasing with  $n$ , which is typically the case in nonparametric problems in general and in confidence band construction in particular. Moreover, approximation error can be shown to be polynomially (in  $n$ ) small under mild conditions.

### APPENDIX B: SOME TECHNICAL TOOLS

**THEOREM B.1.** *Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables taking values in a measurable space  $(S, \mathcal{S})$ . Suppose that  $\mathcal{G}$  is a nonempty, pointwise measurable class of functions on  $S$  uniformly bounded by a constant  $b$  such that there exist constants  $a \geq e$  and  $v > 1$  with  $\sup_Q N(\mathcal{G}, L_2(Q), b\varepsilon) \leq (a/\varepsilon)^v$  for all  $0 < \varepsilon \leq 1$ . Let  $\sigma^2$  be a constant such that  $\sup_{g \in \mathcal{G}} \text{Var}(g) \leq \sigma^2 \leq b^2$ . If  $b^2 v \log(ab/\sigma) \leq n\sigma^2$ , then for all  $t \leq n\sigma^2/b^2$ ,*

$$\mathbb{P}\left[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \{g(\xi_i) - \mathbb{E}[g(\xi_1)]\} \right| > A \sqrt{n\sigma^2 \left\{ t \vee \left( v \log \frac{ab}{\sigma} \right) \right\}} \right] \leq e^{-t},$$

where  $A > 0$  is an absolute constant.



PROOF. This version of Talagrand’s inequality follows from Theorem 3 in [34] combined with a bound on expected values of suprema of empirical processes derived in [15]. See also [41] for the original version of Talagrand’s inequality.  $\square$

Proofs of the following two lemmas can be found in the supplemental material [9].

LEMMA B.1. *Let  $Y := \{Y(t) : t \in T\}$  be a separable, centered Gaussian process such that  $E[Y(t)^2] \leq 1$  for all  $t \in T$ . Let  $c(\alpha)$  denote the  $(1 - \alpha)$ -quantile of  $\|Y\|_T$ . Assume that  $E[\|Y\|_T] < \infty$ . Then  $c(\alpha) \leq E[\|Y\|_T] + \sqrt{2|\log \alpha|}$  and  $c(\alpha) \leq M(\|Y\|_T) + \sqrt{2|\log \alpha|}$  for all  $\alpha \in (0, 1)$  where  $M(\|Y\|_T)$  is the median of  $\|Y\|_T$ .*

LEMMA B.2. *Let  $\mathcal{G}_1$  and  $\mathcal{G}_2$  be  $VC(b_1, a_1, v_1)$  and  $VC(b_2, a_2, v_2)$  type classes, respectively, on a measurable space  $(S, \mathcal{S})$ . Let  $a = (a_1^{v_1} a_2^{v_2})^{1/(v_1+v_2)}$ . Then: (i)  $\mathcal{G}_1 \cdot \mathcal{G}_2 = \{g_1 \cdot g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$  is  $VC(b_1 b_2, 2a, v_1 + v_2)$  type class, (ii)  $\mathcal{G}_1 - \mathcal{G}_2 = \{g_1 - g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$  is  $VC(b_1 + b_2, a, v_1 + v_2)$  type class and (iii)  $\mathcal{G}_1^2 = \{g_1^2 : g_1 \in \mathcal{G}_1\}$  is  $VC(b_1^2, 2a_1, v_1)$  type class.*

APPENDIX C: PROOFS FOR SECTION 2

PROOF OF THEOREM 2.1. The fact that  $a(X) < \infty$  follows from Landau–Shepp–Fernique theorem; see, for example, Lemma 2.2.5 in [14]. In addition, since  $\sup_{t \in T} X_t \geq X_{t_0}$  for any fixed  $t_0 \in T$ ,  $a(X) \geq E[X_{t_0}] = 0$ . We now prove (7).

Since the Gaussian process  $X = (X_t)_{t \in T}$  is separable, there exists a sequence of finite subsets  $T_n \subset T$  such that  $Z_n := \max_{t \in T_n} X_t \rightarrow \sup_{t \in T} X_t =: Z$  a.s. as  $n \rightarrow \infty$ . Fix any  $x \in \mathbb{R}$ . Since  $|Z_n - x| \rightarrow |Z - x|$  a.s. and a.s. convergence implies weak convergence, there exists an at most countable subset  $\mathcal{N}_x$  of  $\mathbb{R}$  such that for all  $\varepsilon \in \mathbb{R} \setminus \mathcal{N}_x$ ,

$$\lim_{n \rightarrow \infty} P(|Z_n - x| \leq \varepsilon) = P(|Z - x| \leq \varepsilon).$$

But by Theorem 3 in [8],

$$P(|Z_n - x| \leq \varepsilon) \leq 4\varepsilon \left( E \left[ \max_{t \in T_n} X_t \right] + 1 \right) \leq 4\varepsilon (a(X) + 1),$$

for all  $\varepsilon \geq 0$ . Therefore,

$$(38) \quad P(|Z - x| \leq \varepsilon) \leq 4\varepsilon (a(X) + 1),$$

for all  $\varepsilon \in \mathbb{R} \setminus \mathcal{N}_x$ . By right continuity of  $P(|Z - x| \leq \cdot)$ , it follows that (38) holds for all  $\varepsilon \geq 0$ . Since  $x \in \mathbb{R}$  is arbitrary, we obtain (7).  $\square$

PROOF OF COROLLARY 2.1. In view of the proof of Theorem 2.1, it suffices to prove the corollary in the case where  $T$  is finite, but then the corollary follows from Comment 5 in [8].  $\square$

APPENDIX D: PROOFS FOR SECTION 3

PROOF OF THEOREM 3.1. Pick any  $f \in \mathcal{F}$ . By the triangle inequality, we have for any  $x \in \mathcal{X}$ ,

$$\frac{\sqrt{n}|\hat{f}_n(x, \hat{l}_n) - f(x)|}{\hat{\sigma}_n(x, \hat{l}_n)} \leq (|Z_{n,f}(x, \hat{l}_n)| + \Delta_{n,f}(\hat{l}_n)) \frac{\sigma_{n,f}(x, \hat{l}_n)}{\hat{\sigma}_n(x, \hat{l}_n)},$$

by which we have

$$\begin{aligned} & P_f(f(x) \in \mathcal{C}_n(x), \forall x \in \mathcal{X}) \\ & \geq P_f(|Z_{n,f}(x, \hat{l}_n)| + \Delta_{n,f}(\hat{l}_n) \leq (\hat{c}_n(\alpha) + c'_n)\hat{\sigma}_n(x, \hat{l}_n)/\sigma_{n,f}(x, \hat{l}_n), \forall x \in \mathcal{X}) \\ & \geq P_f\left(\sup_{x \in \mathcal{X}} |Z_{n,f}(x, \hat{l}_n)| + \Delta_{n,f}(\hat{l}_n) \leq (\hat{c}_n(\alpha) + c'_n)(1 - \varepsilon_{3n})\right) - \delta_{3n} \\ & \geq P_f\left(\sup_{x \in \mathcal{X}} |Z_{n,f}(x, \hat{l}_n)| \leq \hat{c}_n(\alpha)(1 - \varepsilon_{3n}) - c'_n\varepsilon_{3n}\right) - \delta_{3n} - \delta_{4n} \\ & \geq P_f(\|Z_{n,f}\|_{\mathcal{V}_n} \leq \hat{c}_n(\alpha)(1 - \varepsilon_{3n}) - c'_n\varepsilon_{3n}) - \delta_{3n} - \delta_{4n} \\ & \geq P_f(\|Z_{n,f}\|_{\mathcal{V}_n} \leq \hat{c}_n(\alpha)(1 - \varepsilon_{3n}) - u_n\varepsilon_{3n}\sqrt{\log n}) - \delta_{3n} - \delta_{4n} - \delta_{5n}, \end{aligned}$$

where the third line follows from Condition H4, the fourth line from Condition H5, the fifth line from the inequality  $\sup_{x \in \mathcal{X}} |Z_{n,f}(x, \hat{l}_n)| \leq \|Z_{n,f}\|_{\mathcal{V}_n}$  and the sixth line from Condition H6. Further, the probability in the last line above equals (recall that  $W_{n,f} = \|Z_{n,f}\|_{\mathcal{V}_n}$ )

$$\begin{aligned} & P_f(W_{n,f} \leq \hat{c}_n(\alpha)(1 - \varepsilon_{3n}) - u_n\varepsilon_{3n}\sqrt{\log n}) \\ (39) \quad & \geq P_f(W_{n,f} \leq c_{n,f}(\alpha + \tau_n)(1 - \varepsilon_{3n}) - \varepsilon_{2n} - u_n\varepsilon_{3n}\sqrt{\log n}) - \delta_{2n}, \end{aligned}$$

where (39) follows from Condition H3. Now, the probability in (39) is bounded from below by Condition H1 by

$$\begin{aligned} & P_f(W_{n,f}^0 \leq c_{n,f}(\alpha + \tau_n)(1 - \varepsilon_{3n}) - \varepsilon_{1n} - \varepsilon_{2n} - u_n\varepsilon_{3n}\sqrt{\log n}) - \delta_{1n} \\ (40) \quad & \geq P_f(W_{n,f}^0 \leq c_{n,f}(\alpha + \tau_n)) - p_{\bar{\varepsilon}_n}(|G_{n,f}|) - \delta_{1n} \\ (41) \quad & \geq 1 - \alpha - \tau_n - p_{\bar{\varepsilon}_n}(|G_{n,f}|) - \delta_{1n}, \end{aligned}$$

where (40) follows from the definition of the Lévy concentration function  $p_{\bar{\varepsilon}_n}(|G_{n,f}|)$  given that  $\bar{\varepsilon}_n = \varepsilon_{1n} + \varepsilon_{2n} + \varepsilon_{3n}(c_{n,f}(\alpha) + u_n\sqrt{\log n})$ , and (41) follows

since  $c_{n,f}(\cdot)$  is the quantile function of  $W_{n,f}^0$ . Combining these inequalities leads to (13).

To prove (14) and (15), note that  $\delta_n \leq Cn^{-c}$  and  $\tau_n \leq Cn^{-c}$  by Conditions H1 and H3–H6. Further, it follows from Lemma B.1 that  $c_{n,f}(\alpha) \leq E[\|G_{n,f}\|_{\mathcal{V}_n}] + (2|\log \alpha|)^{1/2} \leq C\sqrt{\log n}$ , and so  $\varepsilon_{3n}u_n\sqrt{\log n} \leq C_1n^{-c_1}$  implies that  $\bar{\varepsilon}_{n,f} \leq Cn^{-c}$ . Therefore, (14) and (15) follow from (13) and Condition H2.  $\square$

PROOF OF COROLLARY 3.1. The proof is similar to that of Theorem 3.1. The details are provided in the supplemental material [9].  $\square$

PROOF OF THEOREM 3.2. In this proof,  $c, C > 0$  are constants that depend only on  $c_2, C_2$ , but their values can change at each appearance.

Fix any  $f \in \mathcal{F}$ . Let  $G_{n,f} = \{G_{n,f}(v) : v \in \mathcal{V}_n\}$  be a tight Gaussian random element in  $\ell^\infty(\mathcal{V}_n)$  with mean zero and the same covariance function as that of  $Z_{n,f}$ . Since  $b_n^2\sigma_n^4K_n^4/n \leq C_2n^{-c_2}$ , it follows from Theorem A.1 that we can construct a random variable  $W_{n,f}^0$  such that  $W_{n,f}^0 \stackrel{d}{=} \|G_{n,f}\|_{\mathcal{V}_n}$ , and (9) holds with some  $\varepsilon_{1n}$  and  $\delta_{1n}$  bounded from above by  $Cn^{-c}$ . In addition, inequality  $E[\|G_{n,f}\|_{\mathcal{V}_n}] \leq C\sqrt{\log n}$  follows from Corollary 2.2.8 in [43]. Condition H1 follows. Given Condition H1, Condition H2(b) follows from Corollary 2.1, and Condition H2(a) follows from Condition H2(b).

Consider Condition H4. There exists  $n_0$  such that  $C_2n_0^{-c_2} \leq 1$ . It suffices to verify the condition only for  $n \geq n_0$ . Note that

$$(42) \quad \left| \frac{\hat{\sigma}_n(x, l)}{\sigma_{n,f}(x, l)} - 1 \right| \leq \left| \frac{\hat{\sigma}_n^2(x, l)}{\sigma_{n,f}^2(x, l)} - 1 \right|.$$

Define  $\mathcal{K}_{n,f}^2 := \{g^2 : g \in \mathcal{K}_{n,f}\}$ . Given the definition of  $\hat{\sigma}_n(x, l)$ , the right-hand side of (42) is bounded by

$$(43) \quad \sup_{g \in \mathcal{K}_{n,f}^2} |\mathbb{E}_n[g(X_i)] - E[g(X_1)]| + \sup_{g \in \mathcal{K}_{n,f}} |\mathbb{E}_n[g(X_i)]^2 - E[g(X_1)]^2|.$$

It follows from Lemma B.2 that  $\mathcal{K}_{n,f}^2$  is VC( $b_n^2, 2a_n, v_n$ ) type class. Moreover, for all  $g \in \mathcal{K}_{n,f}^2$ ,

$$E[g(X_i)^2] \leq b_n^2E[g(X_i)] \leq b_n^2\sigma_n^2.$$

Therefore, Talagrand’s inequality (Theorem B.1) with  $t = \log n$ , which can be applied because  $b_n^2K_n/(n\sigma_n^2) \leq b_n^2\sigma_n^4K_n^4/n \leq C_2n^{-c_2} \leq 1$  and  $b_n^2\log n/(n\sigma_n^2) \leq b_n^2K_n/(n\sigma_n^2) \leq 1$  (recall that  $\sigma_n \geq 1$  and  $K_n \geq 1$ ), gives

$$(44) \quad P\left(\sup_{g \in \mathcal{K}_{n,f}^2} |\mathbb{E}_n[g(X_i)] - E[g(X_1)]| > \frac{1}{2}\sqrt{\frac{b_n^2\sigma_n^2K_n}{n}}\right) \leq \frac{1}{n}.$$

In addition,

$$\sup_{g \in \mathcal{K}_{n,f}} |\mathbb{E}_n[g(X_i)]^2 - \mathbb{E}[g(X_1)]^2| \leq 2b_n \sup_{g \in \mathcal{K}_{n,f}} |\mathbb{E}_n[g(X_i)] - \mathbb{E}[g(X_1)]|,$$

so that another application of Talagrand’s inequality yields

$$(45) \quad \mathbb{P}\left(\sup_{g \in \mathcal{K}_{n,f}} |\mathbb{E}_n[g(X_i)]^2 - \mathbb{E}[g(X_1)]^2| > \frac{1}{2} \sqrt{\frac{b_n^2 \sigma_n^2 K_n}{n}}\right) \leq \frac{1}{n}.$$

Given that  $b_n^2 \sigma_n^2 K_n/n \leq b_n^2 \sigma_n^4 K_n^4/n \leq C_2 n^{-c_2}$ , combining (42)–(45) gives Condition H4 with  $\varepsilon_{3n} := (b_n^2 \sigma_n^2 K_n/n)^{1/2}$  and  $\delta_{3n} := 2/n$ .

Finally, we verify Condition H3. There exists  $n_1$  such that  $\varepsilon_{3n_1} \leq 1/2$ . It suffices to verify the condition only for  $n \geq n_1$ , so that  $\varepsilon_{3n} \leq 1/2$ . Define

$$\tilde{\mathbb{G}}_n(x, l) = \tilde{\mathbb{G}}_n(X_1^n, \xi_1^n)(x, l) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \frac{K_l(X_i, x) - \hat{f}_n(x, l)}{\sigma_n(x, l)}$$

and

$$\Delta \mathbb{G}_n(x, l) = \hat{\mathbb{G}}_n(x, l) - \tilde{\mathbb{G}}_n(x, l).$$

In addition, define

$$\widehat{W}_n(x_1^n) := \sup_{(x,l) \in \mathcal{X} \times \mathcal{L}_n} \hat{\mathbb{G}}_n(x_1^n, \xi_1^n)(x, l),$$

$$\widetilde{W}_n(x_1^n) := \sup_{(x,l) \in \mathcal{X} \times \mathcal{L}_n} \tilde{\mathbb{G}}_n(x_1^n, \xi_1^n)(x, l).$$

Consider the set  $S_{n,1}$  of values  $X_1^n$  such that  $|\hat{\sigma}_n(x, l)/\sigma_{n,f}(x, l) - 1| \leq \varepsilon_{3n}$  for all  $(x, l) \in \mathcal{X} \times \mathcal{L}_n$  whenever  $X_1^n \in S_{n,1}$ . The previous calculations show that  $\mathbb{P}_f(X_1^n \in S_{n,1}) \geq 1 - \delta_{3n} = 1 - 2/n$ . Pick and fix any  $x_1^n \in S_{n,1}$ . Then

$$\Delta \mathbb{G}_n(x_1^n, \xi_1^n)(x, l) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \frac{K_l(x_i, x) - \hat{f}_n(x, l)}{\sigma_n(x, l)} \left(\frac{\sigma_n(x, l)}{\hat{\sigma}_n(x, l)} - 1\right)$$

is a Gaussian process with mean zero and

$$\text{Var}(\Delta \mathbb{G}_n(x_1^n, \xi_1^n)(x, l)) = \frac{\hat{\sigma}_n^2(x, l)}{\sigma_n^2(x, l)} \left(\frac{\sigma_n(x, l)}{\hat{\sigma}_n(x, l)} - 1\right)^2 \leq \varepsilon_{3n}^2.$$

Further, the function class

$$\tilde{\mathcal{K}}_{n,f} := \left\{ \frac{K_l(\cdot, x)}{\sigma_n(x, l)} \left(\frac{\sigma_n(x, l)}{\hat{\sigma}_n(x, l)} - 1\right) : (x, l) \in \mathcal{X} \times \mathcal{L}_n \right\}$$

is contained in the function class

$$\left\{ \frac{a K_l(\cdot, x)}{\sigma_n(x, l)} : (x, l, a) \in \mathcal{X} \times \mathcal{L}_n \times [-1, 1] \right\},$$

and hence is  $\text{VC}(b_n, 4a_n, 1 + v_n)$  type class by Lemma B.2. In addition,

$$\begin{aligned} & \mathbb{E}[(\Delta \mathbb{G}_n(x_1^n, \xi_1^n)(x', l') - \Delta \mathbb{G}_n(x_1^n, \xi_1^n)(x'', l''))^2] \\ & \leq \mathbb{E}_n \left[ \left( \frac{K_l(x_i, x')}{\sigma_n(x', l')} \left( \frac{\sigma_n(x', l')}{\hat{\sigma}_n(x', l')} - 1 \right) - \frac{K_l(x_i, x'')}{\sigma_n(x'', l'')} \left( \frac{\sigma_n(x'', l'')}{\hat{\sigma}_n(x'', l'')} - 1 \right) \right)^2 \right], \end{aligned}$$

for all  $x', x'' \in \mathcal{X}$  and  $l', l'' \in \mathcal{L}_n$ , so that covering numbers for the index set  $\mathcal{X} \times \mathcal{L}_n$  with respect to the intrinsic (standard deviation) semimetric induced from the Gaussian process  $\Delta \mathbb{G}_n(x_1^n, \xi_1^n)$  are bounded by uniform covering numbers for the function class  $\tilde{\mathcal{K}}_{n,f}$ . Therefore, an application of Corollary 2.2.8 in [43] gives

$$\mathbb{E} \left[ \sup_{(x,l) \in \mathcal{X} \times \mathcal{L}_n} |\Delta \mathbb{G}_n(x_1^n, \xi_1^n)(x, l)| \right] \leq C \varepsilon_{3n} \sqrt{(1 + v_n) \log \left( \frac{4a_n b_n}{\varepsilon_{3n}} \right)} \leq C n^{-c}.$$

Here the second inequality follows from the definition of  $\varepsilon_{3n}$  above and the following inequalities:

$$\begin{aligned} \sqrt{(1 + v_n) \log \left( \frac{4a_n b_n}{\varepsilon_{3n}} \right)} & \leq \sqrt{(1 + v_n) \left( \log \left( \frac{4a_n b_n}{\sigma_n} \right) + \log \left( \frac{\sigma_n}{\varepsilon_{3n}} \right) \right)} \\ & \leq C \sqrt{K_n} \left( 1 + \sqrt{\log \left( \frac{\sigma_n}{\varepsilon_{3n}} \right)} \right) \\ & \leq C \sqrt{K_n} \left( 1 + \sqrt{\log \left( \frac{n}{b_n^2 K_n} \right)} \right) \\ & \leq C \sqrt{K_n} (1 + \sqrt{\log n}) \leq C K_n, \end{aligned}$$

where in the last line we used  $b_n \geq \sigma_n \geq 1$ , and  $K_n \geq v_n \log n \geq \log n$ . Combining this bound with the Borell–Sudakov–Tsirol’son inequality, and using the inequality

$$|\widehat{W}_n(x_1^n) - \widetilde{W}_n(x_1^n)| \leq \sup_{(x,l) \in \mathcal{X} \times \mathcal{L}_n} |\Delta \mathbb{G}_n(x_1^n, \xi_1^n)(x, l)|,$$

we see that there exists  $\lambda_{1n} \leq C n^{-c}$  such that

$$(46) \quad \mathbb{P}(|\widehat{W}_n(x_1^n) - \widetilde{W}_n(x_1^n)| \geq \lambda_{1n}) \leq C n^{-c},$$

whenever  $x_1^n \in S_{n,1}$ . Further, since  $b_n^2 \sigma_n^4 K_n^4 / n \leq C_2 n^{-c_2}$  and  $b_n \geq \sigma_n \geq 1$ , Theorem A.2 shows that there exist  $\lambda_{2n} \leq C n^{-c}$  and a measurable set  $S_{n,2}$  of values  $X_1^n$  such that  $\mathbb{P}_f(X_1^n \in S_{n,2}) \geq 1 - 3/n$ , and for every  $x_1^n \in S_{n,2}$  one can construct a random variable  $W^0$  such that  $W^0 \stackrel{d}{=} \|G_{n,f}\|_{\mathcal{V}_n}$  and

$$(47) \quad \mathbb{P}(|\widetilde{W}_n(x_1^n) - W^0| \geq \lambda_{2n}) \leq C n^{-c}.$$

Here  $W^0$  may depend on  $x_1^n$ , but  $c, C$  can be chosen in such a way that they depend only on  $c_2, C_2$  (as noted in the beginning).

Pick and fix any  $x_1^n \in S_{n,0} := S_{n,1} \cap S_{n,2}$ , and construct a suitable  $W^0 \stackrel{d}{=} \|G_{n,f}\|_{\mathcal{V}_n}$  for which (47) holds. Then by (46), we have

$$(48) \quad \mathbb{P}(|\widehat{W}_n(x_1^n) - W^0| \geq \lambda_n) \leq Cn^{-c},$$

where  $\lambda_n := \lambda_{1n} + \lambda_{2n}$ . Denote by  $\hat{c}_n(\alpha, x_1^n)$  the  $(1 - \alpha)$ -quantile of  $\widehat{W}_n(x_1^n)$ . Then we have

$$\begin{aligned} \mathbb{P}(\|G_{n,f}\|_{\mathcal{V}_n} \leq \hat{c}_n(\alpha, x_1^n) + \lambda_n) &= \mathbb{P}(W^0 \leq \hat{c}_n(\alpha, x_1^n) + \lambda_n) \\ &\geq \mathbb{P}(\widehat{W}_n(x_1^n) \leq \hat{c}_n(\alpha, x_1^n)) - Cn^{-c} \\ &\geq 1 - \alpha - Cn^{-c}, \end{aligned}$$

by which we have  $\hat{c}_n(\alpha, x_1^n) \geq c_{n,f}(\alpha + Cn^{-c}) - \lambda_n$ . Since  $x_1^n \in S_{n,0}$  is arbitrary and  $\hat{c}_n(\alpha) = \hat{c}_n(\alpha, X_1^n)$ , we see that whenever  $X_1^n \in S_{n,0}$ ,  $\hat{c}_n(\alpha) \geq c_{n,f}(\alpha + Cn^{-c}) - \lambda_n$ . Part (a) of Condition H3 follows from the fact that  $\mathbb{P}_f(X_1^n \in S_{n,0}) \geq 1 - 5/n$  and  $\lambda_n \leq Cn^{-c}$ . Part (b) follows similarly.  $\square$

**Acknowledgments.** The authors would like to thank Emre Barut, Enno Mammen and Richard Nickl for very helpful discussions. We also thank the editors and anonymous referees for their very helpful reviews that helped improve the paper.

## SUPPLEMENTARY MATERIAL

**Supplement to “Anti-concentration and honest, adaptive confidence bands”** (DOI: [10.1214/14-AOS1235SUPP](https://doi.org/10.1214/14-AOS1235SUPP); .pdf). This supplemental file contains additional proofs omitted in the main text, some results regarding nonwavelet projection kernel estimators, and a small-scale simulation study.

## REFERENCES

- [1] BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. [MR0348906](#)
- [2] BIRGÉ, L. (2001). An alternative point of view on Lepski’s method. In *State of the Art in Probability and Statistics (Leiden, 1999)*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **36** 113–133. IMS, Beachwood, OH. [MR1836557](#)
- [3] BISSANTZ, N., DÜMBGEN, L., HOLZMANN, H. and MUNK, A. (2007). Non-parametric confidence bands in deconvolution density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 483–506. [MR2323764](#)
- [4] BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. [MR2988456](#)
- [5] BULL, A. D. (2013). A Smirnov–Bickel–Rosenblatt theorem for compactly-supported wavelets. *Constr. Approx.* **37** 295–309. [MR3019781](#)
- [6] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. [MR2102494](#)
- [7] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2012). Gaussian approximation of suprema of empirical processes. Preprint. Available at [arXiv:1212.6885v2](https://arxiv.org/abs/1212.6885v2).

- [8] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields*. To appear. Available at [arXiv:1301.4807v3](https://arxiv.org/abs/1301.4807v3).
- [9] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Supplement to “Anti-concentration and honest, adaptive confidence bands.” DOI:[10.1214/14-AOS1235SUPP](https://doi.org/10.1214/14-AOS1235SUPP).
- [10] CHERNOZHUKOV, V., LEE, S. and ROSEN, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica* **81** 667–737. [MR3043345](https://arxiv.org/abs/1304.3345)
- [11] CHETVERIKOV, D. (2012). Testing regression monotonicity in econometric models. Available at [arXiv:1212.6757](https://arxiv.org/abs/1212.6757).
- [12] CLAESKENS, G. and VAN KEILEGOM, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31** 1852–1884. [MR2036392](https://arxiv.org/abs/2036392)
- [13] DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics **61**. SIAM, Philadelphia, PA. [MR1162107](https://arxiv.org/abs/1162107)
- [14] DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Univ. Press, Cambridge. [MR1720712](https://arxiv.org/abs/1720712)
- [15] GINÉ, E. and GUILLOU, A. (2001). A law of the iterated logarithm for kernel density estimators in the presence of censoring. *Ann. Inst. Henri Poincaré Probab. Stat.* **37** 503–522.
- [16] GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 907–921. [MR1955344](https://arxiv.org/abs/1955344)
- [17] GINÉ, E., GÜNTÜRK, C. S. and MADYCH, W. R. (2011). On the periodized square of  $L^2$  cardinal splines. *Exp. Math.* **20** 177–188. [MR2821389](https://arxiv.org/abs/2821389)
- [18] GINÉ, E., KOLTCHINSKII, V. and SAKHANENKO, L. (2004). Kernel density estimators: Convergence in distribution for weighted sup-norms. *Probab. Theory Related Fields* **130** 167–198. [MR2093761](https://arxiv.org/abs/2093761)
- [19] 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](https://arxiv.org/abs/2475673)
- [20] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. [MR2604707](https://arxiv.org/abs/2604707)
- [21] 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](https://arxiv.org/abs/2759172)
- [22] GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12** 929–998. [MR0757767](https://arxiv.org/abs/0757767)
- [23] GINÉ, E. and ZINN, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.* **18** 851–869. [MR1055437](https://arxiv.org/abs/1055437)
- [24] HALL, P. (1991). On convergence rates of suprema. *Probab. Theory Related Fields* **89** 447–455. [MR1118558](https://arxiv.org/abs/1118558)
- [25] HALL, P. and HOROWITZ, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *Ann. Statist.* **41** 1892–1921. [MR3127852](https://arxiv.org/abs/3127852)
- [26] HÄRDLE, W., KERKYCHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, Approximation, and Statistical Applications*. Springer, New York. [MR1618204](https://arxiv.org/abs/1618204)
- [27] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. [MR2906872](https://arxiv.org/abs/2906872)
- [28] KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV’s and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131. [MR0375412](https://arxiv.org/abs/0375412)
- [29] LEDOUX, M. and TALAGRAND, M. (1988). Un critère sur les petites boules dans le théorème limite central. *Probab. Theory Related Fields* **77** 29–47. [MR0921817](https://arxiv.org/abs/0921817)
- [30] LEPSKIÍ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.
- [31] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. [MR1015135](https://arxiv.org/abs/1015135)

- [32] LOUNICI, K. and NICKL, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Statist.* **39** 201–231. [MR2797844](#)
- [33] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. [MR1604412](#)
- [34] MASSART, P. (2000). About the constants in Talagrand’s concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–884. [MR1782276](#)
- [35] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. Springer, Berlin. [MR2319879](#)
- [36] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. [MR1762913](#)
- [37] PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. [MR1245301](#)
- [38] RIO, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21–45. [MR1254823](#)
- [39] RUDELSON, M. and VERSHYNIN, R. (2009). Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.* **62** 1707–1739. [MR2569075](#)
- [40] SMIRNOV, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables. *Doklady Akad. Nauk SSSR (N.S.)* **74** 189–191. [MR0037494](#)
- [41] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. [MR1419006](#)
- [42] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. [MR2724359](#)
- [43] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. [MR1385671](#)
- [44] WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York. [MR2172729](#)
- [45] XIA, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 797–811. [MR1649488](#)

V. CHERNOZHUKOV  
 DEPARTMENT OF ECONOMICS AND  
 OPERATIONS RESEARCH CENTER  
 MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
 50 MEMORIAL DRIVE  
 CAMBRIDGE, MASSACHUSETTS 02142  
 USA  
 E-MAIL: [vchern@mit.edu](mailto:vchern@mit.edu)

D. CHETVERIKOV  
 DEPARTMENT OF ECONOMICS  
 UNIVERSITY OF CALIFORNIA, LOS ANGELES  
 BUNCHE HALL, 8283  
 315 PORTOLA PLAZA  
 LOS ANGELES, CALIFORNIA 90095  
 USA  
 E-MAIL: [chetverikov@econ.ucla.edu](mailto:chetverikov@econ.ucla.edu)

K. KATO  
 GRADUATE SCHOOL OF ECONOMICS  
 UNIVERSITY OF TOKYO  
 7-3-1 HONGO, BUNKYO-KU  
 TOKYO 113-0033  
 JAPAN  
 E-MAIL: [kkato@e.u-tokyo.ac.jp](mailto:kkato@e.u-tokyo.ac.jp)