Adaptive Confidence Sets in L^2

Adam D. Bull Richard Nickl

University of Cambridge *

October 7, 2011

Abstract

The problem of constructing nonparametric confidence sets that are adaptive in L^2-loss over a continuous scale of Sobolev classes is considered. Adaptation holds, where possible, with respect to both the radius of the Sobolev ball and its smoothness degree, and over maximal parameter spaces for which adaptation is possible. Two key regimes of parameter constellations are identified: one where full adaptation is possible, and one where adaptation requires critical regions be removed. The phase transition between these regimes is analysed separately. Key ideas needed to derive these results include a general nonparametric minimax test for infinite-dimensional null- and alternative hypotheses, and new lower bound techniques for L^2-adaptive confidence sets.

1 Introduction

The paradigm of adaptive nonparametric inference has developed a fairly complete theory for estimation and testing – we mention the key references [24, 10, 9, 26, 2, 3, 31] – but the theory of adaptive confidence statements has not succeeded to the same extent, and consists in a significant part of negative results that are in a somewhat puzzling contrast to the fact that adaptive estimators exist. The topic of confidence sets is, however, of vital importance, since it addresses the question of whether the accuracy of adaptive estimation can itself be estimated, and to what extent the abundance of adaptive risk bounds and oracle inequalities in the literature are useful for statistical inference.

The parameter spaces for which adaptive confidence statements exist are sensitive to the geometry of the functional space in which the nonparametric problem is embedded: for instance, if the diameter of a confidence set is measured in the uniform norm then the admissible parameter spaces are severely more constrained than in the case where the diameter is measured in a L^2-type metric. Indeed in the L^2-theory – which has been created and significantly advanced in the papers Lepski [25], Hoffmann and Lepski [17],

*Department of Pure Mathematics and Mathematical Statistics, Statistical Laboratory, Wilberforce Road CB30WB Cambridge, UK. Email: a.bull@statslab.cam.ac.uk, r.nickl@statslab.cam.ac.uk.

Juditsky and Lambert-Lacroix [22], Baraud [1], Cai and Low [7] and Robins and van der Vaart [30] – partial positive results have been obtained, see also Theorem 3 below. In contrast in the context of the L^∞-theory Low's [28] by now almost classical negative result (see also [6]) has only recently been revisited: In the papers [14, 18, 23, 5], a 'separation approach' to the problem of adaptive confidence bands was introduced, which attempts to find 'maximal' subsets of the usual parameter spaces of adaptive estimation for which honest confidence statements can be constructed.

In this article we give a complete set of necessary and sufficient conditions for when confidence sets that are adaptive in L^2-diameter exist. We should note that, even though L^2-type confidence sets do not have as clear a geometric interpretation as confidence bands do, they involve the most commonly used loss function in adaptive estimation problems, and so deserve special attention in the theory of adaptive inference.

As a starting point let us illustrate the situation with a simple example of two fixed Sobolev-type classes. Let $X_1, \ldots, X_n$ be i.i.d. with common probability density f contained in the space L^2 of square-integrable functions on $[0, 1]$. Let $\Sigma(r) = \Sigma(r, B)$ be a Sobolev ball of probability densities on $[0, 1]$, of Sobolev-norm radius B – see Section 2.1 for precise definitions – and consider adaptation to the nested models $\Sigma(s) \subset \Sigma(r)$, $s > r$. An adaptive estimator $\hat{f}_n$ exists, achieving the optimal rate $n^{-s/(2s+1)}$ for $f \in \Sigma(s)$ and $n^{-r/(2r+1)}$ otherwise, in L^2-risk; see for instance Theorem 2 below.

A confidence set is a random subset $C_n = C(X_1, \ldots, X_n)$ of L^2, perhaps depending on r, s and for the moment also on B. Define the L^2-diameter of a norm-bounded subset C of L^2 as

$$|C| = \inf \left\{\rho : C \subset \{h : \|h - g\|_2 \leq \rho\} \text{ for some } g \in L^2\right\}, \tag{1}$$

equal to the radius of the smallest L^2-ball containing C. For a metric space (M, d), $f \in M$, $G \subset M$, set, as usual, $d(f, G) = \inf_{g \in G} d(f, g)$, and define, for ρ_n a sequence of nonnegative real numbers, the separated sets

$$\tilde{\Sigma}(r, \rho_n) \equiv \tilde{\Sigma}(r, s, B, \rho_n) = \{f \in \Sigma(r) : \|f - \Sigma(s)\|_2 \geq \rho_n\}.$$

Obviously $\tilde{\Sigma}(r, 0) = \Sigma(r)$, but for $\rho_n > 0$ these sets are proper subsets of $\Sigma(r) \setminus \Sigma(s)$. We are interested in adaptive inference for the model

$$\mathcal{P}_n \equiv \Sigma(s) \cup \tilde{\Sigma}(r, \rho_n).$$

Following [7], we shall say that the confidence set C_n is L^2-adaptive and honest for $\mathcal{P}_n$ if there exists a constant M such that for every $n \in \mathbb{N}$,

$$\sup_{f \in \Sigma(s)} \Pr_f \left\{|C_n| > M n^{-s/(2s+1)}\right\} \leq \alpha', \tag{2}$$

$$\sup_{f \in \tilde{\Sigma}(r, \rho_n)} \Pr_f \left\{|C_n| > M n^{-r/(2r+1)}\right\} \leq \alpha' \tag{3}$$

and if

$$\inf_{f \in \mathcal{P}_n} \Pr_f \{f \in C_n\} \geq 1 - \alpha - r_n \tag{4}$$

where $r_n \to 0$ as $n \to \infty$. We regard the constants α, α' as given 'coverage levels'.

Part Ai) of the following theorem is a main finding of the L^2-theory developed by the authors mentioned above, and follows, for instance, from the results in Robins and van der Vaart [30]. Part B) is a refinement of a negative result in Juditsky and Lambert-Lacroix [22]. Part Aii) is novel and provides a certain converse to part B.

Theorem 1. *Let $0 < \alpha, \alpha' < 1, s > r > 1/2$ and $B > 1$ be given.*
A) An adaptive and honest confidence set for $\tilde{\Sigma}(r, \rho_n) \cup \Sigma(s)$ exists if one of the following conditions is satisfied:
i) $s \leq 2r$ and $\rho_n \geq 0$
ii) $s > 2r$ and

$$\rho_n \geq M n^{-r/(2r+1/2)}$$

for every $n \in \mathbb{N}$ and some constant M that depends on α, α', r, B.
B) If $s > 2r$ and C_n is an adaptive and honest confidence set for $\tilde{\Sigma}(r, \rho_n) \cup \Sigma(s)$, for every $\alpha, \alpha' > 0$, then necessarily

$$\liminf_n \ \rho_n n^{r/(2r+1/2)} > 0.$$

Thus for $s \leq 2r$ adaptive confidence sets exist without any additional restrictions, but for $s > 2r$ one has to separate the two classes as in [18] for the L^∞-case, up to multiplicative constants, by the minimax rate of testing for the composite hypotheses

$$H_0 : f \in \Sigma(s) \ \text{ against } \ H_1 : f \in \tilde{\Sigma}(r, \rho_n),$$

which equals $n^{-r/(2r+1/2)}$. A phase transition occurs at $s = 2r$, when this rate is equal to $n^{-s/(2s+1)}$, the rate of estimation in the submodel H_0.

While the case of two fixed smoothness classes in Theorem 1 is appealing in its conceptual simplicity, it does not describe the typical adaptation problem, where one wants to adapt to a continuous smoothness parameter. As shown in the work of Bull [5], confidence bands that adapt to a continuum of smoothness parameters pose a more difficult problem than in the case of two (or finitely many) classes, and strong qualitative assumptions must be made to make adaptation possible. Bull's results are in the L^∞-setting, and a main point of the present article is to to solve the continuous adaptation problem in the L^2-case. The solution is different from the L^∞-case, and Theorem 1 turns out to have a full analogue for the continuous scale of Sobolev balls $\Sigma(t), t < r$, if the radius B is known. (See Theorems 5 and 6 below.)

The usual practice of 'undersmoothing' in the construction of confidence sets accommodates the fact that B is, unlike in Theorem 1, typically unknown, but incurs a rate-penalty for adaptation. Instead, we shall address the question of simultaneous exact adaptation to the radius B and the smoothness s. We prove that confidence sets that simultaneously adapt to a continuous smoothness parameter and to the unknown radius exist if $s < 2r$, but cannot exist, without additional assumptions, when $s > 2r$. We also analyse the phase transition at $s = 2r$, where separate arguments are needed. We prove in fact the closely related result that for $s > 2r$ even 'dishonest' inference is impossible for the full parameter space of probability densities in the r-Sobolev space, see Theorem 7. *In other words, even asymptotically one has to remove certain subsets of the maximal parameter space if one wants to construct confidence sets that adapt to arbitrary smoothness degrees.* One way to remove is to restrict the space to a fixed ball of known radius as discussed above, but other assumptions come to mind, such as 'self-similarity' conditions employed in Picard and Tribouley [29], Giné and Nickl [14], Kerkyacharian, Nickl and Picard [23] and Bull [5] for confidence intervals and bands. We discuss briefly how this applies in the L^2-setting.

A key technical result needed in the proofs, that seems of independent interest, is a general nonparametric test for composite null-hypotheses that lie in a fixed Sobolev ball. It is based on concentration inequalities for U-statistics, on Talagrand's [32] inequality, and on tools from empirical process theory. It requires an entropy condition on the null-hypothesis (see Proposition 1). Applied to the cases relevant in the present paper this test is minimax in the sense of Ingster [20, 21], a fact that is crucial to establish optimality of our confidence procedures. Another mathematical innovation is that we show how lower bound techniques for 'dishonest' confidence bands from Bull [5] apply to the L^2-situation (Theorem 7) – this amounts to deriving performance limits for confidence procedures under 'pointwise in f' coverage assumptions only, so in a way goes beyond the minimax paradigm of analysing adaptive procedures.

We state all main results other than Theorem 1 above in Section 2, and proofs are given, in a unified way, in Section 4

2 The Setting

2.1 Wavelets and Sobolev-Besov Spaces

Denote by $L^2 := L^2([0,1])$ the Lebesgue space of square integrable functions on $[0,1]$, normed by $\|\cdot\|_2$. For integer s the classical Sobolev spaces are defined as the spaces of functions $f \in L^2$ whose (distributional) derivatives $D^\alpha f, 0 < \alpha \leq s$, all lie in L^2. One can define these spaces, for $s > 0$ any real number, in terms of the natural sequence space isometry of L^2 under an orthonormal basis. We opt here to work with wavelet bases: for index sets $\mathcal{Z} \subset \mathbb{Z}, \mathcal{Z}_l \subset \mathbb{Z}$ and $J_0 \in \mathbb{N}$, let

$$\{\phi_m, 2^{l/2}\psi_k(2^l \cdot) : m \in \mathcal{Z}, k \in \mathcal{Z}_l, l \geq J_0 + 1, l \in \mathbb{N}\}$$

be a compactly supported orthonormal wavelet basis of L^2 of regularity S, and write, as usual, $\psi_{lk} = 2^{l/2}\psi_k(2^l \cdot)$. We shall only consider Cohen-Daubechies-Vial [8] wavelet bases where $|\mathcal{Z}_l| = 2^l, |\mathcal{Z}| \leq c(S) < \infty, J_0 \equiv J_0(S)$. We define, for $\langle f, g\rangle = \int_0^1 fg$ the usual L^2-inner product, and for $0 \leq s < S$, the Sobolev (-type) norms

$$\begin{aligned} \|f\|_{s,2} &:= \max\left(2^{J_0 s}\sqrt{\sum_{k\in\mathcal{Z}}\langle f, \phi_k\rangle^2}, \sup_{l\geq J_0+1} 2^{ls}\sqrt{\sum_{k\in\mathcal{Z}_l}\langle f, \psi_{lk}\rangle^2}\right) \\ &= \max\left(2^{J_0 s}\|\langle f, \phi_\cdot\rangle\|_2, \sup_{l\geq J_0+1} 2^{ls}\|\langle f, \psi_{l\cdot}\rangle\|_2\right) \qquad (5) \end{aligned}$$

where in slight abuse of notation we use the symbol $\|\cdot\|_2$ for the sequence norms on $\ell^2(\mathcal{Z}_l), \ell^2(\mathcal{Z})$ as well as for the usual norm on L^2. Define moreover the Sobolev (-type) spaces $W^s \equiv B^s_{2\infty} = \{f \in L^2 : \|f\|_{s,2} < \infty\}$. We note here that W^s is not the classical Sobolev space – in this case the supremum over $l \geq J_0 + 1$ would have to be replaced by summation over l – but the present definition gives rise to the slightly larger Besov space $B^s_{2\infty}$, which will turn out to be the natural exhaustive class for our results below. We still refer to them as Sobolev spaces for simplicity, and since the main idea is to measure smoothness in L^2. We understand W^s as spaces of continuous functions whenever $s > 1/2$ (possible by standard embedding theorems). We shall moreover set, in abuse of notation, $\phi_k \equiv \psi_{J_0 k}$ (which does not equal $2^{-1/2}\psi_{J_0+1,k}(2^{-1}\cdot)$) in order for the wavelet series of a function $f \in L^2$ to have the compact representation

$$f = \sum_{l=J_0}^{\infty} \sum_{k \in \mathcal{Z}_l} \psi_{lk} \langle \psi_{lk}, f \rangle,$$

with the understanding that $\mathcal{Z}_{J_0} = \mathcal{Z}$. The wavelet projection $\Pi_{V_j}(f)$ of $f \in L^2$ onto the span V_j in L^2 of

$$\{\phi_m, 2^{l/2}\psi_k(2^l\cdot) : m \in \mathcal{Z}, k \in \mathcal{Z}_l, J_0 + 1 \leq l \leq j\}$$

equals

$$K_j(f)(x) \equiv \int_0^1 K_j(x,y) f(y) dy \equiv 2^j \int_0^1 K(2^j x, 2^j y) f(y) dy = \sum_{l=J_0}^{j-1} \sum_{k \in \mathcal{Z}_l} \langle f, \psi_{lk} \rangle \psi_{lk}(x)$$

where $K(x,y) = \sum_k \phi_k(x)\phi_k(y)$ is the wavelet projection kernel.

2.2 Adaptive Estimation in L^2

Let $X_1, \dots, X_n$ be i.i.d. with common density f on $[0,1]$, with joint distribution equal to the first n coordinate projections of the infinite product probability measure $\Pr_f$. Write E_f for the corresponding expectation operator. We shall throughout make the minimal assumption that $f \in W^r$ for some $r > 1/2$, which implies in particular, by Sobolev's lemma, that f is continuous and bounded on $[0,1]$. The adaptation problem arises from the hope that $f \in W^s$ for some s significantly larger than r, without wanting to commit to a particular a priori value of s. In this generality the problem is still not meaningful, since the regularity of f is not only described by containment in W^s, but also by the size of the Sobolev norm $\|f\|_{s,2}$. If one defines, for $0 < s < \infty, 1 \leq B < \infty$, the Sobolev-balls of densities

$$\Sigma(s,B) := \left\{ f : [0,1] \to [0,\infty), \int_T f = 1, \|f\|_{s,2} \leq B \right\}, \tag{6}$$

then Pinsker's minimax theorem (for density estimation) gives, as $n \to \infty$,

$$\inf_{T_n} \sup_{f \in \Sigma(s,B)} E_f \|T_n - f\|_2^2 \sim c(s) B^{2/(2s+1)} n^{-2s/(2s+1)} \tag{7}$$

for some constant $c(s) > 0$ depending only on s, and where the infimum extends over all measurable functions T_n of $X_1, \dots, X_n$ (cf., e.g., the results in Theorem 5.1 in [11]). So any risk bound, attainable uniformly for elements $f \in \Sigma(s, B)$, cannot improve on $B^{2/(2s+1)} n^{-2s/(2s+1)}$ up to multiplicative constants. If s, B are known then constructing estimators that attain this bound is possible, even with the asymptotically exact constant $c(s)$. The adaptation problem poses the question of whether estimators can attain such a risk bound without requiring knowledge of B, s.

The paradigm of adaptive estimation has provided us with a positive answer to this problem, and one can prove the following result.

Theorem 2. *Let $1/2 < r \le R < \infty$ be given. Then there exists an estimator $\hat{f}_n = f(X_1, \dots, X_n, r, R)$ such that, for every $s \in [r, R]$, every $B \ge 1$, and every $n \in \mathbb{N}$,*

$$\sup_{f \in \Sigma(s,B)} E_f \|\hat{f}_n - f\|_2^2 \le c B^{2/(2s+1)} n^{-2s/(2s+1)}$$

for a constant $0 < c < \infty$ that depends only on r, R and on $U \equiv \sup_{f \in \Sigma(r,B)} \|f\|_\infty < \infty$.

More elaborate techniques allow for c to depend only on s, and even to obtain the exact asymptotic minimax ('Pinsker'-) constant, see for instance Theorem 5.1 in [11]. We shall not study exact constants here, mostly to simplify the exposition and to focus on the main problem of confidence statements, but also since exact constants are asymptotic in nature and we prefer to give nonasymptotic bounds.

From a dishonest – 'pointwise in f' – perspective we can conclude from Theorem 2 that adaptive estimation is possible over the full continuous Sobolev scale

$$\bigcup_{s \in [r,R], B \ge 1} \Sigma(s, B) = W^r \cap \left\{ f : [0,1] \to [0, \infty), \int_0^1 f = 1 \right\};$$

for any probability density $f \in W^s, s \in [r, R]$, the single estimator $\hat{f}_n$ satisfies

$$E_f \|\hat{f}_n - f\|_2^2 \le c \|f\|_{s,2}^{2/(2s+1)} n^{-s/(2s+1)},$$

and this result is uniform in any fixed ball of W^s. Since $\hat{f}_n$ depends on neither B nor s we can say that $\hat{f}_n$ adapts to both $s \in [r, R]$ and $B \ge 1$ simultaneously. Our interest here is to understand what remains of this remarkable result if one is interested in adaptive *confidence statements* rather than in risk bounds.

3 Adaptive Confidence Sets for Sobolev Classes

3.1 Honest and Dishonest Inference

We aim to characterise those sets $\mathcal{P}_n$ consisting of probability densities $f \in W^r$ for which we can construct adaptive confidence sets. More precisely, we seek random subsets C_n of L^2 that depend only on known quantities, cover $f \in \mathcal{P}_n$ at least with prescribed probability $1-\alpha$, and have L^2-diameter adaptive with respect to radius and smoothness with prescribed probability at least $1-\alpha'$. Recall the definition of the diameter $|C|$ of a subset of L^2 from (1) above, and of the constant U from Theorem 2.

Definition 1 (L^2-adaptive confidence sets.). *Let $X_1, \dots, X_n$ be i.i.d. on $[0,1]$ with common density f. Let $0 < \alpha, \alpha' < 1$ and $1/2 < r \leq R$ be given and let $C_n = C(X_1, \dots, X_n, r, R, \alpha, \alpha')$ be a random subset of L^2 .*
i) C_n is called L^2-adaptive and honest for a sequence of (nonempty) models $\mathcal{P}_n \subset W^r$ if there exists a constant $L = L(r, R, U)$ such that for every $n \in \mathbb{N}$

$$\sup_{f \in \Sigma(s,B) \cap \mathcal{P}_n} \Pr_f \left\{ |C_n| > L B^{1/(2s+1)} n^{-s/(2s+1)} \right\} \leq \alpha' \quad \text{for every } s \in [r, R], B \geq 1, \tag{8}$$

(the condition being void if $\Sigma(s,B) \cap \mathcal{P}_n$ is empty) and

$$\inf_{f \in \mathcal{P}_n} \Pr_f \{ f \in C_n \} \geq 1 - \alpha - r_n \tag{9}$$

where $r_n \to 0$ as $n \to \infty$.
ii) C_n is called L^2-adaptive and dishonest for a model $\mathcal{P} \subset W^r$ if (8) and (9) in part i) of the definition are replaced by

$$\limsup_n \Pr_f \left\{ |C_n| > L \|f\|_{s,2}^{1/(2s+1)} n^{-s/(2s+1)} \right\} \leq \alpha' \quad \forall f \in W^s \cap \mathcal{P}, s \in [r, R], \tag{10}$$

where now $L = L(r, R, \|f\|_\infty)$, and

$$\liminf_n \Pr_f \{ f \in C_n \} \geq 1 - \alpha \quad \forall f \in \mathcal{P}. \tag{11}$$

respectively.

Some discussion is in order. We shall often only say that C_n is 'adaptive and honest/dishonest for $\mathcal{P}_n/\mathcal{P}$' when no confusion may arise. Note that in Definition 1 the set C_n is not allowed to depend on the radius B, but we require the sharp dependence on B in (8), so that the usual 'undersmoothed', near-adaptive, confidence sets are excluded in our setting. The natural 'maximal' choice in the honest case would be $\mathcal{P}_n = \Sigma(r, B)\ \forall n$ with $B \geq 1$ arbitrary but fixed, whereas in the 'dishonest' case one would ideally want the full parameter space $\mathcal{P} = W^r \cap \{f : [0,1] \to [0, \infty), \int_0^1 f = 1\}$. Dishonest confidence sets are mostly of theoretical interest: they have the uniformity requirements of honesty relaxed to asymptotic 'pointwise in f' statements, so cannot be used for valid asymptotic inference as the index n from which onwards coverage holds depends on f. We study them because they showcase the subtleties of certain situations where dishonest inference is possible but honest inference is not. They further serve as a benchmark to which the limit sets $\lim_n \mathcal{P}_n$ of any sequence of models $\mathcal{P}_n$ for which honest inference can be established should be compared to. Any honest confidence set for a sequence of increasing models $\mathcal{P}_n$ automatically gives rise to 'dishonest' confidence statements for the limit set $\lim_n \mathcal{P}_n$, and given two models it seems reasonable to prefer the one with the larger limit set.

The interval $[r, R]$ describes the range of smoothness parameters one wants to adapt to. Besides the restriction $1/2 < r \leq R < \infty$ the choice of this window of adaptation is arbitrary (although the values of R, r influence the constants). The results below will show that whether L^2-adaptive confidence sets exist – honest or dishonest – depends crucially on whether R exceeds $2r$ or not, with a phase transition at $R = 2r$.

3.2 The Case $R < 2r$.

A first result, the key elements of which have been discovered and discussed in [25, 17, 22, 7, 30], is that L^2-adaptive confidence statements that parallel the situation of Theorem 2 exist without any additional restrictions whatsoever, in the case where $R < 2r$, so that the window of adaptation is $[r, 2r)$. The following theorem is a simple extension of results in Robins and van der Vaart [30] in that it shows that adaptation is possible not only to the smoothness s, but also to the radius B, in particular B does not have to be known for the construction of the procedure.

Theorem 3. *Let $R < 2r$.*
i) An honest and adaptive confidence set in the sense of Definition 1 exists for every fixed Sobolev ball, i.e., for the choice $\mathcal{P}_n \equiv \Sigma(r, B)$ $\forall n$ where $B \geq 1$ is arbitrary.
ii) An adaptive and dishonest confidence set in the sense of Definition 1 exists for the full model $\mathcal{P} = W^r \cap \{f : [0, 1] \to [0, \infty), \int_0^1 f = 1\}$.

The main idea of the proof is that, if $R < 2r$, the squared L^2-risk of $\hat{f}_n$ from Theorem 2 can be estimated at a rate compatible with adaptation, by a suitable U-statistic. Note that the sequence r_n from Definition 1 (bot not C_n) depends on B in the above theorem – one may thus use C_n without any prior choice of parameters, but evaluation of its coverage is still relative to the model $\Sigma(r, B)$.

3.3 The Case $R > 2r$

The obvious question arises as to what happens when the requirement $R < 2r$ is dropped. Let us consider first $R > 2r$, the phase transition $R = 2r$ will be treated separately below.

On the one hand we have, in view of Part B of Theorem 1, an immediate negative result concerning honest inference. Note that we even allow C_n to depend on B in the following theorem.

Theorem 4. *If an honest and adaptive confidence set C_n in the sense of Definition 1 exists for the choice $\mathcal{P}_n \equiv \Sigma(r, B)$ $\forall n$ and some $B > 1$, then necessarily $R \leq 2r$.*

On the other hand we have the following result. While implied by Theorem 3 in case $R < 2r$, it shows that for $R \geq 2r$, perhaps surprisingly, dishonest inference is still possible when an upper bound for the radius B of the Sobolev ball is known.

Theorem 5. *Let $R \geq r$ be arbitrary and let $B \geq 1$ be given. A dishonest and adaptive confidence set $C_n = C(X_1, \ldots, X_n, B, r, R, \alpha, \alpha')$ in the sense of Definition 1 exists for $\mathcal{P} = \Sigma(r, B)$.*

We see that there exists a genuine discrepancy between honest and dishonest adaptive confidence sets in the case $R > 2r$. Moreover Theorem 5 is specific to confidence sets that are L^2-adaptive, and a corresponding result for confidence bands over a continuous scale of Hölder balls cannot be proved, cf. the results in [5]. We wish to understand the mechanisms behind Theorem 5 in more detail.

The proof of Theorem 5 is based on the idea of constructing a subset $\mathcal{P}_n$ of $\Sigma(r, B)$ which grows dense in $\Sigma(r, B)$ and for which honest inference is possible. This approach follows the ideas in Hoffmann and Nickl [18] in the L^∞-case for two fixed Hölder balls, and works as follows in the L^2-setting: Assume without loss of generality that $R = 2Nr$ for some $N \in \mathbb{N}, N > 1$, and define the grid

$$\mathcal{S} = \{s_m\}_{m=1}^N = \{r, 2r, 4r, \ldots, 2(N-1)r\}.$$

Note that $\mathcal{S}$ is independent of n. Since B will be known in what follows we will sometimes write $\Sigma(s)$ for $\Sigma(s, B)$ to expedite notation. Define, for $s \in \mathcal{S} \setminus \{s_N\}$,

$$\tilde{\Sigma}(s, \rho) := \tilde{\Sigma}(s, B, \mathcal{S}, \rho) = \{f \in \Sigma(s) : \|f - \Sigma(t)\|_2 \geq \rho \ \forall t > s, t \in \mathcal{S}\}.$$

We will choose the separation rates

$$\rho_n(s) \sim n^{-s/(2s+1/2)},$$

equal to the minimax rate of testing between $\Sigma(s)$ and any submodel $\Sigma(t)$ for $t \in \mathcal{S}, t > s$, see Section 4.2 below. The resulting model is therefore, for M some positive constant,

$$\mathcal{P}_n(M, \mathcal{S}) = \Sigma(s_N) \bigcup \left(\bigcup_{s \in \mathcal{S} \setminus \{s_N\}} \tilde{\Sigma}(s, M\rho_n(s)) \right).$$

The main idea behind the following theorem, which is a main result of the paper, is to construct a minimax test for the nested hypotheses $\{H_s : f \in \tilde{\Sigma}(s, M\rho_n(s))\}_{s \in \mathcal{S} \setminus \{s_N\}}$ and then use the confidence set from Theorem 3 for the smoothness hypothesis selected by the test.

Theorem 6. *Let $R > 2r$ be arbitrary and let $B \geq 1$ be given. An honest and adaptive confidence set $C_n = C(X_1, \ldots, X_n, B, r, R, \alpha, \alpha')$ in the sense of Definition 1 exists for $\mathcal{P}_n = \mathcal{P}_n(M, \mathcal{S}), n \in \mathbb{N}$, with M a large enough constant depending on α, α'.*

Theorem 6 is optimal in the following sense: First note that, since $\mathcal{S}$ is independent of n, $\mathcal{P}_n(M, \mathcal{S}) \nearrow \Sigma(r)$ as $n \to \infty$, so that the model $\mathcal{P}_n(M, \mathcal{S})$ grows dense in the fixed Sobolev ball, which for known B is the full model. This implies in particular Theorem 5 for $R > 2r$. Another question is whether $\mathcal{P}_n(M, \mathcal{S})$ was taken to grow as fast as possible as a function of n, or in other words, whether a smaller choice of $\rho_n(s)$ would have been possible. The lower bound in Theorem 1 implies that any faster choice for $\rho_n(s)$ makes honest inference impossible. Indeed, if C_n is an honest confidence set over $\mathcal{P}_n(M, \mathcal{S})$ with a faster separation rate $\rho'_n = o(\rho_n(s))$ for some $s \in \mathcal{S} \setminus \{s_N\}$, then we can use C_n

to test $H_0 : f \in \Sigma(s')$ against $H_1 : f \in \tilde{\Sigma}(s, \rho_n')$ for some $s' > 2s$, which by the proof of Theorem 1 gives a contradiction.

The proof of Theorem 6 via testing smoothness hypotheses is strongly tied to knowledge of the radius B, and the question arises whether adaptation to unknown B is also possible, as in the case $R < 2r$, in particular if, as a consequence, dishonest inference is possible over the full model W^r for arbitrary $R \geq r$. The answer to this question is negative, as the following lower bound for dishonest confidence sets shows.

Theorem 7. *Fix $0 < \alpha < 1/2$ and let $s \geq r$ be arbitrary. A random subset C_n of L^2 cannot satisfy*
i) $\liminf_n \Pr_f(f \in C_n) \geq 1 - \alpha$ *for every probability density $f \in W^r$*
ii) $|C_n| = O_{\Pr_f}(r_n)$ *for every probability density $f \in W^s$*
at any rate $r_n = o(n^{-r/(2r+1/2)})$.

Since $n^{-R/(2R+1)} = o(n^{-r/(2r+1/2)})$ for $R > 2r$ we cannot possibly hope for an adaptive dishonest confidence set for all of W^r. In fact Theorem 7 reveals substantial limitations for L^2-adaptive inference over the full Sobolev scale with general R, r: It implies that *any* sequence of models $\mathcal{P}_n$ over which honest adaptive inference is possible *must asymptotically remove* elements from W^r. One way to remove from W^r was given in Theorem 6 by restricting the radius, another one is discussed in Subsection 3.5.

3.4 The Phase Transition at $R = 2r$

In the critical case $R = 2r$ the situation can be described as follows. If B is known, honest confidence sets that are L^2-adaptive exist, but simultaneous adaptation to unknown B is impossible. Dishonest inference is, however, possible for the full Sobolev space, illustrating further the discrepancy between honest and dishonest inference.

Theorem 8. *Let $R = 2r$.*
a) A confidence set C_n that is honest and adaptive in the sense of Definition 1 for $\mathcal{P}_n = \Sigma(r, B)\ \forall n$ for arbitrary $B > 1$ does not exist.
b) Let $B \geq 1$ be given. An honest and adaptive confidence set $C_n = C(X_1, \dots, X_n, B, r, R, \alpha, \alpha')$ in the sense of Definition 1 exists for $\mathcal{P}_n = \Sigma(r, B)\ \forall n$.
c) A dishonest and adaptive confidence set in the sense of Definition 1 exists for the full model $\mathcal{P} = W^r \cap \{f : [0,1] \to [0, \infty), \int_0^1 f = 1\}$.

3.5 Self-Similarity Conditions

Alternative ways to restrict W^r, which may be practically more relevant, are given in [29, 14, 23, 5]. The authors instead restrict to 'self-similar' functions, whose regularity is similar at large and small scales. As the results [14, 23, 5] prove adaptation in L^∞, they naturally imply adaptation also in L^2; however, in the more favourable L^2 case, their assumptions are stronger than necessary. Define the truncated wavelet expansion

$$f_{i,j} := \sum_{l=i}^{j} \sum_{k \in \mathcal{Z}_l} \langle \psi_{lk}, f \rangle \psi_{lk},$$

using the notation from Section 2.1. For $J_0 \le J \le K$, $\varepsilon \in (0,1)$, consider the class of functions

$$\Sigma(t,J,K,\varepsilon) = \left\{ f \in W^t : f \text{ a density, } \min\left(\|f_{j_0,J}\|_{t,2}, \|f_{K,\infty}\|_{t,2}\right) \ge \varepsilon \|f\|_{t,2} \right\}.$$

Using similar arguments to Bull [5], one may construct confidence sets which are honest and adaptive over classes

$$\mathcal{P} = \bigcup_{s\in[r,R]} \Sigma(s,J,K,\varepsilon),$$

for any $0 < r \le R$, and large enough K. These classes are again smaller than the full space W^r, and require the choice of parameters J, K and ε. The functions we exclude, however, are now those whose norm is hard to estimate, rather than those whose norm is merely large.

4 Proofs

4.1 Some Concentration Inequalities

Let $X_i, i = 1,2,\dots$, be the coordinates of the product probability space $(S,\mathcal{S},P)^{\mathbb{N}}$, where P is any probability measure on $(S,\mathcal{S})$, $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ the empirical measure, E expectation under $P^{\mathbb{N}} \equiv \Pr$. For M any set and $H : M \to \mathbb{R}$, set $\|H\|_M = \sup_{m\in M}|H(m)|$. We also write $Pf = \int_S f dP$ for $f : S \to \mathbb{R}$.

The following Bernstein-type inequality for canonical U-statistics of order two is due to Giné, Latala and Zinn [13], with refinements about the numerical constants in Houdré and Reynaud-Bouret [19]: Let $R(x,y)$ be a symmetric real-valued function defined on $S \times S$, such that $ER(X,x) = 0$ for all x, and let

$$\Lambda_1^2 = \frac{n(n-1)}{2} ER(X_1,X_2)^2,$$

$$\Lambda_2 = n \sup\{E[R(X_1,X_2)\zeta(X_1)\xi(X_2)] : E\zeta^2(X_1) \le 1, E\xi^2(X_1) \le 1\},$$

$$\Lambda_3 = \|nER^2(X_1,\cdot)\|_\infty^{1/2}, \quad \Lambda_4 = \|R\|_\infty.$$

Let moreover $U_n^{(2)}(R) = \frac{2}{n(n-1)}\sum_{i<j} R(X_i,X_j)$ be the corresponding degenerate U-statistic of order two. Then, there exists a universal constant $0 < C < \infty$ such that for all $u > 0$ and $n \in \mathbb{N}$:

$$\Pr\left(\frac{n(n-1)}{2}|U_n^{(2)}(R)| > C(\Lambda_1 u^{1/2} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2)\right) \le 6\exp\{-u\}. \qquad (12)$$

We will also need Talagrand's [32] inequality for empirical processes. Let $\mathcal{F}$ be a countable class of measurable functions on S that take values in $[-1/2,1/2]$, or, if $\mathcal{F}$ is P-centered, in $[-1,1]$. Let $\sigma \le 1/2$, or $\sigma \le 1$ if $\mathcal{F}$ is P-centered, and V be any two numbers satisfying

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

Bousquet's [4] version of Talagrand's inequality then states: For every $u > 0$,

$$\Pr\left\{\left\|\sum_{i=1}^{n}(f(X_i) - Pf)\right\|_{\mathcal{F}} \geq E\left\|\sum_{i=1}^{n}(f(X_i) - Pf)\right\|_{\mathcal{F}} + u\right\} \leq \exp\left(-\frac{u^2}{2V + \frac{2}{3}u}\right). \quad (13)$$

A consequence of this inequality, derived in Section 3.1 in [16], is the following. If $S = [0,1]$, P has bounded Lebesgue density f on S, and $f_n(j) = \int_0^1 K_j(\cdot, y)P_n(y)$, then for M large enough, every $j \geq 0, n \in \mathbb{N}$ and some positive constants c, c',

$$\Pr\left(\|f_n(j) - Ef_n(j)\|_2 > M\sqrt{\|f\|_\infty \frac{2^j}{n}}\right) \leq c' e^{-cM^2 2^j}. \quad (14)$$

4.2 A General Purpose Test for Composite Nonparametric Hypotheses

In this subsection we construct a general test for composite nonparametric null hypotheses that lie in a fixed Sobolev ball, under assumptions only on the entropy of the null-model. While of independent interest, the result will be a key step in the proofs of Theorems 1 and 6, where minimax optimal tests between two nested Sobolev-balls are required.

Let $X, X_1, \dots, X_n$ be i.i.d. with common probability density f on $[0,1]$, let Σ be any subset of a fixed Sobolev ball $\Sigma(t) = \Sigma(t, B)$ for some $t > 1/2$ and consider testing

$$H_0 : f \in \Sigma \ \text{ against } H_1 : f \in \Sigma(t) \setminus \Sigma, \|f - \Sigma\|_2 \geq \rho_n, \quad (15)$$

where $\rho_n \geq 0$ is a sequence of nonnegative real numbers. For $\{\psi_{lk}\}$ a S-regular wavelet basis, $S > t$, $J_n \geq J_0$ a sequence of positive integers such that $2^{J_n} \simeq n^{1/(2t+1/2)}$ and for $g \in \Sigma$, define the U-statistic

$$T_n(g) = \frac{2}{n(n-1)} \sum_{i<j} \sum_{l=J_0}^{J_n - 1} \sum_{k \in \mathcal{Z}_l} (\psi_{lk}(X_i) - \langle \psi_{lk}, g\rangle)(\psi_{lk}(X_j) - \langle \psi_{lk}, g\rangle) \quad (16)$$

and, for τ_n some thresholds to be chosen below, the test statistic

$$\Psi_n = 1\left\{\inf_{g \in \Sigma} |T_n(g)| > \tau_n\right\}. \quad (17)$$

We shall neglect measurability issues here, but Proposition 1 below holds without measurability by using outer expectation. In the cases relevant below where Σ equals $\Sigma(s)$, a Sobolev ball contained in $\Sigma(t)$, measurability of the infimum in (17) can be established by standard compactness/continuity arguments.

We shall prove a bound on the sum of the type-one and type-two errors of this test under some entropy conditions on Σ, more precisely, on the class of functions

$$\mathcal{G}(\Sigma) = \bigcup_{J > J_0} \left\{\sum_{l=J_0}^{J-1} \sum_{k \in \mathcal{Z}_l} \psi_{lk}(\cdot)\langle \psi_{lk}, g\rangle : g \in \Sigma\right\}.$$

Recall the usual covering numbers $N(\varepsilon, \mathcal{G}, L^2(P))$ and bracketing metric entropy numbers $N_{[]}(\varepsilon, \mathcal{G}, L^2(P))$ for classes $\mathcal{G}$ of functions and probability measures P on $[0,1]$ (e.g., [33, 34]).

Definition 2. *Say that Σ is s-regular if one of the following conditions is satisfied for some fixed finite constants A and every $0 < \varepsilon < A$:*
a) For any probability measure Q on $[0,1]$ (and A independent of Q) we have

$$\log N(\varepsilon, \mathcal{G}(\Sigma), L^2(Q)) \leq (A/\varepsilon)^{1/s}.$$

b) For P such that $dP = f d\lambda$ with Lebesgue density $f : [0,1] \to [0, \infty)$ we have

$$\log N_{[]}(\varepsilon, \mathcal{G}(\Sigma), L^2(P)) \leq (A/\varepsilon)^{1/s}.$$

Note that a ball $\Sigma(s, B)$ satisfies this condition for the given $s, 1/2 < s < S$, since any element of $\mathcal{G}(\Sigma(s, B))$ has $\|\cdot\|_{s,2}$-norm no more than B, and since $\log N(\varepsilon, \Sigma(s, B), \|\cdot\|_\infty) \leq (A/\varepsilon)^{1/s}$, see, e.g., p.506 in [27].

Proposition 1. *Let*

$$\tau_n = L d_n \max(n^{-2s/(2s+1)}, n^{-2t/(2t+1/2)}), \quad \rho_n^2 = \frac{L_0}{L} \tau_n$$

for real numbers $1 \leq d_n \leq d(\log n)^\gamma$ and positive constants L, L_0, γ, d. Let the hypotheses H_0, H_1 be as in (15), the test Ψ_n as in (17), and assume Σ is s-regular for some $s > 1/2$. Then for $L = L(B, t, S)$, $L_0 = L_0(L, B, t, S)$ large enough and every $n \in \mathbb{N}$ there exist constants $c_i, i = 1, \dots, 3$ depending only on L, L_0, t, B such that

$$\sup_{f \in H_0} E_f \Psi_n + \sup_{f \in H_1} E_f(1 - \Psi_n) \leq c_1 e^{-d_n^2} + c_2 e^{-c_3 n \rho_n^2}.$$

Proof. 1) We first control the type-one errors. Since $f \in H_0 = \Sigma$ we see

$$E_f \Psi_n = \Pr{}_f \left\{ \inf_{g \in \Sigma} |T_n(g)| > \tau_n \right\} \leq \Pr{}_f \left\{ |T_n(f)| > \tau_n \right\}. \tag{18}$$

$T_n(f)$ is a U-statistic with kernel

$$R_f(x, y) = \sum_{l=J_0}^{J_n - 1} \sum_{k \in \mathcal{Z}_l} (\psi_{lk}(x) - \langle \psi_{lk}, f \rangle)(\psi_{lk}(y) - \langle \psi_{lk}, f \rangle),$$

which satisfies $ER_f(x, X_1) = 0$ for every x, since $E_f(\psi_{lk}(X) - \langle \psi_{lk}, f \rangle) = 0$ for every k, l. Consequently $T_n(f)$ is a degenerate U-statistic of order two, and we can apply inequality (12) to it, which we shall do with $u = d_n^2$. We thus need to bound the constants $\Lambda_1, \dots, \Lambda_4$ occurring in inequality (12) in such a way that, for L large enough,

$$\frac{2C}{n(n-1)}(\Lambda_1 d_n + \Lambda_2 d_n^2 + \Lambda_3 d_n^3 + \Lambda_4 d_n^4) \leq L d_n n^{-2t/(2t+1/2)} \leq \tau_n, \tag{19}$$

which is achieved by the following estimates, noting that $n^{-2t/(2t+1/2)} \simeq 2^{J_n/2}/n$.

First, by standard U-statistic arguments, we can bound $ER_f^2(X_1, X_2)$ by the second moment of the uncentred kernel, and thus, using orthonormality of ψ_{lk},

$$\begin{aligned} ER_f^2(X_1, X_2) &\leq \int\int \left(\sum_{k,l} \psi_{lk}(x)\psi_{lk}(y)\right)^2 f(x)f(y)dxdy \\ &\leq \|f\|_\infty^2 \sum_{l=J_0}^{J_n-1} \sum_{k \in \mathcal{Z}_l} \int_0^1 \psi_{lk}^2(x)dx \int_0^1 \psi_{lk}^2(y)dy \\ &\leq C(S)2^{J_n}\|f\|_\infty^2 \end{aligned}$$

for some constant $C(S)$ that depends only on the wavelet basis. We obtain $\Lambda_1^2 \leq C(S)n(n-1)2^{J_n}\|f\|_\infty^2/2$ and it follows, using

$$\sup_{f \in \Sigma(t,B)} \|f\|_\infty \leq C(B,t), \tag{20}$$

that for L large enough and every n,

$$\frac{2C\Lambda_1 d_n}{n(n-1)} \leq C(S,B,t)\frac{2^{J_n/2}d_n}{n} \leq \tau_n/4.$$

For the second term note that, using the Cauchy-Schwarz inequality and that K_j is a projection operator

$$\begin{aligned} \left|\int\int \sum_{l=J_0}^{J_n-1} \sum_{k \in \mathcal{Z}_l} \psi_{lk}(x)\psi_{lk}(y)\zeta(x)\xi(y)f(x)f(y)dxdy\right| &= \left|\int K_{J_n}(\zeta f)(y)\xi(y)f(y)dy\right| \\ &\leq \|K_{J_n}(\zeta f)\|_2\|\xi f\|_2 \leq \|f\|_\infty^2, \end{aligned}$$

and similarly

$$|E[E_{X_1}[K_{J_n}(X_1, X_2)]\zeta(X_1)\xi(X_2)]| \leq \|f\|_\infty^2, \ |EK_{J_n}(X_1, X_2)| \leq \|f\|_\infty^2.$$

Thus

$$E[R_f(X_1, X_2)\zeta(X_1)\xi(X_2)] \leq 4\|f\|_\infty^2$$

so that, using (20),

$$\frac{2C\Lambda_2 d_n^2}{n(n-1)} \leq \frac{C'(B,t)d_n^2}{n} \leq \tau_n/4$$

again for L large enough and every n.

For the third term, using the decomposition $R_f(x_1, x) = (r(x_1, x) - E_{X_1}r(X, x)) + (E_{X,Y}r(X,Y) - E_Y r(x_1, Y))$ for $r(x,y) = \sum_{k,l}\psi_{lk}(x)\psi_{lk}(y)$, the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and again orthonormality, we have that for every $x \in \mathbb{R}$,

$$n|E_{X_1}R_f^2(X_1, x)| \leq 2n\left[\|f\|_\infty \sum_{l=J_0}^{J_n-1} \sum_{k \in \mathcal{Z}_l} \psi_{lk}^2(x) + \|f\|_\infty\|\Pi_{V_{J_n}}(f)\|_2^2\right]$$

so that, using $\|\psi_{lk}\|_\infty \leq d2^{l/2}$, again for L large enough and by (20),

$$\frac{2C\Lambda_3 d_n^3}{n(n-1)} \leq C''(B,t)\frac{2^{J_n/2}d_n^3}{n}\frac{1}{\sqrt{n}} \leq \tau_n/4.$$

Finally, we have $\Lambda_4 = \|R_f\|_\infty \leq c2^{J_n}$ and hence

$$\frac{2C\Lambda_4 d_n^4}{n(n-1)} \leq C'\frac{2^{J_n}d_n^4}{n^2} \leq \tau_n/4,$$

so that we conclude for L large enough and every $n \in \mathbb{N}$, from inequality (12),

$$\Pr_f\{|T_n(f)| > \tau_n\} \leq 6\exp\left\{-d_n^2\right\} \tag{21}$$

which completes the bound for the type-one errors in view of (18).

2) We now turn to the type-two errors. In this case, for $f \in H_1$

$$E_f(1-\Psi_n) = \Pr_f\left\{\inf_{g\in\Sigma}|T_n(g)| \leq \tau_n\right\}. \tag{22}$$

and the typical summand of $T_n(g)$ has Hoeffding-decomposition

$$\begin{aligned}
&(\psi_{lk}(X_i) - \langle\psi_{lk},g\rangle)(\psi_{lk}(X_j) - \langle\psi_{lk},g\rangle)\\
&= (\psi_{lk}(X_i) - \langle\psi_{lk},f\rangle + \langle\psi_{lk},f-g\rangle)(\psi_{lk}(X_j) - \langle\psi_{lk},f\rangle + \langle\psi_{lk},f-g\rangle)\\
&= (\psi_{lk}(X_i) - \langle\psi_{lk},f\rangle)(\psi_{lk}(X_j) - \langle\psi_{lk},f\rangle))\\
&\quad + (\psi_{lk}(X_i) - \langle\psi_{lk},f\rangle)\langle\psi_{lk},f-g\rangle + (\psi_{lk}(X_j) - \langle\psi_{lk},f\rangle)\langle\psi_{lk},f-g\rangle\\
&\quad + \langle\psi_{lk},f-g\rangle^2
\end{aligned}$$

so that by the triangle inequality, writing

$$L_n(g) = \frac{2}{n}\sum_{i=1}^{n}\sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}(\psi_{lk}(X_i) - \langle\psi_{lk},f\rangle)\langle\psi_{lk},f-g\rangle \tag{23}$$

for the linear terms, we conclude

$$\begin{aligned}
|T_n(g)| &\geq \sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}\langle\psi_{lk},f-g\rangle^2 - |T_n(f)| - |L_n(g)|\\
&= \|\Pi_{V_{J_n}}(f-g)\|_2^2 - |T_n(f)| - |L_n(g)| \qquad (24)
\end{aligned}$$

for every $g \in \Sigma$.

We can find random $g_n^* \in \Sigma$ such that $\inf_{g\in\Sigma}|T_n(g)| = |T_n(g_n^*)|$. (If the infimum is not attained the proof below requires obvious modifications; for the case $\Sigma = \Sigma(s,B), s > t$, relevant below, the infimum can be shown to be attained at a measurable minimiser by standard continuity and compactness arguments.) We bound the probability in (22), using (24), by

$$\Pr_f\left\{|L_n(g_n^*)| > \frac{\|\Pi_{V_{J_n}}(f-g_n^*)\|_2^2-\tau_n}{2}\right\} + \Pr_f\left\{|T_n(f)| > \frac{\|\Pi_{V_{J_n}}(f-g_n^*)\|_2^2-\tau_n}{2}\right\}.$$

Now by the standard approximation bound (cf. (5)) and since $g_n^* \in \Sigma \subset \Sigma(t)$,

$$\|\Pi_{V_{J_n}}(f-g_n^*)\|_2^2 \geq \inf_{g\in\Sigma}\|f-g\|_2^2 - c(B)2^{-2J_nt} \geq 4\tau_n \tag{25}$$

for L_0 large enough depending only on B and the choice of L from above. We can thus bound the sum of the last two probabilities by

$$\Pr_f\{|L_n(g_n^*)| > \|\Pi_{V_{J_n}}(f-g_n^*)\|_2^2/4\} + \Pr_f\{|T_n(f)| > \tau_n\}.$$

For the second degenerate part the proof of Step 1 applies, as only boundedness of f was used there. In the linear part somewhat more care is necessary. We have

$$\Pr_f\{|L_n(g_n^*)| > \|\Pi_{V_{J_n}}(f-g_n^*)\|_2^2/4\} \leq \Pr_f\left\{\sup_{g\in\Sigma}\frac{|L_n(g)|}{\|\Pi_{V_{J_n}}(f-g)\|_2^2} > \frac{1}{4}\right\}. \tag{26}$$

Note that the variance of the linear process from (23) can be bounded, for fixed $g \in \Sigma$, using independence and orthonormality, by

$$\begin{aligned} Var_f(|L_n(g)|) &\leq \frac{4}{n}\int\left(\sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}\psi_{lk}(x)\langle\psi_{lk}, f-g\rangle\right)^2 f(x)dx \\ &\leq \frac{4\|f\|_\infty}{n}\sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}\int\psi_{lk}^2(x)dx\cdot\langle\psi_{lk}, f-g\rangle^2 \\ &\leq \frac{4\|f\|_\infty\|\Pi_{V_{J_n}}(f-g)\|_2^2}{n} \end{aligned} \tag{27}$$

so that the supremum in (26) is one of a self-normalised ratio-type empirical process. Such processes can be controlled by slicing the supremum into shells of almost constant variance, cf. Section 5 in [33] or [12]. Define, for $g \in \Sigma$,

$$\sigma^2(g) := \|\pi_{V_{J_n}}(f-g)\|_2^2 \geq \|f-g\|_2^2 - c(B)2^{-2J_nt} \geq c\rho_n^2,$$

the inequality holding for L_0 large enough and some $c > 0$. Define moreover, for $m \in \mathbb{Z}$, the class of functions

$$\mathcal{G}_{m,J_n} = \left\{2\sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}\psi_{lk}(\cdot)\langle\psi_{lk}, f-g\rangle : g\in\Sigma, \sigma^2(g) \leq 2^{m+1}\right\},$$

which is uniformly bounded by a constant multiple of $\|f\|_{t,2} + \sup_{g \in \Sigma(t,B)} \|g\|_{t,2} \leq 2B$ in view of (5) and since $t > 1/2$. Then clearly, in the notation of Subsection 4.1,

$$\sup_{g \in \Sigma : \sigma^2(g) \leq 2^{m+1}} |L_n(g)| = \|P_n - P\|_{\mathcal{G}_{m,J_n}}$$

and we bound the last probability in (26) by

$$\begin{aligned} &\Pr_f \left\{ \max_{m \in \mathbb{Z}: c'\rho_n^2 \leq 2^m \leq C} \sup_{g \in \Sigma: 2^m \leq \sigma^2(g) \leq 2^{m+1}} \frac{|L_n(g)|}{\sigma^2(g)} > \frac{1}{4} \right\} \\ &\leq \sum_{m \in \mathbb{Z}: c'\rho_n^2 \leq 2^m \leq C} \Pr_f \left\{ \sup_{g \in \Sigma : \sigma^2(g) \leq 2^{m+1}} |L_n(g)| > 2^{m-2} \right\} \qquad (28) \\ &\leq \sum_{m \in \mathbb{Z}: c'\rho_n^2 \leq 2^m \leq C} \Pr_f \left\{ \|P_n - P\|_{\mathcal{G}_{m,J_n}} - E\|P_n - P\|_{\mathcal{G}_{m,J_n}} > 2^{m-2} - E\|P_n - P\|_{\mathcal{G}_{m,J_n}} \right\} \end{aligned}$$

where we may take $C < \infty$ as $\Sigma \subset \Sigma(t)$ is bounded in L^2, and where c' is a positive constant such that $c'\rho_n^2 \leq 2^m \leq c\rho_n^2$ for some $m \in \mathbb{Z}$. We bound the expectation of the empirical process. Both the uniform and the bracketing entropy condition for $\mathcal{G}(\Sigma)$ carry over to $\cup_{J \geq 0} \mathcal{G}_{J,m}$ since translation by f preserves the entropy. Using the standard entropy-bound plus chaining moment inequality (3.5) in Theorem 3.1 in [12] in case a) of Definition 2, and the second bracketing entropy moment inequality in Theorem 2.14.2 in [34] in case b), together with the variance bound (27) and with (20), we deduce

$$E\|P_n - P\|_{\mathcal{G}_{m,J_n}} \leq C \left(\sqrt{\frac{2^m}{n}} (2^m)^{-1/4s} + \frac{(2^m)^{-1/2s}}{n} \right). \qquad (29)$$

We see that

$$2^{m-2} - E\|P_n - P\|_{\mathcal{G}_k} \geq c_0 2^m$$

for some fixed c_0 precisely when 2^m is of larger magnitude than $(2^m)^{\frac{1}{2} - \frac{1}{4s}} n^{-1/2} + (2^m)^{-1/2s} n^{-1}$, equivalent to $2^m \geq c'' n^{-2s/(2s+1)}$ for some $c'' > 0$, which is satisfied since $2^m \geq c'\rho_n^2 \geq c'' n^{-2s/(2s+1)}$ if L_0 is large enough, by hypothesis on ρ_n. We can thus rewrite the last probability in (28) as

$$\sum_{m \in \mathbb{Z}: c'\rho_n^2 \leq 2^m \leq C} \Pr_f \left\{ n\|P_n - P\|_{\mathcal{G}_{m,J_n}} - nE\|P_n - P\|_{\mathcal{G}_{m,J_n}} > c_0 n 2^m \right\}.$$

To this expression we can apply Talagrand's inequality (13), noting that the supremum over $\mathcal{G}_{m,J_n}$ can be realised, by continuity, as one over a countable subset of Σ, and since Σ is uniformly bounded by $\sup_{f \in \Sigma(t)} \|f\|_\infty \leq U \equiv U(t,B)$. Renormalising by U and using (13), (27), (29) we can bound the expression in the last display, up to multiplicative constants, by

$$\begin{aligned} \sum_{m \in \mathbb{Z}: c'\rho_n^2 \leq 2^m \leq C} \exp\left\{ -c_1 \frac{n^2 (2^m)^2}{n2^m + nE\|P_n - P\|_{\mathcal{G}_{m,J_n}} + n2^m} \right\} &\leq \sum_{m \in \mathbb{Z}: c'\rho_n^2 \leq 2^m \leq C} e^{-c_2 n 2^m} \\ &\leq c_3 e^{-c_4 n \rho_n^2} \end{aligned}$$

since $2^m \geq c'\rho_n^2 >> n^{-1}$, which completes the proof. □

4.3 Proof of Theorem 2

Proof. We construct a standard Lepski type estimator: Choose integers $j_{\min}, j_{\max}$ such that $J_0 \leq j_{\min} < j_{\max}$,

$$2^{j_{\min}} \simeq n^{1/(2R+1)} \quad \text{and} \quad 2^{j_{\max}} \simeq n^{1/(2r+1)}$$

and define the grid

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

Let $f_n(j) \equiv f_n(j, \cdot) = \int_0^1 K_j(\cdot, y) dP_n(y)$ be a linear wavelet estimator based on wavelets of regularity $S > R$. To simplify the exposition we assume here that $\|f\|_\infty$ (or U) is known, otherwise the result follows from the same proof, with $\|f\|_\infty$ replaced by $\|f_n(j_{\max})\|_\infty$, a consistent estimator that satisfies sufficiently tight exponential error bounds (cf., e.g., [15]), and whose accuracy depends on U. Set

$$\bar{j}_n = \min\left\{ j \in \mathcal{J} : \|f_n(j) - f_n(l)\|_2^2 \leq C(S)(\|f\|_\infty \vee 1)\frac{2^l}{n} \quad \forall l > j, l \in \mathcal{J} \right\} \tag{30}$$

where $C(S)$ is a large enough constant, to be chosen below, in dependence of the wavelet basis. The adaptive estimator is $\hat{f}_n = f_n(\bar{j}_n)$. We shall need the standard estimates

$$E\|f_n(j) - Ef_n(j)\|_2^2 \leq D\frac{2^j}{n} := D\sigma^2(j,n) \tag{31}$$

and, for $f \in W^s, s \in [r, R]$,

$$\|Ef_n(j) - f\|_2 \leq 2^{-js} D' \|f\|_{s,2} := B(j, f) \tag{32}$$

for constants D, D' that depend only on the wavelet basis and on r, R. Define $j^* := j^*(f)$ by

$$j^* = \min\left\{ j \in \mathcal{J} : B(j,f) \leq \sqrt{D}\sigma(j,n) \right\}$$

so that, for every $f \in \Sigma(s, B)$ and $D'' = D''(D, D')$

$$D^{-1}B^2(j^*, f) \leq \sigma^2(j^*, n) \leq D''\|f\|_{s,2}^{2/(2s+1)} n^{-2s/(2s+1)} \leq D'' B^{2/(2s+1)} n^{-2s/(2s+1)}. \tag{33}$$

We will consider the cases $\{\bar{j}_n \leq j^*\}$ and $\{\bar{j}_n > j^*\}$ separately. First, by the definition of $\bar{j}_n, j^*$ and (31), (32), (33),

$$\begin{aligned} E\,\|f_n(\bar{j}_n) - f\|_2^2\, I_{\{\bar{j}_n \leq j^*\}} &= E\left(\|f_n(\bar{j}_n) - f_n(j^*)\|_2^2 + E\|f_n(j^*) - f\|_2^2 \right) I_{\{\bar{j}_n \leq j^*\}} \\ &\leq C(S)(\|f\|_\infty \vee 1)\frac{2^{j^*}}{n} + C'\sigma^2(j^*, n) \leq C'' B^{2/(2s+1)} n^{-2s/2s+1} \end{aligned}$$

which is the desired bound. On the event $\{\bar{j}_n > j^*\}$ we have, using (31) and the definition of j^*,

$$\begin{aligned} E\|f_n(\bar{j}_n) - f\|_2 I_{\{\hat{j}_n > j^*\}} &\leq \sum_{j \in \mathcal{J}: j > j^*} \left(E\|f_n(j) - f\|_2^2\right)^{1/2} \left(EI_{\{\hat{j}_n = j\}}\right)^{1/2} \\ &\leq \sum_{j \in \mathcal{J}: j > j^*} C''' \sigma(j,n) \cdot \sqrt{\Pr_f\{\hat{j}_n = j\}} \\ &\leq C'''' \sum_{j \in \mathcal{J}: j > j^*} \sqrt{\Pr_f\{\hat{j}_n = j\}} \end{aligned}$$

since $\sigma(j_{\max}, n)$ is bounded in n. Now pick any $j \in \mathcal{J}$ so that $j > j^*$ and denote by j^- the previous element in the grid (i.e. $j^- = j - 1$). One has, by definition of $\bar{j}_n$,

$$\Pr_f\{\bar{j}_n = j\} \leq \sum_{l \in \mathcal{J}: l \geq j} \Pr_f \left\{ \left\|f_n(j^-) - f_n(l)\right\|_2 > \sqrt{C(S)(\|f\|_\infty \vee 1)\frac{2^l}{n}} \right\}, \tag{34}$$

and we observe that, by the triangle inequality,

$$\left\|f_n(j^-) - f_n(l)\right\|_2 \leq \left\|f_n(j^-) - f_n(l) - Ef_n(j^-) + Ef_n(l)\right\|_2 + B(j^-, f) + B(l, f),$$

where,

$$B(j^-, f) + B(l, f) \leq 2B(j^*, f) \leq c\sigma(j^*, n) \leq c'\sigma(l, n)$$

by definition of j^* and since $l > j^- \geq j^*$. Consequently, the probability in (34) is bounded by

$$\Pr_f \left(\left\|f_n(j^-) - f_n(l) - Ef_n(j^-) + Ef_n(l)\right\|_2 > \left(\sqrt{C(S)(\|f\|_\infty \vee 1)} - c'\right)\sigma(l, n) \right), \tag{35}$$

and by inequality (14) above this probability is bounded by a constant multiple of e^{-2^l} if we choose $C(S)$ large enough. This gives the overall bound

$$\sum_{l \in \mathcal{J}: l \geq j} c'' e^{-d2^l} \leq d' e^{-d'' 2^{j_{\min}}},$$

which is smaller than a constant multiple times $B^{1/(2s+1)} n^{-s/(2s+1)}$, uniformly in $s \in [r, R], n \in \mathbb{N}$ and for $B \geq 1$, by definition of $j_{\min}$. This completes the proof. □

4.4 Proof of Theorem 3

Proof. Suppose for simplicity that the sample size is $2n$, and split the sample into two halves with index sets $\mathcal{S}^1, \mathcal{S}^2$, of equal size n, write E_1, E_2 for the corresponding expectations, and $E = E_1 E_2$. Let $\hat{f}_n = f_n(\bar{j}_n)$ be the adaptive estimator from the proof of Theorem 2 based on the sample $\mathcal{S}^1$. One shows by a standard bias-variance decomposition, for every $f \in W^r$, using $\bar{j}_n \in \mathcal{J}$ and $\|K_j(f)\|_{r,2} \leq \|f\|_{r,2}$ since K_j is a projection operator, that for every $\varepsilon > 0$ there exists a finite positive constant $B' = B'(\varepsilon, B)$ satisfying

$$\inf_{f \in \Sigma(r,B)} \Pr_f(\|\hat{f}_n\|_{r,2} \le B') \ge 1 - \varepsilon.$$

It therefore suffices to prove the theorem on the event $\{\|\hat{f}_n\|_{r,2} \le B'\}$. For a wavelet basis of regularity $S > R$ and for $J_n \ge J_0$ a sequence of integers such that $2^{J_n} \simeq n^{1/(2r+1/2)}$, define the U-statistic

$$U_n(\hat{f}_n) = \frac{2}{n(n-1)} \sum_{i<j,i,j \in \mathcal{S}^2} \sum_{l=J_0}^{J_n-1} \sum_{k \in \mathcal{Z}_l} (\psi_{lk}(X_i) - \langle \psi_{lk}, \hat{f}_n \rangle)(\psi_{lk}(X_j) - \langle \psi_{lk}, \hat{f}_n \rangle) \quad (36)$$

which has expectation

$$E_2 U_n(\hat{f}_n) = \sum_{l=J_0}^{J_n-1} \sum_{k \in \mathcal{Z}_l} \langle \psi_{lk}, f - \hat{f}_n \rangle^2 = \|\Pi_{V_{J_n}}(f - \hat{f}_n)\|_2^2.$$

Using Chebychev's inequality and that, by definition of the norm (5)

$$\sup_{h \in \Sigma(r,b)} \|\Pi_{V_{J_n}}(h) - h\|_2^2 \le c(b) 2^{-2J_n r}$$

for every $0 < b < \infty$ and some finite constant $c(b)$, we deduce

$$\begin{aligned}
&\inf_{f \in \Sigma(r,B)} \Pr_{f,2} \left\{ U_n(\hat{f}_n) - \|f - \hat{f}_n\|_2^2 \ge -(c(B) + c(B'))2^{-2J_n r} - z(\alpha)\tau_n(f) \right\} \\
&\ge \inf_{f \in \Sigma(r,B)} \Pr_{f,2} \left\{ U_n(\hat{f}_n) - \|\Pi_{V_{J_n}}(f - \hat{f}_n)\|_2^2 \ge -z(\alpha)\tau_n(f) \right\} \\
&\ge 1 - \sup_{f \in \Sigma(r,B)} \frac{Var_2(U_n(\hat{f}_n) - E_2 U_n(\hat{f}_n))}{(z(\alpha)\tau_n(f))^2}.
\end{aligned}$$

We now show that the last quantity is greater than or equal to $1 - z(\alpha)^{-2} \ge 1 - \alpha$ for quantile constants $z(\alpha)$ and with

$$\tau_n^2(f) = \frac{C(S) 2^{J_n} \|f\|_\infty^2}{n(n-1)} + \frac{4\|f\|_\infty}{n} \|\Pi_{V_{J_n}}(f - \hat{f}_n)\|^2,$$

which in turn gives the honest confidence set under Pr

$$C_n(\|f\|_\infty, B) = \left\{ f \in \Sigma(r,B) : \|f - \hat{f}_n\|_2 \le \sqrt{z_\alpha \tau_n(f) + U_n(\hat{f}_n) + (c(B) + c(B'))2^{-2J_n r}} \right\}. \quad (37)$$

We shall comment on the role of the constants $\|f\|_\infty, c(B)$ at the end of the proof. To establish the last claim, note that the Hoeffding decomposition for the centered U-statistic with kernel

$$R(x,y) = \sum_{l=J_0}^{J_n-1} \sum_{k \in \mathcal{Z}_l} (\psi_{lk}(x) - \langle \psi_{lk}, \hat{f}_n \rangle)(\psi_{lk}(y) - \langle \psi_{lk}, \hat{f}_n \rangle)$$

is (cf. the proof of Theorem 4.1 in [30])

$$U_n(\hat{f}_n) - E_2 U_n(\hat{f}_n) = \frac{2}{n}\sum_{i=1}^{n}(\pi_1 R)(X_i) + \frac{2}{n(n-1)}\sum_{i<j}(\pi_2 R)(X_i, X_j) \equiv L_n + D_n$$

where

$$(\pi_1 R)(x) = \sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}(\psi_{lk}(x) - \langle\psi_{lk}, f\rangle)\langle\psi_{lk}, f - \hat{f}_n\rangle$$

and

$$(\pi_2 R)(x,y) = \sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}(\psi_{lk}(x) - \langle\psi_{lk}, f\rangle)(\psi_{lk}(y) - \langle\psi_{lk}, f\rangle)$$

The variance of $U_n(\hat{f}_n) - E_2 U_n(\hat{f}_n)$ is the sum of the variances of the two terms in the Hoeffding decomposition. For the linear term we bound the variance $Var_2(L_n)$ by the second moment, using orthonormality of the ψ_{lk}s,

$$\frac{4}{n}\int\left(\sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}\psi_{lk}(x)\langle\psi_{lk}, \hat{f}_n - f\rangle\right)^2 f(x)dx \le \frac{4\|f\|_\infty}{n}\sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}\langle\psi_{lk}, \hat{f}_n - f\rangle^2,$$

which equals the second term in the definition of $\tau_n^2(f)$. For the degenerate term we can bound $Var_2(D_n)$ analogously by the second moment of the uncentered kernel (cf. after (19)), i.e., by

$$\frac{2}{n(n-1)}\int\left(\sum_{l=J_0}^{J_n-1}\sum_{k\in\mathcal{Z}_l}\psi_{lk}(x)\psi_{lk}(y)\right)^2 f(x)dxf(y)dy \le \frac{C(S)2^{J_n}\|f\|_\infty^2}{n(n-1)},$$

using orthonormality and the cardinality properties of $\mathcal{Z}_l$.

The so constructed confidence set has an adaptive expected maximal diameter: Let $f \in \Sigma(s,B)$ for some $s \in [r,R]$ and some $B \ge 1$. The nonrandom terms are of order

$$\sqrt{c(B)+c(B')}2^{-J_n r} + \|f\|_\infty^{1/2} 2^{J_n/4} n^{-1/2} \le C(S,B,r,B')n^{-r/(2r+1/2)}$$

which is $o(n^{-s/(2s+1)})$ since $s \le R < 2r$. The random component of $\tau_n(f)$ has order $\|f\|_\infty^{1/4} n^{-1/4} E_1\|\Pi_{V_{J_n}}(\hat{f}_n - f)\|_2^{1/2}$ which is also $o(n^{-s/(2s+1)})$ for $s < 2r$, since $\Pi_{V_{J_n}}$ is a projection operator and in view of adaptivity of $\hat{f}_n$ established in Theorem 2. Moreover, by Theorem 2 and again the projection properties,

$$EU_n(\hat{f}_n) = E_1\|\Pi_{V_{J_n}}(\hat{f}_n - f)\|_2^2 \le E_1\|\hat{f}_n - f\|_2^2 \le cB^{2/(2s+1)}n^{-2s/(2s+1)},$$

where c depends on $\|f\|_\infty$. The term in the last display is the leading term in our bound for the diameter of the confidence set, and shows that C_n adapts to both B and s in the sense of Definition 1, using Markov's inequality.

The confidence set $C_n(\|f\|_\infty, B)$ is not feasible if B is unknown, so in particular under the assumptions of Theorem 3, but C_n independent of $\|f\|_\infty, B$ can be constructed as follows: If B is not known, we replace $c(B)+c(B')$ in the definition of (37) by a divergent sequence of positive real numbers c_n, which can still be accommodated in the diameter estimate from the last paragraph since $n^{-2r/(2r+1/2)}c_n$ is still $o(n^{-2s/(2s+1)})$ as long as $s \le R < 2r$ for c_n diverging slowly enough (e.g., like $\log n$). Define thus the confidence set

$$C_n = \left\{ f \in \Sigma(r) : \|f - \hat{f}_n\|_2 \le \sqrt{z_\alpha \tau_n(f) + U_n(\hat{f}_n) + c_n 2^{-2Jr}} \right\}, \tag{38}$$

with $\|f\|_\infty$ replaced by $\|f_n(j_{\max})\|_\infty$ (as at the beginning of the proof of Theorem 2) in all expressions where $\|f\|_\infty$ occurs. The 'honest' part of Theorem 3 then holds for C_n, using that $\|f_n(j_{\max})\|_\infty$ concentrates around $\|f\|_\infty$ (arguing as in [15]), with accuracy of concentration depending on $\|f\|_\infty \le U$. For the dishonest part let f be such that $\|f\|_{s,2} < \infty$, and apply the 'honest' result with $B = \|f\|_{s,2}$. □

4.5 Proof of Theorem 1

Proof. That an L^2-adaptive confidence set exists when $s \le 2r$ follows from Theorem 3; The case $s < 2r$ is immediate, and the case $s = 2r$ follows using the confidence set (37). This set is feasible since B and thus also an upper bound for $\|f\|_\infty$ (cf. (20)) is known under the hypothesis of the theorem, and adaptive since $n^{-r/(2r+1/2)} = n^{-s/(2s+1)}$ for $s = 2r$.

For part Aii we use the test Ψ_n from Proposition 1 with $\Sigma = \Sigma(s), t = r$, – cf. also the remark after Definition 2 – and define a confidence ball as follows. Take $\hat{f}_n = f_n(\bar{j}_n)$ to be the adaptive estimator from the proof of Theorem 2, and let, for $0 < L' < \infty$,

$$C_n = \begin{cases} \{f \in \Sigma(r) : \|f - \hat{f}_n\|_2 \le L' n^{-s/(2s+1)}\} & \text{if } \Psi_n = 0 \\ \{f \in \Sigma(r) : \|f - \hat{f}_n\|_2 \le L' n^{-r/(2r+1)}\} & \text{if } \Psi_n = 1 \end{cases}$$

We first prove that C_n is honest for $\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$ if we choose L' large enough. For $f \in \Sigma(s)$ we have from Theorem 2, by Markov's inequality,

$$\begin{aligned} \inf_{f \in \Sigma(s)} \Pr{}_f(f \in C_n) &\ge 1 - \sup_{f \in \Sigma(s)} \Pr{}_f\left(\|\hat{f}_n - f\|_2 > L' n^{-s/(2s+1)}\right) \\ &\ge 1 - \frac{n^{s/(2s+1)}}{L'} \sup_{f \in \Sigma(s)} E_f \|\hat{f}_n - f\|_2 \\ &\ge 1 - \frac{c(B, s, r)}{L'} \end{aligned}$$

which can be made greater than $1 - \alpha$ for any $\alpha > 0$ by choosing L' large enough depending only on B, α, r, s. When $f \in \tilde{\Sigma}(r, \rho_n)$, using again Markov's inequality

$$\inf_{f \in \tilde{\Sigma}(r, \rho_n)} \Pr{}_f(f \in C_n) \ge 1 - \frac{\sup_{f \in \Sigma(r)} E_f \|\hat{f}_n - f\|_2}{L' n^{-r/(2r+1)}} - \sup_{f \in \tilde{\Sigma}(r, \rho_n)} \Pr{}_f(\Psi_n = 0).$$

The first subtracted term can be made smaller than $\alpha/2$ for L' large enough as before. The second subtracted term can also be made less than $\alpha/2$ in view of Proposition 1 for suitable M, d_n large enough but bounded in n. This proves that C_n is honest. We now turn to adaptivity of C_n: By the definition of C_n we always have $|C_n| \leq L'n^{-r/(2r+1)}$, so the case $f \in \tilde{\Sigma}(r, \rho_n)$ is proved. If $f \in \Sigma(s)$ then using Proposition 1 again, for M, d_n large enough depending on α' but bounded in n,

$$\Pr_f(|C_n| > L'n^{-s/(2s+1)}) = \Pr_f(\Psi_n = 1) \leq \alpha',$$

which completes the proof of part A.

To prove part B of Theorem 1 we argue by contradiction and assume that the limit inferior equals zero. We then pass to a subsequence of n for which the limit is zero, and still denote this subsequence by n. Let $f_0 \equiv 1 \in \Sigma(s)$, suppose C_n is adaptive and honest for $\Sigma(s) \cup \tilde{\Sigma}(r, \rho_n)$ for every α, α', and consider testing

$$H_0 : f = f_0 \quad \text{against} \quad H_1 : f \in \tilde{\Sigma}(r, \rho_n)$$

where $\rho_n \leq \epsilon n^{-r/(2r+1/2)}$ for any $\epsilon > 0$, and every $n \geq n(\epsilon)$ large enough. Since $s > 2r$ we can find a sequence ρ_n' such that $n^{-s/(2s+1)} = o(\rho_n') = o(\rho_n)$ for every $\epsilon > 0$. Accept H_0 if $C_n \cap \tilde{\Sigma}(r, \rho_n)$ is empty and reject otherwise, formally

$$\Psi_n = 1\{C_n \cap \tilde{\Sigma}(r, \rho_n) \neq \emptyset\}.$$

The type-one errors of this test satisfy

$$\begin{aligned} E_{f_0}\Psi_n &= \Pr_{f_0}\left\{C_n \cap \tilde{\Sigma}(r, \rho_n) \neq \emptyset\right\} \\ &\leq \Pr_{f_0}\{f_0 \in C_n, |C_n| \geq \rho_n'\} + \Pr_{f_0}\{f_0 \notin C_n\} \\ &\leq \alpha + \alpha' + r_n \to \alpha + \alpha' \end{aligned}$$

as $n \to \infty$ by the hypothesis of coverage and adaptivity of C_n. The type-two errors satisfy, by coverage of C_n, as $n \to \infty$

$$E_f(1 - \Psi_n) = \Pr_f\{C_n \cap \tilde{\Sigma}(r, \rho_n) = \emptyset\} \leq \Pr_f\{f \notin C_n\} \leq \alpha + r_n \to \alpha,$$

uniformly in $f \in \tilde{\Sigma}(r, \rho_n)$. We conclude that this test satisfies

$$\limsup_n \left[E_{f_0}\Psi_n + \sup_{f \in H_1} E_f(1 - \Psi_n)\right] \leq 2\alpha + \alpha'$$

for arbitrary $\alpha, \alpha' > 0$. Since balls in $W^s = B^s_{2\infty}([0,1])$ contain Hölder balls we have a contradiction to Theorem 1i in [20] which states that the limit inferior of the term in brackets in the last display, even with an infimum over all tests, exceeds a fixed positive constant. Conclude that for such a sequence of ρ_n an adaptive and honest confidence set cannot exist. □

4.6 Proof of Theorem 6

Let us use the shorthand $\tilde{\Sigma}_n(s)$ for $\tilde{\Sigma}(s, \rho_n(s))$. For $i = 1, \dots, N$, let $\Psi(i)$ be the test from (17) with $\Sigma = \Sigma(s_{i+1})$ and $t = s_i$, cf. also the remark after Definition 2. Starting from the largest model we first test $H_0 : f \in \Sigma(s_2)$ against $H_1 : f \in \tilde{\Sigma}_n(s_1)$, accepting H_0 if $\Psi(1) = 0$. If H_0 is rejected we set $\hat{s}_n = s_1 = r$, otherwise we proceed to test $H_0 : f \in \Sigma(s_3)$ against $H_1 : f \in \tilde{\Sigma}_n(s_2)$ using $\Psi(2)$ and iterating this procedure downwards we define $\hat{s}_n$ to be the first element s_i in $\mathcal{S}$ for which $\Psi(i) = 1$ rejects. If no rejection occurs we set $\hat{s}_n$ equal to s_N, the last element in the grid.

For $f \in \mathcal{P}_n(M, \mathcal{S})$ define the unique $s_{i_0} := s_{i_0}(f) = \{s \in \mathcal{S} : f \in \tilde{\Sigma}_n(s)\}$. We now show that for M large enough

$$\sup_{f \in \mathcal{P}_n(M,\mathcal{S})} \Pr_f(\hat{s}_n \neq s_{i_0}(f)) < \max(\alpha, \alpha')/2. \tag{39}$$

Indeed, if $\hat{s}_n < s_{i_0}$ then the test $\Psi(i)$ has rejected for some $i < i_0$. In this case $f \in \tilde{\Sigma}_n(s_{i_0}) \subset \Sigma(s_{i_0}) \subseteq \Sigma(s_{i+1})$ for every $i < i_0$, and thus,

$$\begin{aligned} \Pr_f(\hat{s}_n < s_{i_0}) &= \Pr_f\left(\bigcup_{i<i_0} \{\Psi(i) = 1\}\right) \leq \sum_{i<i_0} \sup_{f \in \Sigma(s_{i+1})} E_f \Psi(i) \\ &\leq C(N) e^{-cd_n^2} < \max(\alpha, \alpha')/2 \end{aligned}$$

by Proposition 1 for constants M, d_n large enough but bounded in n. On the other hand if $\hat{s}_n > s_{i_0}$ (ignoring the trivial case $s_{i_0} = s_N$) then $\Psi(i_0)$ has accepted despite $f \in \tilde{\Sigma}_n(s_{i_0})$. Thus

$$\Pr_f(\hat{s}_n > s_{i_0}) \leq \sup_{f \in \tilde{\Sigma}_n(s_{i_0})} E_f(1 - \Psi(i_0)) \leq C e^{-cd_n^2} \leq \max(\alpha, \alpha')/2$$

again by Proposition 1, for M, d_n large enough.

Denote now by $C_n(s_i)$ the confidence set (37) constructed in the proof of Theorem 3 with r there being s_i and $R = 2s_i = s_{i+1}$ such that the asymptotic coverage level is $\alpha/2$ for any $f \in \Sigma(s_i)$, and set $C_n = C_n(\hat{s}_n)$. We use the knowledge of B and thus an upper bound for $\|f\|_\infty$ under the hypotheses of the theorem. We then have, from the proof of Theorem 3, for $f \in \tilde{\Sigma}_n(s_{i_0}) \subset \Sigma(s_{i_0})$,

$$\Pr_f(f \in C_n(\hat{s}_n)) \geq \Pr_f(f \in C_n(s_{i_0})) - \alpha/2 \geq 1 - \alpha.$$

Moreover, if $f \in \Sigma(s)$ for $s \in [s_{i_0}, s_{i_0+1})$ and also for $s \in [s_N, R]$ if $s_{i_0} = s_N$, the expected diameter of C_n satisfies, by the estimates in the proof of Theorem 3

$$\Pr_f(|C_n(\hat{s}_n)| > C n^{-s/(2s+1)}) \leq \Pr_f(|C_n(s_{i_0})| > C n^{-s/(2s+1)}) + \alpha'/2 < \alpha'$$

for C large enough, so that this band is adaptive as well.

4.7 Proof of Theorem 7

Proof. Suppose such C_n exists. For $m = 0, 1, 2, \dots, \infty$, we will construct functions $f_0 = 1$,

$$f_m = 1 + \varepsilon \sum_{i=1}^{m} \sum_{k \in \mathcal{Z}_{j_i}} 2^{-j_i(r+1/2)} \beta_{ik} \psi_{j_i k}.$$

We will choose $j_1, j_2, \dots \in \mathbb{N}$ satisfying $j_i/j_{i-1} \geq 1 + 1/2r$, and $\beta_{ik} = \pm 1$ at random. Pick $\varepsilon > 0$ small enough that $\|f_m - f_{m-1}\|_\infty \leq 2^{-(m+1)}$ for all $m < \infty$, and any choice of j_i, β_{ik}. Then

$$f_m = 1 + \sum_{i=1}^{m} (f_i - f_{i-1}) \geq \tfrac{1}{2},$$

and $\int f_m = \langle 1, f_m \rangle = 1$, so the f_m are densities. By (5), $f_m \in W^r$, and for $m < \infty$, also $f_m \in W^s$. We will further choose a subsequence n_m so that, for $\delta = \frac{1}{5}(1 - 2\alpha)$,

$$\sup_m \Pr_{f_\infty}(f_\infty \in C_{n_m}) \leq 1 - \alpha - \delta,$$

contradicting our assumptions on C_n.

Inductively, suppose we have defined f_{m-1}, n_{m-1}. For $n_m > n_{m-1}$ and $D > 0$ large, we have:

1. $\Pr_{f_{m-1}}(f_{m-1} \notin C_{n_m}) \leq \alpha + \delta$; and
2. $\Pr_{f_{m-1}}(|C_{n_m}| \geq D r_{n_m}) \leq \delta$.

Setting $T_n = 1(\exists\, f \in C_n : \|f - f_{m-1}\|_2 \geq 2 D r_n)$, we then have

$$\Pr_{f_{m-1}}(T_{n_m} = 1) \leq \Pr_{f_{m-1}}(f_{m-1} \notin C_{n_m}) + \Pr_{f_{m-1}}(|C_{n_m}| \geq D r_{n_m}) \leq \alpha + 2\delta. \tag{40}$$

We claim it is possible to choose j_m, β_{mk} and n_m so that also, for $m > 1$,

$$3 D r_{n_m} \leq \|f_m - f_{m-1}\|_2 \leq \tfrac{1}{4} \|f_{m-1} - f_{m-2}\|_2, \tag{41}$$

and for any further choice of j_i, β_{ik},

$$\Pr_{f_\infty}(T_{n_m} = 0) \geq 1 - \alpha - 4\delta. \tag{42}$$

We may then conclude that, since all further choices will satisfy (41),

$$\|f_\infty - f_{m-1}\|_2 \geq \|f_m - f_{m-1}\|_2 - \sum_{i=m+1}^{\infty} \|f_i - f_{i-1}\|_2 \geq 2 D r_{n_m},$$

so

$$\Pr_{f_\infty}(f_\infty \in C_{n_m}) \leq \Pr_{f_\infty}(T_{n_m} = 1) \leq \alpha + 4\delta = 1 - \alpha - \delta$$

as required.

It remains to verify the claim. For $j \geq (1+1/2r)j_{m-1}$, $\beta_k = \pm 1$, set

$$g_\beta = \varepsilon 2^{-j(r+1/2)} \sum_{k \in \mathcal{Z}_j} \beta_k \psi_{jk},$$

and $f_\beta = f_{m-1} + g_\beta$. Allowing $j \to \infty$, set $n \sim C 2^{j(2r+1/2)}$, for $C > 0$ to be determined. Then

$$\|g_\beta\|_2 = \varepsilon 2^{-jr} \approx n^{-r/(2r+1/2)},$$

so for j large enough, f_β satisfies (41) with any choice of β.

The density of $X_1, \ldots, X_n$ under f_β, w.r.t. under f_{m-1}, is

$$Z_\beta = \prod_{i=1}^n \frac{f_\beta}{f_{m-1}}(X_i).$$

Set $Z = 2^{-j} \sum_\beta Z_\beta$, so $E_{f_{m-1}}[Z] = 1$, and

$$\begin{aligned}
E_{f_{m-1}}[Z^2] &= 2^{-2j} \sum_{\beta,\beta'} \prod_{i=1}^n E_{f_{m-1}} \left[\frac{f_\beta f_{\beta'}}{f_{m-1}^2}(X_i) \right] \\
&= 2^{-2j} \sum_{\beta,\beta'} \left\langle \frac{f_\beta}{\sqrt{f_{m-1}}}, \frac{f_{\beta'}}{\sqrt{f_{m-1}}} \right\rangle^n \\
&= 2^{-2j} \sum_{\beta,\beta'} \left(1 + \left\langle \frac{g_\beta}{\sqrt{f_{m-1}}}, \frac{g_{\beta'}}{\sqrt{f_{m-1}}} \right\rangle \right)^n \\
&\leq 2^{-2j} \sum_{\beta,\beta'} (1 + 2\langle \beta, \beta' \rangle)^n \\
&= E[(1 + \varepsilon^2 2^{1-j(2r+1)} Y)^n],
\end{aligned}$$

where $Y = \sum_{i=1}^{2^j} R_i$, for i.i.d. Rademacher random variables R_i,

$$\begin{aligned}
&\leq E[\exp(n\varepsilon^2 2^{1-j(2r+1)} Y)] \\
&= \cosh \left(D 2^{-j/2} (1 + o(1)) \right)^{2^j},
\end{aligned}$$

as $j \to \infty$, for some $D > 0$,

$$\begin{aligned}
&= \left(1 + D^2 2^{-j} (1 + o(1)) \right)^{2^j} \\
&\leq \exp \left(D^2 (1 + o(1)) \right) \\
&\leq 1 + \delta^2,
\end{aligned}$$

for j large, C small. Hence $E_{f_{m-1}}[(Z-1)^2] \leq \delta^2$, and we obtain

$$\begin{aligned}
\Pr_{f_{m-1}}(T_n = 1) + \max_\beta \Pr_{f_\beta}(T_n = 0) &\geq \Pr_{f_{m-1}}(T_n = 1) + 2^{-j} \sum_\beta \Pr_{f_\beta}(T_n = 0) \\
&= 1 + E_{f_{m-1}}[(Z-1) 1(T_n = 0)] \\
&\geq 1 - \delta.
\end{aligned}$$

Set $f_m = f_\beta$, for β maximizing this expression. The density of $X_1, \ldots, X_n$ under f_∞, w.r.t. under f_m, is

$$Z' = \prod_{i=1}^{n} \frac{f_\infty}{f_m}(X_i).$$

Now, $E_{f_m}[Z'] = 1$, and

$$\|f_\infty - f_m\|_2^2 = \sum_{i=m+1}^{\infty} \varepsilon^2 2^{-2j_i r} \leq E' 2^{-2j_{m+1} r} \leq E' 2^{-j(2r+1)},$$

for some constant $E' > 0$, so similarly

$$\begin{aligned} E_{f_m}[Z'^2] &\leq (1 + 2\|f_\infty - f_m\|_2^2)^n \\ &\leq (1 + E' 2^{1-j(2r+1)})^n \\ &\leq \exp(E' n 2^{1-j(2r+1)}) \\ &= \exp\left(F 2^{-j/2}(1 + o(1))\right), \end{aligned}$$

for some $F > 0$,

$$\leq 1 + \delta^2,$$

for j large. Hence $E_{f_m}[(Z' - 1)^2] \leq \delta^2$, and

$$\begin{aligned} \Pr_{f_{m-1}}(T_n = 1) + \Pr_{f_\infty}(T_n = 0) &= \Pr_{f_{m-1}}(T_n = 1) + E_{f_m}[Z' 1(T_n = 0)] \\ &\geq 1 - \delta + E_{f_m}[(Z' - 1) 1(T_n = 0)] \\ &\geq 1 - 2\delta. \end{aligned}$$

If we take $j_m = j$, $n_m = n$ large enough also that (40) holds, then f_∞ satisfies (42), and our claim is proved. □

4.8 Proof for Theorem 8

Proof. Part b) follows from the proof of Theorem 3, as the confidence set (37) can be used also for $R = 2r$ when B is known. Part c) is proved as follows: We use the standard estimate

$$E\left(\|f_n(j)\|_{r,2}^2 - \|Ef_n(j)\|_{r,2}^2\right)^2 \leq C(2^{(2r+1/2)j}/n)^2,$$

for some constant $C > 0$. Then for any $2^{J_n'} = o(n^{1/(2r+1/2)})$, by Chebyshev's inequality $\|f_n(J_n')\|_{r,2}$ is a consistent estimate of $\|Ef_n(J_n')\|_{r,2}$. Fix $\delta > 0$, and set

$$\hat{B}_n := (1 + \delta)\|f_n(J_n')\|_{r,2}.$$

If also $J_n' \to \infty$, then for n large enough depending on f, with high probability,

$$\|f\|_{r,2} \leq \hat{B}_n \leq (1 + 2\delta)\|f\|_{r,2},$$

and we may thus replace B with $\hat{B}_n$ in the proof of Part b). C_n is then a dishonest confidence set for the full model $\mathcal{P}$.

To prove part a), suppose C_n exists. Set $f_0 = 1$, and

$$f_1 = 1 + B2^{-j(r+1/2)} \sum_{k \in \mathcal{Z}_j} \beta_{jk}\psi_{jk},$$

for $B > 0$, $j > j_0$, and $\beta_{jk} = \pm 1$ to be determined. Having chosen B, we will pick j large enough that $f_1 \geq \frac{1}{2}$. Since $\int f_1 = \langle f_1, 1 \rangle = 1$, f_1 is then a density.

Set $\delta = \frac{1}{4}(1 - 2\alpha)$. As $f_0 \in \Sigma(R, 1)$, for n and L large we have:

1. $\mathrm{Pr}_{f_0}(f_0 \notin C_n) \leq \alpha + \delta$; and
2. $\mathrm{Pr}_{f_0}(|C_n| \geq Ln^{-R/(2R+1)}) \leq \delta$.

Setting $T_n = 1(\exists\, f \in C_n : \|f - f_0\|_2 \geq 2Ln^{-R/(2R+1)})$, we then have

$$\mathrm{Pr}_{f_0}(T_n = 1) \leq \alpha + 2\delta,$$

as in the proof of Theorem 7.

For a constant $C = C(\delta) > 0$ to be determined, set $B = (3L)^{2R+1}C^{-R}$. Allowing $j \to \infty$, set $n \sim CB^{-2}2^{j(R+1/2)}$. Then

$$\|f_1 - f_0\|_2 = B2^{-jr} \simeq 3Ln^{-R/(2R+1)},$$

so for j large, $\|f_1 - f_0\|_2 \geq 2Ln^{-R/(2R+1)}$. Arguing as in the proof of Theorem 7, the density Z of f_1 w.r.t. f_0 has second moment

$$\begin{aligned} E_{f_0}[Z^2] &\leq \cosh(nB^2 2^{1-j(2r+1)})^{2^j} \\ &= \cosh(C2^{1-j/2}(1 + o(1)))^{2^j} \\ &= (1 + C^2 2^{2-j}(1 + o(1)))^{2^j} \\ &\leq \exp(4C^2(1 + o(1))) \\ &\leq 1 + \delta^2, \end{aligned}$$

for $C(\delta)$ small, j large. Hence

$$\mathrm{Pr}_{f_0}(T_n = 1) + \max_{\beta} \mathrm{Pr}_{f_1}(T_n = 0) \geq 1 - \delta.$$

and for all j (and n) large enough, we obtain, for suitable β,

$$\mathrm{Pr}_{f_1}(f_1 \in C_n) \leq \mathrm{Pr}_{f_1}(T_n = 1) \leq \alpha + 3\delta = 1 - \alpha - \delta.$$

Since $f_1 \in \Sigma(r, B + 1)$ for all n, β_{jk} this contradicts the definition of C_n. □

References

[1] Y. Baraud. Confidence balls in Gaussian regression. *Ann. Statist.*, 32(2):528–551, 2004.

[2] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[3] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[4] O. Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 213–247. Birkhäuser, Basel, 2003.

[5] A.D. Bull. Honest adaptive confidence bands and selfsimilar functions. *preprint, available at arxiv.org*, 2011.

[6] T. T. Cai and M. G. Low. An adaptation theory for nonparametric confidence intervals. *Ann. Statist.*, 32(5):1805–1840, 2004.

[7] T. T. Cai and M. G. Low. Adaptive confidence balls. *Ann. Statist.*, 34(1):202–228, 2006.

[8] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.

[9] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[10] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995.

[11] S. Efromovich. Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Statist.*, 36(3):1127–1155, 2008.

[12] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216, 2006.

[13] E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

[14] E. Giné and R. Nickl. Confidence bands in density estimation. *Ann. Statist.*, 38:1122–1170, 2010a.

[15] E. Giné and R. Nickl. Adaptive estimation of the distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli*, 16:1137–1163, 2010b.

[16] E. Giné and R. Nickl. Rates of contraction for posterior distributions in l^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.*, 2011. to appear.

[17] M. Hoffmann and O.V. Lepski. Random rates in anisotropic regression. *Ann. Statist.*, 30(2):325–396, 2002. With discussions and a rejoinder by the authors.

[18] M. Hoffmann and R. Nickl. On adaptive inference and confidence bands. *Ann. Statist.*, 2011. to appear.

[19] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

[20] Yu. I. Ingster. A minimax test of nonparametric hypotheses on the density of a distribution in L_p metrics. *Teor. Veroyatnost. i Primenen.*, 31(2):384–389, 1986.

[21] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.*, 2(2):85–114, 1993.

[22] A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Math. Methods Statist.*, 12(4):410–428 (2004), 2003.

[23] G. Kerkyacharian, R. Nickl, and D. Picard. Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probability Theory and Related Fields*, 2011. to appear.

[24] O. V. Lepski. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990.

[25] O. V. Lepski. How to improve the accuracy of estimation. *Math. Methods Statist.*, 8(4):441–486 (2000), 1999.

[26] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997.

[27] G. G. Lorentz, M. v. Golitschek, and Y. Makovoz. *Constructive approximation.* Springer-Verlag, Berlin, 1996. Advanced problems.

[28] M. G. Low. On nonparametric confidence intervals. *Ann. Statist.*, 25(6):2547–2554, 1997.

[29] D. Picard and K. Tribouley. Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, 28(1):298–335, 2000.

[30] J. Robins and A.W. van der Vaart. Adaptive nonparametric confidence sets. *Ann. Statist.*, 34(1):229–253, 2006.

[31] V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498, 1996.

[32] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[33] S. A. van de Geer. *Applications of empirical process theory.* Cambridge, 2000.

[34] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.