Optimal Inference with a Multidimensional Multiscale Statistic

Pratyay (Ashley) Datta

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Optimal Inference with a Multidimensional Multiscale Statistic

Pratyay (Ashley) Datta

We observe a stochastic process Y on $[0,1]^d$ ($d \geq 1$) satisfying $dY(t) = n^{1/2} f(t)dt + dW(t)$, $t \in [0,1]^d$, where $n \geq 1$ is a given scale parameter ('sample size'), W is the standard Brownian sheet on $[0,1]^d$ and $f \in L_1([0,1]^d)$ is the unknown function of interest. We propose a multivariate multiscale statistic in this setting and prove that the statistic attains a subexponential tail bound; this extends the work of 'Dumbgen and Spokoiny (2001)' who proposed the analogous statistic for $d = 1$. In the process, we generalize Theorem 6.1 of 'Dumbgen and Spokoiny (2001)' about stochastic processes with sub-Gaussian increments on a pseudometric space, which is of independent interest. We use the proposed multiscale statistic to construct optimal tests (in an asymptotic minimax sense) for testing $f = 0$ versus (i) appropriate Hölder classes of functions, and (ii) alternatives of the form $f = \mu_n \mathbb{I}_{B_n}$, where B_n is an axis-aligned hyperrectangle in $[0,1]^d$ and $\mu_n \in \mathbb{R}$; μ_n and B_n unknown. In Chapter 3 we use this proposed multiscale statistics to construct honest confidence bands for multivariate shape-restricted regression including monotone and convex functions.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Prof. Lutz Dümbgen, Prof. Sumit Mukherjee, Prof. Rajarshi Mukherjee and Prof. Somabha Mukherjee for several helpful discussions. I would also like to thank Prof. Bodhisattva Sen and Prof. Tian Zheng for their guidance and patience during my studies at Columbia university.

Dedication

This is for my Mom and Dad, who have given me unending support through the my best and my worst.

Chapter 1: Multivariate Multiscale Statistics

1.1 Introduction

Let us consider the following continuous multidimensional white noise model:

$$Y(t) = \sqrt{n} \int_0^{t_1} \cdots \int_0^{t_d} f(s_1, \ldots, s_d)\, ds_d \ldots ds_1 + W(t), \tag{1.1}$$

where $t := (t_1, \ldots, t_d) \in [0,1]^d$, $d \geq 1$, $\{Y(t_1, \ldots, t_d) : (t_1, \ldots, t_d) \in [0,1]^d\}$ is the observed data, $f \in L_1([0,1]^d)$ is the unknown (regression) function of interest, $W(\cdot)$ is the unobserved d-dimensional Brownian sheet (see Definition A.1.1), and n is a known scale parameter. Estimation and inference in this model is closely related to that of (multivariate) nonparametric regression based on sample size n; see e.g., [1] and [2]. We work with this white noise model as this formulation is more amiable to rescaling arguments; see e.g., [3], [4], [5].

In this paper we develop *optimal* tests (in an asymptotic minimax sense) based on our proposed *multidimensional multiscale statistic* (i.e., $d \geq 1$) for testing:

(i) $f = 0$ versus a Hölder class of functions with unknown degree of smoothness;

(ii) $f = 0$ against alternatives of the form $f = \mu_n \mathbb{I}_{B_n}$, where B_n is an unknown hyperrectangle in $[0,1]^d$ with sides parallel to the coordinate axes (i.e., axis-aligned) and $\mu_n \in \mathbb{R}$ is unknown.

Scenario (i) arises quite often in nonparametric regression where the goal is to test whether the underlying f is 0 versus $f \neq 0$ with unknown smoothness; see e.g., [6], [7], [8], [9] and the references therein. Our proposed multiscale statistic, which extends the work of [4], that considered the analogous statistic for $d = 1$, leads to rate optimal detection in this problem under the uniform metric. Moreover, with the knowledge of the smoothness of the underlying f, we construct a *asymptotically minimax test* which even attains the exact separation constant (see Section 2.0.1 for formal definitions and related concepts).

Setting (ii) is a prototypical problem in signal detection — an unknown (constant) signal spread over an unknown hyperrectangular region — and the goal is to detect the presence of such a signal; see e.g., [10], [11, 12], [13], [14], [15], [16], [17], [18] for a plethora of examples and applications. Compared to the several minimax rate optimal tests that have been proposed in the literature for this problem (see e.g., [11], [13] and [18]), our proposed multiscale test leads to simultaneous optimal detection of signals both at small and large scales. It may be mentioned in this regard that [14] proposed a test that leads to optimal detection of hyperrectangles when the responses are Bernoulli variables. Also recently, [19], using a completely differently approach, proved minimax optimality over hyperrectangles in the general setting of inverse problems.

We first motivate and introduce our multiscale statistic below (Section 1.1.1) and briefly describe the asymptotic minimax testing framework. Our main optimality results are discussed in Section 2.0.1.

1.1.1 Multiscale statistic when $d \geq 1$

To motivate our multiscale statistic let us first look at the following testing problem:

$$H_0 : f = 0 \quad \text{versus} \quad H_1 : f \neq 0 \in \mathbb{H}_{\beta,L}, \tag{1.2}$$

where $\mathbb{H}_{\beta,L}$ is the Hölder class of function with parameters $\beta > 0$ and $L > 0$. For $\beta \in (0, 1]$ and $L > 0$ the Hölder class $\mathbb{H}_{\beta,L}$ is defined as

$$\mathbb{H}_{\beta,L} := \left\{ f \in L_1([0,1]^d) : |f(x) - f(y)| \leq L \, \|x - y\|^{\beta} \text{ for all } x, y \in [0,1]^d \right\}. \tag{1.3}$$

For $\beta > 1$ the Hölder class $\mathbb{H}_{\beta,L}$ is defined similarly; see Definition A.1.2.

Our multiscale statistic is based on the idea of *kernel averaging*. Suppose that $\psi : \mathbb{R}^d \to \mathbb{R}$ is a measurable function such that:

(i) ψ is 0 outside $[-1,1]^d$;

(ii) $\psi \in L_2(\mathbb{R}^d)$, i.e., $\int_{\mathbb{R}^d} \psi^2(x)dx < \infty$;

(iii) ψ is of bounded Hardy-Krause (HK)-variation (see Definition A.1.3 in the Appendix) and

(iv) $\int_{\mathbb{R}^d} \psi(x)dx > 0$.

We call such a function a *kernel*. For any $h := (h_1, \ldots, h_d) \in (0, 1/2]^d$ we define

$$A_h := \{t \in \mathbb{R}^d : h_i \leq t_i \leq 1 - h_i \quad \text{for } i = 1, \ldots, d\}. \tag{1.4}$$

For any $t \in A_h$ we define the centered (at t) and scaled kernel function $\psi_{t,h} : [0,1]^d \to \mathbb{R}$ as

$$\psi_{t,h}(x) := \psi\left(\frac{x_1 - t_1}{h_1}, \ldots, \frac{x_d - t_d}{h_d}\right), \quad \text{for } x = (x_1, \ldots, x_d) \in [0,1]^d. \tag{1.5}$$

Here $h \in (0, 1/2]^d$ is the smoothing bandwidth and $t \in A_h$ ensures that the scaled kernel function $\psi_{t,h}$ is zero outside $[0,1]^d$. For a fixed $t \in A_h$ we can construct a kernel estimator $\hat{f}_h(t)$ of $f(t)$ based on the data process $Y(\cdot)$ as

$$\hat{f}_h(t) := \frac{1}{n^{1/2}(\Pi_{i=1}^d h_i)\langle \mathbb{I}, \psi \rangle} \int_{[0,1]^d} \psi_{t,h}(x) dY(x),$$

where for any functions $g_1, g_2 \in L_2(\mathbb{R}^d)$, define $\langle g_1, g_2 \rangle := \int_{\mathbb{R}^d} g_1(x) g_2(x) dx$. Also define $\mathbb{I} : [-1,1]^d \to \mathbb{R}$ such that $\mathbb{I}(x) := 1$ for all $x \in [-1,1]^d$ and 0 otherwise. We consider the *normalized* version of the above kernel estimator $\hat{f}_h(t)$:

$$\hat{\Psi}(t,h) := \frac{1}{(\Pi_{i=1}^d h_i)^{1/2} \, \|\psi\|} \int_{[0,1]^d} \psi_{t,h}(x) dY(x), \tag{1.6}$$

where $\|\psi\|^2 := \int_{\mathbb{R}^d} \psi^2(x)dx < \infty$. We can use $\hat{\Psi}(t,h)$ to test

$$H_0 : f(t) = 0 \quad \text{versus} \quad H_1 : f(t) \neq 0$$

where we would reject the null hypothesis for extreme values of $\hat{\Psi}(t,h)$. So, a naive approach to testing (1.2) could be to consider $\sup_{t\in A_h}|\hat{\Psi}(t,h)|$. As this test statistic crucially depends on the choice of the smoothing bandwidth vector h, an approach that bypasses the choice of the tuning parameter h and also combines information at various bandwidths (scales) would be to consider the test statistic

$$\sup_{h>0}\sup_{t\in A_h}|\hat{\Psi}(t,h)|, \tag{1.7}$$

where $h>0$ is a short-hand for $h\in(0,1/2]^d$. However, under the null hypothesis (1.2)

$$\sup_{h>0}\sup_{t\in A_h}|\hat{\Psi}(t,h)| = \infty \qquad \text{almost surely (a.s.)}$$

as, for a fixed scale h, $\sup_{t\in A_h}|\hat{\Psi}(t,h)| = O_p(\sqrt{2\log(1/(2^d h_1 \dots h_d))})$; see e.g., [20]. Thus, to use the above approach to construct a valid test for (1.2) we need to put the test statistics $\sup_{t\in A_h}|\hat{\Psi}(t,h)|$ at different scales (i.e., h) in the same footing — this leads to the following definition of the *multiscale statistic* in d-dimensions:

$$T(Y,\psi) := \sup_{h\in(0,1/2]^d}\sup_{t\in A_h}\frac{|\hat{\Psi}(t,h)| - \Gamma(2^d h_1 \dots h_d)}{D(2^d h_1 \dots h_d)} \tag{1.8}$$

where $\Gamma, D : (0,1] \to [0,\infty)$ are two functions defined as

$$\Gamma(r) := (2\log(1/r))^{1/2} \tag{1.9}$$

and

$$D(r) := (\log(e/r))^{-1/2}\log\log(e^e/r); \tag{1.10}$$

see [4]. In Theorem 2.1.1, a main result in this paper, we show that the above multivariate multiscale statistic $T(Y,\psi)$ is well-defined and is a subexponential random variable for any kernel function ψ satisfying (i)-(iv) above, when $f\equiv 0$. This result immediately extends the main result of [4, Theorem 2.1] beyond $d=1$. Although there has been several proposals that extend the definition and the optimality properties of the multiscale statistic of [4] beyond $d = 1$ (see e.g., [14], [16], [18]) we believe that our approach has the closest resemblance to [4]. Further, the exact form of $T(Y, \psi)$ leads to optimal tests for (1.2) and other alternatives (see [18] for more details).

To show the subexponentiality of the proposed multiscale statistic $T(W, \psi)$ we prove a general result about a stochastic process with sub-Gaussian increments on a pseudometric space which may be of independent interest (see Theorem 2.1.2). This result mirrors [4, Theorem 6.1] but improves it in two ways: Firstly it assumes a weaker condition on the packing numbers of the pseudometric space on which the stochastic process is defined, and secondly it proves the subexponentiality (instead of just the finiteness) of the supremum of the process. This weaker condition on the packing numbers is crucial to the proof of Theorem 2.1.1; see Remark 2.1.1 where we compare our result with [4, Theorem 6.1]. Moreover, Lemma 2.1.1 gives a bound on the packing numbers of the pertinent (to our application) pseudometric space, which we believe is also new; see Remarks 2.1.2 and 2.1.3 where we compare our result with some relevant recent papers.

Chapter 2: Optimality in Testing Problems

2.0.1 Optimality of the multiscale statistic

Before we describe our main results let us first introduce the asymptotic minimax hypothesis testing framework. There is an extensive literature on nonparametric testing of the simple hypothesis $\{0\}$. As a starting point we refer the readers to [21]. In the nonparametric setting it is usually assumed that f belongs to a certain class of functions $\mathbb{F}$ and its distance from the null function $f = 0$ is defined by a seminorm $|\cdot|$. In this setting, given $\alpha \in (0, 1)$, the goal is to find a level α test ϕ_n (i.e., $\mathbb{E}_0[\phi_n(Y)] \leq \alpha$) such that

$$\inf_{g\in\mathbb{F}:|g|\geq\delta\rho_n} \mathbb{E}_g[\phi_n(Y)] \tag{2.1}$$

is as large as possible for some $\delta > 0$ and $\rho_n > 0$ where $\rho_n \to 0$ as $n \to \infty$ (ρ_n is a function of the sample size n); in the above notation $\mathbb{E}_g$ denotes expectation under the alternative function g. However, it can be shown that given $\mathbb{F}$ and $|\cdot|$, the constants δ and ρ_n cannot be chosen arbitrarily if one wants to have a statistically meaningful framework (see the survey papers [22], [23], [24] for $d = 1$ and [9] for $d > 1$). It turns out that if $\delta\rho_n$ is too small then it is not possible to test the null hypothesis with nontrivial asymptotic power (i.e., the infimum in (2.1) cannot be strictly larger than $\alpha + o(1)$). On the other hand if $\delta\rho_n$ is very large many procedures can test $f \equiv 0$ with significant power (i.e., the infimum in (2.1) goes to 1 as $n \to \infty$). Note that at first glance it may seem like the detection boundary $\delta\rho_n$ may depend on the level of the test α, but as long as $\alpha \in (0, 1)$ the detection boundary generally turns out to be independent of α; see the survey papers by [22], [23], [24] for details. In our case also the detection boundary is independent of α as illustrated in Theorems 2.2.1 and 2.2.2.

The hypothesis testing problem then reduces to: (a) Finding the largest possible $\delta\rho_n$ such that no test can have nontrivial asymptotic power (i.e., under the alternative f such that $|f| \leq \delta\rho_n$, the asymptotic power is less than or equal to the level α), and (b) trying to construct test procedures that can detect signals f, with $|f| > \delta\rho_n$, with considerable power (power going to 1 as $n \to \infty$). More specifically, δ and ρ_n are defined such that $\delta\rho_n$ is the largest for which, for all $\epsilon > 0$, we have

$$\limsup_{n\to\infty} \sup_{\phi_n} \inf_{g\in\mathbb{F}:|g|\geq(1-\epsilon)\delta\rho_n} \mathbb{E}_g[\phi_n(Y)] \leq \alpha,$$

where the supremum is taken over all sequence of level α tests ϕ_n. In this case ρ_n is called the *minimax rate of testing* and δ is called the *exact separation constant* (see [7], [25] for more details about minimax testing). On the other hand, we want to find a test $\tilde{\phi}_n$ such that

$$\lim_{n\to\infty} \inf_{g\in\mathbb{F}:|g|\geq(1+\epsilon)\delta\rho_n} \mathbb{E}_g[\tilde{\phi}_n(Y)] = 1.$$

In such a scenario, $\tilde{\phi}_n$ is called an *asymptotically minimax test*. Here we would also like to point out that if there exists a test $\hat{\phi}_n$ and a constant $\hat{\delta} > \delta$ such that

$$\lim_{n\to\infty} \inf_{g\in\mathbb{F}:|g|\geq\hat{\delta}\rho_n} \mathbb{E}_g[\hat{\phi}_n(Y)] = 1$$

then the test $\hat{\phi}_n$ is called a *rate optimal test*.

In Section 2.2 we show that our proposed multiscale statistic yields an asymptotically minimax test for the following scenarios:

(i) (Optimality for Hölderian alternatives). Consider testing hypothesis (1.2). If

$$\|f\|_\infty \geq c_*(1+\epsilon_n)(\log(en)/n)^{\frac{\beta}{2\beta+d}},$$

where f belongs to the Hölder class $\mathbb{H}_{\beta,L}$ with $\beta > 0$ and $L > 0$, $\|f\|_\infty := \sup_{x\in[0,1]^d} |f(x)|$ denotes the sup-norm of f, and c_* is a constant (defined explicitly in Theorem 2.2.1), we show that we can construct a level α test based on the multiscale statistic (1.8) that has power converging to

1, as $n \longrightarrow \infty$, provided ϵ_n does not go to 0 too fast (see Theorem 2.2.1 for the exact order of ϵ_n). We note that this multiscale statistic would require the knowledge of β but not of L.

Moreover, we show that if $\|f\|_\infty \leq c_*(1-\epsilon_n)(\log(en)/n)^{\beta/2\beta+d}$ no test of level $\alpha \in (0,1)$ can have nontrivial asymptotic power; see Theorem 2.2.1 for the details. This shows that our proposed multiscale test is asymptotically minimax with rate of testing $\rho_n = (\log(en)/n)^{\beta/(2\beta+d)}$ and exact separation constant $\delta = c_*$. As far as we are aware this is the first instance of an asymptotically minimax test for the Hölder class $\mathbb{H}_{\beta,L}$ when $d > 1$ (under the supremum norm). Moreover, if the smoothness β of the Hölder class $\mathbb{H}_{\beta,L}$ is unknown (but $\beta \leq 1$) then we can still construct a rate optimal test for this problem; see Proposition 2.2.1 for the details.

(ii) (Optimality for detecting signals at large/small scales). Consider testing the hypothesis

$$H_0 : f = 0 \quad \text{versus} \quad H_1 : f = \mu_n \mathbb{I}_{B_n}, \tag{2.2}$$

where $\mu_n \neq 0 \in \mathbb{R}$ and

$$B_n \equiv B_\infty(t^{(n)}, h^{(n)}) := \{x \in [0,1]^d : |x_i - t_i^{(n)}| < h_i^{(n)} \text{ for all } i = 1,\dots,d\}$$

are unknown, for some $h^{(n)} \in (0,1/2]^d$ and $t^{(n)} \in A_{h^{(n)}}$, and $\mathbb{I}_{B_n}$ denotes the indicator of the hyperrectangle B_n. First, consider the scenario $\liminf_{n\to\infty} |B_n| > 0$ where $|B_n|$ denote the Lebesgue measure of B_n. Then, if $\lim_{n\to\infty} \sqrt{n}|\mu_n| \longrightarrow +\infty$, we can construct a level α test based on the multiscale statistic (1.8) that has power converging to 1 as $n \longrightarrow \infty$; see Theorem 2.2.2. Further, we show that, if $\limsup_{n\to\infty} \sqrt{n}|\mu_n| < \infty$, no test of level α can detect the alternative with power going to 1. Thus, the multiscale test is optimal for detecting signals on large scales.

On the other hand, let us now consider the case $\lim_{n\to\infty} |B_n| = 0$. If

$$|\mu_n|\sqrt{n|B_n|} \geq (1+\epsilon_n)\sqrt{2\log(1/|B_n|)}, \qquad \text{for all } n,$$

we can construct a test of level α, based on the proposed multiscale statistic, that has power converging to 1 as $n \to \infty$, provided ϵ_n does not go to 0 too fast (see Theorem 2.2.2). Furthermore, we can show that if

$$|\mu_n|\sqrt{n|B_n|} = (1 - \epsilon_n)\sqrt{2\log(1/|B_n|)}, \qquad \text{for all } n,$$

no test can detect the signal reliably with nontrivial power (i.e., for any level α test ϕ_n there exists a signal f_n of the above described strength such that ϕ_n will fail to detect f_n with asymptotic probability at least $1 - \alpha$); see Theorem 2.2.2 for the details. This shows that our multiscale test is asymptotically minimax for signals at small scales.

2.0.2 Literature review and connection to existing works

Our multiscale statistic (1.8) can be thought of as a penalized scan statistic, as it is based on the maximum of an ensemble of local test statistics $|\hat{\Psi}(t, h)|$, penalized and properly scaled. Scan-type procedures have received much attention in the literature over the past few decades. Examples of such procedures can be found in [26], [27], [28], [29], [30], [31] etc. All the above mentioned papers consider $d = 1$ and no penalization term (like $\Gamma(\cdot)$ in our case) was used. Asymptotic properties of the scan statistic have been studied expensively. In [30] and [32] the authors give asymptotic approximations of the distribution of the scan statistic when $d = 1$. For $d = 2$, similar results can be found in [10], [31], [33], among others. Recently in [34] the authors give exact asymptotics for the scan statistic for any dimension d.

In all of the above papers it is noted that the scan statistic is dominated by small scales; this creates a problem for detecting large scale signals. One common proposal to fix this problem is to modify the scan statistic so that instead of the maximum over all scales we look at the maximum over scales that are in an appropriate interval containing the true scale of the signal; see e.g., [30], [34]. In particular, the last two papers show that if the extent of the signal is of a certain order ($\log n$) then this approach leads to power comparable to an oracle. An obvious drawback with the above approach is that we need to have some prior knowledge on which scales the signal(s) may be present. In contrast, our multiscale method does not require any such knowledge. [19] used a multiple testing procedure to obtain optimal detection in both large and small hyperrectangles in the general setting of inverse problems. Our approach, in fact, can also be seen as a form of multiple testing procedure.

Another approach that has been proposed to optimally detect signals on both large and small scales is to use different critical values (of the scan statistic) to test for signals at different scales separately (see e.g., [14], [16]) and use multiple testing procedures (see [35] and the references within) to calibrate the method. Here we would like to note that most methods, including our multiscale approach, that try to detect signals optimally for both large and small scales suffers from a loss of power in either small or large compared to methods that are fine tuned for either scales. Our method sacrifices power at small scales (compared to the unpenalized scan statistic) in favor of optimal detection at all scales.

Conceptually, our work is most related to that of [4], where the authors proposed our multiscale statistic for $d = 1$. Thus, our work can be thought of as a generalization of [4] to multidimension ($d > 1$).

2.1 Multidimensional multiscale statistic

Let us first recall the definition of the multivariate multiscale statistic $T(Y, \psi)$ given in (1.8). The following theorem, our main result in this section, shows that the multiscale statistic $T(Y, \psi)$ is well-defined and attains a subexponential tail bound for any kernel function ψ; see Appendix A.3 for a proof.

Theorem 2.1.1 *Let ψ be a kernel function satisfying (i)-(iv) in the Introduction. For a positive vector $h := (h_1, \dots, h_d) > 0$, let A_h be as defined in* (1.4)*. For $t \in A_h$, let $\psi_{t,h}(\cdot)$ and $\hat{\Psi}(t, h)$ be as defined in* (1.5) *and* (1.6)*, respectively. Consider the statistic $T(W, \psi)$ as defined in* (1.8)*, where $W(\cdot)$ is the standard Brownian sheet on $[0, 1]^d$. Then, almost surely, $T(W, \psi) < \infty$, i.e., $T(W, \psi)$ is a tight random variable. Moreover, there exists constants c_0 and c_1 depending on the kernel ψ such that $\mathbb{P}(T(W, \psi) > u) \leq c_0 \exp(-u/c_1)$ for all $u > 0$.*

Theorem 2.1.1 immediately extends the main result of [4, Theorem 2.1] beyond $d = 1$. The proof of the above theorem crucially relies on the following two results. We first introduce some notation.

Definition 2.1.1 (Packing number) *For any pseudometric space $(\mathscr{F}, \rho)$ and $\epsilon > 0$, the packing number $N(\epsilon, \mathscr{F})$ is defined as the supremum of the number of elements in $\mathscr{F}'$ where $\mathscr{F}' \subseteq \mathscr{F}$ and for all $a \neq b \in \mathscr{F}'$ we have $\rho(a, b) > \epsilon$.*

We will prove Theorem 2.1.1 as a consequence of the following more general result about stochastic processes with sub-Gaussian increments on some pseudometric space (see Section A.2 for its proof).

Theorem 2.1.2 *Let X be a stochastic process on a pseudometric space $(\mathscr{F}, \rho)$ with continuous sample paths. Suppose that the following three conditions hold:*

(a) There is a function $\sigma : \mathscr{F} \to (0, 1]$ and a constant $K \geq 1$ such that

$$\mathbb{P}\big(X(a) > \sigma(a)\eta\big) \leq K \exp(-\eta^2/2) \qquad \forall\, \eta > 0,\ \forall\, a \in \mathscr{F}.$$

Moreover, $\sigma^2(b) \leq \sigma^2(a) + \rho^2(a, b)$, $\forall\, a, b \in \mathscr{F}$.

(b) For some constants $L, M \geq 1$,

$$\mathbb{P}\big(|X(a) - X(b)| > \rho(a, b)\eta\big) \leq L \exp(-\eta^2/M) \quad \forall\, \eta > 0,\ \forall\, a, b \in \mathscr{F}.$$

(c) For some constants $A, B, V, p > 0$,

$$N((\delta u)^{1/2}, \{a \in \mathscr{F} : \sigma^2(a) \leq \delta\}) \leq A u^{-B} \delta^{-V} (\log(e/\delta))^p \ \ \forall\, u, \delta \in (0, 1].$$

Then the random variable

$$S(X) := \sup_{a \in \mathscr{F}} \frac{X^2(a)/\sigma^2(a) - 2V \log(1/\sigma^2(a))}{\log\log(e^e/\sigma^2(a))} \tag{2.3}$$

is subexponential. More precisely, $\mathbb{P}(S(X) > u) \leq \xi_1 \exp(-u/\xi_2)$ *for all* $u > 0$*, for some* $\xi_1, \xi_2 > 0$ *depending only on the constants* K, L, M, A, B, p *and* V*.*

Remark 2.1.1 (Connection to [4]) *A similar result to Theorem 2.1.2 above appears in [4, Theorem 6.1]. However note that there is a subtle and important difference: The bound on the packing number in (c) of Theorem 2.1.2 involves the additional logarithmic factor* $(\log(e/\delta))^p$ *which is not present in [4, Theorem 6.1]. In fact, we show that even with this additional logarithmic factor, the random variable* $S(X)$*, defined in* (2.3)*, involves the same penalization term* $2V \log(1/\sigma^2(a))$ *as in [4, Theorem 6.1]. Hence, we can think of Theorem 2.1.2 as a generalization of [4, Theorem 6.1]. Here we would also like to point out that our result improves [4, Theorem 6.1] by proving the subexponentiality of the random variable* $S(X)$ *instead of just its finiteness.*

To apply Theorem 2.1.2 to prove Theorem 2.1.1 we need to define a suitable pseudometric space $(\mathscr{F}, \rho)$ and a stochastic process, and verify that conditions (a)-(c) in Theorem 2.1.2 hold. In that vein, let us define the set

$$\mathscr{F} := \left\{(t, h) \in \mathbb{R}^d \times (0, 1/2]^d : h_i \leq t_i \leq 1 - h_i, \text{ for all } i = 1, 2, \ldots, d\right\}$$

with the following pseudometric

$$\rho^2((t, h), (t', h')) := |B_\infty(t, h) \triangle B_\infty(t', h')|, \qquad \text{for } (t, h), (t', h') \in \mathscr{F},$$

where $B_\infty(t, h) := \Pi_{i=1}^d (t_i - h_i, t_i + h_i)$, $A \triangle B := (A \cap B^c) \cup (A^c \cap B)$ denotes the symmetric difference of the sets A and B, and $|A|$ denotes the Lebesgue measure of the set A. Also, define

$$\sigma^2(t, h) := |B_\infty(t, h)| = 2^d \Pi_{i=1}^d h_i, \qquad \text{for } (t, h) \in \mathscr{F}.$$

The following important result shows that indeed for the above defined pseudometric space $(\mathscr{F}, \rho)$ condition (c) of Theorem 2.1.1 holds; see Section A.2.1 for its proof.

Lemma 2.1.1 *Let $\mathscr{F}, \rho(\cdot,\cdot)$ and $\sigma(\cdot)$ be as described above. Then, for all $u, \delta \in (0, 1]$,*

$$N\left((u\delta)^{1/2}, \{(t,h) \in \mathscr{F} : \sigma^2(t,h) \le \delta\}\right) \le Ku^{-2d}\delta^{-1}(\log(e/\delta))^{d-1} \tag{2.4}$$

for some constant K depending only on d.

Remark 2.1.2 *Here we would like to point out that Lemma 2.1.1 shows that condition (c) of Theorem 2.1.2 holds with $B = 2d$, $p = d - 1$ and most importantly for $V = 1$, which was also the case when $d = 1$ (as shown in [4]). An equivalent result for $d = 2$ is proved in [14, Theorem 1].*

Remark 2.1.3 (Connection to [36]) *Note that a similar multiscale statistic, as in* (1.8) *without the* $\log\log(e^e/(2^d h_1 \ldots h_d))$ *multiplier in the denominator, has been proposed in [36] where the subexponentiality of their statistic was also proved. Here we would like to point out the main differences between the two papers. Translated to our setting, [36] scans over hyperrectangles such that each side is greater than a prespecified number* $(1/L)$*, whereas our multiscale statistic* (1.8) *scans over hyperrectangles of any length. As our multiscale statistic scans over hyperrectangles of any length we can optimally test for signals distributed over hyperrectangles on any scale, which would not be possible for the test statistic proposed in [36]; see Section 2.2.2 for more details.*

Compare the numerator of our multiscale statistic (1.8) with the multiscale statistic proposed in [18, Equation (6)]. Translated to our setting, in [18] the authors propose a penalization term $\Gamma_V(2^d h_1 \ldots h_d)$ where $\Gamma_V : (0, 1] \to (0, \infty)$ is defined as

$$\Gamma_V(r) := (2V \log(1/r))^{1/2}.$$

In [18, Section 1.1] the authors also recommend to choose the constant V in the penalization term Γ_V as small as possible for optimal testing. [18, Example 2.3] recommend choosing $V = 1$ by appealing to Lemma 2.1.1 of our paper. The following proposition shows that indeed $V = 1$ is the smallest possible permissible value; see Appendix A.4.1 for a proof.

Proposition 2.1.1 *Suppose $V < 1$. Let Γ_V and $\mathscr{F}$ be as defined above. Then we have*

$$\sup_{(t,h)\in\mathscr{F}} |\hat{\Psi}(t,h)| - \Gamma_V(2^d h_1 \dots h_d) = \infty \quad a.s.$$

Thus, $\sup_{(t,h)\in\mathscr{F}} \frac{|\hat{\Psi}(t,h)|-\Gamma_V(2^d h_1 \dots h_d)}{D(2^d h_1 \dots h_d)} = \infty \quad a.s.$

2.2 Optimality of the multiscale statistic in testing problems

In this section we prove that we can construct tests based on the multiscale statistic that are optimal for testing (1.2) and (2.2). For both the testing problems we can define a multiscale test based on kernel ψ as follows: Let

$$\kappa_{\alpha,\psi} = \inf\{c \in \mathbb{R} : \mathbb{P}(T(W,\psi) > c) \leq \alpha\},$$

where W is the standard Brownian sheet on $[0,1]^d$. For notational simplicity we would denote $\kappa_{\alpha,\psi}$ by κ_α from now on.

For testing (1.2) and (2.2) a test of level α can be defined as follows:

$$\text{Reject } H_0 \qquad \text{if and only if} \qquad T(Y,\psi) > \kappa_\alpha.$$

Let us call this testing procedure the multiscale test. Although any kernel ψ can be used to construct the above test, in Sections 2.2.1 and 2.2.2 we show that specific choices of the kernel function ψ lead to asymptotically minimax tests.

2.2.1 Optimality against Hölder classes of functions

Let us recall the definition of the Hölder class of functions $\mathbb{H}_{\beta,L}$, for $\beta \in (0,1]$ and $L > 0$, as in (1.3); see Definition A.1.2 for the formal definition of $\mathbb{H}_{\beta,L}$ for any $\beta > 0$. Let $\psi_\beta : \mathbb{R}^d \to \mathbb{R}$, for

$0 < \beta < \infty$, be the unique solution of the following optimization problem:

$$\text{Minimize } \|\psi\| \text{ over all } \psi \in \mathbb{H}_{\beta,1} \text{ with } \psi(0) \geq 1. \tag{2.5}$$

Elementary calculations show that for $0 < \beta \leq 1$, we have

$$\psi_\beta(x) = (1 - \|x\|^\beta)\mathbb{I}(\|x\| \leq 1);$$

see Appendix A.4.2 for a proof. For $\beta > 1$, ψ_β can be calculated numerically. We consider the kernel ψ_β, for $\beta > 0$, described above and state our first optimality result for testing (1.2); see Appendix A.5.1 for a proof.

Theorem 2.2.1 *Let $T_\beta \equiv T(Y, \psi_\beta)$ be the multiscale statistic defined in* (1.8) *with kernel ψ_β, for $0 < \beta < \infty$. Define*

$$\rho_n := \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$$

and

$$c_* \equiv c_*(\beta, L) := \left(\frac{2dL^{d/\beta}}{(2\beta + d)\left\|\psi_\beta\right\|^2}\right)^{\frac{\beta}{2\beta+d}}.$$

Then, for arbitrary $\epsilon_n > 0$ with $\epsilon_n \longrightarrow 0$ and $\epsilon_n\sqrt{\log n} \longrightarrow \infty$ as $n \longrightarrow \infty$, the following hold:

(a) For any arbitrary sequence of tests ϕ_n with level α for testing (1.2)*, we have*

$$\limsup_{n\to\infty} \inf_{g\in\mathbb{H}_{\beta,L}:\|g\|_\infty=(1-\epsilon_n)c_*\rho_n} \mathbb{E}_g[\phi_n(Y)] \leq \alpha;$$

(b) for $J_n := [(c_*\rho_n/L)^{1/\beta}, 1 - (c_*\rho_n/L)^{1/\beta}]^d$, we have

$$\lim_{n\to\infty} \inf_{g\in\mathbb{H}_{\beta,L}:\|g\|_{J_n,\infty}\geq(1+\epsilon_n)c_*\rho_n} \mathbb{P}_g(T_\beta > \kappa_\alpha) = 1$$

where $\|g\|_{J_n,\infty} := \sup_{t\in J_n} |g(t)|$.

The above result generalizes [4, Theorem 2.2] beyond $d = 1$. Theorem 2.2.1 can be interpreted as follows: (a) for every test ϕ_n there exists a function with supremum norm $(1 - \epsilon_n)c_*\rho_n$ which cannot be detected with nontrivial asymptotic power; whereas (b) when we restrict to functions with signal strengths (i.e., supremum norm in the interior of $[0, 1]^d$) just a bit larger than the above threshold, our proposed multiscale test is able to detect every such function with asymptotic power 1. In this sense our proposed test is optimal in detecting departures from the zero function for Hölder classes $\mathbb{H}_{\beta,L}$. We note here that to calculate T_β we need the knowledge of β but we do not need to know L.

If β is unknown, but is less than or equal to 1, we can use T_1 as a test statistic for testing (1.2). Although the resulting test is not asymptotically minimax, the test is still rate optimal. The following result formalizes this; see Appendix A.5.1 for its proof.

Proposition 2.2.1 *Consider testing* (1.2) *where $\beta \leq 1$ is unknown. Let us recall the definition of ψ_1 in* (2.5)*. Let $T_1 \equiv T(Y, \psi_1)$ be the multiscale statistic defined in* (1.8) *with kernel ψ_1. Define*

$$\rho_n := \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$$

and let M be any constant such that $M > \left(\frac{2dL^{d/\beta}\|\psi_1\|^2}{(2\beta+d)\langle\psi_1,\psi_\beta\rangle^2}\right)^{\frac{\beta}{2\beta+d}}$. Let $J_n := [(M\rho_n/L)^{1/\beta}, 1 - (M\rho_n/L)^{1/\beta}]^d$. Then we have

$$\lim_{n\to\infty} \inf_{g\in\mathbb{H}_{\beta,L}:\|g\|_{J_n,\infty}\geq M\rho_n} \mathbb{P}_g(T > \kappa_\alpha) = 1$$

where κ_α is the $(1 - \alpha)$ quantile of the multiscale statistic $T(Y, \psi_1)$ under the null hypothesis.

Remark 2.2.1 *Instead of using the test statistic T_β if we use the test statistic*

$$T_\beta^\star := \sup_{h\in(0,1/2]^d} \sup_{t\in A_h} \left[|\hat{\Psi}(t, h)| - \Gamma(2^d h_1 \ldots h_d)\right] \tag{2.6}$$

with the kernel ψ_β, then the same conclusions as that of Theorem 2.2.1 and Proposition 2.2.1 would *hold. Thus the multiscale statistic $T^\star_\beta$ is also optimal against Hölderian alternatives.*

2.2.2 Optimality against axis-aligned hyperrectangular signals

In Theorem 2.2.1 we proved the optimality of the multiscale test when the supremum norm of the signal is large. A natural question that arises next is: "What if the signal is not peaked but distributed evenly on some subset of $[0,1]^d$?". To answer this question we look at the testing problem (2.2), and establish below the optimality of our multiscale test in this setting (see Appendix A.5.2 for a proof of Theorem 2.2.2). Note that when $d = 1$ similar optimality results are known for the multiscale statistic; see [17, Theorem 2.6] and [16, Section 4]. For $d > 1$ see [14] for a similar optimality result when the response variable is Bernoulli. For $h = (h_1, \dots, h_d) \in (0, 1/2]^d$, let us first define

$$\mathscr{B}_h := \{B \subseteq [0,1]^d : B = \Pi_{i=1}^d (t_i - h_i, t_i + h_i) \text{ for some } t = (t_1, \dots, t_d) \in A_h\}.$$

Theorem 2.2.2 *Let $T \equiv T(Y, \psi_0)$ where $\psi_0 = \mathbb{I}_{[-1,1]^d}$. Let $f_n = \mu_n \mathbb{I}_{B_n}$ where B_n is an axis-aligned hyperrectangle and let $|B_n|$ denote the Lebesgue measure of the set B_n. Then we have the following results:*

(a) Suppose that $\liminf_{n\to\infty} |B_n| > 0$. Let ϕ_n be any test of level $\alpha \in (0,1)$ for (2.2). Then, for any $f_n = \mu_n \mathbb{I}_{B_n}$ such that $\limsup_n |\mu_n|\sqrt{n|B_n|} < \infty$, we have

$$\limsup_{n\to\infty} \mathbb{E}_{f_n}[\phi_n(Y)] < 1.$$

Moreover, for the proposed multiscale test based on T, we have

$$\lim_{n\to\infty} \inf_{f_n : \lim |\mu_n|\sqrt{n|B_n|} = \infty} \mathbb{P}_{f_n}(T > \kappa_\alpha) = 1.$$

(b) *Now let us look at the case* $\lim_{n\to\infty}|B_n| = 0$. *Let* $h_n = (h_{1,n},\ldots,h_{d,n}) \in (0,1/2]^d$ *be any sequence of points such that* $\lim_{n\to\infty}\Pi_{i=1}^d h_{i,n} \to 0$. *Let*

$$\mathcal{G}_n^- := \{f_n = \mu_n \mathbb{I}_{B_n} : |\mu_n|\sqrt{n|B_n|} = (1-\epsilon_n)\sqrt{2\log(1/|B_n|)}, B_n \in \mathscr{B}_{h_n}\}$$

with $\epsilon_n \to 0$ *and* $\epsilon_n\sqrt{2\log(1/|B_n|)} \to \infty$. *(Here we have omitted the dependence of* h_n *in the notation* $\mathcal{G}_n^-$*). If* ϕ_n *be any test of level* $\alpha \in (0,1)$ *for* (2.2) *then we have*

$$\limsup_{n\to\infty} \inf_{f_n\in\mathcal{G}_n^-} \mathbb{E}_{f_n}[\phi_n(Y)] \le \alpha.$$

Moreover, let

$$\mathcal{G}_n^+ := \{f_n = \mu_n \mathbb{I}_{B_n} : |\mu_n|\sqrt{n|B_n|} \ge (1+\epsilon_n)\sqrt{2\log(1/|B_n|)}, B_n \in \mathscr{B}_{h_n}\}.$$

Then for our multiscale test we have

$$\lim_{n\to\infty} \inf_{f_n\in\mathcal{G}_n^+} \mathbb{P}_{f_n}(T > \kappa_\alpha) = 1.$$

Remark 2.2.2 *If we use the test statistic* $T^\star$*, as defined in* (2.6) *(with the kernel* ψ_0*), instead of* T *in Theorem 2.2.2, the optimality results described in the theorem still hold.*

Our first result in Theorem 2.2.2 shows that as long as $\liminf_{n\to\infty}|B_n| > 0$, for any test to have power converging to 1 we need to have $\lim|\mu_n|\sqrt{n|B_n|} = \infty$, in which case our multiscale test achieves asymptotic power 1. Thus our multiscale test is optimal for detecting large scale signals. The next result can be interpreted as follows: (i) For signals with small spatial extent (i.e., $\lim_{n\to\infty}|B_n| = 0$) if the signal strength is too small ($|\mu_n|\sqrt{n|B_n|} \le (1-\epsilon_n)\sqrt{2\log(1/|B_n|)}$) no test can detect the signal reliably with nontrivial probability (i.e., for every test ϕ_n there exist a signal such that ϕ_n will fail to detect it with probability $1-\alpha+o(1)$); (ii) on the other hand, if the signal strength is a bit larger than the threshold (i.e., the exact separation constant) described above our multiscale test will detect the signal with asymptotic power 1. This shows that our multiscale test achieves optimal detection for signals with small spatial footprint. We would like to emphasize here that by using the same exact test (using the same kernel ψ_0) we are able to optimally detect both large and small scale signals. In [19], the authors used a multiple testing method to achieve optimal detection in both large and small scale hyperrectangles.

Remark 2.2.3 *We would like to point out that the proofs for the minimax lower bound that have been derived for the two scenarios in Theorems 2.2.1 and 2.2.2 follow the standard techniques that have been used in [22], [23], [24], [7], [4], [11], [9], [12], [17], etc. Note that although all the above cited papers have similar proof techniques there is quite some variation in the strength of their results. Our results and proofs most closely follow that of [4].*

Comparison with the scan and average likelihood ratio statistics when $d = 1$

When $d = 1$ there exists an extensive literature on the optimal detection threshold for signals of the form $f_n = \mu_n \mathbb{I}_{B_n}$, where now $B_n \subseteq [0, 1]$ is an interval. In [16] the authors compare the performance of the scan statistic (i.e., the statistic (1.7) in the discrete setup with $\psi = \mathbb{I}_{[-1,1]}$) and the average likelihood ratio (ALR) statistic (which is the discrete analogue of $\int_0^{1/2} \int_h^{1-h} \exp[|\hat{\Psi}(t, h)|^2/2]\,dt\,dh$); see Section 4 for a description and comparison of the two competing methods with our multiscale test when $d = 2$.

When $\liminf_{n\to\infty} |B_n| > 0$ the scan statistic can only detect the signal, with asymptotic power 1, when $|\mu_n|\sqrt{n} \geq (1+\epsilon_n)\sqrt{2\log n}$, whereas the ALR statistic (and the proposed multiscale statistic) can detect the signal whenever we have $|\mu_n|\sqrt{n} \to \infty$ (which is a less stringent condition). Note that $|\mu_n|\sqrt{n} \to \infty$ is also required for any test to detect the signal with asymptotic power 1. This shows that the scan statistic is not optimal for detecting large scale signals.

On the other hand if $\lim_{n\to\infty} |B_n| = 0$, the scan statistic can detect the signal if $|\mu_n|\sqrt{n|B_n|} \geq (1 + \epsilon_n)\sqrt{2\log n}$ whereas the ALR statistic can detect the signal when $|\mu_n|\sqrt{n|B_n|} \geq \sqrt{2}(1 + \epsilon_n)\sqrt{2\log(1/|B_n|)}$. The optimal detection threshold in this scenario is $|\mu_n|\sqrt{n|B_n|} \geq (1+\epsilon_n)\sqrt{2\log(1/|B_n|)}$, which is attained by the multiscale statistic. Thus that scan statistic is optimal in detecting signals only when $|B_n| = O(1/n)$. The ALR statistic requires the signal to be at least $\sqrt{2}$ times the (detectable) threshold. This shows that neither the standard scan or the ALR is able to achieve the optimal threshold for detecting small scale signals.

[17, Theorem 2.6] shows the optimality of the multiscale statistic (which is a modification of the scan statistic) in detecting signals in both cases when $d = 1$. In [37] and [16] the authors propose a condensed ALR statistic which, much like the multiscale statistic, is able to attain the optimal threshold for detection in both regimes of B_n. As far as we are aware the condensed ALR statistic has not been extended beyond $d = 1$ and therefore whether it achieves the optimal threshold for $d > 1$ is not known. In summary, Theorem 2.2.2 shows that our multidimension multiscale test is asymptotically minimax even when $d > 1$.

2.2.3 The discrete analogue of the multiscale statistic

Although thus far we have defined and analyzed the multiscale statistic arising from a continuous white noise model, in real applications we have to invariably deal with a discrete analogue of this problem. In this subsection we briefly describe this discrete setting and comment on the applicability of our results.

Let us start with the connection to nonparametric regression on gridded design. Let $x_1, \dots, x_n \in \mathbb{R}^d$ be an enumeration of the $m \times \cdots \times m$ uniform grid $G^m := \{1/m, 2/m, \dots, (m-1)/m, 1\}^d$ where $m^d = n$. Let us look at the following nonparametric regression problem:

$$Y_i = f(x_i) + \epsilon_i, \qquad \text{for } i = 1, \dots, n \tag{2.7}$$

where $f : [0,1]^d \to \mathbb{R}$ is the unknown regression function and ϵ_i's are i.i.d. standard normal random variables. For a kernel function $\psi : \mathbb{R}^d \to \mathbb{R}$ and $h, t \in G^m$, such that $t - h, t + h \in G^m$ we can define a kernel estimator $\hat{f}_h$ of f as

$$\hat{f}_h(t) = \frac{\sum_{i:x_i \in B_\infty(t,h)} Y_i \, \psi\big((x_i - t)/h\big)}{\sum_{i:x_i \in B_\infty(t,h)} \psi\big((x_i - t)/h\big)}$$

where by $(u_1, \ldots, u_d)/(h_1, \ldots, h_d)$ we mean the vector $(u_1/h_1, \ldots, u_d/h_d)$. We can also define the standardized kernel estimator as

$$\hat{\Psi}_n(t,h) = \frac{\sum_{i:x_i \in B_\infty(t,h)} Y_i\, \psi\big((x_i - t)/h\big)}{\sqrt{\sum_{i:x_i \in B_\infty(t,h)} \psi^2\big((x_i - t)/h\big)}}.$$

Then the multiscale statistic for this regression problem reduces to

$$T_n(Y, \psi) := \sup_{h \in G^m : t-h, t+h \in G^m} \sup_{t \in G^m} \frac{|\hat{\Psi}_n(t,h)| - \Gamma\left(|B_\infty(t,h) \cap G^m|\right)}{D\left(|B_\infty(t,h) \cap G^m|\right)} \tag{2.8}$$

where $|B_\infty(t,h) \cap G^m|$ now denotes the number of elements in $B_\infty(t,h) \cap G^m$; $\Gamma(\cdot)$ and $D(\cdot)$ are defined in (1.9) and (1.10) respectively. Note that $T(Y, \psi)$ (as defined in (1.8)) stochastically dominates $T_n(Y, \psi)$ and thus $T_n(Y, \psi)$ is well-defined and finite a.s.

Let us now comment on the computation of the discrete multiscale statistic. Observe that a naive approach to computing $T_n(Y, \psi)$ will involve taking the maximum over $O(n^2) \equiv O(m^{2d})$ rectangles. This can indeed be prohibitive for n large. A natural idea is to consider a well chosen subset of all possible hyperrectangles when taking the supremum; we refer the reader to [14] where such a suitably rich collection (of the order of $O(n \log n)$) of hyperrectangles is proposed and analyzed. We believe that such an approximation of the multiscale statistic will still preserve its optimality properties (up to logarithmic factors in the rates).

Chapter 3: Confidence bands for multivariate shape-restricted functions

3.1 Introduction

The area of shape-restricted regression is concerned with nonparametric estimation of a regression function under natural shape constraints such as monotonicity, convexity, unimodality/quasiconvexity, etc. This particular field in the statistical literature has a long history dating back to influential papers such as [38, 39, 40, 41, 42]; also see [43, 44, 45, 46] for book length treatments on this topic. Indeed, such shape constraints arise naturally in various contexts: isotonic regression methods are widely employed in many real-life applications ranging from predicting ad click–through rates [47] to gene–gene interaction search [48]; convex regression arises in productivity analysis [49], efficient frontier methods [50], in stochastic control [51], etc. In the recent years there has been much activity on estimation of such shape-constrained regression functions with multivariate predictors; see e.g., [52, 53, 54, 55]. A primary focus of many of the recent papers has been on consistent estimation of the unknown regression function via derivation of finite sample risk bounds quantifying the performance of estimation; see [56, 57, 58, 59, 60, 61, 62, 54, 63]. The more intricate problem of carrying out statistical inference in these multivariate shape-constrained problems, e.g., construction of confidence sets, is vastly unexplored.

In this paper we consider construction of honest confidence sets for shape-constrained regression problems with multiple covariates with special emphasis to: (i) coordinate-wise isotonic functions, and (ii) convex functions. Our proposed methodology, which extends the ideas of Dümbgen [64] to multiple dimensions, yields asymptotically optimal confidence sets that possess various adaptivity properties. In particular, for scenarios (i) and (ii) above, we prove spatial and local adaptivity of our confidence bands with respect to the smoothness of the underlying function and its intrinsic dimensionality. Our confidence bands are constructed using a multidimensional multiscale statistic developed in [65], which in turn was inspired by the one-dimensional multiscale statistic proposed and studied in [66].

Let us first reiterate our continuous multidimensional white noise model:

$$Y(t) = \sqrt{n} \int_0^{t_1} \dots \int_0^{t_d} f(s_1, \dots, s_d)\, ds_d \dots ds_1 + W(t), \tag{3.1}$$

where $t := (t_1, \dots, t_d) \in [0, 1]^d$, $d \geq 1$, $\{Y(t_1, \dots, t_d) : (t_1, \dots, t_d) \in [0, 1]^d\}$ is the observed data, $f \in L_1([0, 1]^d)$ is the unknown (regression) function of interest,

Given a function class $\mathcal{F} \subset L_1([0, 1]^d)$ (e.g., $\mathcal{F}$ can be the class of coordinate-wise isotonic functions or the class of convex functions defined on $[0, 1]^d$), our goal is to construct an honest confidence band $(\hat{\ell}, \hat{u})$ for the true function $f \in \mathcal{F}$, i.e., find functions $\hat{\ell}$ and $\hat{u}$ depending on the observed data $Y(\cdot)$ (see (1.1)) that satisfy:

$$\mathbb{P}_f \left(\hat{\ell} \leq f \leq \hat{u} \right) \geq 1 - \alpha, \qquad \text{for all } \ f \in \mathcal{F}, \quad \text{for all } \ n \geq 1, \tag{3.2}$$

and for some $\alpha \in (0, 1)$. Here, by $\mathbb{P}_f(\cdot)$ we mean probability computed when the true function is f in (1.1). Although nonparametric estimation of an unknown regression/density function based on smoothness assumptions using techniques such as kernels, splines and wavelets are abundant in the literature (see e.g., [67, 68, 69, 70, 71, 72, 73, 74, 75, 76]), it is known that fully adaptive inference for certain smoothness function classes is not possible without making qualitative assumptions of some kind on the parameter space; see e.g., [77, Theorem 8.3.11] (also see e.g., [78], [79]).

Indeed, shape constrained functions satisfy a *two-sided bias inequality* (see (3.6) below) — a crucial assumption made in this paper for our proposed method — which enables the construction of adaptive inference procedures. Further, in many real-life applications, justifying smoothness assumptions (e.g., involving quantitative bounds on the gradient) is often impractical, and in many such situations qualitative assumptions like shape constraints are available. For example, in many economic applications — such as estimation of production and utility functions — it is more reasonable to assume monotonicity and/or concavity on the shape of the underlying function, rather than making quantitative assumptions on the smoothness of such functions (see e.g., [80, 81, 82, 83, 84, 49]).

Non-asymptotic confidence bands under such shape-constraints on the true function are available in the literature but only for one-dimensional function estimation problems (see e.g., [85, 86, 87, 88]). Dümbgen [64] derived asymptotically optimal confidence bands for the true regression function under shape constraints such as monotonicity and convexity in the continuous univariate white noise model, based on multiscale tests introduced in [66].

Generalizing the approach in [64], we construct a multiscale statistic in the multidimensional setting (also see [65]) which can be written as a supremum of a local weighted average of the response $Y(\cdot)$ with weights determined by a kernel function, parametrized by a vector of smoothing bandwidths and the centers of the kernel function (see (3.4) below). These multiscale local averages are appropriately penalized to have them in the same footing, so that the random fluctuations of the kernel estimators can be bounded uniformly in the bandwidth parameters, i.e., the supremum statistic remains finite almost surely (see [65]). Further, working with the supremum avoids the delicate choice of tuning parameters (smoothing bandwidths). Using the definition of the supremum statistic T (see (3.4) below) and the two-sided bias condition (see (3.5)), one can easily obtain pointwise lower and upper bounds for the function f in terms of $T(\psi^\ell)$ and $T(-\psi^u)$, where ψ^ℓ and ψ^u denote appropriately chosen kernel functions that make the corresponding kernel estimators satisfy the bias condition (3.5) (see Theorem 3.2.3). These upper and lower bounds are random variables depending on $T(\psi^\ell)$ and $T(-\psi^u)$, and hence, one can further bound them with high probability using their respective quantiles, which yields the lower and upper confidence bands for the true function f. Theorem 3.2.1 shows that by choosing the $(1-\alpha)^{\text{th}}$ quantile of the statistic $\max\{T(\psi^\ell), T(-\psi^u)\}$, we can guarantee at least $1-\alpha$ coverage of our constructed confidence band, i.e., (3.2) holds.

Our proposed confidence band automatically adapts to the underlying smoothness of the true function f. In particular, for coordinate-wise isotonic and multivariate convex functions, we show that our constructed confidence band is adaptive with respect to a number of attributes, such as the Hölder smoothness of the underlying function (see Theorem 3.3.1) and the intrinsic dimensionality of the multivariate function, i.e., the number of variables/coordinates it truly depends on (see Theorem 3.3.2). The confidence band also exhibits spatial and local adaptivity (as shown in Theorems 3.2.2 and 3.3.3). To elaborate on the spatial adaptivity property, we show that, as a consequence of Theorem 3.2.2, if the true function is monotone and constant in an open neighborhood, or convex and affine in an open neighborhood, then our constructed confidence band achieves the parametric ($n^{-1/2}$) rate of convergence, uniformly on that neighborhood. Adaptivity with respect to the Hölder smoothness of the true function follows from the fact that the confidence band is constructed in such a way, that the expression for its width involves controlling local variations of the function f in small neighborhoods of width given by the bandwidth parameters. This, by the way, is also the crucial step behind showing spatial and local adaptivities. The variations of f within these small neighborhoods in turn, adapt to its Hölder smoothness/spatial properties, which thus gives rise to adaptivity with respect to the latter.

Our proposed confidence bands have width that diminishes (with n) at the minimax rate on some non-vanishing neighborhood of every point $t_0 \in (0,1)^d$. Moreover, for a coordinate-wise locally strictly increasing function, Theorem 3.4.1 shows that the width of our proposed confidence band also attains the minimax constant up to a multiplicative constant, which is given by the geometric mean of the gradient of the true function f at t_0. Analogously, Theorem 3.4.2 shows that we have a similar minimax property for multivariate convex functions.

The rest of the paper is organized as follows. In Section 3.2.1, we introduce some notation that will be necessary for stating the main results and analyses presented in this paper. In Section 3.2.2, we give the construction of our confidence bands using our multiscale statistic; in Section 3.2.3 we discuss the choice of kernels necessary for this construction, under the natural shape constraints of monotonicity and convexity. Section 3.3 is devoted to proving several adaptivity properties of our constructed confidence band. In Section 3.4, we show that our proposed confidence band is optimal in a certain sense. The proofs of our results are provided in the Appendix.

3.2 Confidence bands for multivariate shape-restricted functions

3.2.1 Definitions and Notation

We now present some notation that we will use throughout the paper.

Notation 1 *We will denote by $\mathcal{F}_1$ the class of all coordinate-wise increasing functions $f : [0,1]^d \to \mathbb{R}$, i.e., functions f that satisfy:*

$$f(x_1, \ldots, x_d) \leq f(y_1, \ldots, y_d) \quad \textit{iff} \quad x_i \leq y_i \, \textit{for all } 1 \leq i \leq d,$$

and by $\mathcal{F}_2$ the class of all convex functions $f : [0,1]^d \to \mathbb{R}$, i.e. functions f that satisfy:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \textit{for all } x, y \in \mathbb{R}^d \textit{ and } \lambda \in [0,1].$$

Notation 2 *For measurable functions $g, h : \mathbb{R}^d \to \mathbb{R}$, we define:*

$$\langle g, h \rangle_{|B} := \int_B g(x)h(x)\, dx \quad \textit{and} \quad \|g\|_{|B} := \sqrt{\langle g, g \rangle_B}\,.$$

When $B := \mathbb{R}^d$, we just drop this subscript.

Notation 3 *For vectors $a, b \in \mathbb{R}^d$, we define:*

$$a \star b := (a_1 b_1, \ldots, a_d b_d).$$

In this section we construct adaptive and optimal confidence bands $(\hat{\ell}, \hat{u})$ for $f : [0,1]^d \to \mathbb{R}$ when f is known to be shape-constrained (e.g., f is (multivariate) isotonic/convex), and our data is generated according to (1.1).

Let us first recall the definitions of A_h in (1.4) and $\psi_{t,h}(x)$ in (1.5). For a fixed $t \in A_h$ we construct a kernel estimator $\hat{f}_h(t)$ of $f(t)$ as:

$$\hat{f}_h(t) := \frac{1}{\sqrt{n}(\prod_{i=1}^d h_i)\langle 1, \psi\rangle} \int_{[0,1]^d} \psi_{t,h}(x)\, dY(x). \tag{3.3}$$

Elementary calculations show that

$$\begin{aligned} \mathbb{E}(\hat{f}_h(t)) &= \frac{\int_{[0,1]^d} \psi_{t,h}(x) f(x)\, dx}{(\prod_{i=1}^d h_i)\langle 1, \psi\rangle} = \frac{\langle f(t + h \star \cdot), \psi\rangle}{\langle 1, \psi\rangle} \\ \mathrm{Var}(\hat{f}_h(t)) &= \frac{\|\psi_{t,h}\|^2}{n(\prod_{i=1}^d h_i)^2 \langle 1, \psi\rangle^2} = \frac{\|\psi\|^2}{n(\prod_{i=1}^d h_i)\langle 1, \psi\rangle^2}. \end{aligned}$$

The main idea of our approach is to notice that the random fluctuations for these kernel estimators can be bounded uniformly in h. To accomplish this, we look at the following multiscale statistic (with kernel ψ):

$$\begin{aligned} T(\pm\psi) &= \sup_{h \in I} \sup_{t \in A_h} \left(\pm \frac{\int_{[0,1]^d} \psi_{t,h}(x)\, dW(x)}{(\prod_{i=1}^d h_i)^{1/2} \|\psi\|} - \Gamma(2^d \prod_{i=1}^d h_i) \right) \\ &= \sup_{h \in I} \sup_{t \in A_h} \left(\pm \frac{\hat{f}_h(t) - \mathbb{E}(\hat{f}_h(t))}{\mathrm{Var}^{1/2}(\hat{f}_h(t))} - \Gamma(2^d \prod_{i=1}^d h_i) \right) \end{aligned} \tag{3.4}$$

where $\Gamma(r) := (2 \log(e/r))^{1/2}$.

3.2.2 Proposed confidence band

We assume that the unknown f belongs to the function class $\mathcal{F}$ (which could be $\mathcal{F}_1$ or $\mathcal{F}_2$). In fact, the results in this section are valid for any function class that satisfies the following *two-sided bias condition*. In particular, we assume that we can find kernels ψ^ℓ and ψ^u such that the corresponding kernel estimators $\hat{f}_h^\ell$ and $\hat{f}_h^u$ (see (1.1.1)) satisfy:

$$\mathbb{E}(\hat{f}_h^\ell(t)) \le f(t) \le \mathbb{E}(\hat{f}_h^u(t)) \quad \text{for all } h \in I,\ t \in A_h \text{ and } f \in \mathcal{F}. \tag{3.5}$$

We will show later that the above condition holds for the function classes $\mathcal{F}_1$ and $\mathcal{F}_2$ (see Section 3.2.3); in fact, it holds for most shape-constrained function classes.

In view of (3.5) and the definition of T, we have the following for all $h \in I$ and $t \in A_h$:

$$\begin{aligned} f(t) &= \left\{f(t) - \mathbb{E}(\hat{f}^l_h(t))\right\} + \left\{\mathbb{E}(\hat{f}^l_h(t)) - \hat{f}^l_h(t)\right\} + \hat{f}^l_h(t) \\ &\geq \hat{f}^\ell_h(t) - \frac{\|\psi^\ell\| \left(T(\psi^\ell) + \Gamma(2^d \prod_{i=1}^d h_i)\right)}{\langle 1, \psi^\ell \rangle (n \prod_{i=1}^d h_i)^{1/2}} \qquad (3.6) \end{aligned}$$

and similarly,

$$f(t) \leq \hat{f}^u_h(t) + \frac{\|\psi^u\| \left(T(-\psi^u) + \Gamma(2^d \prod_{i=1}^d h_i)\right)}{\langle 1, \psi^u \rangle (n \prod_{i=1}^d h_i)^{1/2}}. \qquad (3.7)$$

Now, if κ_α denotes the $(1-\alpha)^{\text{th}}$ quantile of the statistic:

$$T^* := \max\{T(\psi^\ell), T(-\psi^u)\},$$

then in view of (3.6) and (3.7), we can define a $1-\alpha$ confidence band for f as $[\hat{\ell}, \hat{u}]$, where:

$$\hat{\ell}(t) := \sup_{h \in I:\ t \in A_h} \left\{\hat{f}^\ell_h(t) - \frac{\|\psi^\ell\| \left(\kappa_\alpha + \Gamma(2^d \prod_{i=1}^d h_i)\right)}{\langle 1, \psi^\ell \rangle (n \prod_{i=1}^d h_i)^{1/2}}\right\}, \qquad (3.8)$$

$$\hat{u}(t) := \inf_{h \in I:\ t \in A_h} \left\{\hat{f}^u_h(t) + \frac{\|\psi^u\| \left(\kappa_\alpha + \Gamma(2^d \prod_{i=1}^d h_i)\right)}{\langle 1, \psi^u \rangle (n \prod_{i=1}^d h_i)^{1/2}}\right\}. \qquad (3.9)$$

In view of (3.6) and (3.7), we have:

$$\begin{aligned} \mathbb{P}_f\left(\hat{\ell}(t) \leq f(t) \leq \hat{u}(t) \text{ for all } t \in [0,1]^d\right) &\geq \mathbb{P}\left(T(\psi^\ell) \leq \kappa_\alpha\ ,\ T(-\psi^u) \leq \kappa_\alpha\right) \\ &= \mathbb{P}\left(T^* \leq \kappa_\alpha\right) = 1 - \alpha. \qquad (3.10) \end{aligned}$$

This shows that $[\hat{\ell}, \hat{u}]$ is indeed a confidence band with guaranteed coverage probability $1-\alpha$ for all $n \geq 1$, which we state formally below.

Theorem 3.2.1 *For kernels ψ^ℓ and ψ^u satisfying* (3.5)*, we have:*

$$\mathbb{P}_f\left(\hat{\ell}(t) \leq f(t) \leq \hat{u}(t) \text{ for all } t \in [0,1]^d\right) \geq 1-\alpha \tag{3.11}$$

for all $f \in \mathcal{F}$ and for all $n \geq 1$.

The above theorem shows that for any function class $\mathcal{F}$ for which the two-sided bias bounds (3.5) hold, our approach yields an honest finite sample confidence band for any $f \in \mathcal{F}$. It is natural to ask if the above constructed band is conservative in nature. In the following result, we show that if for some function $f \in \mathcal{F}$, the function $-f$ also belongs to $\mathcal{F}$, then our confidence band has *exact* coverage at f.

Proposition 3.2.1 *Suppose $f, -f \in \mathcal{F}$ and* (3.5) *holds. Then, our* $1-\alpha$ *confidence band* $[\hat{\ell}, \hat{u}]$ *has exact coverage probability* $1-\alpha$*, i.e.,*

$$\mathbb{P}_f\left(\hat{\ell}(t) \leq f(t) \leq \hat{u}(t) \text{ for all } t \in [0,1]^d\right) = 1-\alpha.$$

Proposition 3.2.1 shows that if $f \in \mathcal{F}_1$ is a constant or $f \in \mathcal{F}_2$ is an affine function, then the coverage probability of our confidence band is exact. We will now see that for certain functions that exhibit “simple" structure locally (for example, $f \in \mathcal{F}_1$ is locally constant or $f \in \mathcal{F}_2$ is locally affine), our confidence band exhibits adaptive rates, in particular it can shrink at the parametric $n^{-1/2}$ rate locally.

Theorem 3.2.2 *In addition to assuming* (3.5)*, suppose that the true f satisfies*

$$\mathbb{E}(\hat{f}^{\ell}_{\varepsilon_n \mathbf{1}_d}(t)) = f(t) = \mathbb{E}(\hat{f}^{u}_{\varepsilon_n \mathbf{1}_d}(t)) \tag{3.12}$$

for some $\varepsilon_n > 0$ and all $t \in D$ for some $D \subseteq A_{\varepsilon_n \mathbf{1}_d}$. Then, for $\varepsilon_n \equiv \varepsilon$ for some constant ε and

$\varepsilon_n = (\log(en))^{-\frac{1}{d}}$, we have respectively,

$$\sup_{t \in D} \ (\hat{u}(t) - \hat{\ell}(t)) \leq K_\varepsilon n^{-1/2} \left(\kappa_\alpha + T(\psi_k^u) + T(-\psi_k^\ell) \right)$$

for some constant $K_\varepsilon > 0$ depending on $\varepsilon, \psi_k^u, \psi_k^\ell$, and

$$\sup_{t \in D} \ (\hat{u}(t) - \hat{\ell}(t)) \leq K \rho_n \left(1 + \frac{\kappa_\alpha + T(\psi_k^u) + T(-\psi_k^\ell)}{\sqrt{\log \log(en)}} \right)$$

for some constant $K > 0$ depending on ψ_k^u, ψ_k^ℓ, where $\rho_n := (\log(en) \log \log(en)/n)^{1/2}$.

Under the bias assumption (3.5), a sufficient condition for (3.12) to hold is if both f and $-f$, restricted to the set D, belong to the class $\mathcal{F}$. In particular, the above result shows that if the true function $f \in \mathcal{F}_1$ is locally constant at a point, or if the true function $f \in \mathcal{F}_2$ is locally affine in a fixed neighborhood, then our confidence band automatically adapts to this structure, and shrinks at the parametric rate $n^{-1/2}$ uniformly on this neighborhood. In fact, a similar result also holds on shriking neighborhoods of radius $(\log(en))^{-1/d}$, modulo the fact that the rate of convergence now suffers an inflation by a logarithmic factor in n.

3.2.3 Choice of kernels for function classes $\mathcal{F}_1$ and $\mathcal{F}_2$

As we have mentioned in the Introduction, the two prime examples of shape-constrained function classes are: (1) the class of all d-dimensional coordinate-wise increasing functions $\mathcal{F}_1$, and (2) the class of all d-dimensional convex functions $\mathcal{F}_2$. In this subsection, we construct kernels ψ^ℓ and ψ^u for each of the function classes $\mathcal{F}_1$ and $\mathcal{F}_2$, that satisfy (3.5). This would immediately imply that we can construct honest confidence bands for these function classes that satisfy (3.11). Note that for construction of the confidence bands with a guaranteed coverage probability, we do not require any specific choice of kernels, as long as they satisfy (3.5). However, we are going to work with some specific choices of kernels such that the corresponding confidence bands exhibit certain optimality properties. For the class $\mathcal{F}_1$ of all coordinate-wise increasing functions, we define:

$$\psi_1^u(x) := \left(1 - \sum_{i=1}^d x_i\right) \mathbb{1}_{x\in[0,\infty)^d,\ \sum_{i=1}^d x_i \le 1} \quad \text{and} \quad \psi_1^\ell(x) := \left(1 + \sum_{i=1}^d x_i\right) \mathbb{1}_{x\in(-\infty,0]^d,\ \sum_{i=1}^d x_i \ge -1} \tag{3.13}$$

and for the class $\mathcal{F}_2$ of all convex functions, we define:

$$\psi_2^u(x) := (1 - \|x\|^2)\mathbb{1}_{\|x\|\le 1} \quad \text{and} \quad \psi_2^\ell(x) := \left(1 - \frac{2d+4}{d+1}\|x\| + \frac{d+3}{d+1}\|x\|^2\right) \mathbb{1}_{\|x\|\le 1}. \tag{3.14}$$

Note that ψ_2^ℓ can take negative values, too. Theorem 3.2.3 is proved in Section 4.6.3.

Theorem 3.2.3 *Let $\hat{f}_{h,k}^\ell$ and $\hat{f}_{h,k}^u$ denote the kernel estimators corresponding to the kernels ψ_k^ℓ and ψ_k^u, for $k \in \{1, 2\}$. Then,* (3.5) *holds for the function classes $\mathcal{F}_1$ and $\mathcal{F}_2$.*

3.3 Adaptivity of the confidence band

In this section we show that the width of our confidence band $[\hat{\ell}, \hat{u}]$ (see (3.8) and (3.9)) adapts to the smoothness and the *intrinsic dimension* of the true function f. Let us first define the rate of convergence for a confidence band as follows:

We say that the confidence band $\{[\ell(t), u(t)] : t \in [0, 1]^d\}$, with coverage probability $1 - \alpha$, has rate of convergence γ_n on a set $A_n \subseteq [0, 1]^d$ for a class $\mathcal{G}$ of functions if

$$\inf_{f\in\mathcal{G}} \mathbb{P}_f \left(\sup_{t\in A_n}(u(t) - \ell(t)) \le \Delta\gamma_n\right) \ge 1 - \alpha, \qquad \text{for all } n$$

where $\Delta > 0$ is a constant not depending on n (but may depend on α and $\mathcal{G}$). Clearly we want the rate of convergence to be as small as possible.

3.3.1 Adaptivity with respect to the smoothness of the underlying function

Let us first define the notion of Hölder smoothness of a function $f : [0, 1]^d \to \mathbb{R}$.

Definition 3.3.1 *For every fixed $\beta > 0$ and $L > 0$, the Hölder class $\mathbb{H}_{\beta,L}$ on $[0, 1]^d$ is defined as* *the set of all functions $f : [0,1]^d \to \mathbb{R}$ that have all partial derivatives of order $\lfloor \beta \rfloor$ (defined as the largest integer strictly less than β) on $[0,1]^d$, and satisfy:*

$$\sum_{k \in \mathbb{N}^d : \|k\|_1 \leq \lfloor \beta \rfloor} \sup_{x \in [0,1]^d} \left| \frac{\partial^{\|k\|_1} f(x)}{\partial x_1^{k_1} \dots \partial x_d^{k_d}} \right| \leq L$$

and

$$\sum_{k \in \mathbb{N}^d : \|k\|_1 = \lfloor \beta \rfloor} \left| \frac{\partial^{\|k\|_1} f(y)}{\partial x_1^{k_1} \dots \partial x_d^{k_d}} - \frac{\partial^{\|k\|_1} f(z)}{\partial x_1^{k_1} \dots \partial x_d^{k_d}} \right| \leq L \|y - z\|^{\beta - \lfloor \beta \rfloor} \quad \textit{for all } y, z \in [0,1]^d \ .$$

The following theorem shows that the rate of convergence of our confidence band $[\hat{\ell}, \hat{u}]$ for the class $\mathbb{H}_{\beta,L} \cap \mathcal{F}_k$ $(k = 1, 2)$ is $(\log n/n)^{\beta/(2\beta+d)}$.

Theorem 3.3.1 *Suppose that for some $k \in \{1, 2\}$, $f \in \mathcal{F}_k \cap \mathbb{H}_{\beta,L}$ with $k - 1 < \beta \leq k$ and $L > 0$. Then there exist a constant $\Delta > 0$ depending only on $L, \beta, \psi^\ell, \psi^u$ such that*

$$\sup_{t \in A_{\varepsilon_n \mathbf{1}_d}} \left(\hat{u}(t) - \hat{\ell}(t) \right) \leq \Delta \varepsilon_n^\beta \left(1 + \frac{\kappa_\alpha + T(-\psi^\ell) + T(\psi^u)}{(\log(en))^{1/2}} \right)$$

where $\varepsilon_n := (\log(en)/n)^{1/(2\beta+d)}$, and $\mathbf{1}_d$ is the d-dimensional vector of all ones. This, in particular, implies that

$$\inf_{f \in \mathcal{F}_i \cap \mathbb{H}_{\beta,L}} \mathbb{P} \left(\sup_{t \in A_{\varepsilon_n \mathbf{1}_d}} (\hat{u}(t) - \hat{\ell}(t)) \leq \Delta \left[\frac{\log(en)}{n} \right]^{\frac{\beta}{2\beta+d}} \right) \geq 1 - \alpha, \quad \textit{for all } n,$$

for some constant $\Delta > 0$ depending only on $L, \beta, \psi^\ell, \psi^u, \alpha$. Here we would like to point out that ψ^ℓ, ψ^u depend on the choice of the function class $\mathcal{F}_1, \mathcal{F}_2$.

Theorem 3.3.1 is proved in Section 4.6.4. Its proof starts by bounding the pointwise deviation of the upper (and lower) band of our constructed confidence set from the true function f, in terms of the inner product of the variation of f in a small neighborhood. The variation of f over this neighborhood can then be bounded in terms of appropriate powers of the smoothing bandwidth, using the Hölder smoothness of f. The rate of convergence obtained in Theorem 3.3.1 is minimax in the class $\mathbb{H}_{\beta,L}$ (see [89]).

3.3.2 Adaptivity with respect to the intrinsic dimension

The intrincic dimension of a function refers to the number of variables it actually depends on.

Definition 3.3.2 *The intrinsic dimension* $\dim(f)$ *of a function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* k *iff:*

1. *there exist* $1 \leq i_1 < i_2 < \ldots < i_k \leq d$ *and a function* $g : \mathbb{R}^k \to \mathbb{R}$ *such that* $f(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_k})$ *for all* $x \in \mathbb{R}^d$*, and*
2. f *is not a function of* $(x_s)_{s \in S}$ *for any strict subset* S *of* $\{i_1, \ldots, i_k\}$.

It can be verified easily from Definition 3.3.2 that the intrinsic dimension of a function f is unique. We will now show that our confidence band $[\hat{\ell}, \hat{u}]$ adapts to the intrinsic dimension of the true function f.

Theorem 3.3.2 *Suppose that for some* $j \in \{1, 2\}$, $f \in \mathcal{F}_j \cap \mathbb{H}_{\beta,L}$ *with* $j - 1 < \beta \leq j$ *and* $L > 0$. *Let* $\dim(f) = k$ *and suppose that* $f(x) = g(x_{i_1}, \ldots, x_{i_k})$ *for some* $1 \leq i_1 < \ldots < i_k \leq d$ *and some function* $g : [0, 1]^k \to \mathbb{R}$. *Then, for every* $\varepsilon > 0$, *there exists a constant* $\Delta > 0$ *depending only on* $L, \beta, \psi^\ell, \psi^u, \varepsilon$ *such that:*

$$\sup_{t \in A_{\boldsymbol{\varepsilon}_{n,i_1,\ldots,i_k}}} \left(\hat{u}(t) - \hat{\ell}(t)\right) \leq \Delta \rho_{n,k} \left(1 + \frac{\kappa_\alpha + T(-\psi^\ell) + T(\psi^u)}{(\log(en))^{1/2}}\right)$$

where $\rho_{n,k} := (\log(en)/n)^{\beta/(2\beta+k)}$, $\varepsilon_n := \varepsilon \rho_{n,k}^{1/\beta}$ *and* $\boldsymbol{\varepsilon}_{n,i_1,\ldots,i_k}$ *is the* d*-dimensional vector with the* $i_1^{\text{th}}, \ldots, i_k^{\text{th}}$ *entries all equal to* ε_n *and all other entries equal to* ε.

Theorem 3.3.2 is proved in Section 4.6.5. It states that the rate of convergence of our confidence band when the function f lies in $\mathcal{F}_j \cap \mathbb{H}_{\beta,L}$ with $j-1 < \beta \leq j$ and $L > 0$, is $(\log(en)/n)^{\beta/(2\beta+\dim(f))}$. This is a highly desirable property, as the rate of convergence should depend only on the variables which actually affect the true function, and not on the redundant variables that the function does not vary with, something that one does not get directly from the statement of Theorem 3.3.1.

3.3.3 Local Adaptivity

Our results in Section 3.3.1 showed that our confidence bands achieve the optimal rate of convergence when the true function is globally Hölder smooth. In this section we show that our adaptivity results hold when the true function is locally smooth. Specifically, we will look at the behavior of $\hat{u}(t_0) - \hat{\ell}(t_0)$ for a fixed $t_0 \in (0,1)^d$.

Theorem 3.3.3 *Suppose that $f \in \mathcal{F}_k$ $(k = 1,2)$, and that there exists $\varepsilon > 0$ such that f is Hölder smooth on $\bar{B}_\infty(t_0, \varepsilon)$ with smoothness parameter $\beta \in (k-1, k]$ and $L > 0$. Then there exists a constant K depending on $\beta, L, \psi^\ell, \psi^u$ such that:*

$$\hat{u}(t_0) - \hat{\ell}(t_0) \leq K\rho_n \left(1 + \frac{\kappa_\alpha + T(-\psi^\ell) + T(\psi^u)}{(\log(en))^{1/2}}\right)$$

where $\rho_n = (\log(en)/n)^{\beta/(2\beta+d)}$. Note that this implies that

$$\mathbb{P}_f\left(\hat{u}(t_0) - \hat{\ell}(t_0) \leq \Delta\Big[\frac{\log(en)}{n}\Big]^{\frac{\beta}{2\beta+d}}\right) \geq 1 - \alpha, \quad \textit{for all } n,$$

for some constant $\Delta > 0$ depending only on $L, \beta, \psi^l, \psi^u, \alpha$.

Remark 3.3.1 (On the proof of Theorem 3.3.3) *For proving Theorem 3.3.3, first one needs to observe that for $h = \varepsilon\mathbf{1}_d$, we have $\|h \star x\|_\infty = \varepsilon\|x\|_\infty \leq \varepsilon$, and hence, $t_0 + h \star x \in \bar{B}_\infty(t_0, \varepsilon)$. The rest of the proof follows exactly as the proof of Theorem 3.3.1, on noting that one only needs the Hölder smoothness assumption for bounding the terms $|f(t_0 + h \star x) - f(t_0)|$ and $\|\nabla f(t_0 + h \star \xi_x) - \nabla f(t_0)\|$ for some ξ_x lying in the segment joining $\mathbf{0}$ and x, and consequently, it is enough to have Hölder smoothness on $\bar{B}_\infty(t_0, \varepsilon)$ only.*

3.4 Optimality of the Confidence Band

In this section, we prove that our proposed confidence band (3.8), (3.9) is optimal in a certain sense. Our results extend Theorem 4.2 in [64]. In order to state our result, we need some notation.

For a function $g : \mathbb{R}^d \to \mathbb{R}$ and $U \subseteq \mathbb{R}^d$, define:

$$\|g\|_U := \sup_{x \in U} |g(x)|.$$

We first state our optimality result for the class of monotone functions $\mathcal{F}_1$.

Theorem 3.4.1 *Let $f \in \mathcal{F}_1$ be a continuously differentiable function in an open neighborhood U of $t_0 \in (0,1)^d$ such that*

$$L_1 = L_1[f, t_0] := \left[\prod_{i=1}^{d} \frac{\partial}{\partial x_i} f(x)\Big|_{x=t_0} \right]^{1/d} > 0.$$

Define

$$\rho_n := \left(\frac{\log(en)}{n} \right)^{\frac{1}{2+d}} \quad \textit{and} \quad \Delta^{(z)} := \left(\left(\frac{d+2}{2d} \right) \|\psi^z\|^2 \right)^{-\frac{1}{2+d}}$$

where z stands for u and ℓ corresponding to kernels ψ_1^u and ψ_1^ℓ (respectively) as defined in (3.13). *Then we have the following:*

(a) Let (ℓ, u) be any confidence band such that, for some $\alpha \in (0,1)$,

$$\mathbb{P}_f\big(\ell(t) \le f(t) \le u(t) \textit{ for all } t \in [0,1]^d\big) \ge 1 - \alpha \quad \textit{for all } f \in \mathcal{F}_1.$$

Then, for any $\epsilon > 0$,

$$\liminf_{n\to\infty} \mathbb{P}_f \left(\|f - \ell\|_U \ge (1-\epsilon)\Delta^{(\ell)} L_1^{\frac{d}{2+d}}[t_0]\rho_n \right) \ge 1 - \alpha,$$
$$\liminf_{n\to\infty} \mathbb{P}_f \left(\|u - f\|_U \ge (1-\epsilon)\Delta^{(u)} L_1^{\frac{d}{2+d}}[t_0]\rho_n \right) \ge 1 - \alpha.$$

(b) Moreover, let $(\hat{\ell}, \hat{u})$ be the confidence band with coverage probability $1 - \alpha$ as defined in

(3.8) *and* (3.9)*, with kernels as in* (3.13)*. Then, for any* $\epsilon > 0$*, we have*

$$\lim_{n\to\infty} \mathbb{P}_f \left((f - \hat{\ell})(t_0) \leq (1+\epsilon)\Delta^{(\ell)} L_1^{\frac{d}{2+d}}[f, t_0]\rho_n \right) = 1,$$
$$\lim_{n\to\infty} \mathbb{P}_f \left((\hat{u} - f)(t_0) \leq (1+\epsilon)\Delta^{(\ell)} L_1^{\frac{d}{2+d}}[f, t_0]\rho_n \right) = 1.$$

Theorem 3.4.1 is proved in Section 4.6.6. It states that the *length* any confidence band of $f \in \mathcal{F}_1$ with guaranteed coverage probability $1-\alpha$, is at least $(\log n/n)^{1/(2+d)}$ upto a constant factor. Further, this optimal length is achieved by our constructed confidence band, again up to a constant multiplicative factor. We can, in particular, exactly compute the constants $\Delta^{(\ell)}$ and $\Delta^{(u)}$ appearing in the above result. For example, for $d = 2$, $\Delta^{(\ell)} = \Delta^{(u)} \approx 1.86121$.

Remark 3.4.1 (On Theorem 3.4.1) *Note that the asymptotic probabilities in the upper bound results in part (b) of Theorem 3.4.1 do not depend on* α*, unlike the corresponding lower bound probabilities in part (a). This can be understood from the fact that the random variable in part (a) is a supremum of pointwise deviations of the lower and upper confidence bands from the true function over a neighborhood, unlike the corresponding random variable in part (b), and hence, is intrinsically 'larger'. To draw a simple analogy, note that the* α^{th} *quantile* $q_\alpha^{(n)}$ *of the maximum of a sequence of i.i.d. Gaussians* $Z_1, \ldots, Z_n$ *satisfies* $\mathbb{P}(\max_{1\leq i\leq n} Z_i \geq q_\alpha^{(n)}) = 1-\alpha$*, but since* $q_\alpha^{(n)}$ *is of the order* $\sqrt{\log n}$*,* $\mathbb{P}(Z_1 \leq q_\alpha^{(n)}) = 1 - o(1)$*.*

Next, we state our optimality result for the class of convex functions $\mathcal{F}_2$.

Theorem 3.4.2 *Let* $f \in \mathcal{F}_2$ *be a twice continuously differentiable function in an open neighborhood* U *of* $t_0 \in (0,1)^d$ *such that*

$$L_2 = L_2[f, t_0] := \det(H(t_0))^{1/d} > 0$$

where $H(t_0) \in \mathbb{R}^{d\times d}$ *denotes the Hessian of* f *at* t_0*, and* $\det(\cdot)$ *denotes the determinant operator.*

Define

$$\rho_n := \left(\frac{\log(en)}{n}\right)^{\frac{2}{4+d}} \quad \textit{and} \quad \Delta^{(z)} := \left(\frac{d+4}{2d}\sqrt{\frac{2(d+3)}{d+1}}\|\psi^z\|^2\right)^{-\frac{2}{4+d}}$$

where z stands for u and ℓ corresponding to kernels ψ_2^u and ψ_2^ℓ (respectively) as defined in (3.14). *Then for any confidence band (ℓ, u) such that, for some $\alpha \in (0,1)$,*

$$\mathbb{P}_f\big(\ell(t) \le f(t) \le u(t) \textit{ for all } t \in [0,1]^d\big) \ge 1-\alpha \quad \textit{for all } f \in \mathcal{F}_2,$$

we have for any $\epsilon > 0$,

$$\liminf_{n\to\infty} \mathbb{P}_f\left(\|f-\ell\|_U \ge (1-\epsilon)\Delta^{(\ell)} L_2^{\frac{d}{4+d}}[t_0]\rho_n\right) \ge 1-\alpha,$$
$$\liminf_{n\to\infty} \mathbb{P}_f\left(\|u-f\|_U \ge (1-\epsilon)\Delta^{(u)} L_2^{\frac{d}{4+d}}[t_0]\rho_n\right) \ge 1-\alpha.$$

The above result gives a lower bound on the maximal (local) deviation of any honest confidence band (for the class of convex functions) around the true function. We can, in particular, exactly compute the constants $\Delta^{(\ell)}$ and $\Delta^{(u)}$ appearing in the above result. For example, for $d = 2$, $\Delta^{(\ell)} \approx 1.464067$ and $\Delta^{(u)} \approx 0.70385$.

Remark 3.4.2 (On the proofs of Theorems 3.4.1 and 3.4.2) *The proofs of the lower bound results for both Theorems 3.4.1 and 3.4.2 involve the following main ideas. As a first step, one constructs a grid with spacing corresponding to the bandwidth vector h starting from the center point t_0, and for each such grid point t, defines a function f_t by perturbing the true function f by an amount proportional to the kernel function corresponding to t and h. The proportionality constant c depends on the particular shape constraint, and involves either the minimum entry of the gradient of f at t_0 or the minimum eigenvalue of its Hessian at t_0, depending on whether the true function is monotone or convex, respectively. The second step is to show that all these perturbed functions satisfy the corresponding shape constraint. As a next step, one shows that the deviation of the upper and lower limits of any honest confidence band with coverage probability* $1-\alpha$ *can be lower bounded by the proportionality constant c with probability at least* $1-\alpha$ *minus some* *remainder term, that depends on the perturbed function. One then argues that this remainder term is asymptotically negligible, by expressing it in terms of an average of the likelihood ratio between the measures at the perturbed and the true function, and applying the Cameron-Martin-Girsanov theorem in stochastic calculus to evaluate this likelihood ratio. As a final step, several parameters are tuned appropriately to control the proportionality constant c so as to obtain the optimal constant.*

3.5 Construction and adaptivity of the confidence band under additive models

In this section we try to construct a confidence band for our unknown function f as described in (1.1) under the additional assumption that f is of the form

$$f(x_1, \ldots, x_d) = \mu + \sum_{i=1}^{d} f_i(x_i) \tag{3.15}$$

where for all $i = 1, \ldots, d$ $f_i : [0,1] \to \mathbb{R}$ is non-decreasing or convex and $\mu \in \mathbb{R}$. For identifiability we also assume that for all $i = 1, \ldots, d$ we have

$$\int_0^1 f_i(y)\, dy = 0.$$

We do assume that $f \in \mathcal{F}$ for some shape restricted function class $\mathcal{F}$ like non-decreasing or convex. We construct the confidence band under the additive model as follows: First estimate μ by

$$\hat{\mu} := \frac{1}{\sqrt{n}} \int_{[0,1]^d} dY.$$

Elementary calculations show that $\mathbb{E}(\hat{\mu}) = \mu$ and $\mathrm{Var}(\hat{\mu}) = 1/n$. Now we will construct the confidence band for $f_i(t_i)$ using the kernel $\psi : [-1,1] \to \mathbb{R}$. Fix $h_i > 0$ and suppose $t_i \in [h_i, 1 - h_i]$. Now we can estimate $f_i(t_i)$ by

$$\hat{f}^{(i)}_{h_i}(t_i) := \frac{1}{n^{1/2} h \langle 1, \psi \rangle} \int_{[0,1]^d} \psi^{(i)}_{t_i,h_i}(x) dY(x) - \hat{\mu}$$

where $\hat{\psi}^{(i)}_{t_i,h_i}(x) = \psi((x_i - t_i)/h_i)$. Elementary calculations can easily show that

$$\mathbb{E}(\hat{f}^{(i)}_{h_i}(t_i)) = \frac{\int_0^1 \psi^{(i)}_{t,h}(x)f(x)dx}{h\langle 1, \psi\rangle}$$

and

$$\mathrm{Var}(\hat{f}^{(i)}_{h_i}(t_i) + \hat{\mu}) = \frac{\|\psi\|^2}{nh\langle 1, \psi\rangle^2}.$$

Now the main idea that we will use to construct our confidence bands is that the random fluctuations can be bounded uniformly i.e., we claim that

$$T^{(i)}(\pm\psi) := \sup_{h>0} \sup_{t\in[h,1-h]} \left(\frac{\hat{f}^{(i)}_{h_i}(t_i) - \mathbb{E}(\hat{f}^{(i)}_{h_i}(t_i))}{\mathrm{Var}^{1/2}(\hat{f}^{(i)}_{h_i}(t_i) + \hat{\mu})} - \Gamma(2h) \right) < \infty.$$

The above assertion can be easily proven from the fact that

$$\sup_{h>0} \sup_{t\in[h,1-h]} \left(\frac{\hat{f}^{(i)}_{h_i}(t_i) + \hat{\mu} - \mathbb{E}(\hat{f}^{(i)}_{h_i}(t_i)) - \mu}{\mathrm{Var}^{1/2}(\hat{f}^{(i)}_{h_i}(t_i) + \hat{\mu})} - \Gamma(2h) \right) < \infty$$

and $\hat{\mu} - \mu$ is a normal variable (i.e., finite almost surely). As we have done previously we will choose a kernels ψ^u and ψ^l such that the bias of the estimators are controlled i.e., we need

$$\mathbb{E}(\hat{f}^{(i),(u)}_{h_i}(t_i)) \geq f_i(t_i) \geq \mathbb{E}(\hat{f}^{(i),(l)}_{h_i}(t_i)) \text{ for all } h > 0, t_i \in [h, 1-h], f \in \mathcal{F}. \tag{3.16}$$

Now let $\kappa^{(i)}_\alpha$ be the $(1 - \alpha)$ quantile of the combined statistic

$$T^{\star,(i)} := \max(T^{(i)}(\psi^l), T^{(i)}(-\psi^u)).$$

Hence by similar argument as used in Theorem 3.2.1 we can construct the optimal band for f_i as

$$\hat{\ell}_i(t_i) = \sup_{h>0:t\in A_h} \left\{ \hat{f}^{(i)}_{h_i}(t_i) - \frac{\|\psi^l\| \left(\kappa^{(i)}_\alpha + \Gamma(2h)\right)}{\langle 1, \psi^l\rangle (nh)^{1/2}} \right\}$$

and

$$\hat{u}_i(t_i) = \inf_{h>0:t\in A_h} \left\{ \hat{f}^{(i)}_{h_i}(t_i) + \frac{\|\psi^u\| \left(\kappa^{(i)}_\alpha + \Gamma(2h)\right)}{\langle 1, \psi^u \rangle (nh)^{1/2}} \right\}.$$

The confidence band for the the combined function f can also be defined in a similar manner. Let κ^a_α be the $(1-\alpha)$ quantile of the combined statistic

$$T^{\star,a} := \max \left\{ \hat{\mu} + \sum_{i=1}^{d} T^{(i)}(\psi^l), \hat{\mu} + \sum_{i=1}^{d} T^{(i)}(-\psi^u) \right\}.$$

Then the confidence band for the function f is given by

$$\hat{\ell}^a(t) = \sup_{h>0:t\in A_h} \left\{ \hat{\mu} + \sum_{i=1}^{d} \hat{f}^{(i)}_{h_i}(t_i) - \frac{\|\psi^l\| \left(\kappa^a_\alpha + \Gamma(2h)\right)}{\langle 1, \psi^l \rangle (nh)^{1/2}} \right\}$$

and

$$\hat{u}^a(t) = \inf_{h>0:t\in A_h} \left\{ \hat{\mu} + \sum_{i=1}^{d} \hat{f}^{(i)}_{h_i}(t_i) + \frac{\|\psi^u\| \left(\kappa^a_\alpha + \Gamma(2h)\right)}{\langle 1, \psi^u \rangle (nh)^{1/2}} \right\}.$$

Our next Theorem shows that our bands $\hat{\ell}_i$ and $\hat{u}_i$ and $\hat{\ell}^a$ and $\hat{u}^a$ are actually honest confidence bands for the respective functions when the function is shape-restricted.

Theorem 3.5.1 *Let f satisfy* (3.15) *and $f \in \mathcal{F}$. Suppose that we observe the stochastic process Y as given in* (1.1)*. Suppose that we can find kernels that satisfy* (3.16)*. Let $(\hat{\ell}^a, \hat{u}^a) : [0,1]^d \to \mathbb{R}\times\mathbb{R}$ and $(\hat{u}_i, \hat{\ell}_i) : [0,1] \to \mathbb{R}\times\mathbb{R}$ be defined as above. Then*

$$\mathbb{P}_f \left(\hat{\ell}^a(t) \le f(t) \le \hat{u}^a(t) \quad \textit{for all } t \in [0,1]^d \right) \ge 1-\alpha, \qquad \textit{for all } f \in \mathcal{F}$$

and

$$\mathbb{P}_f \left(\hat{\ell}_i(t) \le f_i(t) \le \hat{u}_i(t) \quad \textit{for all } t \in [0,1] \right) \ge 1-\alpha, \qquad \textit{for all } f \in \mathcal{F}$$

3.5.1 Adaptivity under additive models

Our next theorems show that for the additive model ((3.15)) our confidence bands $\hat{\ell}_i$ and $\hat{u}_i$ $\hat{\ell}^a(t)$ and $\hat{u}^a(t)$ achieves the optimal rate of convergence (i.e., $(\log n/n)^{\beta/2\beta+1}$) for the additive model. Here we would like to point out that for the construction of the confidence band we needed the additional knowledge that the function is additive. Without this knowledge our generic confidence band (i.e., $\hat{\ell}$ and $\hat{u}$) do not attain the required optimal rate of convergence but attains the rate of $(\log n/n)^{\beta/2\beta+d}$.

Theorem 3.5.2 *Suppose that f is additive as defined in* (3.15). *Also assume that the function $f_i \in \mathbb{H}_{\beta,L}$ where $0 < \beta \leq 1$ is f_i is non-decreasing or $1 < \beta \leq 2$ if f_i is convex and $L > 0$. Let $\delta_n := (\log(en)/n)^{1/2\beta+1}$. Then there exist a constant K depending on L, β, ψ^l, ψ^u only such that*

$$\sup_{t\in(\delta_n,1-\delta_n)} (\hat{u}_i(t) - \hat{\ell}_i(t)) \leq K\left(\frac{\log(en)}{n}\right)^{\frac{\beta}{2\beta+1}}\left(1 + \frac{\kappa_\alpha^a + W_i}{\log^{1/2}(en)}\right)$$

for all n, where $W_i := T^{(i)}(-\psi^l) + T^{(i)}(\psi^u)$. Note that W_i is an almost sure finite random variable.

Theorem 3.5.3 *Suppose that f is additive as defined in* (3.15). *Also assume that the function $f \in \mathbb{H}_{\beta,L}$ where $0 < \beta \leq 1$ is f is non-decreasing or $1 < \beta \leq 2$ if f is convex and $L > 0$. Let $\delta_n := (\log(en)/n)^{1/2\beta+1}$. Then there exist a constant K depending on L, β, ψ^l, ψ^u only such that*

$$\sup_{t\in A_{(\delta_n,\ldots,\delta_n)}} (\hat{u}^a(t) - \hat{\ell}^a(t)) \leq K\left(\frac{\log(en)}{n}\right)^{\frac{\beta}{2\beta+1}}\left(1 + \frac{\kappa_\alpha^a + W}{\log^{1/2}(en)}\right)$$

for all n, where $W := 2(\hat{\mu} - \mu) + \sum_{i=1}^d \left(T^{(i)}(-\psi^l) + T^{(i)}(\psi^u)\right)$. Note that W is an almost sure finite random variable.

Note that the above theorem implies that for any $\gamma \in (0,1)$ there exist a constant K_γ depending on $\gamma, \beta, L, \psi^l, \psi^u$ such that

$$\mathbb{P}\left(\sup_{t\in A_{(\delta_n,\ldots,\delta_n)}} (\hat{u}^a(t) - \hat{\ell}^a(t)) \leq K_\gamma(\log(en)/n)^{\beta/(2\beta+1)}\right) \geq 1 - \gamma.$$

The construction of the above confidence bands require the knowledge that the function is additive and either non-decreasing or convex; no knowledge of the smoothness of the function is required.

Our next two theorems extend our local adaptivity properties to the case of additive models as well.

Theorem 3.5.4 *Suppose $f \in \mathcal{F}_1$ or $\mathcal{F}_2$. Also assume that f satisfies* (3.15). *Fix $i \in \{1, 2, \ldots, d\}$ and $t_0 \in (0, 1)$. Suppose there exists $\epsilon > 0$ such that the function f_i is Hölder smooth on $(t_0 - \epsilon, t_0 + \epsilon) \subset [0, 1]$ with smoothness parameter $\beta \in (j - 1, j]$ if $f \in \mathcal{F}_j$ for $j = 1, 2$ and $L > 0$. Then for some constant K depending on β, L, ψ^l, ψ^u we have*

$$\hat{u}_i(t_0) - \hat{\ell}_i(t_0) \leq K\rho_n \left(1 + \frac{\kappa_\alpha + T^{(i)}(-\psi^l) + T^{(i)}(\psi^u)}{(\log(en))^{1/2}}\right)$$

where $\rho_n = (\log(en)/n)^{\beta/(2\beta+1)}$. Note that this implies that

$$\mathbb{P}_f \left(\hat{u}_i(t_0) - \hat{\ell}_i(t_0) \leq \Delta \left[\frac{\log(en)}{n}\right]^{\frac{\beta}{2\beta+1}}\right) \geq 1 - \alpha, \quad \textit{for all } n,$$

for some constant Δ depending only on $L, \beta, \psi^l, \psi^u, \alpha$.

Theorem 3.5.5 *Suppose $f \in \mathcal{F}_1$ or $\mathcal{F}_2$. Also assume that f satisfies* (3.15). *Fix $t_0 := (t_0^1, \ldots, t_0^d) \in (0, 1)^d$. Suppose there exists $\epsilon > 0$ such that the functions f_i's are Hölder smooth on $(t_0^i - \epsilon, t_0^i + \epsilon) \subset [0, 1]$ with smoothness parameter $\beta \in (j - 1, j]$ if $f \in \mathcal{F}_j$ for $j = 1, 2$ and $L > 0$ for all $i = 1, \ldots, d$. Note here we have assumed that all the component functions f_i have the same smoothness parameter β close to t_0. Then for some constant K depending on β, L, ψ^l, ψ^u we have*

$$\hat{u}^a(t_0) - \hat{\ell}^a(t_0) \leq K\rho_n \left(1 + \frac{\kappa_\alpha + W}{(\log(en))^{1/2}}\right)$$

where $\rho_n = (\log(en)/n)^{\beta/(2\beta+1)}$ and $W = 2(\hat{\mu} - \mu) + \sum_{i=1}^d \left(T^{(i)}(-\psi^l) + T^{(i)}(\psi^u)\right)$. Note that this *implies that*

$$\mathbb{P}_f\left(\hat{u}^a(t_0) - \hat{\ell}^a(t_0) \leq \Delta\left[\frac{\log(en)}{n}\right]^{\frac{\beta}{2\beta+1}}\right) \geq 1-\alpha, \quad \textit{for all } n,$$

for some constant Δ *depending only on* $L, \beta, \psi^l, \psi^u, \alpha$.

Remark 3.5.1 *This section shows that as long as we have apriori knowledge that the function that we are trying to estimate is additive, the problem basically boils down to multiple one-dimensional problems where we are just taking* $d-1$*th order integrals. What our results show is that this simplification does not result in any loss in terms of adaptivity of the functions. We can expect the same rate of convergence as if we were only trying to solve a one-dimensional problem. The results for both global and local adaptivity goes through in this case. The results can be proven by the same techniques that we have used throughout the chapter. Look at appendix for more information about the techniques.*

Chapter 4: Simulation studies

In this section we demonstrate the performance of the multiscale testing procedure described in Section 2.2 and compare it with other competing methods through simulation studies. For computational tractability, we choose $d = 2$ and replace the continuous white noise model (1.1) with its discrete analogue (2.7). For the simulations we have used the kernel function $\psi = \mathbb{I}_{[-1,1]^d}$. In Table 4.1 we give the empirical 0.95-quantile of the multiscale statistic $T_n(W, \psi)$ (see (2.8)) for different values of $n = m^2$; the computation of the empirical quantiles were based on 3000 replications. Observe that the empirical quantiles seem to stabilize as m increases beyond 100. Figure 4.1 shows the empirical distribution function estimates of $T_n(W, \psi)$ for different values of n, based on 3000 replications.

In Tables 4.2 and 4.3 we compare the powers of the multiscale test, a test based on a scan-statistic, and the ALR test (see [16] for the details). Formally, we consider testing (2.2) against alternatives of the form $H_1 : f = \mu_n \mathbb{I}_{B_n}$, for both small and large scale signals (B_n). We briefly describe the above two competing procedures. For $m \geq 1$, let $\mathscr{B}$ be the set of all axis-aligned rectangles on $[0, 1]^2$ with corner points in the following grid:

$$\mathscr{B} := \left\{ \left(\frac{i_1}{m}, \frac{i_2}{m} \right] \times \left(\frac{j_1}{m}, \frac{j_2}{m} \right] : 0 \leq i_1 < i_2 \leq m, 0 \leq j_1 < j_2 \leq m \right\}.$$

Critical values			
m	95% quantile	m	95% quantile
25	3.02	75	3.27
40	3.12	100	3.31
50	3.18	125	3.32
60	3.22	150	3.30*

Table 4.1: Critical values $\kappa_{0.05}$ for different $n = m^2$.
**Note that 0.95 quantiles necessarily increase as n increases. But in our simulations the 0.95 quantile for* $n = 150^2$ *turned out to be slightly less than that of* $n = 125^2$ *due to sampling variability.*

$k = 1$			
μ	Scan	Multiscale	ALR
3.5	0.23	0.08	0.07
4.0	0.34	0.13	0.08
4.5	0.50	0.18	0.08
5.0	0.71	0.30	0.08
5.5	0.86	0.53	0.09

$k = 4$			
μ	Scan	Multiscale	ALR
1.00	0.22	0.14	0.11
1.20	0.43	0.31	0.30
1.35	0.60	0.48	0.44
1.50	0.74	0.55	0.52
1.65	0.86	0.72	0.61

$k = 18$			
μ	Scan	Multiscale	ALR
0.20	0.15	0.21	0.19
0.30	0.49	0.68	0.67
0.35	0.65	0.80	0.82
0.40	0.80	0.90	0.89

$k = 40$			
μ	Scan	Multiscale	ALR
0.040	0.15	0.32	0.31
0.043	0.30	0.56	0.54
0.047	0.45	0.78	0.78
0.050	0.68	0.94	0.95

Table 4.2: Power of the scan, the multiscale and the ALR tests for $m = 40$ (i.e., $n = 40^2$) as μ changes.

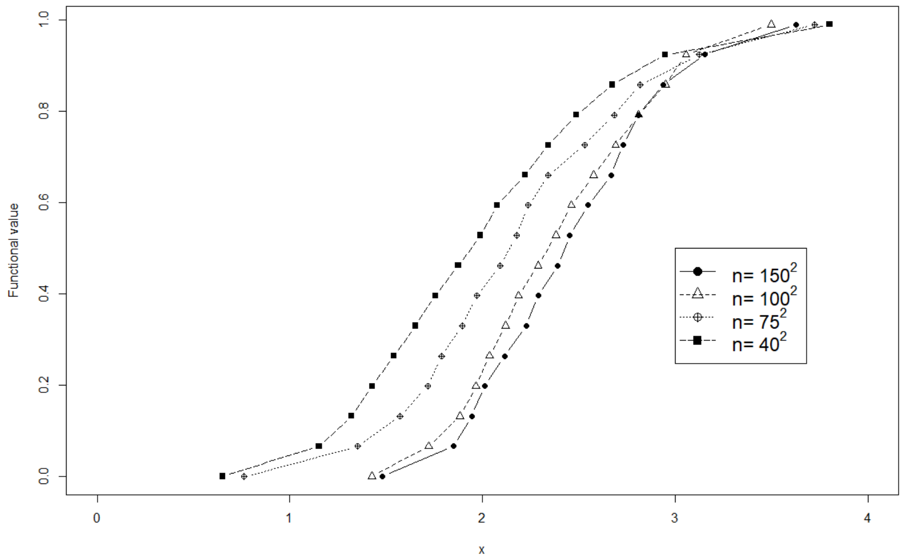

Figure 4.1: The empirical distribution functions of the multiscale statistic for different values of n.

For every $B \in \mathscr{B}$ define

$$\hat{\Psi}(B) := \frac{1}{\sqrt{|B|}} \sum_{(i/m, j/m) \in B} Y\left(\frac{i}{m}, \frac{j}{m}\right).$$

Note that $\hat{\Psi}(\cdot)$ is the discrete analogue of the normalized kernel estimator as defined in (1.6). The scan test statistic (see [90, Chapter 5]) for this problem is defined as

$$M_n := \max_{B \in \mathscr{B}} |\hat{\Psi}(B)|.$$

The ALR test statistic (see [13]) is defined as

$$A_n := \frac{1}{\binom{m+1}{2}^2} \sum_{B \in \mathscr{B}} \exp(\hat{\Psi}(B)^2/2).$$

The scan test (ALR test) rejects the null hypothesis if the observed M_n (A_n) exceeds the 0.95-quantile for M_n (A_n) under the null hypothesis. In Tables 4.2 and 4.3 we compare the performances of the three procedures where μ denotes the signal strength, and k/m denotes the length of each side of the square signal B_n (i.e., B_n is a square of size $k/m \times k/m$). The power of the tests were calculated using 1000 replications. In each replication the location of the square signal B_n was chosen randomly.

We make the following observations. For both the cases ($m = 40$ and 100) when the signal is at the smallest scale, e.g., $k = 1$, the scan statistic outperforms everything else. However, when $m = 100$, even in relatively small scales, e.g., $k = 8$ (i.e., about 0.6% of the observations contain the signal) our multiscale test starts to outperform the scan test. Note that in this setting (small scales) the ALR performs the worst. As the spatial extent of the signal increases, our multiscale procedure and the ALR procedure starts performing favorably whereas the performance of the scan statistics deteriorates. Thus, the simulation experiments corroborate our theoretical findings.

In the next part of our simulation studies we construct confidence bands for shape restricted regression function for (2.7). At first we consider the regression function $f(x_1, x_2) = x_1 + x_2$, In our simulation studies we have data on a 50×50 grid on $[0, 1]^2$ and we have assumed that

$k = 1$			
μ	Scan	Multiscale	ALR
4.5	0.34	0.11	0.06
5.0	0.52	0.28	0.06
5.5	0.75	0.43	0.09
6.0	0.95	0.61	0.13

$k = 8$			
μ	Scan	Multiscale	ALR
0.25	0.08	0.17	0.07
0.30	0.35	0.46	0.13
0.35	0.60	0.72	0.22
0.40	0.82	0.96	0.50

$k = 30$			
μ	Scan	Multiscale	ALR
0.040	0.07	0.22	0.22
0.050	0.17	0.42	0.45
0.055	0.42	0.74	0.75
0.060	0.58	0.93	0.96

$k = 100$			
μ	Scan	Multiscale	ALR
0.014	0.08	0.42	0.42
0.018	0.17	0.62	0.63
0.020	0.22	0.84	0.86
0.025	0.45	0.96	0.95

Table 4.3: Power of the scan, the multiscale and the ALR tests for $m = 100$ (i.e., $n = 100^2$) as μ changes.

the underlying regression function is isotonic. Figure 4.2 shows the constructed confidence band. Here we would like to point our the confidence band achieves the smallest width in the center of the rectangle and the width gets larger as we move towards the sides which is expected because of smaller number of datapoints close by.

In our second simulation study, we construct confidence bands for the function $f(x_1, x_2) = \mathbb{I}(x_1 \geq 0.5)$; see Figure 4.3. We can clearly see the local adaptivity of our bands in action. On the regions where the function is constant (regions where x_1 is away from 0.5) we see that the confidence band has significantly less width.

We have also constructed confidence bands under the assumption that the regression function is convex. Figure 4.4 shows the constructed confidence band for the regression function $f(x_1, x_2) = |x_1 - 0.5|$, whereas Figure 4.5 is for the regression function $f(x_1, x_2) = 40((x_1 - 0.5)^2 + (x_2 - 0.5)^2)$. Both of these simulations show that the upper band is much smoother that the lower band in the convex regression case. We also see that the lower confidence band is closer to the actual function in the center of the rectangle than on the sides. Table 4.4 also shows that at least 95% of cases our confidence bands do encompass the actual regression function, which is expected as we have guaranteed converge even in the finite sample case; see Theorem 3.2.1 for details.

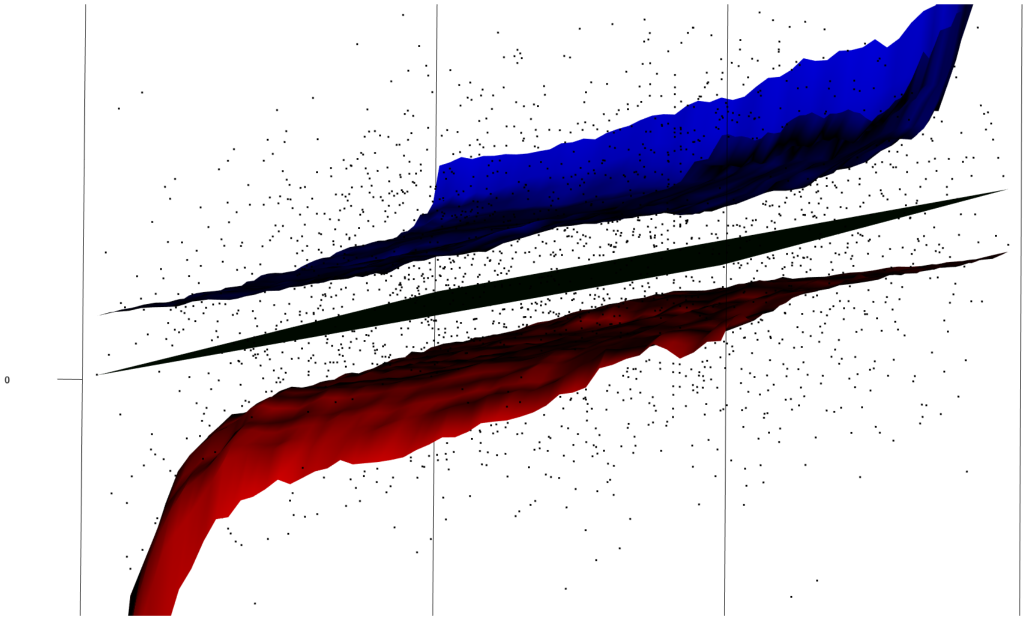

Figure 4.2: Confidence band for the function $f(x_1, x_2) = x_1 + x_2$ assuming f is isotonic and sample size $n = 50^2$

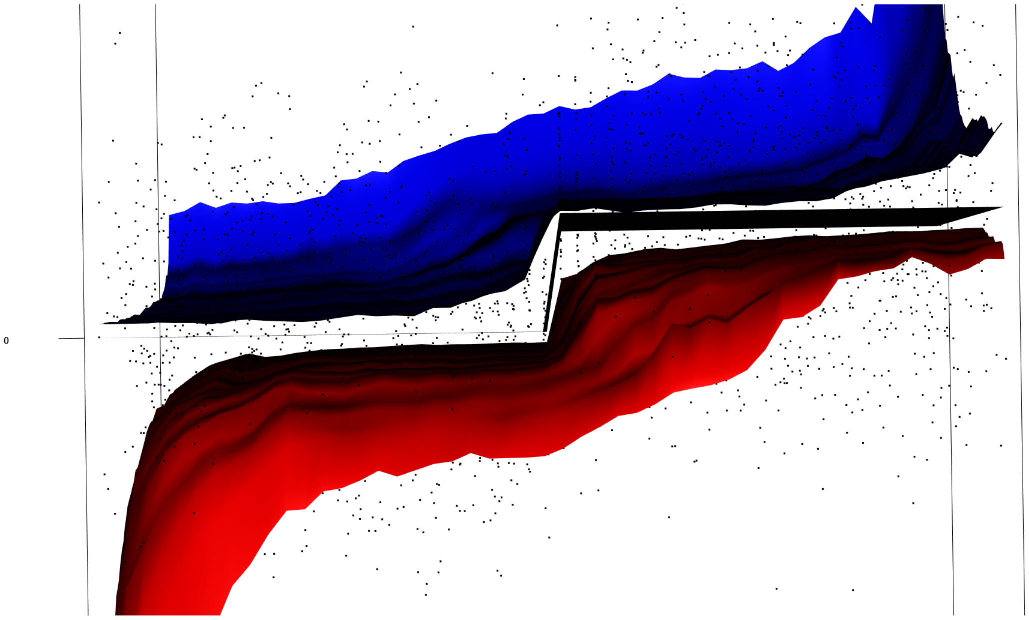

Figure 4.3: Confidence band for the function $f(x_1, x_2) = \mathbb{I}(x_1 \geq 0.5)$ assuming f is isotonic and sample size $n = 50^2$

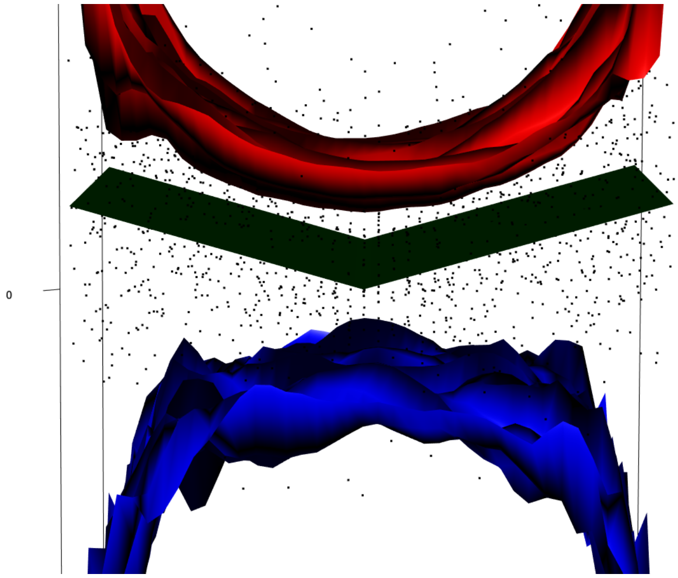

Figure 4.4: Confidence band for the function $f(x_1, x_2) = |x_1 - 0.5|$ assuming f is convex and sample size $n = 40^2$

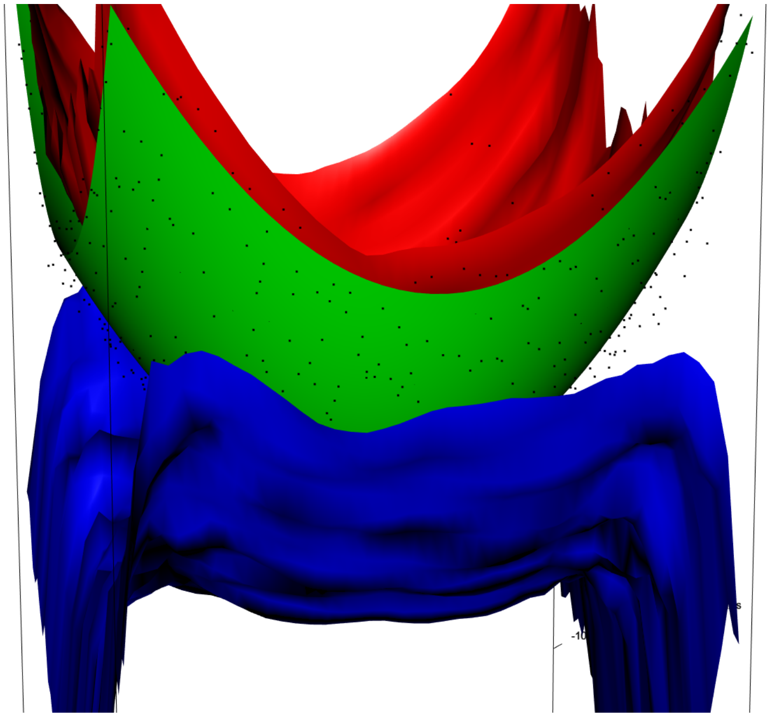

Figure 4.5: Confidence band for the function $f(x_1, x_2) = 40((x_1 - 0.5)^2 + (x_2 - 0.5)^2)$ assuming f is convex and sample size $n = 40^2$

$f(x_1, x_2)$	Category	Coverage Probability
0	isotonic	0.95
$x_1 + X_2$	isotonic	0.97
$20(x_1 + X_2)$	isotonic	1.00
$\mathbb{I}(x_1 \geq 0.5)$	isotonic	0.97
0	convex	0.95
$x_1 + X_2$	convex	0.96
$10(x_1 + X_2)$	convex	0.95
$(x_1 - 0.5)^2 + (x_2 - 0.5)^2$	convex	0.98

Table 4.4: Coverage probability of our constructed confidence bands for different regression functions

Conclusion or Epilogue

In this dissertation we have proposed a multidimensional multiscale statistic in the continuous white noise model and used this statistic to construct asymptotically minimax tests for testing $f = 0$ against (i) Hölder classes of functions; and (ii) alternatives of the form $f = \mu_n \mathbb{I}_{B_n}$, where B_n is an unknown axis-aligned hyperrectangle in $[0, 1]^d$ and $\mu_n \in \mathbb{R}$ is unknown. However, there are many open questions in this area. We briefly delineate a few of them below and in the process describe some important papers in related areas of research.

We have shown that for the Hölder class $\mathbb{H}_{\beta,L}$, if the smoothness parameter β is known, we can construct an asymptotically minimax test. However, if β is unknown (and $\beta \leq 1$) we can only construct a rate optimal test. A natural question that arises is whether a test can be constructed that is asymptotically minimax (for the Hölder class of functions with the supremum norm) without the knowledge of the smoothness parameter β (and $L > 0$); see [91, Section 1.3]. Another interesting question would be to try to extend our results to other smoothness classes like Sobolev/Besov classes; in [25] the authors gave the minimax rate of testing for Sobolov class, but no test was proposed that achieves the exact separation constant.

Note that we have shown that our multiscale test is asymptotically minimax for detecting the presence of a signal on an axis-aligned hyperrectangle in $[0, 1]^d$. One obvious extension of our work would be to correctly identify the hyperrectangle on which the signal is present. Further, we could go beyond hyperrectangles and try to identify signals that are present on some other geometric structures $A \subset [0, 1]^d$ (i.e., $f = \mu \mathbb{I}_A$ where A is not necessarily an axis-aligned hyperrectangle). Examples of such geometric structures could be: (i) A is an hyperrectangle which is not necessarily axis-aligned, (ii) A is a d-dimensional ellipsoid, (iii) $A = \bigcup_{i=1}^{k} A_i$ where each $A_i \subseteq [0, 1]^d$ is an (axis-aligned) hyperrectangle, etc. [17] and the references therein investigated the problem of finding change points in $d = 1$ which can be thought of as detection of multiple intervals. In [11] the authors use the scan statistic to detect regions in $\mathbb{R}^d$ where the underlying function is non-zero. [12] considers the problem of finding a cluster of signals (not necessarily rectangular) in a network using the scan statistic. Although the method they propose achieves the optimal boundary for detection, it requires the knowledge of whether the signal shape is "thick" or "thin". For hyperrectangles this refers to whether or not the minimum side length is of order $\log n/n$ or not. We believe that the multiscale statistic, with proper modifications, can be used to find asymptotically minimax/rate optimal tests in such problems.

In our white noise model (1.1) we assume that the distribution of the response variables is (homogeneous and independent) Gaussian. Similar questions about signal detection can be asked when the response is non-Gaussian; see e.g., [14], [37], [92], [18], etc. In [93] the authors looked at the problem of detecting change points under heterogeneous variance of the response variable (when $d = 1$). [94] looked at this problem where the error distribution is known to be symmetric (when $d = 1$). A multiscale approach could be used to tackle such problems as well. Here we note that [14] studied a similar problem where the response variable is binary when $d > 1$.

Several interesting applications of the multiscale approach exist when $d = 1$ (following the seminal paper of [4]): In [95] the authors propose a multiscale test statistic to make inference about a probability density on the real line given i.i.d. observations; [96] use multiscale methods to make inference in a deconvolution problem; [37] use multiscale methods to detect a jump in the intensity of a Poisson process; [97] and [98] use multiscale approaches to make inferences about multivariate densities in deconvolution problems, etc. We believe that our extension beyond $d = 1$ will also lead to several interesting multidimensional applications.

One such application we have looked at in Chapter 3. We have used this multidimensional statistics to find confidence band for both co-ordinatewise monotone and convex functions. Our constructed confidence bands are honest, adaptive with respect to the smoothness of the underlying function (both globally and locally). We have also shown that the function is adaptive to the intrinsic dimension of the function. We have also shown that our constructed confidence bands not only achieves the optimal rate of convergence but also the optimal constant. We have also demonstrated a method of constructing confidence bands in the special case of additive functions, where the problem essentially boils down to solving multiple one-dimensional problems and attains the same rate of convergence corresponding to $d = 1$.

References

[1] L. D. Brown and M. G. Low, “Asymptotic equivalence of nonparametric regression and white noise,” *Ann. Statist.*, vol. 24, no. 6, pp. 2384–2398, 1996.

[2] M. Reiß, “Asymptotic equivalence for nonparametric regression with multivariate and random design,” *Ann. Statist.*, vol. 36, no. 4, pp. 1957–1982, Aug. 2008.

[3] D. L. Donoho and M. G. Low, “Renormalization exponents and optimal pointwise rates of convergence,” *Ann. Statist.*, vol. 20, no. 2, pp. 944–970, 1992.

[4] L. Dümbgen and V. G. Spokoiny, “Multiscale testing of qualitative hypotheses,” *Ann. Statist.*, vol. 29, no. 1, pp. 124–152, 2001.

[5] A. V. Carter, “A continuous Gaussian approximation to a nonparametric regression in two dimensions,” *Bernoulli*, vol. 12, no. 1, pp. 143–156, 2006.

[6] O. V. Lepski, “On asymptotically exact testing of nonparametric hypotheses,” Université catholique de Louvain, Tech. Rep., 1993.

[7] O. V. Lepski and A. B. Tsybakov, “Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point,” *Probab. Theory Related Fields*, vol. 117, no. 1, pp. 17–48, 2000.

[8] J. L. Horowitz and V. G. Spokoiny, “An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative,” *Econometrica*, vol. 69, no. 3, pp. 599–631, 2001.

[9] Y. I. Ingster and T. Sapatinas, “Minimax goodness-of-fit testing in multivariate nonparametric regression,” *Math. Methods Statist.*, vol. 18, no. 3, pp. 241–269, 2009.

[10] J. Glaz and Z. Zhang, “Multiple window discrete scan statistics,” *J. Appl. Stat.*, vol. 31, no. 8, pp. 967–980, 2004.

[11] E. Arias-Castro, D. L. Donoho, and X. Huo, “Near-optimal detection of geometric objects by fast multiscale methods,” *IEEE Trans. Inform. Theory*, vol. 51, no. 7, pp. 2402–2425, 2005.

[12] E. Arias-Castro, E. J. Candès, and A. Durand, “Detection of an anomalous cluster in a network,” *Ann. Statist.*, vol. 39, no. 1, pp. 278–304, 2011.

[13] H. P. Chan, “Detection of spatial clustering with average likelihood ratio test statistics,” *Ann. Statist.*, vol. 37, no. 6B, pp. 3985–4010, 2009.

[14] G. Walther, “Optimal and fast detection of spatial clusters with scan statistics,” *Ann. Statist.*, vol. 38, no. 2, pp. 1010–1033, 2010.

[15] C. Butucea and Y. I. Ingster, “Detection of a sparse submatrix of a high-dimensional noisy matrix,” *Bernoulli*, vol. 19, no. 5B, pp. 2652–2688, 2013.

[16] H. P. Chan and G. Walther, “Detection with the scan and the average likelihood ratio,” *Statist. Sinica*, vol. 23, no. 1, pp. 409–428, 2013.

[17] K. Frick, A. Munk, and H. Sieling, “Multiscale change point inference,” *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 76, no. 3, pp. 495–580, 2014, With 32 discussions by 47 authors and a rejoinder by the authors.

[18] C. König, A. Munk, and F. Werner, “Multidimensional multiscale scanning in exponential families: Limit theory and statistical consequences,” *Ann. Statist.*, vol. 48, no. 2, pp. 655–678, 2020.

[19] K. Proksch, F. Werner, and A. Munk, “Multiscale scanning in inverse problems,” *Ann. Statist.*, vol. 46, no. 6B, pp. 3569–3602, 2018.

[20] E. Giné and A. Guillou, “Rates of strong uniform consistency for multivariate kernel density estimators,” *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 38, no. 6, pp. 907–921, 2002, En l’honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

[21] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models* (Lecture Notes in Statistics). Springer-Verlag, New York, 2003, vol. 169, pp. xiv+453, ISBN: 0-387-95531-3.

[22] Y. I. Ingster, “Asymptotically minimax hypothesis testing for nonparametric alternatives. I,” *Math. Methods Statist.*, vol. 2, no. 2, pp. 85–114, 1993.

[23] Y. I. Ingster, “Asymptotically minimax hypothesis testing for nonparametric alternatives. II,” *Math. Methods Statist.*, vol. 2, no. 3, pp. 171–189, 1993.

[24] Y. I. Ingster, “Asymptotically minimax hypothesis testing for nonparametric alternatives. III,” *Math. Methods Statist.*, vol. 2, no. 4, pp. 249–268, 1993.

[25] Y. Ingster and N. Stepanova, “Estimation and detection of functions from anisotropic Sobolev classes,” *Electron. J. Stat.*, vol. 5, pp. 484–506, 2011.

[26] D. Siegmund and E. S. Venkatraman, “Using the generalized likelihood ratio statistic for sequential detection of a change-point,” *Ann. Statist.*, vol. 23, no. 1, pp. 255–271, 1995.

[27] M. Kulldorff, “A spatial scan statistic,” *Comm. Statist. Theory Methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[28] D. Siegmund and B. Yakir, “Tail probabilities for the null distribution of scanning statistics,” *Bernoulli*, vol. 6, no. 2, pp. 191–213, 2000.

[29] T. Jiang, “Maxima of partial sums indexed by geometrical structures,” *Ann. Probab.*, vol. 30, no. 4, pp. 1854–1892, 2002.

[30] J. I. Naus and S. Wallenstein, “Multiple window and cluster size scan procedures,” *Methodol. Comput. Appl. Probab.*, vol. 6, no. 4, pp. 389–400, 2004.

[31] G. Haiman and C. Preda, “Estimation for the distribution of two-dimensional discrete scan statistics,” *Methodol. Comput. Appl. Probab.*, vol. 8, no. 3, pp. 373–381, 2006.

[32] V. Pozdnyakov, J. Glaz, M. Kulldorff, and J. M. Steele, “A martingale approach to scan statistics,” *Ann. Inst. Statist. Math.*, vol. 57, no. 1, pp. 21–37, 2005.

[33] X. Wang and J. Glaz, “Variable window scan statistics for normal data,” *Comm. Statist. Theory Methods*, vol. 43, no. 10-12, pp. 2489–2504, 2014.

[34] J. Sharpnack and E. Arias-Castro, “Exact asymptotics for the scan statistic and fast alternatives,” *Electron. J. Stat.*, vol. 10, no. 2, pp. 2641–2684, 2016.

[35] P. Hall and J. Jin, “Properties of higher criticism under strong dependence,” *Ann. Statist.*, vol. 36, no. 1, pp. 381–402, 2008.

[36] J. Sharpnack, “Learning patterns for detection with multiscale scan statistics,” in *Proceedings of the 31st Conference On Learning Theory*, S. Bubeck, V. Perchet, and P. Rigollet, Eds., ser. Proceedings of Machine Learning Research, vol. 75, PMLR, 2018, pp. 950–969.

[37] C. Rivera and G. Walther, “Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics,” *Scand. J. Stat.*, vol. 40, no. 4, pp. 752–769, 2013.

[38] H. D. Brunk, “Maximum likelihood estimates of monotone parameters,” *Ann. Math. Statist.*, vol. 26, pp. 607–616, 1955.

[39] H. D. Brunk, “Estimation of isotonic regression,” in *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, Cambridge Univ. Press, London, 1970, pp 177–197.

[40] C. Hildreth, “Point estimates of ordinates of concave functions,” *J. Amer. Statist. Assoc.*, vol. 49, pp. 598–619, 1954.

[41] B. L. S. Prakasa Rao, "Estimation of a unimodal density," *Sankhyā Ser. A*, vol. 31, pp. 23–36, 1969.

[42] P. Groeneboom, G. Jongbloed, and J. A. Wellner, "Estimation of a convex function: Characterizations and asymptotic theory," *Ann. Statist.*, vol. 29, no. 6, pp. 1653–1698, 2001.

[43] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney, 1972, pp. xii+388.

[44] T. Robertson, F. T. Wright, and R. L. Dykstra, *Order restricted statistical inference* (Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics). John Wiley & Sons, Ltd., Chichester, 1988, pp. xx+521, ISBN: 0-471-91787-7.

[45] P. Groeneboom and J. A. Wellner, *Information bounds and nonparametric maximum likelihood estimation* (DMV Seminar). Birkhäuser Verlag, Basel, 1992, vol. 19, pp. viii+126, ISBN: 3-7643-2794-4.

[46] P. Groeneboom and G. Jongbloed, *Nonparametric estimation under shape constraints* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, New York, 2014, vol. 38, pp. xi+416, Estimators, algorithms and asymptotics, ISBN: 978-0-521-86401-5.

[47] H. B. McMahan *et al.*, "Ad click prediction: A view from the trenches," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1222–1230.

[48] R. Luss, S. Rosset, and M. Shahar, "Efficient regularized isotonic regression with application to gene-gene interaction search," *Ann. Appl. Stat.*, vol. 6, no. 1, pp. 253–283, 2012.

[49] G. Allon, M. Beenstock, S. Hackman, U. Passy, and A. Shapiro, "Nonparametric estimation of concave production technologies by entropic methods," *Journal of Applied Econometrics*, vol. 22, no. 4, pp. 795–816, 2007.

[50] T. Kuosmanen and A. L. Johnson, "Data envelopment analysis as nonparametric least-squares regression," *Operations Research*, vol. 58, no. 1, pp. 149–160, 2010.

[51] A. Keshavarz, Y. Wang, and S. Boyd, "Imputing a convex objective function," in *2011 IEEE international symposium on intelligent control*, IEEE, 2011, pp. 613–619.

[52] E. Seijo and B. Sen, "Nonparametric least squares estimation of a multivariate convex regression function," *Ann. Statist.*, vol. 39, no. 3, pp. 1633–1657, 2011.

[53] X. Chen, Q. Lin, and B. Sen, “On degrees of freedom of projection estimators with applications to multivariate nonparametric regression,” *J. Amer. Statist. Assoc.*, vol. 115, no. 529, pp. 173–186, 2020.

[54] Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth, “Isotonic regression in general dimensions,” *Ann. Statist.*, vol. 47, no. 5, pp. 2440–2471, 2019.

[55] H. Deng and C.-H. Zhang, “Isotonic regression in multi-dimensional spaces and graphs,” *Ann. Statist.*, vol. 48, no. 6, pp. 3672–3698, 2020.

[56] M. Meyer and M. Woodroofe, “On the degrees of freedom in shape-restricted regression,” *Ann. Statist.*, vol. 28, no. 4, pp. 1083–1104, 2000.

[57] C.-H. Zhang, “Risk bounds in isotonic regression,” *Ann. Statist.*, vol. 30, no. 2, pp. 528–555, 2002.

[58] A. Guntuboyina and B. Sen, “Global risk bounds and adaptation in univariate convex regression,” *Probab. Theory Related Fields*, vol. 163, no. 1-2, pp. 379–411, 2015.

[59] S. Chatterjee, A. Guntuboyina, and B. Sen, “On risk bounds in isotonic and other shape restricted regression problems,” *Ann. Statist.*, vol. 43, no. 4, pp. 1774–1800, 2015.

[60] S. Chatterjee, A. Guntuboyina, and B. Sen, “On matrix estimation under monotonicity constraints,” *Bernoulli*, vol. 24, no. 2, pp. 1072–1100, 2018.

[61] A. Guntuboyina and B. Sen, “Nonparametric shape-restricted regression,” *Statist. Sci.*, vol. 33, no. 4, pp. 568–594, 2018.

[62] S. Chatterjee and J. Lafferty, “Adaptive risk bounds in unimodal regression,” *Bernoulli*, vol. 25, no. 1, pp. 1–25, 2019.

[63] G. Kur, F. Gao, A. Guntuboyina, and B. Sen, “Convex regression in multidimensions: Suboptimality of least squares estimators,” *arXiv preprint arXiv:2006.02044*, 2020.

[64] L. Dümbgen, “Optimal confidence bands for shape-restricted curves,” *Bernoulli*, vol. 9, no. 3, pp. 423–449, 2003.

[65] P. Datta and B. Sen, “Optimal inference with a multidimensional multiscale statistic,” *Electron. J. Statist.*, vol. 15, no. 2, pp. 5203–5244, 2021.

[66] L. Dümbgen and V. G. Spokoiny, “Multiscale testing of qualitative hypotheses,” *Ann. Statist.*, vol. 29, no. 1, pp. 124–152, 2001.

[67] M. P. Wand and M. C. Jones, *Kernel smoothing* (Monographs on Statistics and Applied Probability). Chapman and Hall, Ltd., London, 1995, vol. 60, pp. xii+212, ISBN: 0-412-55270-1.

[68] I. M. Johnstone, “Wavelets and the theory of non-parametric function estimation,” *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1760, pp. 2475–2493, 1999.

[69] L. D. Brown, T. T. Cai, and H. H. Zhou, “Robust nonparametric estimation via wavelet median regression,” *Ann. Statist.*, vol. 36, no. 5, pp. 2055–2084, 2008.

[70] J. Hart, *onparametric Smoothing and Lack-of-Fit Tests*. Springer, New York, 1997.

[71] H. Lian, K. Zhao, and S. Lv, “Projected spline estimation of the nonparametric function in high- dimensional partially linear models for massive data,” *Ann. Statist.*, vol. 47, no. 5, pp. 2922–2949, 2019.

[72] C. Gu, *Smoothing spline ANOVA models* (Springer Series in Statistics). Springer-Verlag, New York, 2002, pp. xiv+289, ISBN: 0-387-95353-1.

[73] J. Fan, *Local Polynomial Modelling and Its Applications* (Monographs on Statistics and Applied Probability). Chapman and Hall, 1996, vol. 66, p. 360, ISBN: 9780203748725.

[74] G. Wahba, *Spline Models for Observational Data* (CBMS-NSF Regional Conference Series in Applied Mathematics). Society for Industrial and Applied Mathematics, 1990, vol. 59, pp. XII + 169, ISBN: 978-0-898712-44-5.

[75] D. L. Donoho and I. M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[76] D. L. Donoho and I. M. Johnstone, “Adapting to unknown smoothness via wavelet shrinkage,” *JASA*, vol. 90, no. 432, pp. 1200–1224, 1995.

[77] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, New York, 2016, vol. [40], pp. xiv+690, ISBN: 978-1-107-04316-9.

[78] D. L. Donoho, “One-sided inference about functionals of a density,” *Ann. Statist.*, vol. 16, pp. 1390–1420, 1988.

[79] T. T. Cai and M. G. Low, “An adaptation theory for nonparametric confidence intervals,” *Ann. Statist.*, vol. 32, no. 5, pp. 1805–1840, 2004.

[80] H. R. Varian, *Microeconomic analysis*. WW Norton, 1992.

[81] H. R. Varian, *Intermediate Microeconomics, a modern approach*, Eighth. Macmillan & Company, 2010.

[82] R. Chambers, "Duality, the output effect, and applied comparative statics," *American Journal of Agricultural Economics*, vol. 64, no. 1, pp. 152–156, 1982.

[83] S. Mukherjee, R. Patra, A. Johnson, and H. Morita, "Least squares estimation of a monotone quasiconvex regression function," *arXiv preprint arXiv:2003.04433*, 2021.

[84] X. Chen, V. Chernozhukov, I. Fernández-Val, S. Kostyshak, and Y. Luo, "Shape-enforcing operators for point and interval estimators," *arXiv:1809.01038v3*, 2018.

[85] L. Dümbgen, "New goodness-of-fit tests and their application to nonparametric confidence sets," *Ann. Statist.*, vol. 26, no. 1, pp. 288–314, 1998.

[86] P. Davies, "Data features," *Statistica Neerlandica*, vol. 49, pp. 185–245, 1995.

[87] N. W. Hengartner and P. B. Stark, "Finite-sample confidence envelopes for shape-restricted densities," *Ann. Statist.*, vol. 23, no. 2, pp. 525–550, 1995.

[88] J. Freyberger and B. Reeves, "Inference under shape restrictions," *Social Science Research Network*, 2018.

[89] A. B. Tsybakov, *Introduction to nonparametric estimation* (Springer Series in Statistics). Springer, New York, 2009, pp. xii+214, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats, ISBN: 978-0-387-79051-0.

[90] J. Glaz, J. I. Naus, and S. Wallenstein, *Scan statistics*. Springer, 2011.

[91] P. Ji and M. Nussbaum, "Sharp minimax adaptation over Sobolev ellipsoids in nonparametric testing," *Electron. J. Stat.*, vol. 11, no. 2, pp. 4515–4562, 2017.

[92] H. P. Chan and G. Walther, "Optimal detection of multi-sample aligned sparse signals," *Ann. Statist.*, vol. 43, no. 5, pp. 1865–1895, 2015.

[93] F. Pein, H. Sieling, and A. Munk, "Heterogeneous change point inference," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 79, no. 4, pp. 1207–1227, 2017.

[94] A. Rohde, "Adaptive goodness-of-fit tests based on signed ranks," *Ann. Statist.*, vol. 36, no. 3, pp. 1346–1374, 2008.

[95] L. Dümbgen and G. Walther, "Multiscale inference about a density," *Ann. Statist.*, vol. 36, no. 4, pp. 1758–1785, 2008.

[96] J. Schmidt-Hieber, A. Munk, and L. Dümbgen, "Multiscale methods for shape constraints in deconvolution: Confidence statements for qualitative features," *Ann. Statist.*, vol. 41, no. 3, pp. 1299–1328, 2013.

[97] K. Eckle, N. Bissantz, and H. Dette, "Multiscale inference for multivariate deconvolution," *Electron. J. Stat.*, vol. 11, no. 2, pp. 4179–4219, 2017.

[98] K. Eckle, N. Bissantz, H. Dette, K. Proksch, and S. Einecke, "Multiscale inference for a multivariate density with applications to X-ray astronomy," *Ann. Inst. Statist. Math.*, vol. 70, no. 3, pp. 647–689, 2018.

[99] E. Wong and M. Zakai, "An extension of stochastic integrals in the plane," *Ann. Probab.*, vol. 5, no. 5, pp. 770–778, 1977.

[100] D. Khoshnevisan, *Multiparameter processes* (Springer Monographs in Mathematics). Springer-Verlag, New York, 2002, pp. xx+584, An introduction to random fields, ISBN: 0-387-95459-7.

[101] P. E. Protter, *Stochastic integration and differential equations* (Stochastic Modelling and Applied Probability). Springer-Verlag, Berlin, 2005, vol. 21, pp. xiv+419, Second edition. Version 2.1, Corrected third printing, ISBN: 3-540-00313-4.

[102] C. Aistleitner and J. Dick, "Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality," *Acta Arith.*, vol. 167, no. 2, pp. 143–171, 2015.

[103] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes* (Springer Series in Statistics). Springer-Verlag, New York, 1996, pp. xvi+508, With applications to statistics, ISBN: 0-387-94640-3.

[104] Z. Kabluchko and A. Munk, "Exact convergence rate for the maximum of standardized Gaussian increments," *Electron. Commun. Probab.*, vol. 13, pp. 302–310, 2008.

Appendix A: Proofs of our main results

A.1 Some useful concepts

In this subsection we formally define some technical concepts that we use in this paper.

Definition A.1.1 (Brownian sheet) *By a d-dimensional Brownian sheet we mean a mean-zero Gaussian process $\{W(t) : t \in [0,1]^d\}$ with covariance*

$$Cov(W(t_1, \ldots, t_d), W(s_1, \ldots, s_d)) = \Pi_{i=1}^d \min(t_i, s_i),$$

for $(t_1, \ldots, t_d), (s_1, \ldots, s_d) \in [0,1]^d$. The Brownian sheet is the d-dimensional counterpart of the standard Brownian motion; see e.g., [99], [100, Chapter 5] for detailed properties of the Brownian sheet. See Appendix A.1.1 for some important properties of the Brownian sheet used in our proofs.

A.1.1 Properties of Brownian Sheet

In the following we give some useful properties of the Brownian sheet $W(\cdot)$.

- If $g \in L_2([0,1]^d)$ then $\int g dW := \int_{[0,1]^d} g(t) dW(t) \sim N(0, \|g\|^2)$.
- If $g_1, g_2 \in L_2([0,1]^d)$ then $\text{Cov}\left(\int g_1 dW, \int g_2 dW\right) = \int_{[0,1]^d} g_1(t) g_2(t) dt$.
- *Cameron-Martin-Girsanov Theorem for Brownian sheet:* Let us state the simplest version of the Cameron-Martin-Girsanov Theorem that we will use in this paper (see [101, Chapter 3] for detailed discussion about change of measure and the result).

  Assume $f \in L_1([0,1]^d)$ and let $\{W(t) : t \in [0,1]^d\}$ be a standard Brownian sheet. Let Ω be the set of all real-valued continuous functions defined on $[0,1]^d$. Let P denote the measure on Ω induced by the Brownian sheet $\{W(t) : t \in [0,1]^d\}$ and let Q denote the measure induced by $\{Y(t) : t \in [0,1]^d\}$ where $Y(t)$ is defined as in (1.1). Then Q is absolutely continuous with respect to P and the Radon-Nikodym derivative is given by

$$\frac{dQ}{dP}(Y) = \exp\left(\sqrt{n}\int f dW - \frac{n}{2}\|f\|^2\right).$$

This, in turn, implies that for any measurable function ϕ we have

$$\mathbb{E}_Q\left(\phi(Y)\right) = \mathbb{E}_P\left(\phi(Y)\frac{dQ}{dP}(Y)\right).$$

Definition A.1.2 (Hölder Function) *Fix $\beta > 0$ and $L > 0$. Let $\lfloor\beta\rfloor$ be the largest integer which is strictly less than β and for $k = (k_1, k_2, \ldots, k_d) \in \mathbb{N}^d$ set $\|k\|_1 := \sum_{i=1}^d k_i$. The Hölder class $\mathbb{H}_{\beta,L}$ on $[-1,1]^d$ is the set of all functions $f : [-1,1]^d \to \mathbb{R}$ having all partial derivatives of order $\lfloor\beta\rfloor$ on $[-1,1]^d$ such that*

$$\sum_{0\le\|k\|_1\le\lfloor\beta\rfloor}\sup_{x\in[0,1]^d}\left|\frac{\partial^{\|k\|_1}f(x)}{\partial x_1^{k_1}\ldots\partial x_d^{k_d}}\right| \le L$$

and

$$\sum_{\|k\|_1=\lfloor\beta\rfloor}\left|\frac{\partial^{\|k\|_1}f(y)}{\partial x_1^{k_1}\ldots\partial x_d^{k_d}} - \frac{\partial^{\|k\|_1}f(z)}{\partial x_1^{k_1}\ldots\partial x_d^{k_d}}\right| \le L\|y-z\|^{\beta-\lfloor\beta\rfloor} \quad \forall y,z\in[-1,1]^d.$$

See Appendix A.1.1 for an important property of Hölder classes of functions useful in our proofs.

Properties of Hölder functions

One of the most important properties of $\mathbb{H}_{\beta,L}$ that we will use is the following: If $f \in \mathbb{H}_{\beta,1}$ then, for any $h = (h_1, \ldots, h_d) > 0$ and $t \in A_h$,

$$g(x_1,\ldots,x_d) := L\min(h)^\beta f\left(\frac{x_1-t_1}{h_1},\ldots,\frac{x_d-t_d}{h_d}\right) \in \mathbb{H}_{\beta,L}$$

where $\min(h) := \min_{i=1,\ldots,d} h_i$. The proof of the above result follows directly from the definition of Hölder functions.

Definition and Properties of Hardy-Krause variation

The notion of bounded variation for a function $f : \mathbb{R}^d \to \mathbb{R}$, where $d \geq 2$, is more involved than when $d = 1$. In fact there is no unique notion of bounded variation for a function when $d \geq 2$. Below we describe the notion of Hardy and Krause variation as given in [102], which suffices for our purpose.

Definition A.1.3 (Hardy-Krause variation) *Let $f : [-1, 1]^d \to \mathbb{R}$ be a measurable function. Let $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$ be elements of $[-1, 1]^d$ such that $a < b$ (coordinate-wise). We introduce the d-dimensional difference operator $\Delta^{(d)}$ which assigns to the axis-aligned box $A := [a, b]$ a d-dimensional quasi-volume*

$$\Delta^{(d)}(f; A) = \sum_{j_1=0}^{1} \cdots \sum_{j_d=0}^{1} (-1)^{j_1+\cdots+j_d} f(b_1 + j_1(a_1 - b_1), \ldots, b_d + j_d(a_d - b_d)).$$

Let $m_1, \ldots, m_d \in \mathbb{N}$. For $s = 1, \ldots, d$, let $-1 =: x_0^{(s)} < x_1^{(s)} < \cdots < x_{m_s}^{(s)} := 1$ be a partition of $[-1, 1]$ and let P be a partition of $[-1, 1]^d$ which is given by

$$\mathsf{P} := \left\{ [x_{l_1}^{(1)}, x_{l_1+1}^{(1)}] \times \cdots \times [x_{l_d}^{(d)}, x_{l_d+1}^{(d)}] \,:\, l_s = 0, 1, \ldots, m_s - 1, \text{ for } s = 1, \ldots, d \right\}.$$

Then the variation of f on $[-1, 1]^d$ in the sense of Vitali is given by

$$V^{(d)}(f; [-1, 1]^d) := \sup_{\mathsf{P}} \sum_{A \in \mathsf{P}} |\Delta^{(d)}(f; A)|$$

where the supremum is extended over all partitions of $[-1, 1]^d$ into axis-parallel boxes generated by d one-dimensional partitions of $[-1, 1]$. For $1 \leq s \leq d$ and $1 \leq i_1 < \ldots < i_s \leq d$, let $V^{(s)}(f; i_1, \ldots, i_s; [-1, 1]^d)$ denote the s-dimensional variation in the sense of Vitali of the restriction of f to the face

$$U_d^{(i_1,\ldots,i_s)} = \left\{ (x_1, \ldots, x_d) \in [-1, 1]^d : x_j = 1 \text{ for all } j \neq i_1, \ldots, i_s \right\}$$

of $[-1,1]^d$*. Then the variation of* f *on* $[-1,1]^d$ *in the sense of Hardy and Krause anchored at 1, abbreviated by HK-variation, is given by*

$$TV(f) := \sum_{i=1}^{d} \sum_{1\leq s\leq d} V^{(s)}(f; i_1, \ldots, i_s; [-1,1]^d).$$

We say a function f *has bounded HK-variation if* $TV(f) < \infty$*.*

The main property of a bounded HK-variation function that we will need in this paper is stated below.

Remark A.1.1 *If* f *is a right continuous function on* $[-1,1]^d$ *which has bounded HK-variation then there exists a unique signed Borel measure* ν *on* $[-1,1]^d$ *for which*

$$f(x) = \nu([-1,x]), \quad x \in [-1,1]^d;$$

A.2 Proof of Theorem 2.1.2

In the following proofs K would be used to denote a generic constant whose value would change from line to line.

For every $v > 0$, we define

$$\Gamma(X,v) := \sup_{a,b\in\mathscr{F},\rho(a,b)\leq v} |X(a) - X(b)|.$$

For simplicity we divide the proof in three steps.

Step 1: In this step we will prove that

$$\mathbb{P}\big(\Gamma(X,v) > \eta\big) \leq K \exp\left(-\frac{\eta^2}{Kv^2\log(e/v)}\right) \quad \forall \eta > 0 \text{ and } v \in (0,1], \tag{A.1}$$

where $K > 0$ is a positive constant not depending on v. We will prove the above result by introducing the notion of Orlicz norm. Let $\lambda : \mathbb{R}_+ \to \mathbb{R}$ be a nondecreasing convex function with $\lambda(0) = 0$. For any random variable X the Orlicz norm $\|X\|_\lambda$ is defined as

$$\|X\|_\lambda = \inf\left\{C > 0 : \mathbb{E}\lambda\left(\frac{|X|}{C}\right) \le 1\right\}.$$

The Orlicz norm is of interest to us as any Orlicz norm easily yields a bound on the tail probability of a random variable i.e., $\mathbb{P}(|X| > x) \le [\lambda(x/\|X\|_\lambda)]^{-1}$, for all $x \in \mathbb{R}$ (see [103, Page 96] for a simple proof). Let us define $\lambda(x) := \exp(x^2) - 1$, $x > 0$. Hence,

$$\mathbb{P}\big(|X| > x\big) \le \min\left\{1, \frac{1}{\exp(x^2/\|X\|_\lambda^2) - 1}\right\} \le 2 \times \exp(-x^2/\|X\|_\lambda^2). \tag{A.2}$$

Hence, it is enough to bound the Orlicz norm of $\Gamma(X, v)$. A bound on the Orlicz norm of $\Gamma(X, v)$ can be shown by appealing to [103, Theorem 2.2.4] which we state below.

Lemma A.2.1 *Let $\lambda : \mathbb{R}_+ \to \mathbb{R}$ be a convex, nondecreasing, non-zero function with $\lambda(0) = 0$ and for some constant $c > 0$, $\limsup_{x,y\to\infty} \frac{\lambda(x)\lambda(y)}{\lambda(cxy)} < \infty$. Let $\{X_a, a \in \mathscr{F}\}$ be a separable stochastic process with*

$$\|X_a - X_b\|_\lambda \le C\rho(a, b) \textit{ for all } a, b \in \mathscr{F}$$

for some pseudometric ρ on $\mathscr{F}$ and constant C. Then for any $\zeta, v > 0$,

$$\|\Gamma(X, v)\|_\lambda \le K\left[\int_0^\zeta \lambda^{-1}(N(\epsilon, \mathscr{F}))d\epsilon + v\lambda^{-1}(N^2(\zeta, \mathscr{F}))\right]$$

for some constant K depending only on λ and C.

We apply the above lemma with $\lambda(x) := \exp(x^2) - 1$ (i.e., $\lambda^{-1}(y) = \sqrt{\log(1 + y)})$. Note that condition (b) of Theorem 2.1.2 directly implies that $\|X_a - X_b\|_\lambda \le C\rho(a, b)$ by an application of [103, Lemma 2.2.1].

By taking $\delta = 1, \epsilon = u^{1/2}$, condition (c) of Theorem 2.1.2 yields $N(\epsilon, \mathscr{F}) \le A\epsilon^{-2B}$. Thus, Lemma A.2.1 gives (with $\zeta = v$)

$$\|\Gamma(X, v)\|_\lambda \le K \left[\int_0^v \sqrt{\log(1 + A\epsilon^{-2B})} d\epsilon + v\sqrt{\log(1 + A^2 v^{-4B})} \right].$$

The expression on the right side of the above display can be easily shown to be less than or equal to $Kv\sqrt{\log(e/v)}$ for some constant K. This result along with an application of (A.2) with $\Gamma(X, v)$ instead of X imply

$$\mathbb{P}\big(\Gamma(X, v) > \eta\big) \le K \exp\left(-\frac{\eta^2}{Kv^2 \log(e/v)} \right) \qquad \text{for all } \eta > 0,\ 0 < v \le 1,$$

for some constant K.

Step 2: Let us define $\mathscr{F}\delta := \{a \in \mathscr{F} : \delta/2 < \sigma^2(a) \le \delta\}$, for $\delta \in (0, 1]$, and

$$\Pi(\delta) := \mathbb{P}\left(\frac{X^2(a)}{\sigma^2(a)} > 2V \log(\frac{1}{\delta}) + S \log\log(\frac{e^e}{\delta}) \text{ for some } a \in \mathscr{F}(\delta) \right) \tag{A.3}$$

for $S \ge 4p + 1$. In this step we will prove that

$$\Pi(\delta) \le K \exp((K - S/K) \log\log(e^e/\delta))$$

for some constant K.

Fix $u < 1/2$. Let $\mathscr{F}(\delta, u)$ be a $\sqrt{u\delta}$-packing set of $\mathscr{F}\delta)$. By our assumption the cardinality of $\mathscr{F}(\delta, u)$ is less than or equal to $Au^{-B}\delta^{-V}(\log(e/\delta))^p$. Fix $a \in \mathscr{F}(\delta)$. From the definition of $\mathscr{F}(\delta, u)$ we can associate $\hat{a} \in \mathscr{F}(\delta, u)$ (corresponding to $a \in \mathscr{F}(\delta)$) such that $\rho^2(a, \hat{a}) \le u\delta$. Using assumption (a) of Theorem 2.1.2 we have

$$\sigma^2(a) \ge \sigma^2(\hat{a}) - u\delta \ge \sigma^2(\hat{a})(1 - 2u) \tag{A.4}$$

where the last inequality follows from the fact that $\hat{a} \in \mathscr{F}\delta)$ (thus $\sigma^2(\hat{a}) > \delta/2$).

We want to study the event

$$\frac{X^2(a)}{\sigma^2(a)} > r \tag{A.5}$$

for some $r > 0$. Obviously, for any $\lambda \in (0, 1)$, either (i) $|X(a) - X(\hat{a})|^2 > \lambda^2 X^2(a)$ or (ii) $|X(a) - X(\hat{a})|^2 \leq \lambda^2 X^2(a)$ (which, in particular implies $|X(\hat{a})| \geq (1 - \lambda)|X(a)|$). The above two cases reduce to:

$$\Gamma(X, (u\delta)^{1/2})^2 \geq |X(a) - X(\hat{a})|^2 > \lambda^2 X^2(a) \geq \lambda^2 r\sigma^2(a) \geq \lambda^2 r\frac{\delta}{2} \tag{A.6}$$

(here the first inequality follows from the definition of $\Gamma(X, (u\delta)^{1/2})$ and the third inequality follows from condition (A.5)), and

$$X^2(\hat{a}) \geq (1 - \lambda)^2 X^2(a) \geq (1 - \lambda)^2 r\sigma^2(a) \geq (1 - \lambda)^2 r(1 - 2u)\sigma^2(\hat{a}) \tag{A.7}$$

(here the second inequality follows from (A.5) and last inequality follows from (A.4)). Therefore, for any $r > 0$,

$$\begin{aligned} \Pi_r(\delta) \quad &:= \quad \mathbb{P}\left(\frac{X^2(a)}{\sigma^2(a)} > r \text{ for some } a \in \mathscr{F}\delta\right) \\ &\leq \quad \mathbb{P}\left(\Gamma(X, (u\delta)^{1/2})^2 > \lambda^2\delta r/2\right) \\ &\qquad\qquad + \sum_{\hat{a}\in\mathscr{F}(\delta,u)} \mathbb{P}\left(X^2(\hat{a})/\sigma^2(\hat{a}) > (1 - \lambda)^2 r(1 - 2u)\right) \end{aligned}$$

where we have used the fact that if $X^2(a)/\sigma^2(a) > r$ for some $a \in \mathscr{F}$, then either (A.6) holds or (A.7) is satisfied for some $\hat{a} \in \mathscr{F}(\delta, u)$. The first term on the right side of the above display can be bounded by appealing to (A.1) with $\eta = \sqrt{\lambda^2\delta r/2}$ and $v = \sqrt{u\delta}$ and the second term can be bounded by using conditions (a) and (c) of Theorem 2.1.2. Hence we get

$$
\begin{aligned}
\Pi_r(\delta) &\leq K \exp\left(-\frac{\lambda^2 \delta r/2}{Ku\delta \log(e/\sqrt{u\delta})}\right) \\
&\qquad\qquad + Au^{-B}\delta^{-V}\big(\log(\frac{e}{\delta})\big)^p \exp\left(-\frac{(1-\lambda)^2 r(1-2u)}{2}\right) \\
&\leq K\Big[\exp\left(-\frac{\lambda^2 r}{Ku\log(e/(u\delta))}\right) \\
&\qquad + \exp\big(B\log(1/u) + V\log(1/\delta) + p\log\log(e/\delta) + ur - (1/2-\lambda)r\big)\Big]. \qquad \text{(A.8)}
\end{aligned}
$$

Fix $S \geq 8p+1$ and set

$$
r := 2V\log(1/\delta) + S\log\log\big(\frac{e^e}{\delta}\big)
$$

and

$$
\lambda := \frac{1}{r}\Big((S/4)\log\log(e^e/\delta) - p\log\log(e/\delta)\Big).
$$

Observe that $r > 1$ and $0 < \lambda < 1/4$. Moreover, we have

$$
(1/2-\lambda)r = V\log(1/\delta) + p\log\log(e/\delta) + (S/4)\log\log(e^e/\delta).
$$

Putting these values in (A.8) gives us

$$
\begin{aligned}
\Pi(\delta) \equiv \Pi_r(\delta) \quad \leq \quad & K\left[\exp\left(-\frac{(S-4p)^2(\log\log(e^e/\delta))^2}{Kur\log(e/(u\delta))}\right)\right. \\
& \qquad \left. + \exp\Big(B\log(1/u) + ur - (S/4)\log\log(e^e/\delta)\Big)\right] \qquad \text{(A.9)}
\end{aligned}
$$

where we have used the fact that $\lambda^2 r^2 = ((S/4)\log\log(e^e/\delta) - p\log\log(e/\delta))^2 \geq (S-4p)^2(\log\log(e^e/\delta))^2/16$. Now, let us pick

$$
u := \frac{S}{8r\log(e/\delta)} < \frac{1}{2}.
$$

Then we have $\frac{1}{u} \leq K\log^2(e/\delta)$ for some constant K. Let us consider the two terms on the right side of (A.9) separately. For the first term, using $ur = S[\log(e/\delta)]^{-1}/8$, and that $\frac{1}{u} \leq K\log^2(e/\delta)$, we have

$$
\begin{aligned}
\frac{(S-4p)^2(\log\log(e^e/\delta))^2}{Kur\log(e/(u\delta))} &= \frac{8(S-8p+16p^2/S)(\log\log(e^e/\delta))^2\log(e/\delta)}{K\big(\log(e/\delta)+\log(u^{-1})\big)} \\
&\geq (S-8p)\Big(\frac{(\log\log(e^e/\delta))^2\log(e/\delta)}{K\big(\log(e/\delta)+\log K+2\log\log(e/\delta)\big)}\Big) \\
&\geq (1/K')(S-8p)(\log\log(e^e/\delta)).
\end{aligned}
$$

Here the last inequality follows from the following fact: As

$$
\tau(\delta) := \frac{(\log\log(e^e/\delta))\log(e/\delta)}{K\big(\log(e/\delta)+\log K+2\log\log(e/\delta)\big)} \to \infty, \qquad \text{as } \delta \to 0,
$$

we can find a lower bound $K' > 0$ such that $\tau(\delta) \geq 1/K'$ for all $\delta \in (0,1]$.

For the second term on the right side of (A.9) we have

$$
\begin{aligned}
& B\log(1/u) + ur - (S/4)\log\log(e^e/\delta) \\
\leq\ & B\log K + 2B\log\log(e/\delta) + S/8 - (S/4)\log\log(e^e/\delta) \\
\leq\ & B\log K + 2B\log\log(e/\delta) - (S/8)\log\log(e^e/\delta) \\
\leq\ & B\log K + (2B - S/8)\log\log(e^e/\delta).
\end{aligned}
$$

Thus, both the terms on the right side of (A.9) have the form $K\exp[(C - S/K')\log\log(e^e/\delta)]$ for some constants $K, C, K' > 0$. Putting these values in (A.9) gives us, for suitable constant $K > 0$, we get

$$
\Pi(\delta) \leq K\exp\left((K - S/K)\log\log(e^e/\delta)\right).
$$

Step 3: In this step we will prove that as $S \to \infty$

$$
\mathbb{P}\left(\frac{X^2(a)}{\sigma^2(a)} > 2V\log(1/\sigma^2(a)) + S\log\log\left(\frac{e^e}{\sigma^2(a)}\right) \text{ for some } a \in \mathscr{F}\right) \to 0.
$$

First let us define

$$\tilde{\Pi}(\delta) := \mathbb{P}\left(\frac{X^2(a)}{\sigma^2(a)} > 2V\log(1/\sigma^2(a)) + S\log\log\left(\frac{e^e}{\sigma^2(a)}\right) \text{ for some } a \in \mathscr{F}(\delta)\right).$$

Comparing with (A.3) we can see that for any $\delta \in (0, 1]$,

$$\tilde{\Pi}(\delta) \leq \Pi(\delta)$$

as: If $a \in \mathscr{F}(\delta)$ then $\sigma^2(a) \leq \delta$ and $x \longmapsto 2V\log(1/x) + S\log\log(e^e/x)$ is a decreasing function of x. Hence, we have

$$\tilde{\Pi}(\delta) \leq K\exp\left((K - S/K)\log\log(e^e/\delta)\right).$$

Therefore, for $S > 0$ such that $S/K > K + 1$ (as $\mathscr{F} = \bigcup_{l\geq 0}\mathscr{F}(2^{-l})$),

$$\begin{aligned}
&\mathbb{P}\left(\frac{X^2(a)}{\sigma^2(a)} > 2V\ \log(1/\sigma^2(a)) + S\log\log\left(\frac{e^e}{\sigma^2(a)}\right) \text{ for some } a \in \mathscr{F}\right) \\
&\qquad \leq \sum_{l=0}^{\infty}\tilde{\Pi}(2^{-l}) \\
&\qquad \leq K\sum_{l=0}^{\infty}\exp((K - S/K)\log\log(e^e 2^l)) \\
&\qquad = K\sum_{l=0}^{\infty}(e + l\log 2)^{-(S/K-K)} \\
&\qquad \leq K\sum_{j=2}^{\infty} j^{-(S/K-K)}.
\end{aligned}$$

Note that the last term can be further upper bounded as

$$K\sum_{j=2}^{\infty} j^{-(S/K-K)} \leq K\int_2^{\infty} x^{-(S/K-K)}dx \leq \frac{K\,2^{-(S/K-K)+1}}{(S/K-K)-1} \leq \xi_1\exp(-S/\xi_2)$$

for some constants ξ_1 and ξ_2 depending only on the constants K, L, M, A, B, p, V. This proves that $S(X) := \sup_{a\in\mathscr{F}}\frac{X^2(a)/\sigma^2(a) - 2V\log(1/\sigma^2(a))}{\log\log(e^e/\sigma^2(a))}$ is a subexponential random variable.

A.2.1 Proof of Lemma 2.1.1

First let us define the following sets:

$$\begin{aligned}
\mathscr{F}_{\delta,(l_1,\ldots,l_d)} &:= \Big\{(t,h)\in\mathscr{F} : \delta/2 < \sigma^2(t,h) \le \delta,\ 2^{l_i-1} < \frac{h_i}{\delta^{1/d}} \le 2^{l_i}, \\
&\qquad \forall\, i = 1,\ldots,d\Big\} \text{ for some } (l_1,\ldots,l_d)\in\mathbb{Z}^d, \\
\mathscr{F}(\delta) &:= \Big\{(t,h)\in\mathscr{F} : \delta/2 < \sigma^2(t,h) \le \delta\Big\}.
\end{aligned}$$

We note that $\mathscr{F}_{\delta,(l_1,\ldots,l_d)}$ is empty unless we have

$$\text{(i)} \quad l_i \le (1/d)\log_2(1/\delta) \qquad \text{for all } i = 1,\ldots,d;$$

(this restriction is a consequence of the fact that $h_i \le 1/2$) and

$$\text{(ii)} \quad -(d+1) < \sum_{i=1}^{d} l_i \le 0$$

(this restriction is a consequence of the fact that $\delta/2 < \sigma^2(t,h) \le \delta$).

Step 1: First, we will show that for any $(l_1,\ldots,l_d)\in\mathbb{Z}^d$, and $\delta, u \in (0,1]$,

$$N\left((u\delta)^{1/2}, \mathscr{F}_{\delta,(l_1,\ldots,l_d)}\right) \le K u^{-2d}\delta^{-1}. \tag{A.10}$$

Let $\mathscr{F}'$ be a subset of $\mathscr{F}_{\delta,(l_1,\ldots,l_d)}$ such that for any two elements $(t,h),(t',h')\in\mathscr{F}'$ we have

$$\rho^2((t,h),(t',h')) > u\delta. \tag{A.11}$$

Our aim is to show that

$$|\mathscr{F}'| \le K u^{-2d}\delta^{-1},$$

for some constant K independent of $(l_1, \ldots, l_d)$, u and δ. If $\mathscr{F}_{\delta,(l_1,\ldots,l_d)}$ is empty then the assertion is trivial. So assume that $\mathscr{F}_{\delta,(l_1,\ldots,l_d)}$ is non-empty which imposes bounds on the l_i's as shown above.

Let us define the following partition of $[0,1]^d$ into disjoint hyperrectangles:

$$R := \Big\{ M_{(i_1,\ldots,i_d)} \cap [0,1]^d : M_{(i_1,\ldots,i_d)} := \Pi_{k=1}^{d} \Big((i_k - 1)\frac{u\delta^{\frac{1}{d}}2^{l_k}}{c}, i_k \frac{u\delta^{\frac{1}{d}}2^{l_k}}{c} \Big], \\ 1 \le i_k \le \lceil c u^{-1} \delta^{-\frac{1}{d}} 2^{-l_k} \rceil \Big\}$$

where we take $c := d4^d$. We would like to point out that in the above definition when $i_k = 1$, for any $k = 1, \ldots, d$, by $\big((i_k - 1)c^{-1}u\delta^{1/d}2^{l_k}, i_k c^{-1}u\delta^{1/d}2^{l_k}\big]$ we mean the closed interval $\big[0, c^{-1}u\delta^{1/d}2^{l_k}\big]$. Observe that all the sets in R are disjoint and moreover $\bigcup_{M \in R} M = [0,1]^d$. Observe that

$$\begin{aligned} 2^{l_i - 1}\delta^{1/d} < h_i \le 1/2 \;\; \Rightarrow \;\; 2^{l_i}\delta^{1/d} < 1 \;\; &\Rightarrow \;\; cu^{-1}\delta^{-1/d}2^{-l_i} > 1 \\ &\Rightarrow \;\; \lceil cu^{-1}\delta^{-1/d}2^{-l_i} \rceil \le 2cu^{-1}\delta^{-1/d}2^{-l_i}. \end{aligned}$$

Hence we can easily see that

$$|R| = \Pi_{i=1}^{d} \lceil cu^{-1}\delta^{-1/d}2^{-l_i} \rceil \le 2^d c^d u^{-d} \delta^{-1} 2^{-\sum_{i=1}^{d} l_i} \le 2^{2d+1} c^d u^{-d} \delta^{-1}.$$

Here the last inequality follows from the fact that $\sum_{i=1}^{d} l_i \ge -(d+1)$. Let us define the following set:

$$R_2 := \Big\{ (M_{\underset{\sim}{i}}, M_{\underset{\sim}{i}'}) \in R \times R : \exists\, (t,h) \in \mathscr{F}' \text{ s.t. } t - h \in M_{\underset{\sim}{i}} \text{ and } t + h \in M_{\underset{\sim}{i}'} \Big\}.$$

Note that if $(t,h) \in \mathscr{F}'$ then $h_k \le 2^{l_k}\delta^{1/d}$ for all $k = 1, \ldots, d$. This implies that if $(M_{\underset{\sim}{i}}, M_{\underset{\sim}{i}'}) \in R_2$, where $\underset{\sim}{i} = (i_1, \ldots, i_d)$ and $\underset{\sim}{i}' = (i'_1, \ldots, i'_d)$, then

$$(i'_k - i_k) \le (1 + 2cu^{-1}), \qquad \text{for all } \; k = 1, \ldots, d, \tag{A.12}$$

as (i) $(i'_k - 1)u\delta^{1/d}2^{l_k}c^{-1} \leq t_k + h_k$, and (ii) $i_k u\delta^{\frac{1}{d}}2^{l_k}c^{-1} \geq t_k - h_k$. Thus for each hyperrectangle $M_{\underset{\sim}{i}} \in R$ the number of hyperrectangles $M_{\underset{\sim}{i'}} \in R$ such that $(M_{\underset{\sim}{i}}, M_{\underset{\sim}{i'}}) \in R_2$ is less than or equal to $(1 + 2cu^{-1})^d \leq 4^d c^d u^{-d}$. Hence we have

$$|R_2| \leq |R| \times 4^d c^d u^{-d} \leq 2^{4d+1}c^{2d}u^{-2d}\delta^{-1} \leq d^{2d}2^{4d^2+4d+1}u^{-2d}\delta^{-1}.$$

Thus, our proof will be complete if we can show that $|R_2| = |\mathscr{F}'|$. From the definition of R_2 and the fact that elements in R are disjoint it is easy to observe that $|R_2| \leq |\mathscr{F}'|$.

Therefore, the only thing left to show is that $|\mathscr{F}'| \leq |R_2|$. Let us assume the contrary, i.e., $|R_2| < |\mathscr{F}'|$. This implies that there exist two elements (t, h) and $(t', h') \in \mathscr{F}'$ and $(M_{\underset{\sim}{i}}, M_{\underset{\sim}{i'}}) \in R_2$ such that both $t - h$ and $t' - h'$ belong to $M_{\underset{\sim}{i}}$ and, also, $t + h$ and $t' + h'$ belong to $M_{\underset{\sim}{i'}}$. Let us first define the following two hyperrectangles:

$$B_1 := \Pi_{k=1}^d (i_k - 1, i'_k] \times c^{-1}u\delta^{1/d}2^{l_k} \quad \text{and} \quad B_2 := \Pi_{k=1}^d (i_k, i'_k - 1] \times c^{-1}u\delta^{1/d}2^{l_k}.$$

Our goal is to show that

$$B_\infty(t, h) \triangle B_\infty(t', h') \subseteq B_1 \setminus B_2 \tag{A.13}$$

which is implied by the following two assertions:

(1) $B_\infty(t, h) \cup B_\infty(t', h') \subseteq B_1$ and

(2) $B_2 \subseteq B_\infty(t, h) \cap B_\infty(t', h')$.

See Figure A.1 for a visual illustration of (A.13) when $d = 2$. Now, as $t - h \in M_{\underset{\sim}{i}}$, this implies $t_k - h_k \geq (i_k - 1)c^{-1}u\delta^{1/d}2^{l_k}$, for all $k = 1, \ldots, d$. Also $t + h \in M_{\underset{\sim}{i'}}$ implies that $t_k + h_k \leq i'_k c^{-1}u\delta^{1/d}2^{l_k}$, for all $k = 1, \ldots, d$. Therefore, $B_\infty(t, h) = \Pi_{i=1}^d (t_i - h_i, t_i + h_i) \subseteq B_1$. A similar argument shows that $B_\infty(t', h') \subseteq B_1$. Hence assertion (1) above holds.

Now as $t - h \in M_{\underset{\sim}{i}}$, we have $t_k - h_k \leq i_k c^{-1}u\delta^{1/d}2^{l_k}$, for all $k = 1, \ldots, d$. Also $t + h \in M_{\underset{\sim}{i'}}$ implies that $t_k + h_k \geq (i'_k - 1)c^{-1}u\delta^{1/d}2^{l_k}$, for all $k = 1, \ldots, d$. Hence we have $B_2 \subseteq B_\infty(t, h)$. A

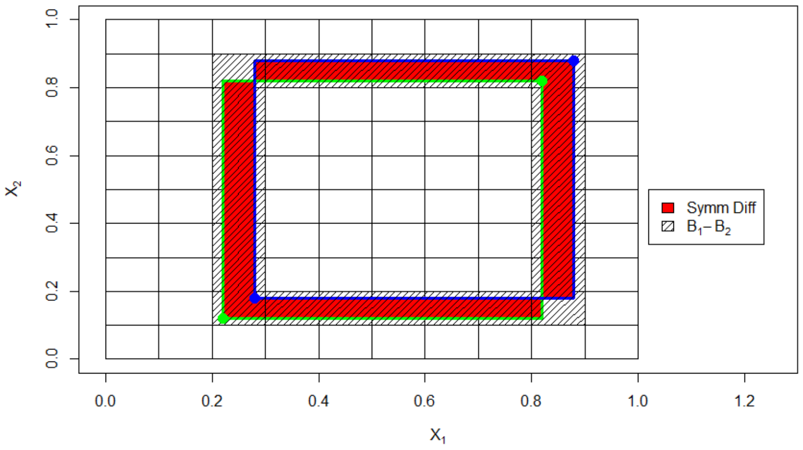

Figure A.1: The figure shows how the symmetric difference of the hyperrectangles $B_\infty(t, h)$ (denoted by the green border) and $B_\infty(t', h')$ (denoted by the blue border) is contained in the set $B_1 \backslash B_2$ (denoted by the shaded region).

similar argument shows that $B_2 \subseteq B_\infty(t', h')$. Therefore, assertion (2) is also satisfied. Now let us define the following set

$$I := \big\{ \underset{\sim}{j} = (j_1, \ldots, j_d) \in \mathbb{N}^d \quad : \quad j_k \in (i_k - 1, i'_k], \text{ for all } k = 1, \ldots, d, \\ \exists\, l \in \{1, \ldots, d\} \text{ such that } j_l = i_l \text{ or } i'_l \big\}.$$

Clearly, using (A.12),

$$|I| \leq 2d(2 + 2cu^{-1})^{d-1}.$$

Also see that $w = (w_1, \ldots, w_d) \in B_1 \setminus B_2$ if and only if

(1) for every $k = 1, \ldots, d$, we have $w_k \in \big(i_k - 1, i'_k\big] \times c^{-1}u\delta^{1/d}2^{l_k}$ (this is true as $w \in B_1$),

(2) there exists $l \in \{1, 2, \ldots, d\}$ such that either $w_l \in \big(i_l - 1, i_l\big] \times c^{-1}u\delta^{1/d}2^{l_l}$ or $w_l \in \big(i'_l - 1, i'_l\big] \times c^{-1}u\delta^{1/d}2^{l_l}$ (this is true as $w \notin B_2$ implies that there exist l such that $w_l \notin (i_l, i'_l - 1] \times c^{-1}u\delta^{1/d}2^{l_l}$ and $w \in B_1$ implies that $w_l \in (i_l - 1, i'_l] \times c^{-1}u\delta^{1/d}2^{l_l}$).

Therefore, we see that

$$B_1 \setminus B_2 = \bigcup_{\underset{\sim}{j} \in I} M_{\underset{\sim}{j}}.$$

Also, note that, $|M_{\underset{\sim}{j}}| \leq u^d \delta c^{-d} 2^{\sum_{i=1}^d l_i} \leq u^d \delta c^{-d}$ for all $\underset{\sim}{j}$. Therefore, using (A.13) and the fact that $c = d4^d$, we easily see that

$$\rho^2((t,h),(t',h')) \leq |B_1 \setminus B_2| \leq 2d(2+2cu^{-1})^{d-1}\frac{u^d\delta}{c^d} \leq 2^d d(1+c^{-1})^{d-1}\frac{u\delta}{c} < u\delta$$

which contradicts (A.11). This proves that two elements of $\mathscr{F}'$ cannot correspond to the same pair of hyperrectangles $(M_{\underset{\sim}{i}}, M_{\underset{\sim}{i'}}) \in R_2$. Hence we have proved (A.10).

Step 2: In this part of the proof we show that

$$N\left((u\delta)^{1/2}, \mathscr{F}(\delta)\right) \leq K u^{-2d} \delta^{-1} (\log(e/\delta))^{d-1}. \tag{A.14}$$

Let us define the set

$$S := \left\{(l_1, \ldots, {}_d) \in \mathbb{Z}^d : -(d+1) < \sum_{k=1}^d l_k \leq 0 \text{ and } l_k \leq \frac{1}{d}\log_2(1/\delta) \;\forall\; k = 1, \ldots, d\right\}.$$

Now it can be easily seen that $l := (l_1, \ldots, {}_d) \in S$ implies $l_k \geq -(d+1) - (d-1)(1/d)\log_2(1/\delta)$, for all $k = 1, \ldots, d$. This shows that each l_k can only take at most $(d+2) + \log_2(1/\delta) \leq (d+2) + \log(1/\delta)\log_2(e) \leq d + 2(\log(e/\delta))$ many values. This shows that

$$|S| \leq (d+1)(d+2\log(e/\delta))^{d-1} \leq (d+2)^d(\log(e/\delta))^{d-1}.$$

Note that the power of $(d + 2\log(e/\delta))$ in the above display is $d-1$ because if we fix the values of $l_1, l_2, \ldots, l_{d-1}$ then l_d can only take at most $(d+1)$ values such that $(l_1, l_2, \ldots l_d) \in S$ (as $\sum_{k=1}^d l_k$

can take at most $d + 1$ distinct values). Also note that

$$\mathscr{F}(\delta) \subseteq \bigcup_{l \in S} \mathscr{F}_{\delta,l}.$$

The above representation of $\mathscr{F}(\delta)$ along with the trivial fact that $N(\epsilon, \bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} N(\epsilon, A_i)$ gives us (A.14).

Step 3: In this step we will complete the proof of Lemma 2.1.1. We want control the $\sqrt{u\delta}$-packing number of the set $\{(t, h) \in \mathscr{F} : \sigma^2(t, h) \leq \delta\}$ which can be decomposed in the following way: for $u \in (0, 1]$,

$$\{(t, h) \in \mathscr{F} : \sigma^2(t, h) \leq \delta\} = \left(\bigcup_{l=0}^{\lfloor 1+\log_2(1/u) \rfloor} \mathscr{F}(\delta 2^{-l}) \right) \cup \{a \in \mathscr{F} : \sigma^2(a) \leq u\delta/2\}.$$

Now we can control the $\sqrt{u\delta}$-packing number of each of the above sets. First observe that $N((u\delta)^{1/2}, \{(t, h) \in \mathscr{F} : \sigma^2(t, h) \leq u\delta/2\}) = 1$. Also, for any $u \in (0, 2)$ and $\delta \in (0, 1]$ we have

$$N((u\delta)^{1/2}, \mathscr{F}(\delta)) \leq N((u\delta/2)^{1/2}, \mathscr{F}(\delta)) \leq K u^{-2d} \delta^{-1} (\log(e/\delta))^{d-1} \tag{A.15}$$

for some constant K. Putting $\delta \leftarrow \delta/2^l$ and $u \leftarrow 2^l u$ for $0 \leq l \leq \lfloor 1 + \log_2(1/u) \rfloor$ in (A.15) we get

$$N((u\delta)^{1/2}, \mathscr{F}(\delta 2^{-l})) \leq K 2^{-(2d-1)l} u^{-2d} \delta^{-1} (\log(e/\delta))^{d-1}.$$

Now from the trivial fact that $N(\epsilon, \bigcup_{i=1}^m A_i) \le \sum_{i=1}^m N(\epsilon, A_i)$ we get

$$\begin{aligned}
& N\left(\sqrt{u\delta}, \{(t,h) \in \mathscr{F} : \sigma^2(t,h) \le \delta\}\right) \\
\le \quad & \sum_{l=0}^{\lfloor 1+\log_2(1/u) \rfloor} N\left(\sqrt{u\delta}, \mathscr{F}(\delta 2^{-l})\right) + N\left(\sqrt{u\delta}, \{(t,h) \in \mathscr{F} : \sigma^2(t,h) \le u\delta/2\}\right) \\
\le \quad & 1 + Ku^{-2d}\delta^{-1}(\log(e/\delta))^{d-1} \sum_{l=0}^{\infty} 2^{-(2d-1)l} \\
\le \quad & 1 + 2Ku^{-2d}\delta^{-1}(\log(e/\delta))^{d-1} \;\le\; (2K+1)u^{-2d}\delta^{-1}(\log(e/\delta))^{d-1},
\end{aligned}$$

which proves Lemma 2.1.1.

A.3 Proof of Theorem 2.1.1

We use Theorem 2.1.2 to prove Theorem 2.1.1. Let us recall the definitions of $\mathscr{F}, \sigma$ and ρ as introduced just before Lemma 2.1.1 in the main article. Without loss of generality we assume that $\|\psi\| = 1$. For $h \in (0, 1/2]^d$, let us define the stochastic process

$$X(t,h) := 2^{d/2}(h_1 h_2 \dots h_d)^{1/2}\hat{\Psi}(t,h) = 2^{d/2}\int \psi_{t,h}(x)dW(x), \qquad t \in A_h,$$

where $W(\cdot)$ is the standard Brownian sheet on $[0,1]^d$. This defines a centered Gaussian process with $\mathrm{Var}\big(X(t,h)\big) = \sigma^2(t,h)$. Also by a standard calculation on the variance we have $\mathrm{Var}\big(X(t,h) - X(t',h')\big) \le 2^d TV^2(\psi)\rho^2((t,h),(t',h'))$ when the function ψ has finite HK-variation. Note that when ψ satisfy average Hölder condition with parameters $\gamma > 1/2$ and L we have $\mathrm{Var}\big(X(t,h) - X(t',h')\big) \le 2^d dL\rho^2((t,h),(t',h'))$. As $X(t,h)$ and $X(t,h) - X(t',h')$ have normal distributions this shows that conditions (a) and (b) of Theorem 2.1.2 are satisfied. Condition (c) is also satisfied because of Lemma 2.1.1. Thus, by an application of Theorem 2.1.2 we have

$$\mathbb{P}\left(\sup_{0<h\le 1/2}\sup_{t\in A_h} \frac{\hat{\Psi}^2(t,h) - 2\log(1/2^d h_1 h_2 ... h_d)}{\log\log(e^e/2^d h_1 h_2 ... h_d)} < S\right) \ge 1 - \xi_1 \exp(-S/\xi_2)$$

for some constants ξ_1 and ξ_2 and large enough S.

For notational simplicity, let us define $\kappa_1 := 2\log(1/\sigma^2(t,h))$ and $\kappa_2 := 2\sqrt{2}S\log\log(e^e/\sigma^2(t,h))$. Therefore,

$$\begin{aligned}
& \mathbb{P}\left(|\hat{\Psi}(t,h)| \le \sqrt{2\log\left(\frac{1}{\sigma^2(t,h)}\right)} + S\left(\frac{\log\log(e^e/\sigma^2(t,h))}{\log^{\frac{1}{2}}(1/\sigma^2(t,h))}\right)\forall(t,h) \in \mathscr{F}\right) \\
= \; & \mathbb{P}\left(|\hat{\Psi}(t,h)| \le \kappa_1^{1/2} + \kappa_1^{-1/2}\kappa_2/2 \quad \forall\ (t,h) \in \mathscr{F}\right) \\
= \; & \mathbb{P}\left(\hat{\Psi}(t,h)^2 \le \left(\kappa_1^{1/2} + \kappa_1^{-1/2}\kappa_2/2\right)^2 \forall(t,h) \ \in \mathscr{F}\right) \\
\ge \; & \mathbb{P}\left(\hat{\Psi}(t,h)^2 \le \kappa_1 + \kappa_2 \quad \forall(t,h) \in \mathscr{F}\right) \\
= \; & \mathbb{P}\left(\sup_{t,h\in\mathscr{F}} \frac{\hat{\Psi}^2(t,h) - 2\log(1/2^d h_1 h_2 ... h_d)}{\log\log(e^e/2^d h_1 h_2 ... h_d)} < 2\sqrt{2}S\right) \\
\ge \; & 1 - \xi_1 \exp\left(-\frac{2\sqrt{2}S}{\xi_2}\right).
\end{aligned}$$

A.4 Proof of some smaller Results

A.4.1 Proof of Proposition 2.1.1

The proof of this result follows from the following result. Suppose that $Z_1, \dots, Z_n$ are i.i.d. standard normal random variables. Then, we know that

$$\frac{\max_{1\le i\le n} Z_i}{\sqrt{2\log n}} \to 1 \quad \text{a.s.}$$

The above result follows trivially from [104, Theorem 1.1]. Let F_n be the distribution function of $\max_{1\le i\le n} Z_i/\sqrt{2\log n}$, i.e., $F_n(x) := \mathbb{P}(\max_{1\le i\le n} Z_i \le x\sqrt{2\log n})$, for $x \in \mathbb{R}$. Therefore, for every $x < 1$, we have $F_n(x) \to 0$. We want to show that

$$\sup_{(t,h)\in\mathscr{F}} |\hat{\Psi}(t,h)| - \Gamma_V(2^d h_1 \dots h_d) = \infty \quad \text{a.s.}$$

Hence it is enough to show that for every $s \in \mathbb{R}$ we have $\mathbb{P}(\sup_{(t,h)\in\mathscr{F}} |\hat{\Psi}(t,h)| - \Gamma_V(2^d h_1 \dots h_d) < s) = 0$. Fix $m \in \mathbb{N}$. Now,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{(t,h)\in\mathscr{F}} |\hat{\Psi}(t,h)| - \Gamma_V(2^d h_1 \dots h_d) < s\right) \\
\leq\ &\mathbb{P}\left(\sup_{t\in A_{\left(\frac{1}{2m},\dots,\frac{1}{2m}\right)}} \left|\hat{\Psi}\left(t,\left(\frac{1}{2m},\dots,\frac{1}{2m}\right)\right)\right| - \Gamma_V(m^{-d}) < s\right) \\
\leq\ &\mathbb{P}\left(\sup_{t\in A_m^\star} |\hat{\Psi}(t,(2m)^{-1})| - \Gamma_V(m^{-d}) < s\right)
\end{aligned}$$

where $A_m^\star := \{(t_1,\dots,t_d) : t_i = k_i/2m$ for some odd integer $k_i < 2m$, for all $i = 1,\dots,d\}$. Thus, the last term in the above display can be further upper bounded by

$$\mathbb{P}\left(\sup_{t\in A_m^\star} \frac{\hat{\Psi}(t,(2m)^{-1})}{\sqrt{2\log(m^d)}} - \sqrt{V} < \frac{s}{\sqrt{2\log(m^d)}}\right) = F_{m^d}(\sqrt{V} + s/\sqrt{2\log(m^d)}),$$

where we have used the fact that now we are dealing with m^d i.i.d. standard normal random variables. Now, for every $s > 0$, choose m such that $\sqrt{V} + s/\sqrt{2\log(m^d)} < 1 - \epsilon$, for some fixed $\epsilon > 0$. Hence, $F_{m^d}(\sqrt{V} + s/\sqrt{2\log(m^d)}) \leq F_{m^d}(1-\epsilon)$, if m is large enough. As this is true for all large m, taking $m \to \infty$ gives us the desired result.

A.4.2 Solution to (2.5)

Let $\psi \in \mathbb{H}_{\beta,1}$ such that $\psi(0) \geq 1$. Hence by the property of $\mathbb{H}_{\beta,1}$ we have

$$|\psi(x) - \psi(0)| \leq \|x\|^\beta, \qquad \text{for all } x \in \mathbb{R}^d,$$

which implies $\psi(x) \geq 1 - \|x\|^\beta$. Hence, on the set $\|x\| \leq 1$, we have $\psi(x) \geq 1 - \|x\|^\beta \geq 0$. Therefore, we have

$$\int_{\|x\|\leq 1} \psi^2(x)dx \geq \int_{\|x\|\leq 1} (1 - \|x\|^\beta)^2 dx \quad \Rightarrow \quad \|\psi\| \geq \|\psi_\beta\|,$$

where $\psi_\beta(x) = (1-\|x\|^\beta)\mathbb{I}(\|x\| \leq 1)$. Hence the only thing left to prove is that $\psi_\beta \in \mathbb{H}_{\beta,1}$. Suppose that $x, y \in \mathbb{R}^d$ such that $1 \geq \|x\| \geq \|y\|$. Then

$$0 \leq \psi_\beta(y) - \psi_\beta(x) = \|x\|^\beta - \|y\|^\beta \leq (\|x\| - \|y\|)^\beta \leq \|x - y\|^\beta .$$

Here the third inequality follows from the fact that when $\beta \leq 1$ the function $u \mapsto u^\beta$ is a β-Hölder continuous function; the last inequality follows from the triangle inequality. If $x, y \in \mathbb{R}^d$ such that $\|x\| \geq 1 \geq \|y\|$ then we have

$$0 \leq \psi_\beta(y) - \psi_\beta(x) = 1 - \|y\|^\beta \leq (1 - \|y\|)^\beta \leq (\|x\| - \|y\|)^\beta \leq \|x - y\|^\beta .$$

If $x, y \in \mathbb{R}^d$ is such that $\|x\| \geq \|y\| \geq 1$ then the assertion is trivial. Hence we have proved that ψ_β minimizes (2.5).

A.5 Proofs of Optimality of multiscale statistics

The proofs of Theorems 2.2.1 and 2.2.2 depend on the following lemma (stated and proved in [4, Lemma 6.2]).

Lemma A.5.1 *Let $Z_1, Z_2, \ldots$ be a sequence of independent standard normal variables. If $w_m := (1 - \epsilon_m)\sqrt{2 \log m}$ with $\lim_{m\to\infty} \epsilon_m \sqrt{\log m} = \infty$ and $\lim_{m\to\infty} \epsilon_m = 0$ then we have*

$$\lim_{m\to\infty} \mathbb{E} \left| \frac{1}{m} \sum_{i=1}^{m} \exp\left(w_m Z_i - \frac{w_m^2}{2} \right) - 1 \right| = 0.$$

A.5.1 Proof of Theorem 2.2.1

Proof of part (a). For any bandwidth $h = (h_1, \ldots, h_d) \in (0, 1/2]^d$ and $t = (t_1, \ldots, t_d) \in A_h$, let us define the function $g_t : [0, 1]^d \to \mathbb{R}$ as

$$g_t(x) := L \min(h)^\beta \psi_{t,h}^{(\beta)}(x), \quad \text{for } x \in [0, 1]^d,$$

where $\min(h) := \min\{h_1, h_2, \ldots, h_d\}$ and $\psi_{t,h}^{(\beta)}(x_1, \ldots, x_d) = \psi_\beta((x_1 - t_1)/h_1, \ldots, (x_d - t_d)/h_d)$. Elementary calculations show that $g_t \in \mathbb{H}_{\beta,L}$ and $\|g_t\|_\infty = L \min(h)^\beta$. Now let us define the set

$$S := \left\{t \in A_h : t_i = k_i h_i \text{ for some odd integer } k_i, i = 1, \ldots, d\right\}.$$

Let ϕ_n be an arbitrary test for (1.2) with level α. Then,

$$\begin{aligned}
\inf_{g \in \mathbb{H}_{\beta,L}:\|g\|_\infty = L\min(h)^\beta} \mathbb{E}_g[\phi_n(Y)] - \alpha \quad &\leq \min_{g_t : t \in S} \mathbb{E}_{g_t}[\phi_n(Y)] - \mathbb{E}_0[\phi_n(Y)] \\
&\leq |S|^{-1} \sum_{t \in S} \mathbb{E}_{g_t}[\phi_n(Y)] - \mathbb{E}_0[\phi_n(Y)] \\
&\leq \mathbb{E}_0\left[\left(|S|^{-1} \sum_{t \in S} \frac{dP_{g_t}}{dP_0}(Y) - 1\right)\phi_n(Y)\right] \\
&\leq \mathbb{E}_0\left||S|^{-1} \sum_{t \in S} \frac{dP_{g_t}}{dP_0}(Y) - 1\right|. \qquad \text{(A.16)}
\end{aligned}$$

Here P_0 denotes the measure of the process Y under the null hypothesis $f = 0$ and P_{g_t} denotes the measure of Y under the alternative $f = g_t$. Also for $g \in \mathbb{H}_{\beta,L}$, $\frac{dP_g}{dP_0}$ denotes the Radon-Nikodym derivative of the measure P_g with respect to the measure P_0. By Cameron-Martin-Girsanov's Theorem (see [101, Chapter 3] for more details about absolutely continuous measures and Radon-Nikodym derivatives) we get that

$$\log\left(\frac{dP_g}{dP_0}(Y)\right) = \sqrt{n} \int g dW - \frac{n}{2} \|g\|^2 .$$

For $g_t(\cdot) = L \min(h)^\beta \psi_{t,h}^{(\beta)}(\cdot)$, $\sqrt{n} \int g_t dW = \sqrt{n} L \left\|\psi_\beta\right\| \min(h)^\beta \sqrt{\Pi_{i=1}^d h_i} \hat{\Psi}(t, h)$. Observe that $\{Z_t \equiv \hat{\Psi}(t, h)\}_{t \in S}$ are i.i.d. standard normals; note that the independence of the normals arises from the disjoint supports of the functions $\{g_t : t \in S\}$. Let

$$w_n := \sqrt{n} L \left\|\psi_\beta\right\| \min(h)^\beta \sqrt{\Pi_{i=1}^d h_i}.$$

Then $\Gamma_t = \exp(w_n Z_t - \frac{w_n^2}{2})$ and we can write $\frac{dP_{g_t}}{dP_0}(Y) - 1 = \Gamma_t - 1$.

Hence we have $\mathbb{E}_0 \left| |S|^{-1} \sum_{t \in S} \frac{dP_{g_t}}{dP_0}(Y) - 1 \right| = \mathbb{E}_0 \left| |S|^{-1} \sum_{t \in S} \Gamma_t - 1 \right|$. According to Lemma A.5.1 the above term will go to zero if $|S| \to \infty$ and the corresponding w_n's satisfy:

$$\left(1 - \frac{w_n}{\sqrt{2 \log |S|}}\right) \to 0 \qquad \text{and} \qquad \sqrt{\log |S|}\left(1 - \frac{w_n}{\sqrt{2 \log |S|}}\right) \to \infty.$$

Now let us pick

$$h_1 = \ldots = h_d = L^{-\frac{2}{2\beta+d}} ((1 - \epsilon_n)\rho_n)^{1/\beta} \left(\left\|\psi_\beta\right\|^2 (2\beta + d)/2d\right)^{-1/(2\beta+d)} =: \tilde{h}.$$

Then,

$$\begin{aligned} w_n &= \sqrt{n} L \left\|\psi_\beta\right\| L^{-1} ((1 - \epsilon_n)\rho_n)^{\frac{2\beta+d}{2\beta}} \left(\left\|\psi_\beta\right\|^2 (2\beta + d)/2d\right)^{-1/2} \\ &= \sqrt{n}(1 - \epsilon_n)^{1+d/2\beta} \sqrt{\frac{\log n}{n}} \sqrt{(2d/(2\beta + d))} \\ &= \sqrt{(2d/(2\beta + d))}(1 - \epsilon_n)^{1+d/2\beta} \sqrt{\log n}. \end{aligned} \tag{A.17}$$

Also, as $n \to \infty$, $|S|/(\Pi_{i=1}^{d}(1/h_i)) \to 2^{-d}$. Therefore, for a suitable constant K,

$$\begin{aligned} \log |S| / \log n &= (-d \log \tilde{h} - d \log 2 + o(1))/\log n \\ &= [K + o(1) - (d/\beta) \log ((1 - \epsilon_n)\rho_n)]/\log n \\ &= \left(K + o(1) - \frac{d}{\beta} \log(1 - \epsilon_n) + \frac{d}{2\beta + d} \log\left(\frac{n}{\log n}\right)\right)/\log n \\ &\to \frac{d}{2\beta + d} \text{ as } n \to \infty. \end{aligned} \tag{A.18}$$

Also notice that for all large n, $\log |S| / \left(\frac{d}{2\beta+d} \log n\right) < 1$. Combining (A.17) and (A.18), we get

$$\frac{w_n}{\sqrt{2 \log |S|}} = \frac{w_n}{\sqrt{\log n}} \frac{\sqrt{\log n}}{\sqrt{2 \log |S|}} \to 1 \quad \text{as} \quad n \to \infty.$$

Similarly, for suitable constants $K, K' > 0$,

$$\begin{aligned}\sqrt{\log|S|}\left(1-\frac{w_n}{\sqrt{2\log|S|}}\right) &\geq \sqrt{K}\sqrt{\log n}\left(1-(1-\epsilon_n)^{1+d/2\beta}+o(1)\right)\\ &\geq \sqrt{K'}\sqrt{\log n}\,(\epsilon_n+o(1)) \to \infty \quad \text{as } n\to\infty,\end{aligned}$$

as the o(1) term above is positive when n is large. This proves part (a) of Theorem 2.2.1 by noting that $L\min(h)^\beta = (1-\epsilon_n)c_*\rho_n$.

Proof of part (b). Let $\delta \equiv \delta_n := c_*\rho_n$ and $h_{i,n} = (\delta/L)^{1/\beta} =: \tilde{h}_n$ for all $i = 1, 2, \ldots, d$. For notational simplicity, in the following we drop the subscript n. As the term $D(2^d h_1 \ldots h_d)$ is bounded from above, for any $t \in J \equiv J_n$, the probability of rejecting the null hypothesis, $\mathbb{P}_g(T_\beta(Y) > \kappa_\alpha)$, is bounded from below by, for some constant $K > 0$,

$$\begin{aligned}&\mathbb{P}_g\left(|\hat{\Psi}(t,h)| > \Gamma(2^d\tilde{h}^d)+K\right)\\ &= \mathbb{P}_0\left(\left|\hat{\Psi}(t,h)+\sqrt{\frac{n}{\tilde{h}^d}}\left\|\psi_\beta\right\|^{-1}\langle g,\psi_{t,h}^{(\beta)}\rangle\right| > \Gamma(2^d\tilde{h}^d)+K\right)\\ &\geq \mathbb{P}_0\left(-\text{sign}(\langle g,\psi_{t,h}^{(\beta)}\rangle)\hat{\Psi}(t,h) < \sqrt{\frac{n}{\tilde{h}^d}}\frac{|\langle g,\psi_{t,h}^{(\beta)}\rangle|}{\|\psi_\beta\|} - K - \Gamma(2^d\tilde{h}^d)\right)\\ &= \Phi\left(\sqrt{\frac{n}{\tilde{h}^d}}\left\|\psi_\beta\right\|^{-1}|\langle g,\psi_{t,h}^{(\beta)}\rangle| - K - \Gamma(2^d\tilde{h}^d)\right) \qquad \text{(A.19)}\end{aligned}$$

where Φ is the standard normal distribution function. Hence, to prove our claim it suffices to show that

$$(1+\epsilon_n)\max_{t\in J}\sqrt{\frac{n}{\tilde{h}^d}}\left\|\psi_\beta\right\|^{-1}|\langle g,\psi_{t,h}^{(\beta)}\rangle| - \Gamma(2^d\tilde{h}^d) \to \infty$$

uniformly for all $g \in \mathbb{H}_{\beta,L}$ such that $\|g\|_{J,\infty} \geq \delta$. Note that $A_h = J$.

Let g be any such function, and let $t \in J$ be such that $|g(t)| \geq \delta$. Let us assume that $g(t) \geq \delta$; the other case where $g(t) \leq -\delta$ can be handled similarly by looking at $-g$. By construction of ψ_β we have $\delta\psi_{t,h}^{(\beta)} \in \mathbb{H}_{\beta,L}$. Also note that as ψ_β minimizes $\|\psi\|$ in the set $\{\psi \in \mathbb{H}_{\beta,1} : \psi(0) \geq 1\}$,

$\delta\psi_{t,h}^{(\beta)}$ minimizes $\|\psi\|$ in the set $\{\psi \in \mathbb{H}_{\beta,L} : \psi(t) \geq \delta\}$. Note that both g and $\delta\psi_{t,h}^{(\beta)}$ belong to the closed convex set $\{\psi \in \mathbb{H}_{\beta,L} : \psi(t) \geq \delta\}$. As $\delta\psi_{t,h}^{(\beta)}$ is the projection of the zero function onto the above closed convex set, we have

$$|\langle \psi_{t,h}^{(\beta)}, g\rangle| = \delta^{-1}|\langle \delta\psi_{t,h}^{(\beta)}, g\rangle| \geq \delta^{-1}\|\delta\psi_{t,h}^{(\beta)}\|^2 = \delta \left\|\psi_\beta\right\|^2 \tilde{h}^d.$$

Thus,

$$\begin{aligned}
&(1+\epsilon_n)\max_{t\in J}\sqrt{\frac{n}{\tilde{h}^d}}\left\|\psi_\beta\right\|^{-1}|\langle g, \psi_{t,h}^{(\beta)}\rangle| - \Gamma(2^d\tilde{h}^d)\\
\geq\ &(1+\epsilon_n)\left\|\psi_\beta\right\|\delta\sqrt{n\tilde{h}^d} - \Gamma(2^d\tilde{h}^d)\\
=\ &(1+\epsilon_n)\left\|\psi_\beta\right\| c_*\rho_n\sqrt{n}(c_*\rho_n)^{d/2\beta}L^{-d/2\beta} - \Gamma(2^d\tilde{h}^d)\\
=\ &(1+\epsilon_n)\sqrt{\left(\frac{2d}{2\beta+d}\right)\log n} - \sqrt{K + \left(\frac{2d}{2\beta+d}\right)\log\left(\frac{n}{\log n}\right)}\\
\geq\ &\epsilon_n(2d/(2\beta+d))^{1/2}(\log n)^{1/2} + o(1) \to \infty.
\end{aligned}$$

This proves part (b) of Theorem 2.2.1.

Proof of Proposition 2.2.1

Let $h := (\tilde{h}, \ldots, \tilde{h}) \in \mathbb{R}^d$, where $\tilde{h} = (M\rho_n/L)^{1/\beta}$, for M as defined in the statement of the proposition. By the same argument as in (A.19) we have

$$\mathbb{P}_g(T(Y) > \kappa_\alpha) \geq \Phi\left(\sqrt{\frac{n}{\tilde{h}^d}}\|\psi_1\|^{-1}|\langle g, \psi_{t,h}^{(1)}\rangle| - K - \Gamma(2^d\tilde{h}^d)\right).$$

Now we would want to bound $|\langle g, \psi_{t,h}^{(1)}\rangle|$ uniformly for all $g \in \mathbb{H}_{\beta,L}$ such that $\|g\|_{J_n,\infty} \geq M\rho_n$. Without loss of generality, let us assume that $g(t) \geq M\rho_n$ for some $t \in J_n$ and $g \in \mathbb{H}_{\beta,L}$. Then

$$g(x) \geq g(t) - L\|x-t\|^\beta \geq M\rho_n - L\|x-t\|^\beta = M\rho_n\left(1 - \left\|\frac{x-t}{\tilde{h}}\right\|^\beta\right).$$

This shows that if $\|x-t\| \leq \tilde{h}$ then $g(x) \geq 0$. Hence,

$$\begin{aligned}
\langle g, \psi^{(1)}_{t,h} \rangle &\geq \int_{\|x-t\| \leq \tilde{h}} M\rho_n \left(1 - \left\|\frac{x-t}{\tilde{h}}\right\|^{\beta}\right) \left(1 - \left\|\frac{x-t}{\tilde{h}}\right\|\right) dx \\
&= M\rho_n \tilde{h}^d \int_{\|x\| \leq 1} (1 - \|x\|) \left(1 - \|x\|^{\beta}\right) dx \\
&= M\rho_n \tilde{h}^d \langle \psi_{\beta}, \psi_1 \rangle.
\end{aligned}$$

Here the last equality follows as $\psi_\beta(x) = (1 - \|x\|^{\beta})\mathbb{I}(\|x\| \leq 1)$. Also note that

$$\Gamma(2^d \tilde{h}^d) = \sqrt{2d \log\left(\frac{1}{2}\right) + \frac{2d}{\beta}\log\left(\frac{L}{M}\right) + \frac{2d}{2\beta+d}\log\left(\frac{n}{\log n}\right)} \leq \sqrt{\frac{2d}{2\beta+d}\log n}$$

for large n. Therefore, for large n,

$$\begin{aligned}
&\sqrt{\frac{n}{\tilde{h}^d}}\, \|\psi_1\|^{-1} \, |\langle g, \psi^{(1)}_{t,h} \rangle| - K - \Gamma(2^d \tilde{h}^d) \\
\geq \quad & \sqrt{n\tilde{h}^d} M\rho_n \frac{\langle \psi_\beta, \psi_1 \rangle}{\|\psi_1\|} - K - \sqrt{\frac{2d}{2\beta+d}\log n} \\
= \quad & -K + \sqrt{\log n}\left(L^{-d/2\beta} M^{\frac{(d+2\beta)}{2\beta}} \frac{\langle \psi_\beta, \psi_1 \rangle}{\|\psi_1\|} - \sqrt{\frac{2d}{2\beta+d}}\right) \to \infty \text{ as } n \to \infty.
\end{aligned}$$

Here the last equality holds by the choice of M, as

$$\begin{aligned}
\sqrt{n\tilde{h}^d} M\rho_n \frac{\langle \psi_\beta, \psi_1 \rangle}{\|\psi_1\|} &= \sqrt{n} M^{\frac{d}{2\beta}} \rho_n^{\frac{d}{2\beta}} L^{-\frac{d}{2\beta}} M\rho_n \frac{\langle \psi_\beta, \psi_1 \rangle}{\|\psi_1\|} \\
&= \sqrt{\log n}\, L^{-d/2\beta} M^{\frac{(d+2\beta)}{2\beta}} \frac{\langle \psi_\beta, \psi_1 \rangle}{\|\psi_1\|} \\
&> \sqrt{\log n} \sqrt{\frac{2d}{2\beta+d}}.
\end{aligned}$$

Hence $\lim_{n\to\infty} \mathbb{P}_g(T(Y) > \kappa_\alpha) = 1$.

A.5.2 Proof of Theorem 2.2.2

Proof of part (a). Let us suppose that $B_n := B_\infty(t_n, h_n) \subseteq [0,1]^d$ for some $t_n, h_n \in [0,1]^d$. Let us first look at the case when $\liminf_{n\to\infty} |B_n| > 0$. Now assume that the location B_n was known and it was also known that $\mu_n > 0$. In such a scenario the best test statistic would be $\hat{\Psi}(t_n, h_n)$ (with kernel ψ_0) which follows the normal distribution with mean 0 and variance 1, under the null hypothesis. Hence in this case, the UMP test rejects $H_0 : \mu_n = 0$ if $\hat{\Psi}(t_n, h_n) > z_{1-\alpha}$ where $z_{1-\alpha}$ is the $(1-\alpha)$'th quantile of the standard normal distribution. When B_n is not known then, obviously, the power of any level α test ϕ_n is less than the test described above. Hence,

$$\begin{aligned}
\mathbb{E}_{f_n}[\phi_n(Y)] &\leq \mathbb{P}_{\mu_n}\left(\hat{\Psi}(t_n, h_n) \geq z_{1-\alpha}\right) = \mathbb{P}_0\left(\hat{\Psi}(t_n, h_n) + \sqrt{n|B_n|}\mu_n \geq z_{1-\alpha}\right) \\
&= 1 - \Phi\left(z_{1-\alpha} - \sqrt{n|B_n|}\mu_n\right) \\
&\not\to 1 \text{ unless } \mu_n\sqrt{n|B_n|} \to \infty.
\end{aligned}$$

A similar argument can be made when $\mu_n < 0$ as well. Hence the power of any level α test does not go to 1 unless $|\mu_n|\sqrt{n|B_n|} \to \infty$.

Now suppose that $|\mu_n|\sqrt{n|B_n|} \to \infty$. Then we will show that $\lim_{n\to\infty} \mathbb{P}_{f_n}(T > \kappa_\alpha) = 1$. Without loss of generality assume $\mu_n > 0$. Hence,

$$\begin{aligned}
\mathbb{P}_{f_n}(T > \kappa_\alpha) &\geq \mathbb{P}_{f_n}\left(\frac{|\hat{\Psi}(t_n, h_n)| - \Gamma(|B_n|)}{D(|B_n|)} > \kappa_\alpha\right) \\
&= \mathbb{P}_0\left(\left|\hat{\Psi}(t_n, h_n) + \mu_n\sqrt{n|B_n|}\right| - \Gamma(|B_n|) \geq \kappa_\alpha D(|B_n|)\right) \\
&\geq \mathbb{P}_0\left(\left|\hat{\Psi}(t_n, h_n) + \mu_n\sqrt{n|B_n|}\right| \geq K\right) \to 1 \text{ as } \mu_n\sqrt{n|B_n|} \to \infty.
\end{aligned}$$

Here the last inequality follows from the fact that as $\liminf_n |B_n| > 0$, $\Gamma(|B_n|) + \kappa_\alpha D(|B_n|)$ is bounded from above (say, by K) for all large n.

Proof of part (b). Now let us look at the case $\lim |B_n| \to 0$. Let us assume that $|\mu_n|\sqrt{n|B_n|} = (1-\epsilon_n)\sqrt{2\log(1/|B_n|)}$ where $\epsilon_n \to 0$ and also $\epsilon_n\sqrt{2\log(1/|B_n|)} \to \infty$. Without loss of generality also assume that $\mu_n > 0$. Recall that $B_n = B_\infty(t_n, h_n)$ for $h_n = (h_{1,n}, \ldots, h_{d,n}) \in (0, 1/2]^d$. Let us first define the following grid points:

$$G_{h_n} := \left\{ t = (t_1, \ldots, t_d) \in [0,1]^d : t_i = (2k_i - 1)h_{i,n} \text{ for } k_i \in \mathbb{N}, B_\infty(t, h_n) \subseteq [0,1]^d \right\}.$$

Clearly $|G_{h_n}| \leq 1/|B_n|$. Also, as $n \to \infty$, $|G_{h_n}||B_n| \to 1$. For each $t \in G_{h_n}$ define $f_t := \mu_n \mathbb{I}_{B_\infty(t,h_n)}$. Clearly as $|B_n| = |B_\infty(t, h_n)|$, we have $f_t \in \mathcal{G}_n^-$. Let ϕ_n be a test of level α for testing (2.2). Similar arguments as in (A.16) show that

$$\inf_{g \in \mathcal{G}_n^-} \mathbb{E}_g \phi_n(Y) - \alpha \leq \mathbb{E}_0 \left| |G_{h_n}|^{-1} \sum_{t \in G_{h_n}} \frac{dP_{f_t}}{dP_0}(Y) - 1 \right|.$$

Now by an argument similar to that in the proof of Theorem 2.2.1, we have

$$\log\left(\frac{dP_{f_t}}{dP_0}(Y)\right) = \sqrt{n} \int f_t dW - n \|f_t\|^2 /2 = \mu_n \sqrt{n|B_n|} \hat{\Psi}(t, h_n) - \mu_n^2 n |B_n|/2.$$

Also note that the collection of random variables in $\{\hat{\Psi}(t, h_n) : t \in G_{h_n}\}$ are mutually independent. Now putting $w_n = \mu_n \sqrt{n|B_n|} = (1 - \epsilon_n)\sqrt{2 \log(1/|B_n|)}$ and $m = |G_{h_n}|$ we see that

$$\mathbb{E}_0 \left| |G_{h_n}|^{-1} \sum_{t \in G} \frac{dP_{f_t}}{dP_0}(Y) - 1 \right| \to 0$$

if $\epsilon_n \to 0$ and $\epsilon_n \sqrt{\log(1/|B_n|)} \to \infty$, by a direct application of Lemma A.5.1. This proves that

$$\limsup_{n \to \infty} \inf_{f_n \in \mathcal{G}_n^-} \mathbb{E}_{f_n} \phi_n \leq \alpha.$$

Now let us assume that $|\mu_n|\sqrt{n|B_n|} \geq (1 + \epsilon_n)\sqrt{2 \log(1/|B_n|)}$. Without loss of generality also assume that $\mu_n > 0$. A similar argument as in part (a) shows that

$$\begin{aligned}
\mathbb{P}_{f_n}(T > \kappa_\alpha) &\geq \mathbb{P}_{f_n}\left(\frac{|\hat{\Psi}(t_n,h_n)| - \Gamma(|B_n|)}{D(|B_n|)} > \kappa_\alpha\right) \\
&= \mathbb{P}_0\left(\left|\hat{\Psi}(t_n,h_n) + \mu_n\sqrt{n|B_n|}\right| \geq \Gamma(|B_n|) + \kappa_\alpha D(|B_n|)\right) \\
&\geq \mathbb{P}_0\left(\hat{\Psi}(t_n,h_n) \geq \Gamma(|B_n|) + \kappa_\alpha D(|B_n|) - \mu_n\sqrt{n|B_n|}\right) \\
&\geq \mathbb{P}_0\left(\hat{\Psi}(t_n,h_n) \geq -\epsilon_n\sqrt{2\log(1/|B_n|)} + \kappa_\alpha D(|B_n|)\right) \longrightarrow 1
\end{aligned}$$

as $n \longrightarrow \infty$. This completes the proof of Theorem 2.2.2.

4.6 Proofs Associated with Confidence Band Construction

4.6.1 Proof of Proposition 3.2.1

We begin by showing that under the hypothesis of Proposition 3.2.1, we have:

$$\mathbb{E}\hat{f}_h^{\ell}(t) = f(t) = \mathbb{E}\hat{f}_h^{u}(t) \quad \text{for all } t \in A_h. \tag{4.20}$$

Towards this, note that since $f, -f \in \mathcal{F}$, and since $-\hat{f}_h$ is the kernel estimator of $-f$ as defined in (1.1.1) (since for a standard d-dimensional Brownian sheet W, we have $W \overset{D}{=} -W$), we have:

$$\mathbb{E}(\hat{f}_h^{\ell}(t)) \leq f(t) \leq \mathbb{E}(\hat{f}_h^{u}(t)) \quad \text{and} \quad \mathbb{E}(-\hat{f}_h^{\ell}(t)) \leq -f(t) \leq \mathbb{E}(-\hat{f}_h^{u}(t))$$

for all $h \in I,\ t \in A_h$. This proves (4.20).

Next, observe that in view of (3.10), all it requires to complete the proof of Proposition 3.2.1 is to show that:

$$\{\hat{\ell}(t) \leq f(t) \leq \hat{u}(t) \text{ for all } t \in [0,1]^d\} \subseteq \{T(\psi^{\ell}) \leq \kappa_\alpha\ ,\ T(-\psi^{u}) \leq \kappa_\alpha\} \tag{4.21}$$

Towards this, suppose that $\hat{\ell}(t) \leq f(t) \leq \hat{u}(t)$ for all $t \in [0,1]^d$. In view of (4.20), we have:

$$\begin{aligned}
& \hat{\ell}(t) \leq f(t) \text{ for all } t \in [0,1]^d \\
\implies \quad & \hat{\ell}(t) \leq \mathbb{E}\hat{f}_h^{\ell}(t) \text{ for all } t \in [0,1]^d \\
\implies \quad & \hat{f}_h^{\ell}(t) - \frac{\|\psi^{\ell}\| \left(\kappa_\alpha + \Gamma(2^d \prod_{i=1}^d h_i)\right)}{\langle 1, \psi^{\ell}\rangle (n \prod_{i=1}^d h_i)^{1/2}} \leq \mathbb{E}\hat{f}_h^{\ell}(t) \text{ for all } t \in [0,1]^d \text{ and } h \in I \text{ with } t \in A_h \\
\implies \quad & \frac{\hat{f}_h^{\ell}(t) - \mathbb{E}\hat{f}_h^{\ell}(t)}{\sqrt{\mathrm{Var}(\hat{f}_h^{\ell}(t))}} - \Gamma(2^d \prod_{i=1}^d h_i) \leq \kappa_\alpha \text{ for all } t \in [0,1]^d \text{ and } h \in I \text{ with } t \in A_h \\
\implies \quad & T(\psi^{\ell}) \leq \kappa_\alpha.
\end{aligned}$$

Similarly, one has:

$$\begin{aligned}
& \hat{u}(t) \geq f(t) \text{ for all } t \in [0,1]^d \\
\implies \quad & \hat{u}(t) \geq \mathbb{E}\hat{f}_h^{u}(t) \text{ for all } t \in [0,1]^d \\
\implies \quad & \hat{f}_h^{u}(t) + \frac{\|\psi^{u}\| \left(\kappa_\alpha + \Gamma(2^d \prod_{i=1}^d h_i)\right)}{\langle 1, \psi^{u}\rangle (n \prod_{i=1}^d h_i)^{1/2}} \geq \mathbb{E}\hat{f}_h^{u}(t) \text{ for all } t \in [0,1]^d \text{ and } h \in I \text{ with } t \in A_h \\
\implies \quad & \frac{\hat{f}_h^{u}(t) - \mathbb{E}\hat{f}_h^{u}(t)}{\sqrt{\mathrm{Var}(\hat{f}_h^{u}(t))}} + \Gamma(2^d \prod_{i=1}^d h_i) \geq -\kappa_\alpha \text{ for all } t \in [0,1]^d \text{ and } h \in I \text{ with } t \in A_h \\
\implies \quad & T(-\psi^{u}) \leq \kappa_\alpha.
\end{aligned}$$

This proves (4.21) and completes the proof of Proposition 3.2.1.

4.6.2 Proof of Theorem 3.2.2

We will take $h := \varepsilon_n \mathbf{1}_d$ throughout the proof. Note that the hypothesis (3.12) of Proposition 3.2.2 implies that for all $t \in D$,

$$\frac{\langle f(t + h \star \cdot) - f(t), \psi(\cdot)\rangle}{\langle 1, \psi \rangle} = \mathbb{E}\hat{f}_h(t) - f(t) = 0$$

and hence, we can conclude from (4.24) that as long as $t \in D$ and $\varepsilon_n = \varepsilon$,

$$\hat{u}(t) - f(t) \le n^{-1/2} \frac{\|\psi^u\|(\kappa_\alpha + 2\Gamma(2^d\varepsilon^d) + T(\psi^u))}{\langle 1, \psi^u \rangle \varepsilon^{d/2}} \le K_\varepsilon n^{-1/2} \left(\kappa_\alpha + T(\psi^u)\right)$$

for some constant $K_\varepsilon > 0$. The rest of the proof for the case $\varepsilon_n = \varepsilon$ can be completed following the steps of the proof of Theorem 3.3.1.

For the case $\varepsilon_n := (\log(en))^{-\frac{1}{d}}$, we can conclude from (4.24) that for $t \in D$,

$$\begin{aligned} \hat{u}(t) - f(t) &\le K_1 (\log(en))^{1/2} n^{-1/2} \|\psi^u\| \left(\kappa_\alpha + T(\psi^u) + \sqrt{\log\log(en)}\right) / \langle 1, \psi^u \rangle \\ &= K_2 \left(\frac{\log(en)\log\log(en)}{n}\right)^{1/2} \left(\frac{1}{2} + \frac{\kappa_\alpha/2 + T(\psi^u)}{\sqrt{\log\log(en)}}\right) \end{aligned}$$

for some constants $K_1, K_2 > 0$. The bound for $f(t) - \hat{\ell}(t)$ follows similarly, thereby completing the proof of Proposition 3.2.2.

4.6.3 Proof of Theorem 3.2.3

For $k = 1$, we will show that just the facts that ψ_1^u is supported on a subset of $[0, \infty)^d$, ψ_1^ℓ is supported on a subset of $(-\infty, 0]^d$, and they are non-negative, are enough to conclude Theorem 3.2.3. This follows easily from the fact that the coordinate-wise increasing nature of f ensures that:

$$\langle f(t + h \star \cdot), \psi_1^u \rangle \ge f(t) \langle 1, \psi_1^u \rangle \quad \text{and} \quad \langle f(t + h \star \cdot), \psi_1^\ell \rangle \le f(t) \langle 1, \psi_1^\ell \rangle.$$

Next, we consider the case $k = 2$. If we could show that for every convex function $g : \mathbb{R}^d \longrightarrow \mathbb{R}$, we have:

$$\langle g, \psi_2^u \rangle \ge g(0) \langle 1, \psi_2^u \rangle \quad \text{and} \quad \langle g, \psi_2^\ell \rangle \le g(0) \langle 1, \psi_2^\ell \rangle, \tag{4.22}$$

then we would be done, because substituting $g(x) := f(t + h \star x)$ (which is a convex function) in (4.22) will complete the proof. We can also assume, without loss of generality, that $g(0) = 0$, because otherwise we can apply (4.22) on the function $g - g(0)$. In view of all these reductions, we just need to show that $\langle g, \psi_2^u \rangle \geq 0$ and $\langle g, \psi_2^\ell \rangle \leq 0$. The first inequality is a direct consequence of Jensen's inequality, because if U denotes a random vector distributed on the d-dimensional sphere $S_{d-1} := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$, with density at x being proportional to $1 - \|x\|^2$, then there exists a constant $C > 0$ such that:

$$\langle g, \psi_2^u \rangle = C\mathbb{E}g(U) \geq Cg(\mathbb{E}U) = Cg(0) = 0$$

where we used the fact that $\mathbb{E}U = 0$ by symmetry of the distribution of U around 0.

Finally, to prove that $\langle g, \psi_2^\ell \rangle \leq 0$, first note that by convexity of g and the fact that $g(0) = 0$, we have:

$$g(\alpha y) \leq \alpha g(y) \qquad \text{for all} \quad y \in \mathbb{R}^d \text{ and } \alpha \in [0, 1].$$

We can now substitute $\alpha := (d+3)\|x\|/(d+1)$ and $y := (d+1)x/((d+3)\|x\|)$, and have:

$$g\left(\frac{(d+1)x}{(d+3)\|x\|}\right) \geq \frac{(d+1)}{(d+3)\|x\|} g(x) \qquad \text{when} \quad \|x\| \leq \frac{d+1}{d+3} .$$

Similarly, we can substitute $\alpha := (d+1)/((d+3)\|x\|)$ and $y := x$, and have:

$$g\left(\frac{(d+1)x}{(d+3)\|x\|}\right) \leq \frac{(d+1)}{(d+3)\|x\|} g(x) \qquad \text{when} \quad \frac{d+1}{d+3} \leq \|x\| \leq 1 .$$

Moreover, note that $\psi_2^\ell(x) \leq 0$ when $(d+1)/(d+3) \leq \|x\| \leq 1$ and $\psi_2^\ell(x) \geq 0$ when $\|x\| \leq (d+1)/(d+3)$. We have:

$$\begin{aligned}
\langle g, \psi_2^\ell \rangle &= \int_{S_{d-1}} \left(1 - \frac{2d+4}{d+1}\|x\| + \frac{d+3}{d+1}\|x\|^2\right) g(x) dx \\
&\leq \frac{d+3}{d+1} \int_{S_{d-1}} \|x\| \left(1 - \frac{2d+4}{d+1}\|x\| + \frac{d+3}{d+1}\|x\|^2\right) g\left(\frac{(d+1)x}{(d+3)\|x\|}\right) dx .
\end{aligned}$$

At this point, for every $e \in \{-1, 1\}^d$, define:

$$H_e := \{x \in S_{d-1} : e_i x_i \geq 0 \text{ for all } 1 \leq i \leq d\} .$$

Note that $\{H_e\}_{e \in \{-1,1\}^d}$ form the 2^d orthants of $\mathbb{R}^d$, intersected with S_{d-1}. We will show that for all $e \in \{-1, 1\}^d$,

$$\int_{H_e} \|x\| \left(1 - \frac{2d+4}{d+1}\|x\| + \frac{d+3}{d+1}\|x\|^2\right) g\left(\frac{(d+1)x}{(d+3)\|x\|}\right) dx = 0 \tag{4.23}$$

which is enough to complete the proof. Towards this, fix $e \in \{-1, 1\}^d$, and make the following change of variables $x \mapsto y := (y_0, y_1, \ldots, y_{d-1})$ on H_e:

$$y_0 = \|x\| \quad \text{and} \quad y_i = \frac{x_i}{\|x\|} \text{ for all } 1 \le i \le d-1 \, .$$

This transformation is invertible, and we have:

$$x_i = y_0 y_i \text{ for all } 1 \le i \le d-1 \quad \text{and} \quad x_d := y_0 e_d \sqrt{1 - y_1^2 - \ldots - y_{d-1}^2} \, .$$

The Jacobian of this transformation is given by:

$$J(y) = \begin{bmatrix} y_1 & y_0 & 0 & \cdots & 0 \\ y_2 & 0 & y_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e_d \sqrt{1 - \sum_{i=1}^{d-1} y_i^2} & -\frac{y_0 y_1 e_d}{\sqrt{1 - \sum_{i=1}^{d-1} y_i^2}} & -\frac{y_0 y_2 s_d}{\sqrt{1 - \sum_{i=1}^{d-1} y_i^2}} & \cdots & -\frac{y_0 y_{d-1} e_d}{\sqrt{1 - \sum_{i=1}^{d-1} y_i^2}} \end{bmatrix}$$

and hence, we have:

$$|\det(J(y))| = \frac{y_0^{d-1}}{\sqrt{1 - \sum_{i=1}^{d-1} y_i^2}}.$$

Therefore, defining $\tilde{y} := \left(y_1, \ldots, y_{d-1}, e_d\sqrt{1 - y_1^2 - \ldots - y_{d-1}^2}\right)$, we have:

$$\begin{aligned}
& \int_{H_e} \|x\| \left(1 - \frac{2d+4}{d+1}\|x\| + \frac{d+3}{d+1}\|x\|^2\right) g\left(\frac{(d+1)x}{(d+3)\|x\|}\right) dx \\
= \; & \int_{S_{d-2} \cap \prod_{i=1}^{d-1} e_i[0,1]} \frac{g\left(\frac{d+1}{d+3}\,\tilde{y}\right)}{\sqrt{1 - \sum_{i=1}^{d-1} y_i^2}} \int_0^1 y_0^d \left(1 - \frac{2d+4}{d+1} y_0 + \frac{d+3}{d+1} y_0^2\right) \, dy_0 dy_1 \ldots dy_{d-1} \\
= \; & 0 \qquad \text{(since the inner integral is 0)}.
\end{aligned}$$

This proves (4.23) and completes the proof of Theorem 3.2.3.

4.6.4 Proof of Theorem 3.3.1

To begin with, note that for $t \in A_h$, we have

$$\begin{aligned}
& \hat{u}(t) - f(t) \\
\leq \; & \hat{f}_h^u(t) - f(t) + \frac{\|\psi^u\| \left(\kappa_\alpha + \Gamma(2^d \prod_{i=1}^d h_i)\right)}{\langle 1, \psi^u \rangle (n \prod_{i=1}^d h_i)^{1/2}} \\
\leq \; & \frac{\langle f(t + h \star \cdot) - f(t), \psi^u(\cdot) \rangle}{\langle 1, \psi^u \rangle} + \frac{\|\psi^u\| (\kappa_\alpha + 2\Gamma(2^d (\prod_{i=1}^d h_i)) + T(\psi^u))}{\langle 1, \psi^u \rangle (n \prod_{i=1}^d h_i)^{1/2}}. \qquad (4.24)
\end{aligned}$$

Here the last line follows from the inequality

$$T(\psi^u) \geq \frac{\hat{f}_h^u(t) - \langle f(t + h \star \cdot), \psi^u(\cdot) \rangle / \langle 1, \psi^u \rangle}{\|\psi\| \langle 1, \psi^u \rangle^{-1} (n \prod_{i=1}^d h_i)^{-1/2}} - \Gamma(2^d \prod_{i=1}^d h_i).$$

Now, if $f \in \mathbb{H}_{\beta,L} \cap \mathcal{F}_1$ (where $0 < \beta \leq 1$) we have

$$|f(t + h \star x) - f(t)| \leq L \|h \star x\|^\beta$$

Hence, denoting $h := \max\{h_1, \ldots, h_d\}$, we have:

$$\frac{\langle f(t + h \star \cdot) - f(t), \psi^u(\cdot)\rangle}{\langle 1, \psi^u\rangle} \le \frac{L \int_{[-1,1]^d} \|h \star x\|^\beta \psi^u(x) dx}{\langle 1, \psi^u\rangle} \le \frac{Lh^\beta \int_{[-1,1]^d} \|x\|^\beta \psi^u(x) dx}{\langle 1, \psi^u\rangle} := K_1 h^\beta \tag{4.25}$$

where $K_1 := L \int_{[-1,1]^d} \|x\|^\beta \psi^u(x) dx / \langle 1, \psi^u\rangle$. On the other hand, if $f \in \mathbb{H}_{\beta,L} \cap \mathcal{F}_2$ (where $1 < \beta \le 2$) then defining $g(x) := f(t + h \star x)$, we have the following for some ξ_x lying in the segment joining 0 and x:

$$\begin{aligned}
& \langle f(t + h \star \cdot) - f(t), \psi^u(\cdot)\rangle \\
= \; & \int_{[-1,1]^d} x^\top \nabla g(\xi_x) \psi^u(x) dx \\
= \; & \int_{[-1,1]^d} x^\top \left(\nabla g(\xi_x) - \nabla g(0)\right) \psi^u(x) dx \\
\le \; & \int_{[-1,1]^d} \|x\| \|\nabla g(\xi_x) - \nabla g(0)\| \psi^u(x) dx \\
= \; & L \int_{[-1,1]^d} \|x\| \|h \star (\nabla f(t + h \star \xi_x) - \nabla f(t))\| \psi^u(x) dx \\
\le \; & Lh \int_{[-1,1]^d} \|x\| \|\nabla f(t + h \star \xi_x) - \nabla f(t)\| \psi^u(x) dx \\
\le \; & Lh^\beta \int_{[-1,1]^d} \|x\|^\beta \psi^u(x) dx \, .
\end{aligned} \tag{4.26}$$

Hence, in this case also, we have:

$$\frac{\langle f(t + h \star \cdot) - f(t), \psi^u(\cdot)\rangle}{\langle 1, \psi^u\rangle} \le \frac{Lh^\beta \int_{[-1,1]^d} \|x\|^\beta \psi^u(x) dx}{\langle 1, \psi^u\rangle} = K_1 h^\beta .$$

Hence, (4.24) tells us that as long as $t \in A_h$ we have

$$\hat{u}(t) - f(t) \le K_1 h^\beta + \frac{\|\psi^u\|(\kappa_\alpha + 2\Gamma(2^d(\prod_{i=1}^d h_i)) + T(\psi^u))}{\langle 1, \psi^u\rangle (n \prod_{i=1}^d h_i)^{1/2}} . \tag{4.27}$$

Putting $h_1 = h_2 = \ldots = h_d = \varepsilon_n := (\log(en)/n)^{1/(2\beta+d)}$ in (4.27), we get $K_1 h^\beta = K_1 \varepsilon_n^\beta$ and

$$\begin{aligned}\frac{\|\psi^u\|(\kappa_\alpha + 2\Gamma(2^d(\prod_{i=1}^d h_i)) + T(\psi^u))}{\langle 1, \psi^u\rangle (n \prod_{i=1}^d h_i)^{1/2}} &\leq K_2 \frac{\kappa_\alpha + T(\psi^u) + 2\sqrt{2 + 2\log n}}{n^{\beta/(2\beta+d)} \log(en)^{d/(4\beta+2d)}} \\ &\leq K_3 \varepsilon_n^\beta \left(\frac{\kappa_\alpha + T(\psi^u)}{\log^{1/2}(en)} + 1 \right).\end{aligned}$$

for some constants K_2 and K_3 not depending on n. The above two equation tells us that as long as $t \in A_{\varepsilon_n \mathbf{1}_d}$, we have

$$\hat{u}(t) - f(t) \leq K\varepsilon_n^\beta \left(2 + \frac{\kappa_\alpha + T(\psi^u)}{\log^{1/2}(en)} \right) \leq 4K\varepsilon_n^\beta \left(\frac{1}{2} + \frac{\kappa_\alpha/2 + T(\psi^u)}{\log^{1/2}(en)} \right) \tag{4.28}$$

for some constant K not depending on n.

The steps for bounding $f(t) - \hat{\ell}(t)$ are similar, but we point out some differences. First, we have:

$$f(t) - \hat{\ell}(t) \leq \frac{\langle f(t) - f(t + h \star \cdot), \psi^\ell(\cdot)\rangle}{\langle 1, \psi^\ell \rangle} + \frac{\|\psi^\ell\|(\kappa_\alpha + 2\Gamma(2^d(\prod_{i=1}^d h_i)) + T(-\psi^\ell))}{\langle 1, \psi^\ell\rangle (n \prod_{i=1}^d h_i)^{1/2}},$$

which gives:

$$f(t) - \hat{\ell}(t) \leq K_2 h^\beta + \frac{\|\psi^\ell\|(\kappa_\alpha + 2\Gamma(2^d(\prod_{i=1}^d h_i)) + T(-\psi^\ell))}{\langle 1, \psi^\ell\rangle (n \prod_{i=1}^d h_i)^{1/2}}$$

where $K_2 := L \int_{[-1,1]^d} \|x\|^\beta |\psi^\ell(x)| dx / |\langle 1, \psi^\ell \rangle|$. Hence, for some constant K' not depending on n, we have:

$$f(t) - \hat{\ell}(t) \leq K' \varepsilon_n^\beta \left(\frac{1}{2} + \frac{\kappa_\alpha/2 + T(-\psi^\ell)}{\log^{1/2}(en)} \right). \tag{4.29}$$

Theorem 3.3.1 now follows by adding (4.28) and (4.29).

4.6.5 Proof of Theorem 3.3.2

It follows from (4.24) and (4.27) that for $t \in A_{\varepsilon_n, i_1, \ldots, i_k}$, we have:

$$\begin{aligned}
\hat{u}(t) - f(t) &\le K_1 \varepsilon_n^\beta + \frac{\|\psi^u\|(\kappa_\alpha + 2\Gamma(2^d \varepsilon_n^k \varepsilon^{d-k}) + T(\psi^u))}{\langle 1, \psi^u\rangle (n\varepsilon_n^k \varepsilon^{d-k})^{1/2}} \\
&= K_1 \varepsilon^\beta \rho_{n,k} + \frac{\|\psi^u\|(\kappa_\alpha + 2\Gamma(2^d \varepsilon^d (\log(en)/n)^{k/(2\beta+k)}) + T(\psi^u))}{\langle 1, \psi^u\rangle (n\varepsilon^d (\log(en)/n)^{k/(2\beta+k)})^{1/2}} \\
&\le \Delta_1 \rho_{n,k} \left(\frac{1}{2} + \frac{\kappa_\alpha/2 + T(\psi^u)}{(\log(en))^{1/2}} \right)
\end{aligned}$$

for some constants K_1 and $\Delta_1 > 0$. The rest of the proof follows the idea of the proof of Theorem 3.3.1. The only modifications are in (4.25) and (4.26), where one now uses the fact that the function f only depends on k coordinates, and hence, the vector h can now be replaced by its restriction on the $i_1^{\text{th}}, \ldots, i_k^{\text{th}}$ coordinates.

4.6.6 Proof of Theorem 3.4.1

(a) We prove only the bound for $\|f - \ell\|_U$ as the other case can be handled similarly. Thus, we will show that for any level $1-\alpha$ confidence band (ℓ, u) with guaranteed coverage probability for the class $\mathcal{F}_1$, and any $0 < \gamma < 1$, we have

$$\liminf_{n\to\infty} \mathbb{P}_f \left(\|f - \ell\|_U \ge \gamma \Delta^{(\ell)} L_1^{\frac{d}{2+d}}[t_0] \rho_n \right) \ge 1 - \alpha.$$

For notational simplicity, we will abbreviate ψ^ℓ by ψ. By our assumption, f is continuously differentiable on an open neighborhood U of $t_0 \in (0,1)^d$ such that $L_1[f, t_0] := \left[\prod_{i=1}^d \frac{\partial}{\partial x_i} f(x) \Big|_{x=t_0} \right]^{1/d} > 0$. Let us define, for $i = 1, \ldots, d$,

$$M_i^\star := \frac{\partial}{\partial x_i} f(x)|_{x=t_0}.$$

Without loss of generality, let us assume that $M_1^\star \leq M_2^\star \leq \ldots \leq M_d^\star$. Since $\gamma < 1$, we can find $\epsilon > 0$ and $\gamma^* < 1$ such that:

$$\gamma L_1^{d/(2+d)}[t_0] = \gamma^\star (L_1[t_0] - \epsilon)^{d/(2+d)}.$$

Also since f is continuously differentiable on U, we can find $h_0 \in [0,1]^d$ and $\epsilon^\star > 0$ small enough such that

(i) $B_\infty(t_0, h_0) \subset U$,

(ii) $\left[\prod_{i=1}^d (M_i^\star - \epsilon^\star)\right]^{1/d} \geq L_1[t_0] - \epsilon$,

(iii) for all $x \in B_\infty(t_0, h_0)$ we have

$$\frac{\partial}{\partial x_i} f(x) \geq M_i^\star - \epsilon^\star > 0, \qquad \text{for all } i = 1, 2, \ldots, d.$$

Suppose that $h \equiv (h_1, h_2, \ldots, h_d)$ is such that $h_1 \in (0, 1/2]$ and $h_i = h_1 \times \left(\frac{M_1^\star - \epsilon^\star}{M_i^\star - \epsilon^\star}\right)$, for $i = 2, \ldots, d$. Let us define a set of grid points G for bandwidth h as

$$G := \{t = (t_1, \ldots, t_d) : t_i = t_{0i} + h_i(2k_i) \text{ for some integer } k_i, B_\infty(t, h) \subset B_\infty(t_0, h_0)\}$$

where $t_0 = (t_{01}, \ldots, t_{0d})$. For $t \in G$, let

$$f_t := f - h_1(M_1^\star - \epsilon^\star)\psi_{t,h},$$

where $\psi_{t,h}$ is defined in (1.5). We will now show that for every $t \in G$, $f_t \in \mathcal{F}_1$.

Lemma 4.6.1 *$f_t \in \mathcal{F}_1$ for all $t \in G$.*

Proof 1 *Fix $t \in G$. Suppose that $x \leq y$ (coordinatewise) and $x, y \in T_{t,h}$ where*

$$T_{t,h} := \left\{ u = (u_1, \ldots, u_d) \in [0,1]^d : u_i \leq t_i \text{ for all } i \text{ and } \sum_{i=1}^d \left(\frac{u_i - t_i}{h_i}\right) \geq -1 \right\}.$$

Then, for some $\boldsymbol{\xi} \in [x, y] \subset B_\infty(t_0, h_0)$,

$$f(y) - f(x) \;=\; \nabla f(\boldsymbol{\xi})^\top (y - x) \;\geq\; \sum_{i=1}^{d} (M_i^\star - \epsilon^\star)(y_i - x_i). \tag{4.30}$$

Then,

$$-\psi_{t,h}(y) + \psi_{t,h}(x) \;=\; -\sum_{i=1}^{d} \left(\frac{y_i - x_i}{h_i}\right) = -\frac{1}{h_1(M_1^\star - \epsilon^*)} \sum_{i=1}^{d} (M_i^\star - \epsilon^\star)(y_i - x_i). \tag{4.31}$$

It now follows from (4.30) *and* (4.31) *that:*

$$f_t(y) - f_t(x) \geq \sum_{i=1}^{d} (M_i^\star - \epsilon^\star)(y_i - x_i) - h_1(M_1^\star - \epsilon^\star) \left[\frac{1}{h_1(M_1^\star - \epsilon^*)} \sum_{i=1}^{d} (M_i^\star - \epsilon^\star)(y_i - x_i) \right] = 0$$

thereby yielding $f_t(y) \geq f_t(x)$.

Let us now look at the case $x \notin T_{t,h}$, $y \in T_{t,h}$ *and* $x \leq y$. *Define* $a := 1 + \sum_{i=1}^{d} \left(\frac{x_i - t_i}{h_i}\right)$ *and* $b := 1 + \sum_{i=1}^{d} \left(\frac{y_i - t_i}{h_i}\right)$. *By the assumptions on* x, y, *we have* $x \leq y \leq t$ *(coordinatewise). Hence as* $x \notin T_{t,h}$, $a < 0$ *and as* $y \in T_{t,h}$, $b \geq 0$. *Define* $z = \alpha x + (1 - \alpha) y$ *where* $\alpha = b/(b - a)$. *Note that* $\sum_{i=1}^{d} (z_i - t_i)/h_i = -1$ *and* $x \leq z \leq y \leq t$ *which implies that* $z \in T_{t,h}$. *Hence,*

$$f_t(y) \geq f_t(z) = f(z) \geq f(x) = f_t(x).$$

Here, the first inequality follows from the fact that $z, y \in T_{t,h}$, *the third inequality follows from monotonicity of* f *and the second and fourth equality follows from the fact that* $\psi_{t,h}(z) = \psi_{t,h}(x) = 0$.

Now let us look into the case where $x \in T_{t,h}$, $y \notin T_{t,h}$ *and* $x \leq y$. *In this case*

$$f_t(y) = f(y) \geq f(x) \geq f(x) - h_1(M_1^\star - \epsilon^\star)\psi_{t,h}(x) = f_t(x).$$

The only case left is when $x \notin T_{t,h}$, $y \notin T_{t,h}$ *and* $x \leq y$. *In this case* $\psi_{t,h}(x) = \psi_{t,h}(y) = 0$, *hence the monotonicity of* f_t *directly follows from the monotonicity of* f. *This completes the proof of Lemma*

4.6.3.

Now let us define the set

$$A := \{\ell(x) \leq f_t(x) \text{ for all } x \in [0,1]^d, \text{ for some } t \in G\}.$$

Now, since (ℓ, u) is a confidence band for all $f \in \mathcal{F}_1$, and since $f_t \in \mathcal{F}_1$, we have

$$\mathbb{P}_{f_t}(A) \geq 1 - \alpha \text{ for all } t \in G.$$

Hence we have:

$$\mathbb{P}_f\left(\|f - \ell\|_U \geq h_1(M_1^\star - \epsilon^\star)\right) \geq \mathbb{P}_f(A) \geq 1 - \alpha - \min_{t \in G}\left(\mathbb{P}_{f_t}(A) - \mathbb{P}_f(A)\right). \tag{4.32}$$

Here the first inequality follows from the fact that if A happens then there exists $t \in G \subset U$ such that $\ell \leq f_t$, thereby giving:

$$\ell(t) \leq f_t(t) = f(t) - h_1(M_1^\star - \epsilon^\star).$$

Hence it is enough to bound $\min_{t \in G}\left(\mathbb{P}_{f_t}(A) - \mathbb{P}_f(A)\right)$.

$$\begin{aligned}
\min_{t \in G} \mathbb{P}_{f_t}(A) - \mathbb{P}_f(A) &\leq |G|^{-1} \sum_{t \in G}\left(\mathbb{P}_{f_t}(A) - \mathbb{P}_f(A)\right) \\
&= |G|^{-1} \sum_{t \in G} \mathbb{E}_f\left(\left(\frac{d\mathbb{P}_{f_t}}{d\mathbb{P}_f}(Y) - 1\right)\mathbb{I}_A(Y)\right) \\
&\leq \mathbb{E}_f\left||G|^{-1} \sum_{t \in G}\left(\frac{d\mathbb{P}_{f_t}}{d\mathbb{P}_f}(Y) - 1\right)\right|. \qquad (4.33)
\end{aligned}$$

Now by Cameron-Martin-Girsanov's Theorem, we have

$$\log \frac{d\mathbb{P}_{f_t}}{d\mathbb{P}_f}(Y) = n^{1/2} h_1(M_1^\star - \epsilon^\star)\sqrt{\Pi_{i=1}^d h_i}\|\psi\| X_t - n(M_1^\star - \epsilon^\star)^2 h_1^2(\Pi_{i=1}^d h_i)\|\psi\|^2/2$$

where

$$X_t = (\Pi_{i=1}^n h_i)^{-1/2} \|\psi\|^{-1} \int \psi_{t,h} dW$$

with W being the standard Brownian sheet on $[0,1]^d$. Note that here X_t follows a standard normal distribution and for $t \neq t' \in G$, X_t and X_t' are independent. Now let

$$w_n := n^{1/2} h_1 (M_1^\star - \epsilon^\star) \sqrt{\Pi_{i=1}^d h_i} \|\psi\|.$$

At this point, Let us recall the following lemma A.5.1 (stated and proved in Lemma 6.2 in [66]).

Define $\varepsilon_n := 1 - (w_n / \sqrt{2 \log |G|})$. If $\varepsilon_n \to 0$ and $\varepsilon_n \sqrt{\log |G|} \to \infty$ are satisfied, then by Lemma A.5.1 and (4.40), we have the following as $|G| \to \infty$:

$$\min_{t \in G} \mathbb{P}_{f_t}(A) - \mathbb{P}_f(A) \to 0 \tag{4.34}$$

Now let us choose $h_1 = (1 - \epsilon_n) c \rho_n$ where $\rho_n = (\log(en)/n)^{1/(2+d)}$, with $\epsilon_n \to 0$ and $\epsilon_n \sqrt{\log n} \to \infty$ and c is a constant to be chosen later. This implies that

$$\sqrt{2 \log |G|} = (1 - o(1)) \sqrt{\frac{2d}{d+2} \log n}$$

and for large n, $\sqrt{2 \log |G|} < \sqrt{\frac{2d}{d+2} \log n}$. Hence,

$$w_n = (1 - \epsilon_n)^{(2+d)/2} c^{(2+d)/2} \|\psi^\ell\| \frac{(M_1^\star - \epsilon^\star)^{(2+d)/2}}{\prod_{i=1}^d (M_i^\star - \epsilon^\star)^{1/2}} \sqrt{\log(en)}.$$

Let us now put:

$$c = \frac{\prod_{i=1}^d (M_i^\star - \epsilon^\star)^{1/(2+d)}}{(M_1^\star - \epsilon^\star)} \left[\frac{(d+2) \|\psi^\ell\|^2}{2d} \right]^{-1/(2+d)}$$

whence we have:

$$\begin{aligned}\frac{w_n}{\sqrt{2\log|G|}} &\sim (1-\epsilon_n)^{(2+d)/2} c^{(2+d)/2}\|\psi^\ell\| \frac{(M_1^\star-\epsilon^\star)^{(2+d)/2}}{\Pi_{i=1}^d (M_i^\star-\epsilon^\star)^{1/2}}\sqrt{\frac{d+2}{2d}} \\ &= (1-\epsilon_n)^{(2+d)/2} \longrightarrow 1 \text{ as } n\longrightarrow\infty.\end{aligned}$$

Also note that for large n, $(1-w_n/\sqrt{2\log|G|}) > 0$ as $\sqrt{2\log|G|} < \sqrt{\frac{2d}{d+2}\log n}$. Also, note that:

$$\begin{aligned}\sqrt{\log|G|}\left(1-\frac{w_n}{\sqrt{2\log|G|}}\right) &\sim \sqrt{\frac{d}{2+d}}\sqrt{\log n}(1-(1-\epsilon_n)^{1+d/2}) \\ &= \sqrt{\frac{d}{2+d}}\left(\frac{2+d}{2}+o(1)\right)\epsilon_n\sqrt{\log n} \longrightarrow \infty \quad \text{(by assumption).}\end{aligned}$$

Also note that we denoted $\left((d+2)\|\psi^\ell\|^2/2d\right)^{-1/(2+d)}$ as $\Delta^{(\ell)}$ in the statement of the theorem. Hence, by (4.39) and (4.41), we have:

$$\begin{aligned}1-\alpha &\leq \liminf_{n\to\infty}\mathbb{P}_f\left(\|f-\ell\|_U \geq h_1(M_1^\star-\epsilon^\star)\right) \\ &= \liminf_{n\to\infty}\mathbb{P}_f\left(\|f-\ell\|_U \geq (1-\epsilon_n)\rho_n\Delta^{(\ell)}\Pi_{i=1}^d(M_i^\star-\epsilon^\star)^{1/(2+d)}\right) \\ &\leq \liminf_{n\to\infty}\mathbb{P}_f\left(\|f-\ell\|_U \geq \gamma^\star\Pi_{i=1}^d(M_i^\star-\epsilon^\star)^{1/(2+d)}\rho_n\Delta^{(\ell)}\right) \\ &\leq \liminf_{n\to\infty}\mathbb{P}_f\left(\|f-\ell\|_U \geq \gamma^\star(L_1[t_0]-\epsilon)^{d/(2+d)}\rho_n\Delta^{(\ell)}\right) \\ &= \liminf_{n\to\infty}\mathbb{P}_f\left(\|f-\ell\|_U \geq \gamma L_1^{d/(2+d)}[t_0]\rho_n\Delta^{(\ell)}\right).\end{aligned}$$

This completes the proof of part (a).

(b) We again restrict our attention to $(f-\hat{\ell})(t_0)$. The other case can be done by a similar argument. For $i=1,2,\ldots,d$, let us recall

$$M_i^\star = \frac{\partial}{\partial x_i}f(x)|_{x=t_0}.$$

Fix $\epsilon > 0$. As f has continuous derivative on an open neighborhood U of t_0 we can find $\epsilon^\star > 0$

and a hyperrectangle $B_\infty(t_0, h_0)$ small enough such that

$$\sup_{x \in B_\infty(t_0,h_0)} \frac{\partial}{\partial x_i} f(x) \le M_i^\star + \epsilon^\star$$

and

$$[\Pi_{i=1}^d (M_i^\star + \epsilon^\star)]^{1/d} \le (1+\epsilon) L_1[t_0].$$

Recall that we have assumed without loss of generality $0 < M_1^\star \le M_2^\star \le \ldots \le M_d^\star$. Now let $h = (h_1, \ldots, h_d)$ be such that $h_i := \tilde{h} \times \frac{M_1^\star + \epsilon^\star}{M_i^\star + \epsilon^\star}$. Let $M := \left[\Pi_{i=1}^d \frac{M_1^\star+\epsilon^\star}{M_i^\star+\epsilon^\star}\right]^{1/d}$ which implies that $\Pi_{i=1}^d h_i = \tilde{h}^d M^d$. Recall that

$$\hat{\ell}(t_0) = \sup_{h \in I : t_0 \in A_h} \left\{ \hat{f}_h(t_0) - \frac{\|\psi\|}{\langle 1, \psi \rangle (n \Pi_{i=1}^d h_i)^{1/2}} \left(\kappa_\alpha + \Gamma(2^d \Pi_{i=1}^d h_i) \right) \right\}$$

and

$$\begin{aligned} \hat{f}_h(t_0) &= \frac{1}{n^{1/2}(\Pi_{i=1}^d h_i)\langle 1, \psi \rangle} \int_{[0,1]^d} \psi_{t_0,h}(x) dY(x) \\ &= \frac{1}{\langle 1, \psi \rangle} \langle f(t_0 + h \star \cdot), \psi(\cdot) \rangle + \frac{1}{n^{1/2}(\Pi_{i=1}^d h_i)\langle 1, \psi \rangle} \int_{B_\infty(t_0,h)} \psi_{t_0,h}(x) dW(x) \end{aligned}$$

where for $h, x \in \mathbb{R}^d$ we define $h \star x := (h_1 x_1, \ldots, h_d x_d)$.

Now, it follows from the definition of $\hat{\ell}(t_0)$ that if $f(t_0) - \hat{\ell}(t_0) \ge (M_1^\star + \epsilon^\star)\tilde{h}$, then

$$\hat{f}_h(t_0) - \frac{\|\psi\| \left(\kappa_\alpha + \Gamma(2^d M^d \tilde{h}^d) \right)}{n^{1/2} \tilde{h}^{d/2} M^{d/2} \langle 1, \psi \rangle} \le f(t_0) - (M_1^\star + \epsilon^\star)\tilde{h},$$

which can be rewritten as:

$$\begin{aligned} \frac{\int_{B_\infty(t_0,h)} \psi_{t_0,h}(x) dW(x)}{\|\psi\| \tilde{h}^{d/2} M^{d/2}} &\le -\frac{(n \tilde{h}^d M^d)^{1/2}}{\|\psi\|} \langle f(t_0 + h \star \cdot) - f(t_0) + (M_1^\star + \epsilon^\star)\tilde{h}, \psi(\cdot) \rangle \\ &\quad + \Gamma(2^d M^d \tilde{h}^d) + \kappa_\alpha. \end{aligned} \tag{4.35}$$

Lemma 4.6.2 *Suppose $h \in [0,1]^d$ is such that $B_\infty(t,h) \subseteq B_\infty(t_0,h_0)$ where $B_\infty(t_0,h_0), \psi$ are as described above. Then*

$$\langle f(t+h\star\cdot) - f(t) + (M_1^\star + \epsilon^\star)\tilde{h}, \psi(\cdot)\rangle \geq (M_1^\star + \epsilon^\star)\tilde{h}\|\psi\|^2$$

Proof of Lemma 4.6.2: Suppose $B_\infty(t_0,h) \subseteq B_\infty(t_0,h_0)$ holds. Note that

$$\psi(x) = \left(1 + \sum_{i=1}^d x_i\right)\mathbb{I}\left(x \leq 0, \sum_{i=1}^d x_i \geq -1\right)$$

Let $x \in [-1,0]^d$ be such that $\sum_{i=1}^d x_i \geq -1$.

$$\begin{aligned} f(t+h\star x) - f(t) &= (h\star x)^\top \nabla f(\boldsymbol{\xi}) \qquad \text{for some } \boldsymbol{\xi} \in [t, t+h\star x] \\ &\geq \sum_{i=1}^d h_i x_i (M_i^\star + \epsilon^\star) \qquad \text{note that } h\star x \leq 0 \\ &= \tilde{h}(M_1^\star + \epsilon^\star)\sum_{i=1}^d x_i. \end{aligned}$$

Hence on the set $D := \{x \leq 0, \sum_{i=1}^d x_i \geq -1\}$

$$f(t+h\star x) - f(t) + (M_1^\star + \epsilon^\star)\tilde{h} \geq \tilde{h}(M_1^\star + \epsilon^\star)(1 + \sum_{i=1}^d x_i) \geq 0.$$

Hence

$$\langle f(t+h\star\cdot) - f(t) + (M_1^\star + \epsilon^\star)\tilde{h}, \psi(\cdot)\rangle \geq \int_D \tilde{h}(M_1^\star + \epsilon^\star)\left(1 + \sum_{i=1}^d x_i\right)^2 dx = \tilde{h}(M_1^\star + \epsilon^\star)\|\psi\|^2. \quad (4.36)$$

This completes the proof of Lemma 4.6.2.

By (4.35) and (4.36) we get that as long as $B_\infty(t_0,h) \subseteq B_\infty(t_0,h_0)$, $(f - \hat{\ell})(t_0) \geq (M_1^\star + \epsilon^\star)\tilde{h}$ implies that

$$\frac{\int_{B_\infty(t_0,h)} \psi_{t_0,h}(x)dW(x)}{\|\psi\|\tilde{h}^{d/2}M^{d/2}} \leq -\sqrt{n}\tilde{h}^{1+d/2}M^{d/2}(M_1^\star + \epsilon^\star)\|\psi\| + \Gamma(2^d M^d \tilde{h}^d) + \kappa_\alpha.$$

Also note that
$$\frac{\int_{B_\infty(t_0,h)} \psi_{t_0,h}(x)dW(x)}{\|\psi\|\tilde{h}^{d/2}M^{d/2}} \sim N(0,1).$$

Now let us choose
$$\tilde{h} = c(M_1^\star + \epsilon^\star)^{-\frac{2}{2+d}}\rho_n$$

where $\rho_n = (\log(en)/n)^{1/(2+d)}$ and a constant c to be chosen later.

Note that $\rho_n \to 0$ as $n \to \infty$. Hence for large enough n, $B_\infty(t_0, h) \subseteq B_\infty(t_0, h_0)$ (here h depends on n). Also we have

$$\begin{aligned} \Gamma(2^d M^d \tilde{h}^d) &\leq \sqrt{\left(\frac{2d}{2+d}\right)\log n} \qquad \text{for large } n \\ \sqrt{n}\tilde{h}^{1+d/2}M^{d/2}(M_1^\star + \epsilon^\star)\|\psi\| &= M^{d/2}\|\psi\|c^{(d+2)/2}\sqrt{\log(en)}. \end{aligned}$$

Now let us pick
$$c = (1+\epsilon)\left(\frac{(d+2)\|\psi\|^2}{2d}\right)^{-1/(d+2)} M^{-\frac{d}{d+2}}.$$

Note that $\Delta^{(\ell)} = \left(\frac{(d+2)\|\psi\|^2}{2d}\right)^{-1/(d+2)}$ as defined in the statement of the theorem.

Hence for large n we have

$$\begin{aligned} &\mathbb{P}\left(f(t_0) - \hat{\ell}(t_0) \geq (M_1^\star + \epsilon^\star)\tilde{h}\right) \\ \leq\ &\mathbb{P}\left(\frac{\int_{B_\infty(t_0,h)} \psi_{t_0,h}(x)dW(x)}{\|\psi\|\tilde{h}^{d/2}M^{d/2}} \leq -\sqrt{n}\tilde{h}^{1+d/2}M^{d/2}(M_1^\star + \epsilon^\star)\|\psi\| + \Gamma(2^d M^d \tilde{h}^d) + \kappa_\alpha\right) \\ =\ &\Phi\left(-\sqrt{n}\tilde{h}^{1+d/2}M^{d/2}(M_1^\star + \epsilon^\star)\|\psi\| + \Gamma(2^d M^d \tilde{h}^d) + \kappa_\alpha\right) \\ \leq\ &\Phi\left(\kappa_\alpha - \sqrt{\log(en)}\left[M^{d/2}\|\psi\|c^{(d+2)/2} - \sqrt{\frac{2d}{2+d}}\right]\right) \\ =\ &\Phi\left(\kappa_\alpha - \sqrt{\log(en)}\sqrt{\frac{2d}{2+d}}\left[(1+\epsilon)^{\frac{d+2}{2}} - 1\right]\right) \to 0 \text{ as } n \to \infty. \end{aligned}$$

Here Φ denotes the distribution function of standard normal. Hence

$$\lim_{n\to\infty} \mathbb{P}\left((f-\hat{\ell})(t_0) \le (M_1^\star+\epsilon^\star)\tilde{h}\right) = 1. \tag{4.37}$$

Now

$$\begin{aligned}
(M_1^\star+\epsilon^\star)\tilde{h} &= (M_1^\star+\epsilon^\star)c(M_1^\star+\epsilon^\star)^{-\frac{2}{2+d}}\rho_n \\
&= \rho_n(M_1^\star+\epsilon^\star)^{d/(2+d)}(1+\epsilon)\Delta^{(\ell)}M^{-d/(d+2)} \\
&= (1+\epsilon)\Delta^{(\ell)}\rho_n\left(\frac{(M_1^\star+\epsilon^\star)^d}{M^d}\right)^{1/(d+2)} \\
&= (1+\epsilon)\Delta^{(\ell)}\rho_n\left(\Pi_{i=1}^d(M_i^\star+\epsilon^\star)\right)^{1/(d+2)} \\
&\le (1+\epsilon)\Delta^{(\ell)}\rho_n(1+\epsilon)^{d/(d+2)}L_1^{d/(d+2)}[t_0] \\
&= (1+\epsilon)^{(2d+2)/(d+2)}\Delta^{(\ell)}\rho_n L_1^{d/(d+2)}[t_0].
\end{aligned} \tag{4.38}$$

Hence our assertion is proved by (4.37) and (4.38).

4.6.7 Proof of Theorem 3.4.2

Once again, we prove only the bound for $\|f-\ell\|_U$ and the other case can be handled similarly. we will show that for any level $1-\alpha$ confidence band (ℓ,u) with guaranteed coverage probability for the class $\mathcal{F}_2$, and any $0<\gamma<1$, we have

$$\liminf_{n\to\infty} \mathbb{P}_f\left(\|f-\ell\|_U \ge \gamma\Delta^{(\ell)}L_2^{\frac{d}{4+d}}[t_0]\rho_n\right) \ge 1-\alpha.$$

Once again, for notational convenience, we will abbreviate ψ^ℓ by ψ. It is more convenient to solve the problem if we introduce a rotation of the coordinate system, so that the Hessian of f at t_0 with respect to this new changed coordinate system, is diagonal. If $\nabla^2 f(t_0) = PDP^\top$ is the spectral decomposition of the Hessian of f at t_0, then for any point y, we will define $y' := P^\top y$, and for any set $S \subseteq \mathbb{R}^d$, we will define:

$$S' := \{P^\top s : s \in S\} .$$

Further, defining $g(t) := f(Pt)$, we note that $g(t') = f(t)$ for all t (recall our notation that $t' = P^\top t$), and $\nabla^2 g(t_0') = D$.

Recall that by assumption, f is twice continuously differentiable on an open neighborhood U of $t_0 \in (0,1)^d$ such that $L_2[f, t_0] := \det(\nabla^2 f(t_0))^{1/d} > 0$. Hence, g is twice continuously differentiable on the open neighborhood U' of t_0'. Denote the i^{th} smallest eigenvalue of $\nabla^2 f(t_0)$ by $\lambda_i(t_0)$. Since $\gamma < 1$, we can find $\epsilon > 0$ and $\gamma^* < 1$ such that:

$$\gamma L_2^{d/(4+d)}[t_0] = \gamma^\star (L_2[t_0] - \epsilon)^{d/(4+d)}.$$

Using the twice continuous differentiability of g on U', we can find $h_0 \in [0,1]^d$ and $\epsilon^\star > 0$ small enough such that

(i) $B'_\infty(t_0', h_0) \subset U'$,

(ii) $\left[\prod_{i=1}^d (\lambda_i(t_0) - \epsilon^\star)\right]^{1/d} \geq L_2[t_0] - \epsilon$,

(iii) for all $x' \in B'_\infty(t_0', h_0)$, we have

$$\sup_{v \in B_2(0,1)} \left| v^\top \left(\nabla^2 g(x') - \nabla^2 g(t_0') \right) v \right| < \epsilon^*$$

where $B_2(0,1)$ denotes the ball around 0 with ℓ^2 norm 1, and $B'_\infty(y, h) := (B_\infty(Py, h))'$.

Next, let $h \equiv (h_1, h_2, \ldots, h_d)$ be such that $h_1 \in (0, 1/2]$ and $h_i = h_1 \times \sqrt{\frac{\lambda_1(t_0)-\epsilon^\star}{\lambda_i(t_0)-\epsilon^\star}}$, for $i = 2, \ldots, d$. Let us define a set of grid points G for bandwidth h as

$$G := \{t = (t_1, \ldots, t_d) : t_i = t_{0i} + h_i(2k_i) \text{ for some integer } k_i, B_\infty(t, h) \subset B_\infty(t_0, h_0)\}$$

where $t_0 = (t_{01}, \dots, t_{0d})$. For $t' \in G'$, let

$$g_{t'}(x) := g(x) - h_1^2(\lambda_1(t_0) - \epsilon^\star)\psi^* \left(\frac{x_1 - t_1'}{h_1}, \dots, \frac{x_d - t_d'}{h_d}\right)$$

where $\psi^* := (G_A - G_0)(P\cdot)$ with G_A, G_0 as defined as follows:

$$G_A(y) := \frac{\|y\|^2}{2}\mathbb{1}_{\|y\| \le \sqrt{2(d+3)/(d+1)}} \quad \text{and} \quad G_0(y) := \left(-1 + \frac{\sqrt{2}(d+2)}{\sqrt{(d+1)(d+3)}}\|y\|\right)\mathbb{1}_{\|y\| \le \sqrt{2(d+3)/(d+1)}} .$$

We will now show that for every $t' \in G'$, the function $g_{t'} \in \mathcal{F}_2$.

Lemma 4.6.3 *$g_{t'} \in \mathcal{F}_2$ for all $t' \in G'$.*

Proof 2 *Fix $t' \in G'$ and a vector $v \in B_2(0, 1)$. In order to prove Lemma 4.6.3, it suffices to show that the univariate function $h : \mathbb{R} \to \mathbb{R}$ defined as $h(x) := g_{t'}(xv)$ is convex. Towards proving this, take scalars $\alpha > \beta$, such that αv and $\beta v \in B'_\infty(t_0', h_0)$, and define*

$$\phi(x) = \phi_{t',h}(x) := \psi^* \left(\frac{x_1 - t_1'}{h_1}, \dots, \frac{x_d - t_d'}{h_d}\right)$$

Then, we have:

$$\begin{aligned} & h'(\alpha) - h'(\beta) \\ = \;& (\nabla g(\alpha v) - \nabla g(\beta v))^\top v - h_1^2(\lambda_1(t_0) - \epsilon^\star)\left[\nabla\phi(\alpha v) - \nabla\phi(\beta v)\right]^\top v \\ = \;& (\alpha - \beta)\left[v^\top \nabla^2 g(\xi v)v - h_1^2(\lambda_1(t_0) - \epsilon^\star)v^\top\nabla^2\phi(\eta v)v\right] \end{aligned}$$

for some ξ, η lying between α and β. First, note that since $\xi v \in B'_\infty(t_0', h_0)$, we have:

$$v^\top\nabla^2 g(\xi v)v \ge \left[v^\top\nabla^2 g(t_0')v - \epsilon^*\right] = \sum_{i=1}^d (\lambda_i(t_0) - \epsilon^*)v_i^2 .$$

Next, note that $\nabla^2\phi$ is diagonal, with the i^{th} diagonal entry being h_i^{-2}. Hence, we have:

$$h'(\alpha) - h'(\beta) \;\; \geq \;\; (\alpha - \beta)\left[\sum_{i=1}^{d}(\lambda_i(t_0) - \epsilon^*)v_i^2 - (\lambda_1(t_0) - \epsilon^*)\sum_{i=1}^{d}(h_1/h_i)^2 v_i^2\right] = 0.$$

This completes the proof of Lemma 4.6.3.

Now let us define the set

$$A := \{\ell(Px) \leq g_{t'}(x) \text{ for all } x \in ([0,1]^d)', \text{ for some } t' \in G'\}.$$

Now, since (ℓ, u) is a confidence band for all functions in $\mathcal{F}_2$, and since the function $g_{t'}(P^\top \cdot) \in \mathcal{F}_2$, we have

$$\mathbb{P}_{g_{t'}}(A) \geq 1 - \alpha \text{ for all } t' \in G'.$$

Hence, defining $\ell^*(x) := \ell(Px)$, we have:

$$\mathbb{P}_g\left(\|g - \ell^*\|_{U'} \geq h_1^2(\lambda_1(t_0) - \epsilon^\star)\right) \geq \mathbb{P}_g(A) \geq 1 - \alpha - \min_{t' \in G'}\left(\mathbb{P}_{g_{t'}}(A) - \mathbb{P}_g(A)\right). \tag{4.39}$$

Note that the first inequality follows from the fact that if A happens then there exists $t' \in G' \subset U'$ such that $\ell^* \leq g_{t'}$ on $([0,1]^d)'$, thereby giving:

$$\ell^*(t') \leq g_{t'}(t') = g(t') - h_1^2(\lambda_1(t_0) - \epsilon^\star).$$

Hence it is enough to bound $\min_{t' \in G'}\left(\mathbb{P}_{g_{t'}}(A) - \mathbb{P}_g(A)\right)$.

$$\begin{aligned}
\min_{t' \in G'} \mathbb{P}_{g_{t'}}(A) - \mathbb{P}_g(A) \;\; &\leq \;\; |G'|^{-1}\sum_{t' \in G'}\left(\mathbb{P}_{g_{t'}}(A) - \mathbb{P}_g(A)\right) \\
&= \;\; |G'|^{-1}\sum_{t' \in G'}\mathbb{E}_g\left(\left(\frac{d\mathbb{P}_{g_{t'}}}{d\mathbb{P}_g}(Y) - 1\right)\mathbb{I}_A(Y)\right) \\
&\leq \;\; \mathbb{E}_g\left||G'|^{-1}\sum_{t' \in G'}\left(\frac{d\mathbb{P}_{g_{t'}}}{d\mathbb{P}_g}(Y) - 1\right)\right|.
\end{aligned} \tag{4.40}$$

Now by Cameron-Martin-Girsanov's theorem, we have

$$\log \frac{d\mathbb{P}_{g_{t'}}}{d\mathbb{P}_g}(Y) = n^{1/2} h_1^2 (\lambda_1(t_0) - \epsilon^\star) \sqrt{\Pi_{i=1}^d h_i} \|\psi^*\| X_{t'} - \frac{n}{2} (\lambda_1(t_0) - \epsilon^\star)^2 h_1^4 (\Pi_{i=1}^d h_i) \|\psi^*\|^2$$

where

$$X_{t'} = (\Pi_{i=1}^d h_i)^{-1/2} \|\psi^*\|^{-1} \int \phi_{t',h}\, dW$$

with W being the standard Brownian sheet on $[0,1]^d$. Note that here $X_{t'}$ follows a standard normal distribution and for $s' \neq t' \in G'$, $X_{s'}$ and $X_{t'}$ are independent. Now let

$$w_n := n^{1/2} h_1^2 (\lambda_1(t_0) - \epsilon^\star) \sqrt{\Pi_{i=1}^d h_i} \|\psi^*\| \quad \text{and} \quad \varepsilon_n := 1 - (w_n / \sqrt{2 \log |G'|}).$$

If $\varepsilon_n \to 0$ and $\varepsilon_n \sqrt{\log |G'|} \to \infty$ are satisfied, then by Lemma A.5.1 and (4.40), we have the following as $|G'| \to \infty$:

$$\min_{t' \in G'} \mathbb{P}_{g_{t'}}(A) - \mathbb{P}_g(A) \to 0 \tag{4.41}$$

Now let us choose $h_1 = \sqrt{(1 - \epsilon_n) c \rho_n}$ where $\rho_n = (\log(en)/n)^{2/(4+d)}$, with $\epsilon_n \to 0$ and $\epsilon_n \sqrt{\log n} \to \infty$ and c is a constant to be chosen later. This implies that

$$\sqrt{2 \log |G'|} = (1 - o(1)) \sqrt{\frac{2d}{d+4} \log n}$$

and for large n, $\sqrt{2 \log |G'|} < \sqrt{\frac{2d}{d+4} \log n}$. Hence,

$$w_n = [c(1 - \epsilon_n)]^{(4+d)/4} \|\psi^*\| \frac{(\lambda_1(t_0) - \epsilon^\star)^{(4+d)/4}}{\prod_{i=1}^d (\lambda_i(t_0) - \epsilon^\star)^{1/4}} \sqrt{\log(en)}.$$

Let us now put:

$$c := \frac{\prod_{i=1}^d (\lambda_i(t_0) - \epsilon^\star)^{1/(4+d)}}{(\lambda_1(t_0) - \epsilon^\star)} \left[\frac{(d+4)\|\psi^\star\|^2}{2d} \right]^{-2/(4+d)}$$

whence we have:

$$\frac{w_n}{\sqrt{2\log|G'|}} = (1-\epsilon_n)^{(4+d)/4} \to 1 \text{ as } n \to \infty.$$

Also note that for large n, $(1 - w_n/\sqrt{2\log|G|}) > 0$ as $\sqrt{2\log|G|} < \sqrt{\frac{2d}{d+2}\log n}$. Also, note that:

$$\begin{aligned}\sqrt{\log|G'|}\left(1 - \frac{w_n}{\sqrt{2\log|G'|}}\right) &\sim \sqrt{\frac{d}{4+d}}\sqrt{\log n}(1-(1-\epsilon_n)^{1+d/4}) \\ &= \sqrt{\frac{d}{4+d}}\left(\frac{4+d}{4}+o(1)\right)\epsilon_n\sqrt{\log n} \to \infty \quad \text{(by assumption).}\end{aligned}$$

Also denoting $\left((d+4)\|\psi^*\|^2/2d\right)^{-2/(4+d)}$ by $\Delta_*^{(\ell)}$, Hence, by (4.39) and (4.41), we have:

$$\begin{aligned}1-\alpha &\le \liminf_{n\to\infty}\mathbb{P}_g\left(\|g-\ell^*\|_{U'} \ge h_1^2(\lambda_1(t_0)-\epsilon^\star)\right) \\ &= \liminf_{n\to\infty}\mathbb{P}_g\left(\|g-\ell^*\|_{U'} \ge (1-\epsilon_n)\rho_n\Delta_\star^{(\ell)}\Pi_{i=1}^d(\lambda_i(t_0)-\epsilon^\star)^{1/(4+d)}\right) \\ &\le \liminf_{n\to\infty}\mathbb{P}_g\left(\|g-\ell^*\|_{U'} \ge \gamma^\star\rho_n\Delta_\star^{(\ell)}\Pi_{i=1}^d(\lambda_i(t_0)-\epsilon^\star)^{1/(4+d)}\right) \\ &\le \liminf_{n\to\infty}\mathbb{P}_g\left(\|g-\ell^*\|_{U'} \ge \gamma^\star(L_2[t_0]-\epsilon)^{d/(4+d)}\rho_n\Delta_\star^{(\ell)}\right) \\ &= \liminf_{n\to\infty}\mathbb{P}_g\left(\|g-\ell^*\|_{U'} \ge \gamma L_2^{d/(4+d)}[t_0]\rho_n\Delta_\star^{(\ell)}\right).\end{aligned}$$

Now, note that $\|g-\ell^*\|_{U'} = \|f-\ell\|_U$ and $\|\psi^*\|^2 = \sqrt{2(d+3)/(d+1)}\|\psi\|^2$. This completes the proof of Theorem 3.4.2.

4.6.8 Some Technical Lemmas

Lemma 4.6.4 *The function ψ_2^ℓ defined in* (3.14) *satisfies:*

$$\langle 1, \psi_2^\ell\rangle \ge \|\psi_2^\ell\|^2 \quad \textit{and} \quad \langle g, \psi_2^\ell\rangle \ge \|\psi_2^\ell\|^2 - \langle 1, \psi_2^\ell\rangle$$

for all $g \in \mathbb{H}_{2,\sqrt{2(d+3)/(d+1)}}$ whenever $g(0) \ge 0$.

Proof 3 *To begin with, note that* $\psi_2^\ell(x) = (G_A - G_0)\left(\sqrt{2(d+3)/(d+1)}x\right)$*, where*

$$G_A(y) := \frac{\|y\|^2}{2}\mathbb{1}_{\|y\|\leq\sqrt{2(d+3)/(d+1)}} \quad \textit{and} \quad G_0(y) := \left(-1+\frac{\sqrt{2}(d+2)}{\sqrt{(d+1)(d+3)}}\|y\|\right)\mathbb{1}_{\|y\|\leq\sqrt{2(d+3)/(d+1)}} .$$

We will now prove the following claim:

Proposition 4.6.1 *For all* $g \in \mathbb{H}_{2,1}$*,* $G_A - g$ *is convex on the set* $\mathscr{B}_d := \{y : \|y\| \leq \sqrt{2(d+3)/(d+1)}\}$.

For proving Claim 4.6.1, it suffices to show that for every $v \in \mathbb{R}^d$ *such that* $\sum_{i=1}^d v_i \geq 0$*, the function* $f_v : \mathbb{R} \mapsto \mathbb{R}$ *defined as* $f_v(\alpha) := \frac{\|\alpha v\|^2}{2} - g(\alpha v)$ *is convex. Towards this, note that:*

$$f_v'(\alpha) = \alpha\|v\|^2 - v^\top\nabla g(\alpha v) .$$

Now, take any pair (α, β) *such that* $\alpha < \beta$*, and note that:*

$$\begin{aligned}\left|v^\top\nabla g(\alpha v) - v^\top\nabla g(\beta v)\right| &\leq \|v\|\|\nabla g(\alpha v) - \nabla g(\beta v)\| \\ &\leq \|v\|\sum_{i=1}^d |\nabla_i g(\alpha v) - \nabla_i g(\beta v)| \\ &\leq \|v\|\|(\alpha-\beta)v\| = (\beta-\alpha)\|v\|^2 .\end{aligned}$$

The last inequality followed from the fact that $g \in \mathbb{H}_{2,1}$*. Hence, we have:*

$$v^\top\nabla g(\beta v) - v^\top\nabla g(\alpha v) \leq \beta\|v\|^2 - \alpha\|v\|^2 \implies f_v'(\alpha) \leq f_v'(\beta) ,$$

thereby showing that f_v *is convex, and completing the proof of Claim 4.6.1.*

With Claim 4.6.1 in hand, we are now ready to prove Lemma 4.6.4. Defining $\psi := G_A - G_0$*, we have in view of Claim 4.6.1 and* (4.22)*, that for any* $g \in \mathbb{H}_{2,1}$*,*

$$\langle G_A - g, \psi\rangle \leq (G_A - g)(0)\langle 1, \psi\rangle$$

and hence, we have:

$$\begin{aligned}\langle g, \psi\rangle &= \langle G_A, \psi\rangle - \langle G_A - g, \psi\rangle \\ &\geq \langle G_A, \psi\rangle - (G_A - g)(0)\langle 1, \psi\rangle \\ &= \langle G_A, \psi\rangle + g(0)\langle 1, \psi\rangle \\ &= \|\psi\|^2 + \langle G_0, \psi\rangle + g(0)\langle 1, \psi\rangle \\ &= \|\psi\|^2 + (g(0) - 1)\langle 1, \psi\rangle\end{aligned}$$

where the last equality followed from the fact that $\langle G_0 + 1, \psi\rangle = 0$*, which follows by an argument similar to the proof of* (4.23)*. Since* $g(0) \geq 0$*, we conclude that:*

$$\langle g, \psi\rangle \geq \|\psi\|^2 - \langle 1, \psi\rangle \tag{4.42}$$

On putting $g \equiv 0$ *in* (4.42)*, we get:*

$$\langle 1, \psi\rangle \geq \|\psi\|^2 \tag{4.43}$$

Lemma 4.6.4 now follows from (4.42) *and* (4.43) *by a change of variables.*