Electronic Journal of Statistics
Vol. 15 (2021) 5203–5244
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1914

Optimal inference with a multidimensional multiscale statistic*

Pratyay Datta and Bodhisattva Sen

Columbia University
1255 Amsterdam Avenue
New York, NY 10027
e-mail: pd2511@columbia.edu; bodhi@stat.columbia.edu

Abstract: 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 Dümbgen and Spokoiny [11] who proposed the analogous statistic for $d = 1$. In the process, we generalize Theorem 6.1 of Dümbgen and Spokoiny [11] 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.

MSC2020 subject classifications: Primary 62G08, 62G86; Secondary 62C20.
Keywords and phrases: Asymptotic minimax testing, Brownian sheet, kernel estimation, multivariate continuous white noise model, Hölder classes of functions, signal detection on hyperrectangles.

Received March 2021.

1. Introduction

Let us consider the following continuous multidimensional white noise model:

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

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, $W(\cdot)$ is the unobserved d-dimensional Brownian sheet (see Definition 6.1), 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., Brown and Low [4] and Reiß [41]. We work with this white

*Supported by NSF grant DMS-17-12822 and AST-16-14743.

noise model as this formulation is more amiable to rescaling arguments; see e.g., Donoho and Low [10], Dümbgen and Spokoiny [11], Carter [6].

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., Lepski [34], Lepski and Tsybakov [35], Horowitz and Spokoiny [21], Ingster and Sapatinas [26] and the references therein. Our proposed multiscale statistic, which extends the work of Dümbgen and Spokoiny [11], 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 1.2 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., Glaz and Zhang [18], Arias-Castro et al. [3, 2], Chan [7], Walther [50], Butucea and Ingster [5], Chan and Walther [8], Frick et al. [15], König et al. [32] 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., Arias-Castro et al. [3], Chan [7] and König et al. [32]), 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 Walther [50] proposed a test that leads to optimal detection of hyperrectangles when the responses are Bernoulli variables. Also recently, Proksch et al. [39], 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) and briefly describe the asymptotic minimax testing framework. Our main optimality results are discussed in Section 1.2.

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 6.2.

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

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

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

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

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

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, $\sup_{t \in A_h} |\hat{\Psi}(t,h)| = O_p(\sqrt{2\log(1/(2^d h_1 \ldots h_d))})$; see e.g., Giné and Guillou [16]. 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:

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

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 Dümbgen and Spokoiny [11]. In Theorem 2.1, 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 Dümbgen and Spokoiny [11, 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 Dümbgen and Spokoiny [11] beyond $d = 1$ (see e.g., Walther [50], Chan and Walther [8], König et al. [32]) we believe that our approach has the closest resemblance to Dümbgen and Spokoiny [11]. Further, the exact form of $T(Y,\psi)$ leads to optimal tests for (1.2) and other alternatives (see König et al. [32] 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 2.2). This result mirrors Dümbgen and Spokoiny [11, 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; see Remark 2.1 where we compare our result with Dümbgen and Spokoiny [11, Theorem 6.1]. Moreover, Lemma 2.1 gives a bound on the packing numbers of the pertinent (to our application) pseudometric space, which we believe is also new; see Remarks 2.2 and 2.3 where we compare our result with some relevant recent papers.

1.2. 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 Ingster and Suslina [27]. 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., $\mathbb{E}_0[\phi_n(Y)] \le \alpha$) such that

$$\inf_{g \in \mathbb{F}: |g| \ge \delta\rho_n} \mathbb{E}_g[\phi_n(Y)] \tag{1.11}$$

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 Ingster [23], Ingster [24], Ingster [25] for $d = 1$ and Ingster and Sapatinas [26] 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 (1.11) 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 (1.11) 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 Ingster [23], Ingster [24], Ingster [25] for details. In our case also the detection boundary is independent of α as illustrated in Theorems 3.1 and 3.2.

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 $|f| \le \delta\rho_n$, 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

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

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 Lepski and Tsybakov [35], Ingster and Stepanova [22] for more details about minimax testing). On the other hand, we want to find a test $\tilde{\phi}_n$ such that

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

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 3 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 3.1), we show that we can construct a level α test based on the multiscale statistic (1.8) that has power converging to 1, as $n \to \infty$, provided ϵ_n does not go to 0 too fast (see Theorem 3.1 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 3.1 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 3.1 for the details.

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

$$H_0 : f = 0 \quad \text{versus} \quad H_1 : f = \mu_n \mathbb{I}_{B_n}, \tag{1.12}$$

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

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

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 $\lim_{n\to\infty}\sqrt{n}|\mu_n| \to +\infty$, we can construct a level α test based on the multiscale statistic (1.8) that has power converging to 1 as $n \to \infty$; see Theorem 3.2. 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 3.2). 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 3.2 for the details. This shows that our multiscale test is asymptotically minimax for signals at small scales.

1.3. 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 Siegmund and Venkatraman [47], Kulldorff [33], Siegmund and Yakir [48], Jiang [29], Naus and Wallenstein [36], Haiman and Preda [19] 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 Naus and Wallenstein [36] and Pozdnyakov et al. [38] the authors give asymptotic approximations of the distribution of the scan statistic when $d = 1$. For $d = 2$, similar results can be found in Glaz and Zhang [18], Haiman and Preda [19], Wang and Glaz [51], among others. Recently in Sharpnack and Arias-Castro [46] 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., Naus and Wallenstein [36], Sharpnack and Arias-Castro [46]. 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. Proksch et al. [39] 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., Walther [50], Chan and Walther [8]) and use multiple testing procedures (see Hall and Jin [20] 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 Dümbgen and Spokoiny [11], where the authors proposed our multiscale statistic for $d = 1$. Thus, our work can be thought of as a generalization of Dümbgen and Spokoiny [11] to multidimension ($d > 1$).

1.4. Organization of the paper

The proposed multiscale statistic is studied in Section 2. In Section 3 we construct *optimal* tests for: (i) $f = 0$ versus Hölderian alternatives; (ii) $f = 0$ versus alternatives of the form $f = \mu_n \mathbb{I}_{B_n}$, where B_n is an axis-aligned hyperrectangle in $[0,1]^d$ and $\mu_n \in \mathbb{R}$ (both unknown). In Section 3.3 we discuss the discrete analogue of our statistic and the computational issues. We compare the performance of our multiscale based test with other competing methods in Section 4. In Section 5 we discuss some open problems and possible applications/extensions of our work. Section 6 gives the proofs of Lemma 2.1 and Theorem 2.2. The proofs of the other results are relegated to Appendix A.

2. 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.2 for a proof.

Theorem 2.1. *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 immediately extends the main result of Dümbgen and Spokoiny [11, 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 (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 as a consequence of the following more general result about stochastic processes with sub-Gaussian increments on some pseudometric space (see Section 6.2 for its proof).

Theorem 2.2. *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

$$S(X) := \sup_{a \in \mathscr{F}} \frac{X^2(a)/\sigma^2(a) - 2V \log(1/\sigma^2(a))}{\log\log(e^e/\sigma^2(a))} \tag{2.1}$$

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 (Connection to Dümbgen and Spokoiny [11])**.** *A similar result to Theorem 2.2 above appears in Dümbgen and Spokoiny [11, Theorem 6.1]. However note that there is a subtle and important difference: The bound on the packing number in (c) of Theorem 2.2 involves the additional logarithmic factor $(\log(e/\delta))^p$ which is not present in Dümbgen and Spokoiny [11, Theorem 6.1]. In fact, we show that even with this additional logarithmic factor, the random variable $S(X)$, defined in (2.1), involves the same penalization term $2V \log(1/\sigma^2(a))$ as in Dümbgen and Spokoiny [11, Theorem 6.1]. Hence, we can think of Theorem 2.2 as a generalization of Dümbgen and Spokoiny [11, Theorem 6.1]. Here we would also like to point out that our result improves Dümbgen and Spokoiny [11, Theorem 6.1] by proving the subexponentiality of the random variable $S(X)$ instead of just its finiteness.*

To apply Theorem 2.2 to prove Theorem 2.1 we need to define a suitable pseudometric space $(\mathscr{F}, \rho)$ and a stochastic process, and verify that conditions (a)-(c) in Theorem 2.2 hold. In that vein, let us define the set

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

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 holds; see Section 6.3 for its proof.

Lemma 2.1. *Let $\mathscr{F}, \rho(\cdot,\cdot)$ and $\sigma(\cdot)$ be as described above. Then, for all $u, \delta \in (0,1]$,*

$$N\left((u\delta)^{1/2}, \{(t,h) \in \mathscr{F} : \sigma^2(t,h) \le \delta\}\right) \le K u^{-2d} \delta^{-1} (\log(e/\delta))^{d-1} \tag{2.2}$$

for some constant K depending only on d.

Remark 2.2. *Here we would like to point out that Lemma 2.1 shows that condition (c) of Theorem 2.2 holds with $B = 2d$, $p = d-1$ and most importantly for $V = 1$, which was also the case when $d = 1$ (as shown in Dümbgen and Spokoiny [11]). An equivalent result for $d = 2$ is proved in Walther [50, Theorem 1].*

Remark 2.3 (Connection to Sharpnack [45])**.** *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 Sharpnack [45] 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, Sharpnack [45] 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 Sharpnack [45]; see Section 3.2 for more details.*

Compare the numerator of our multiscale statistic (1.8) with the multiscale statistic proposed in König et al. [32, Equation (6)]. Translated to our setting, in König et al. [32] 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 König et al. [32, 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. König et al. [32, Example 2.3] recommend choosing $V = 1$ by appealing to Lemma 2.1 of our paper. The following proposition shows that indeed $V = 1$ is the smallest possible permissible value; see Appendix A.3 for a proof.

Proposition 2.1. *Suppose $V < 1$. Let Γ_V and $\mathscr{F}$ be as defined above. Then we have*

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

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

3. 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 (1.12). 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 (1.12) 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 3.1 and 3.2 we show that specific choices of the kernel function ψ lead to asymptotically minimax tests.

3.1. 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 6.2 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:

$$\text{Minimize } \|\psi\| \text{ over all } \psi \in \mathbb{H}_{\beta,1} \text{ with } \psi(0) \geq 1. \tag{3.1}$$

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 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 3.1. *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

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

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

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

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

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

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

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

The above result generalizes Dümbgen and Spokoiny [11, Theorem 2.2] beyond $d=1$. Theorem 3.1 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.5.2 for its proof.

Proposition 3.1. *Consider testing* (1.2) *where $\beta \le 1$ is unknown. Let us recall the definition of ψ_1 in* (3.1)*. 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

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

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

Remark 3.1. *Instead of using the test statistic T_β if we use the test statistic*

$$T_\beta^\star := \sup_{h\in(0,1/2]^d} \sup_{t\in A_h} \left[|\hat\Psi(t,h)| - \Gamma(2^d h_1 \dots h_d)\right] \tag{3.2}$$

with the kernel ψ_β, then the same conclusions as that of Theorem 3.1 and Proposition 3.1 would hold. Thus the multiscale statistic $T_\beta^\star$ is also optimal against Hölderian alternatives.

3.2. Optimality against axis-aligned hyperrectangular signals

In Theorem 3.1 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 (1.12), and establish below the optimality of our multiscale test in this setting (see Appendix A.5.3 for a proof of Theorem 3.2). Note that when $d = 1$ similar optimality results are known for the multiscale statistic; see Frick et al. [15, Theorem 2.6] and Chan and Walther [8, Section 4]. For $d > 1$ see Walther [50] 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 3.2. *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 (1.12). 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 $h_n = (h_{1,n}, \dots, h_{d,n}) \in (0, 1/2]^d$ 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 (1.12) 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 3.2. *If we use the test statistic $T^\star$, as defined in (3.2) (with the kernel ψ_0), instead of T in Theorem 3.2, the optimality results described in the theorem still hold.*

Our first result in Theorem 3.2 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 Proksch et al. [39], the authors used a multiple testing method to achieve optimal detection in both large and small scale hyperrectangles.

Remark 3.3. *We would like to point out that the proofs for the minimax lower bound that have been derived for the two scenarios in Theorems 3.1 and 3.2 follow the standard techniques that have been used in Ingster [23], Ingster [24], Ingster [25], Lepski and Tsybakov [35], Dümbgen and Spokoiny [11], Arias-Castro et al. [3], Ingster and Sapatinas [26], Arias-Castro et al. [2], Frick et al. [15], 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 Dümbgen and Spokoiny [11].*

3.2.1. 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 Chan and Walther [8] 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 $|\mu_n|\sqrt{n} \ge (1+\epsilon_n)\sqrt{2\log n}$, 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.

Frick et al. [15, 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 Rivera and Walther [42] and Chan and Walther [8] 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 3.2 shows that our multidimension multiscale test is asymptotically minimax even when $d > 1$.

3.3. 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:

$$Y_i = f(x_i) + \epsilon_i, \qquad \text{for } i = 1, \dots, n \tag{3.3}$$

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

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

TABLE 1
Critical values $\kappa_{0.05}$ for different $n = m^2$.

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	$3.30^\star$

* *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.*

Then the multiscale statistic for this regression problem reduces to

$$T_n(Y, \psi) := \sup_{h \in G^m : t-h, t+h \in G^m} \; \sup_{t \in G^m} \frac{|\hat{\Psi}_n(t,h)| - \Gamma\left(|B_\infty(t,h) \cap G^m|\right)}{D\left(|B_\infty(t,h) \cap G^m|\right)} \tag{3.4}$$

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 Walther [50] 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).

4. Simulation studies

In this section we demonstrate the performance of the multiscale testing procedure described in Section 3 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 (3.3). For the simulations we have used the kernel function $\psi = \mathbb{I}_{[-1,1]^d}$. In Table 1 we give the empirical 0.95-quantile of the multiscale statistic $T_n(W, \psi)$ (see (3.4)) 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 1 shows the empirical distribution function estimates of $T_n(W, \psi)$ for different values of n, based on 3000 replications.

In Tables 2 and 3 we compare the powers of the multiscale test, a test based on a scan-statistic, and the ALR test (see Chan and Walther [8] for the details). Formally, we consider testing (1.12) 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

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

$k = 1$			
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

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:

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

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 Glaz et al. [17, Chapter 5]) for this problem is defined as

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

The ALR test statistic (see Chan [7]) 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 2 and 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,

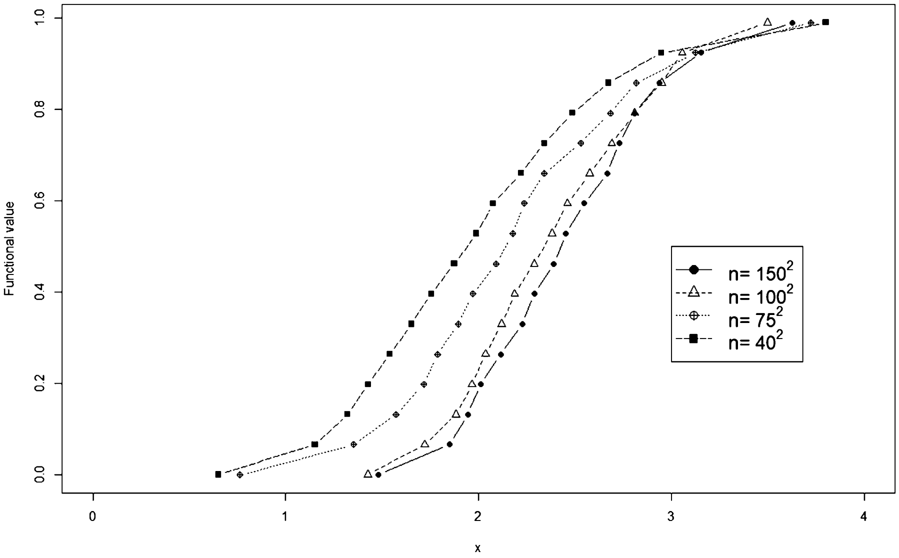

Fig 1. *The empirical distribution functions of the multiscale statistic for different values of n.*

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.

5. Discussion

In this paper 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 Ji and Nussbaum [28, Section 1.3]. Another interesting question would be to try to extend our results to other smoothness classes like Sobolev/Besov classes; in Ingster and

TABLE 3

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

$k = 1$			
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

Stepanova [22] 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. Frick et al. [15] 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 Arias-Castro et al. [3] the authors use the scan statistic to detect regions in $\mathbb{R}^d$ where the underlying function is non-zero. Arias-Castro et al. [2] 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., Walther [50], Rivera and Walther [42], Chan and Walther [9], König et al. [32], etc. In Pein et al. [37] the authors looked at the problem of detecting change points under heterogeneous variance of the response variable (when $d = 1$). Rohde [43] 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 Walther [50] 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 Dümbgen and Spokoiny [11]): In Dümbgen and Walther [12] the authors propose a multiscale test statistic to make inference about a probability density on the real line given i.i.d. observations; Schmidt-Hieber et al. [44] use multiscale methods to make inference in a deconvolution problem; Rivera and Walther [42] use multiscale methods to detect a jump in the intensity of a Poisson process; Eckle et al. [13] and Eckle et al. [14] 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.

6. Proofs of our main result

6.1. Some useful concepts

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

Definition 6.1 (Brownian sheet)**.** *By a d-dimensional Brownian sheet we mean a mean-zero Gaussian process $\{W(t) : t \in [0,1]^d\}$ with covariance*

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

for $(t_1,\dots,t_d),(s_1,\dots,s_d) \in [0,1]^d$. The Brownian sheet is the d-dimensional counterpart of the standard Brownian motion; see e.g., Wong and Zakai [52], Khoshnevisan [31, 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.

Definition 6.2. *Fix $\beta > 0$ and $L > 0$. Let $\lfloor \beta \rfloor$ be the largest integer which is strictly less than β and for $k = (k_1, k_2, \dots, k_d) \in \mathbb{N}^d$ 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*

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

and

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

See Appendix A.1.2 for an important property of Hölder classes of functions useful in our proofs.

6.2. Proof of Theorem 2.2

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

$$\mathbb{P}\big(\Gamma(X, v) > \eta\big) \leq K \exp\left(-\frac{\eta^2}{K v^2 \log(e/v)}\right) \quad \forall\, \eta > 0 \text{ and } v \in (0, 1], \tag{6.1}$$

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

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

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., $\mathbb{P}(|X| > x) \leq [\lambda(x/\|X\|_\lambda)]^{-1}$, for all $x \in \mathbb{R}$ (see van der Vaart and Wellner [49, Page 96] for a simple proof). Let us define $\lambda(x) := \exp(x^2) - 1$, $x > 0$. Hence,

$$\mathbb{P}\big(|X| > x\big) \leq \min\left\{1, \frac{1}{\exp(x^2/\|X\|_\lambda^2) - 1}\right\} \leq 2 \times \exp(-x^2/\|X\|_\lambda^2). \tag{6.2}$$

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 van der Vaart and Wellner [49, Theorem 2.2.4] which we state below.

Lemma 6.1. *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*

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

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

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

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 2.2 directly implies that $\|X_a - X_b\|_\lambda \leq C\rho(a,b)$ by an application of van der Vaart and Wellner [49, Lemma 2.2.1].

By taking $\delta = 1, \epsilon = u^{1/2}$, condition (c) of Theorem 2.2 yields $N(\epsilon, \mathscr{F}) \leq A\epsilon^{-2B}$. Thus, Lemma 6.1 gives (with $\zeta = v$)

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

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 (6.2) with $\Gamma(X,v)$ instead of X imply

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

for some constant K.

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

$$\Pi(\delta) := \mathbb{P}\left(\frac{X^2(a)}{\sigma^2(a)} > 2V\log(\frac{1}{\delta}) + S\log\log(\frac{e^e}{\delta}) \text{ for some } a \in \mathscr{F}(\delta)\right) \tag{6.3}$$

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

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

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 $\rho^2(a,\hat{a}) \leq u\delta$. Using assumption (a) of Theorem 2.2 we have

$$\sigma^2(a) \geq \sigma^2(\hat{a}) - u\delta \geq \sigma^2(\hat{a})(1-2u) \tag{6.4}$$

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

$$\frac{X^2(a)}{\sigma^2(a)} > r \tag{6.5}$$

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:

$$\Gamma(X,(u\delta)^{1/2})^2 \geq |X(a) - X(\hat{a})|^2 > \lambda^2 X^2(a) \geq \lambda^2 r\sigma^2(a) \geq \lambda^2 r\frac{\delta}{2} \tag{6.6}$$

(here the first inequality follows from the definition of $\Gamma(X, (u\delta)^{1/2})$ and the third inequality follows from condition (6.5)), and

$$X^2(\hat{a}) \geq (1-\lambda)^2 X^2(a) \geq (1-\lambda)^2 r\sigma^2(a) \geq (1-\lambda)^2 r(1-2u)\sigma^2(\hat{a}) \qquad (6.7)$$

(here the second inequality follows from (6.5) and last inequality follows from (6.4)). 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 (6.6) holds or (6.7) 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 (6.1) 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 2.2. Hence we get

$$\begin{aligned}\Pi_r(\delta) &\leq K \exp\left(-\frac{\lambda^2\delta r/2}{Ku\delta \log(e/\sqrt{u\delta})}\right) \\ &\qquad\qquad + Au^{-B}\delta^{-V}\left(\log(\frac{e}{\delta})\right)^p \exp\left(-\frac{(1-\lambda)^2 r(1-2u)}{2}\right) \\ &\leq K\Big[\exp\left(-\frac{\lambda^2 r}{Ku\log(e/(u\delta))}\right) \\ &\qquad + \exp\left(B\log(1/u) + V\log(1/\delta) + p\log\log(e/\delta) + ur - (1/2-\lambda)r\right)\Big].\end{aligned} \qquad (6.8)$$

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

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

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 (6.8) gives us

$$\Pi(\delta) \equiv \Pi_r(\delta) \quad \leq \quad K\left[\exp\left(-\frac{(S-4p)^2(\log\log(e^e/\delta))^2}{Kur\log(e/(u\delta))}\right)\right.$$

$$+\exp\Big(B\log(1/u)+ur-(S/4)\log\log(e^e/\delta)\Big)\Bigg] \tag{6.9}$$

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 (6.9) 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 (6.9) 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 (6.9) have the form $K\exp[(C - S/K')\log\log(e^e/\delta)]$ for some constants $K, C, K' > 0$. Putting these values in (6.9) 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 (6.3) 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 $\mathscr{F} = \bigcup_{l\geq 0}\mathscr{F}(2^{-l})$),

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

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. □

6.3. *Proof of Lemma 2.1*

First let us define the following sets:

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

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

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

(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 $(l_1,\dots,l_d) \in \mathbb{Z}^d$, and $\delta, u \in (0,1]$,

$$N\left((u\delta)^{1/2}, \mathscr{F}_{\delta,(l_1,\dots,l_d)}\right) \le K u^{-2d}\delta^{-1}. \tag{6.10}$$

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

$$\rho^2((t,h),(t',h')) > u\delta. \tag{6.11}$$

Our aim is to show that

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

for some constant K independent of $(l_1,\dots,l_d)$, u and δ. If $\mathscr{F}_{\delta,(l_1,\dots,l_d)}$ is empty then the assertion is trivial. So assume that $\mathscr{F}_{\delta,(l_1,\dots,l_d)}$ 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:

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

where we take $c := d4^d$. We would like to point out that in the above definition when $i_k = 1$, for any $k = 1,\dots,d$, 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

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

Hence we can easily see that

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

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 $k = 1,\dots,d$. This implies that if $(M_{\underset{\sim}{i}}, M_{\underset{\sim}{i}'}) \in R_2$, where $\underset{\sim}{i} = (i_1,\dots,i_d)$ and $\underset{\sim}{i}' = (i'_1,\dots,i'_d)$, then

$$(i'_k - i_k) \le (1 + 2cu^{-1}), \qquad \text{for all } \ k = 1,\dots,d, \tag{6.12}$$

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

$$B_\infty(t, h) \triangle B_\infty(t', h') \subseteq B_1 \setminus B_2 \tag{6.13}$$

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 2 for a visual illustration of (6.13) 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 similar argument shows that $B_2 \subseteq B_\infty(t', h')$. Therefore, assertion (2) is also satisfied. Now let us define the following set

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

Clearly, using (6.12),

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

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 $w \in B_1$),

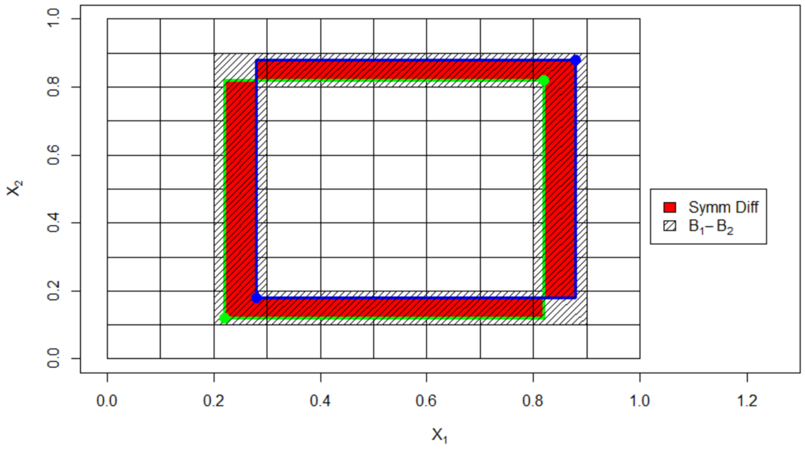

FIG 2. *The figure shows how the symmetric difference of the hyperrectangles $B_\infty(t,h)$ (denoted by the green border) and $B_\infty(t',h')$ (denoted by the blue border) is contained in the set $B_1 \setminus B_2$ (denoted by the shaded region).*

(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 (6.13) 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 (6.11). 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 (6.10).

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

$$N\left((u\delta)^{1/2}, \mathscr{F}(\delta)\right) \leq K u^{-2d} \delta^{-1} (\log(e/\delta))^{d-1}. \qquad (6.14)$$

Let us define the set

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

Now it can be easily seen that $l := (l_1, \dots, l_d) \in S$ implies $l_k \geq -(d+1) - (d-1)(1/d)\log_2(1/\delta)$, for all $k = 1, \dots, d$. 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 $l_1, l_2, \dots, l_{d-1}$ then l_d can only take at most $(d+1)$ values such that $(l_1, l_2, \dots l_d) \in S$ (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 (6.14).

Step 3: In this step we will complete the proof of Lemma 2.1. 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

$$N((u\delta)^{1/2}, \mathscr{F}(\delta)) \leq N((u\delta/2)^{1/2}, \mathscr{F}(\delta)) \leq K u^{-2d} \delta^{-1} (\log(e/\delta))^{d-1} \tag{6.15}$$

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 (6.15) 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 $N(\epsilon, \bigcup_{i=1}^m A_i) \leq \sum_{i=1}^m N(\epsilon, A_i)$ we get

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

which proves Lemma 2.1. □

Appendix A: Proofs

In this Appendix we: (i) Elaborate on some parts of the main text that were deferred so as not to impede the flow of the paper, and (ii) prove some of the results that were stated in the main paper.

A.1. Some useful concepts

In this subsection we discuss some important concepts and properties that are used in the 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 $\mathrm{Cov}\left(\int g_1 dW, \int g_2 dW\right) = \int_{[0,1]^d} g_1(t) g_2(t) dt$.
- *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 Protter [40, 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

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

A.1.2. 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.

A.1.3. 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 Aistleitner and Dick [1], which suffices for our purpose.

Definition A.1 (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

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

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

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

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. *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

We use Theorem 2.2 to prove Theorem 2.1. Let us recall the definitions of $\mathscr{F}, \sigma$ and ρ as introduced just before Lemma 2.1 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 2.2 are satisfied. Condition (c) is also satisfied because of Lemma 2.1. Thus, by an application of Theorem 2.2 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,

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

A.3. Proof of Proposition 2.1

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

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

The above result follows trivially from Kabluchko and Munk [30, 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,

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

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, $F_{m^d}(\sqrt{V} + s/\sqrt{2\log(m^d)}) \le F_{m^d}(1-\epsilon)$, if m is large enough. As this is true for all large m, taking $m \to \infty$ gives us the desired result. □

A.4. Solution to (3.1)

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

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

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

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

where $\psi_\beta(x) = (1 - \|x\|^\beta)\mathbb{I}(\|x\| \le 1)$. 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 $1 \ge \|x\| \ge \|y\|$. Then

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

Here the third inequality follows from the fact that when $\beta \le 1$ 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 $\|x\| \ge 1 \ge \|y\|$ then we have

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

If $x, y \in \mathbb{R}^d$ is such that $\|x\| \ge \|y\| \ge 1$ then the assertion is trivial. Hence we have proved that ψ_β minimizes (3.1). □

A.5. Proofs of Theorems 3.1 and 3.2

The proofs of Theorems 3.1 and 3.2 depend on the following lemma (stated and proved in Dümbgen and Spokoiny [11, Lemma 6.2]).

Lemma A.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 3.1

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,

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

$$\leq \mathbb{E}_0 \left| |S|^{-1} \sum_{t\in S} \frac{dP_{g_t}}{dP_0}(Y) - 1 \right|. \qquad \text{(A.1)}$$

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 Protter [40, 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 $g_t(\cdot) = L\min(h)^\beta \psi_{t,h}^{(\beta)}(\cdot)$, $\sqrt{n}\int g_t dW = \sqrt{n}L\|\psi_\beta\|\min(h)^\beta\sqrt{\Pi_{i=1}^d h_i}\hat{\Psi}(t,h)$. 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

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

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.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

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

Then,

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

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

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

$$\to \quad \frac{d}{2\beta+d} \text{ as } n \to \infty. \tag{A.3}$$

Also notice that for all large n, $\log|S|/\left(\frac{d}{2\beta+d}\log n\right) < 1$. Combining (A.2) and (A.3), 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 $K, K' > 0$,

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

as the o(1) term above is positive when n is large. This proves part (a) of Theorem 3.1 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 $K > 0$,

$$\begin{aligned} &\mathbb{P}_g\left(|\hat{\Psi}(t,h)| > \Gamma(2^d\tilde{h}^d) + K\right) \\ &= \mathbb{P}_0\left(\left|\hat{\Psi}(t,h) + \sqrt{\frac{n}{\tilde{h}^d}}\,\|\psi_\beta\|^{-1}\,\langle g, \psi_{t,h}^{(\beta)}\rangle\right| > \Gamma(2^d\tilde{h}^d) + K\right) \\ &\geq \mathbb{P}_0\left(-\text{sign}(\langle g, \psi_{t,h}^{(\beta)}\rangle)\hat{\Psi}(t,h) < \sqrt{\frac{n}{\tilde{h}^d}}\frac{|\langle g, \psi_{t,h}^{(\beta)}\rangle|}{\|\psi_\beta\|} - K - \Gamma(2^d\tilde{h}^d)\right) \\ &= \Phi\left(\sqrt{\frac{n}{\tilde{h}^d}}\,\|\psi_\beta\|^{-1}\,|\langle g, \psi_{t,h}^{(\beta)}\rangle| - K - \Gamma(2^d\tilde{h}^d)\right) \end{aligned} \tag{A.4}$$

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

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

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

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

Thus,

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

This proves part (b) of Theorem 3.1. □

A.5.2. Proof of Proposition 3.1

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.4) 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,

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

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,

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

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.5.3. Proof of Theorem 3.2

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,

$$\mathbb{P}_{f_n}(T>\kappa_\alpha) \;\geq\; \mathbb{P}_{f_n}\left(\frac{|\hat{\Psi}(t_n,h_n)|-\Gamma(|B_n|)}{D(|B_n|)}>\kappa_\alpha\right)$$

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

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 (1.12). Similar arguments as in (A.1) 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 3.1, we have

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

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.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

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

$$\geq \quad \mathbb{P}_0\left(\hat{\Psi}(t_n, h_n) \geq -\epsilon_n\sqrt{2\log(1/|B_n|)} + \kappa_\alpha D(|B_n|)\right) \to 1$$

as $n \to \infty$. This completes the proof of Theorem 3.2. □

Acknowledgments

The authors would like to thank Lutz Dümbgen, Sumit Mukherjee and Rajarshi Mukherjee for several helpful discussions. We would also like to thank the Editor, Associate Editor and the referees for their helpful comments.

References

[1] Aistleitner, C. and Dick, J. (2015). Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality. *Acta Arith.*, 167(2):143–171. MR3546971

[2] Arias-Castro, E., Candès, E. J., and Durand, A. (2011). Detection of an anomalous cluster in a network. *Ann. Statist.*, 39(1):278–304. MR1331667

[3] Arias-Castro, E., Donoho, D. L., and Huo, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory*, 51(7):2402–2425. MR1748719

[4] Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398. MR1385671

[5] Butucea, C. and Ingster, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688. MR2604703

[6] Carter, A. V. (2006). A continuous Gaussian approximation to a nonparametric regression in two dimensions. *Bernoulli*, 12(1):143–156. MR3217827

[7] Chan, H. P. (2009). Detection of spatial clustering with average likelihood ratio test statistics. *Ann. Statist.*, 37(6B):3985–4010. MR0448555

[8] Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statist. Sinica*, 23(1):409–428.

[9] Chan, H. P. and Walther, G. (2015). Optimal detection of multi-sample aligned sparse signals. *Ann. Statist.*, 43(5):1865–1895.

[10] Donoho, D. L. and Low, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.*, 20(2):944–970.

[11] Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152.

[12] Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Ann. Statist.*, 36(4):1758–1785.

[13] Eckle, K., Bissantz, N., and Dette, H. (2017). Multiscale inference for multivariate deconvolution. *Electron. J. Stat.*, 11(2):4179–4219.

[14] Eckle, K., Bissantz, N., Dette, H., Proksch, K., and Einecke, S. (2018). Multiscale inference for a multivariate density with applications to X-ray astronomy. *Ann. Inst. Statist. Math.*, 70(3):647–689.

[15] Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580. With 32 discussions by 47 authors and a rejoinder by the authors.

[16] Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 38(6):907–921. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

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

[18] Glaz, J. and Zhang, Z. (2004). Multiple window discrete scan statistics. *J. Appl. Stat.*, 31(8):967–980.

[19] Haiman, G. and Preda, C. (2006). Estimation for the distribution of two-dimensional discrete scan statistics. *Methodol. Comput. Appl. Probab.*, 8(3):373–381.

[20] Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence. *Ann. Statist.*, 36(1):381–402.

[21] Horowitz, J. L. and Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, 69(3):599–631.

[22] Ingster, Y. and Stepanova, N. (2011). Estimation and detection of functions from anisotropic Sobolev classes. *Electron. J. Stat.*, 5:484–506.

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

[24] Ingster, Y. I. (1993b). Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Math. Methods Statist.*, 2(3):171–189.

[25] Ingster, Y. I. (1993c). Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Math. Methods Statist.*, 2(4):249–268.

[26] Ingster, Y. I. and Sapatinas, T. (2009). Minimax goodness-of-fit testing in multivariate nonparametric regression. *Math. Methods Statist.*, 18(3):241–269.

[27] Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

[28] Ji, P. and Nussbaum, M. (2017). Sharp minimax adaptation over Sobolev ellipsoids in nonparametric testing. *Electron. J. Stat.*, 11(2):4515–4562.

[29] Jiang, T. (2002). Maxima of partial sums indexed by geometrical structures. *Ann. Probab.*, 30(4):1854–1892.

[30] Kabluchko, Z. and Munk, A. (2008). Exact convergence rate for the maximum of standardized Gaussian increments. *Electron. Commun. Probab.*, 13:302–310.

[31] Khoshnevisan, D. (2002). *Multiparameter processes*. Springer Monographs in Mathematics. Springer-Verlag, New York. An introduction to random fields.

[32] König, C., Munk, A., and Werner, F. (2020). Multidimensional multiscale scanning in exponential families: limit theory and statistical consequences. *Ann. Statist.*, 48(2):655–678.

[33] Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods*, 26(6):1481–1496.

[34] Lepski, O. V. (1993). On asymptotically exact testing of nonparametric hypotheses. Technical report, Université catholique de Louvain.

[35] Lepski, O. V. and Tsybakov, A. B. (2000). Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probab. Theory Related Fields*, 117(1):17–48.

[36] Naus, J. I. and Wallenstein, S. (2004). Multiple window and cluster size scan procedures. *Methodol. Comput. Appl. Probab.*, 6(4):389–400.

[37] Pein, F., Sieling, H., and Munk, A. (2017). Heterogeneous change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(4):1207–1227.

[38] Pozdnyakov, V., Glaz, J., Kulldorff, M., and Steele, J. M. (2005). A martingale approach to scan statistics. *Ann. Inst. Statist. Math.*, 57(1):21–37.

[39] Proksch, K., Werner, F., and Munk, A. (2018). Multiscale scanning in inverse problems. *Ann. Statist.*, 46(6B):3569–3602.

[40] Protter, P. E. (2005). *Stochastic integration and differential equations*, volume 21 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin. Second edition. Version 2.1, Corrected third printing.

[41] Reiß, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.*, 36(4):1957–1982.

[42] Rivera, C. and Walther, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Stat.*, 40(4):752–769.

[43] Rohde, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Ann. Statist.*, 36(3):1346–1374.

[44] Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Ann. Statist.*, 41(3):1299–1328.

[45] Sharpnack, J. (2018). Learning patterns for detection with multiscale scan statistics. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 950–969. PMLR.

[46] Sharpnack, J. and Arias-Castro, E. (2016). Exact asymptotics for the scan statistic and fast alternatives. *Electron. J. Stat.*, 10(2):2641–2684.

[47] Siegmund, D. and Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.*, 23(1):255–271.

[48] Siegmund, D. and Yakir, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli*, 6(2):191–213.

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

[50] Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.*, 38(2):1010–1033.

[51] Wang, X. and Glaz, J. (2014). Variable window scan statistics for normal data. *Comm. Statist. Theory Methods*, 43(10-12):2489–2504.

[52] Wong, E. and Zakai, M. (1977). An extension of stochastic integrals in the plane. *Ann. Probab.*, 5(5):770–778.