Probab. Theory Relat. Fields (2011) 149:515–559
DOI 10.1007/s00440-010-0263-1

Optimal calibration for multiple testing against local inhomogeneity in higher dimension

Angelika Rohde

Received: 5 November 2008 / Revised: 24 September 2009 / Published online: 4 February 2010

Abstract Based on two independent samples $X_1, \ldots, X_m$ and $X_{m+1}, \ldots, X_n$ drawn from multivariate distributions with unknown Lebesgue densities p and q respectively, we propose an exact multiple test in order to identify simultaneously regions of significant deviations between p and q. The construction is built from randomized nearest-neighbor statistics. It does not require any preliminary information about the multivariate densities such as compact support, strict positivity or smoothness and shape properties. The properly adjusted multiple testing procedure is shown to be sharp-optimal for typical arrangements of the observation values which appear with probability close to one. The proof relies on a new coupling Bernstein type exponential inequality, reflecting the non-subgaussian tail behavior of a combinatorial process. For power investigation of the proposed method a reparametrized minimax set-up is introduced, reducing the composite hypothesis "$p = q$" to a simple one with the multivariate mixed density $(m/n)p + (1 - m/n)q$ as infinite dimensional nuisance parameter. Within this framework, the test is shown to be spatially and sharply asymptotically adaptive with respect to uniform loss on isotropic Hölder classes. The exact minimax risk asymptotics are obtained in terms of solutions of the optimal recovery.

Keywords Combinatorial process · Exponential concentration bound · Coupling · Decoupling inequality · Exact multiple test · Nearest-neighbors · Optimal recovery · Sharp asymptotic adaptivity

Mathematics Subject Classification (2000) 62G10 · 62G20

A. Rohde (✉)
Department Mathematik, Universität Hamburg,
Bundesstraße 55, 20146 Hamburg, Germany
e-mail: angelika.rohde@math.uni-hamburg.de

1 Introduction

Given two independent multivariate iid samples

$$X_1, \ldots, X_m \quad \text{and} \quad X_{m+1}, \ldots, X_n$$

with corresponding Lebesgue densities p and q respectively, we are interested in identifying simultaneously subregions of the densities support where p deviates significantly from q at prespecified but arbitrarily chosen level $\alpha \in (0, 1)$. For this aim a multiple test of the composite hypothesis $H_0 : p = q$ versus $H_A : p \neq q$ is proposed, built from a suitable combination of randomized nearest-neighbor statistics. The procedure does not require any preliminary information about the multivariate densities such as compact support, strict positivity or smoothness and shape properties, and it is valid for arbitrary finite sample sizes m and $n - m$. The hierarchical structure of p-values for subsets of deviation between p and q provides insight into the local power of nearest-neighbor classifiers, based on the training set $\{X_1, \ldots, X_n\}$. Thus our method is of interest in particular if the classification error depends strongly on the value of the feature vector, related to recent literature on classification procedures by Belomestny and Spokoiny [2].

There is an extensive amount on literature concerning two-sample problems. Most of it is devoted to the one-dimensional case as there exists the simple but powerful "quantile transformation", allowing for distribution-freeness under the null hypothesis of several test statistics. Starting from the classical univariate mean shift problem (see e.g. [14]), more flexible alternatives as stochastically larger or omnibus alternatives have been investigated for instance by Behnen et al. [1], Neuhaus [26,27], Fan [13], Janic-Wróblewska and Ledwina [18], and Ducharme and Ledwina [7]. Our approach is different in that it aims at spatially adaptive and simultaneous identification of local rather than global deviations. In the above cited literature asymptotic power is discussed against single directional alternatives tending to zero at a prespecified rate, typically formulated by means of the densities $\tilde{p}$ and $\tilde{q}$ corresponding to the transformed observations $\tilde{X}_i = H(X_i)$, where H denotes the mixed distribution function with density $h = (m/n)p + (1 - m/n)q$. Note that the mapping H coincides with the inverse quantile transformation under the null.

For power investigation of our procedure a specific two-sample minimax set-up is introduced. It is based on a reparametrization of (p, q) to a couple (ϕ, h), reducing the composite hypothesis "$p = q$" to the simple one "$\phi \equiv 0$" with the multivariate mixed density h as infinite dimensional nuisance parameter. The reparametrization conceptionally differs from the above described transformation for the univariate situation as it cannot rely on the inverse mixed distribution function. Nevertheless it leads under moderate additional assumptions in that case to the same notion of efficiency. In order to explore the power of our method, the alternative is assumed to be of the form

$$\{(p, q) : (m/n)p + (1 - m/n)q = h,\ \phi \in \mathcal{F},\ \|\phi\| \geq \delta\} \tag{1}$$

for fixed but unknown h, some suitably chosen (semi-)norm $\|\cdot\|$, a constant $\delta > 0$ and a given smoothness class $\mathcal{F}$. For any $\alpha \in (0, 1)$ the quality of a statistical level-α-test

ψ is then quantified by its minimal power $\inf \mathbb{E}_{(p,q)}\psi$, where the infimum is running over all couples (p, q) which are contained in the set (1). It is a general problem that an optimal solution ψ may depend on $\mathcal{F}$ and h. Since the smoothness and shape of a potential difference $p - q$ are rarely known in practice, it is of interest to come up with a procedure which does not depend on these properties but is (almost) as good as if they were known, leading to the notion of minimax adaptive testing as introduced in [36]. Note that here we have however h as an additional infinite dimensional nuisance parameter.

The problem of data-driven testing a simple hypothesis is further investigated for instance by Ingster [17], Eubank and Hart [12], Ledwina [22], Ledwina and Kallenberg [21], Fan [13] and Dümbgen and Spokoiny [10] among others , the two-sample context by Butucea and Tribouley [4]. The idea in common is to combine a family of test statistics corresponding to different values of the smoothing parameters, respectively; see, for instance, Rufibach and Walther [33] for a general criterion of multiscale inference. The closest in spirit to ours is the procedure developed in Dümbgen and Spokoiny [10] within the continuous time Gaussian white noise model and further explored by Dümbgen [9], Dümbgen and Walther [11] and Rohde [32], all concerned with univariate problems. Walther [38] treats the problem of spatial cluster analysis in two dimensions.

The paper is organized as follows. In the subsequent section, a multiple randomization test is introduced, built from a combination of suitably standardized nearest-neighbor statistics. Its calibration relies on a new coupling exponential bound and an appropriate extension of the multiscale empirical process theory. Asymptotic power investigations and adaptivity properties are studied in Sect. 3, where the reparametrized minimax set-up is introduced. It is shown that our procedure is sharply asymptotically adaptive with respect to sup-norm $\|\cdot\|$ on isotropic Hölder classes $\mathcal{F}$, i.e. minimax efficient over a broad range of Hölder smoothness classes simultaneously. The application to local classification is discussed in Sect. 4. The one-dimensional situation is considered separately in Sect. 5 where an alternative approach based on local pooled order statistics is proposed. In that case the statistic does not depend on the observations explicitly but only on their order which in contrast to nearest-neighbor relations is invariant under the quantile transformation. Section 6.1 is concerned with a decoupling inequality and the coupling exponential bounds which are essential for our construction. Both results are of independent theoretical interest. All proofs and auxiliary results about empirical processes are deferred to Sects. 6.2 and 6.3.

2 Combining randomized nearest-neighbor statistics

The procedure below is mainly designed for dimension $d \geq 2$. The univariate case contains a few special features and is considered separately in Sect. 5. Let $\underline{X} := (X_1, \ldots, X_n)'$ and denote by $\mathcal{X}_n$ the pooled set of observations. For any $0 \leq k \leq n-1$, the k'th nearest-neighbor of $X \in \mathcal{X}_n$ with respect to the *Euclidean distance* is denoted by X^k; we define $X^0 := X$. Note that the nearest-neighbors are unique a.s. The

weighted labels are defined as follows

$$\Lambda(X) := \begin{cases} \frac{n}{m} & \text{if } X \text{ is contained in the first sample} \\ -\frac{n}{n-m} & \text{otherwise.} \end{cases}$$

In order to judge about some possible deviation of p from q on a given set $B \in \mathcal{B}^d$, a natural statistic to look at is a standardized version of $\widehat{\mathbb{P}}_n(B) - \widehat{\mathbb{Q}}_n(B)$ or more sophisticated,

$$\int_B k_B(x) \left(d\widehat{\mathbb{P}}_n(x) - d\widehat{\mathbb{Q}}_n(x) \right)$$

for some kernel k_B supported by B, where $\widehat{\mathbb{P}}_n := m^{-1} \sum_{i=1}^m \delta_{X_i}$ and $\widehat{\mathbb{Q}}_n := (n-m)^{-1} \sum_{j=m+1}^n \delta_{X_j}$ denote the empirical measures corresponding to the first and second sample, respectively. Note that the statistic is not distribution-free, and in order to build up a multiple testing procedure several statistics corresponding to different sets B have to be combined in a certain way.

2.1 Local nearest-neighbor statistics

Let $\psi : [0, \infty) \to \mathbb{R}$ denote any kernel of bounded total variation with $\max_{x \in [0,\infty)} |\psi(x)| = \psi(0) = 1$ and $\psi(x) = 0$ for $x > 1$. We introduce the local test statistics

$$\begin{aligned} T_{jkn} &:= \frac{\sqrt{(m/n)(1-m/n)}}{\gamma_{jkn}} \frac{1}{\sqrt{n}} \sum_{i=0}^{k} \psi\left(\frac{\|X_j - X_j^i\|_2}{\|X_j - X_j^k\|_2} \right) \Lambda(X_j^i) \\ &= \frac{\sqrt{(m/n)(1-m/n)}}{\gamma_{jkn}} \sqrt{n} \int \psi\left(\frac{\|X_j - x\|_2}{\|X_j - X_j^k\|_2} \right) \left(d\widehat{\mathbb{P}}_n(x) - d\widehat{\mathbb{Q}}_n(x) \right), \end{aligned}$$

where

$$\gamma_{jkn}{}^2 := \frac{1}{n-1} \sum_{i=0}^{n-1} \left[\psi\left(\frac{\|X_j - X_j^i\|_2}{\|X_j - X_j^k\|_2} \right) - \frac{1}{n} \sum_{l=0}^{n-1} \psi\left(\frac{\|X_j - X_j^l\|_2}{\|X_j - X_j^k\|_2} \right) \right]^2 .$$

Every T_{jkn} is some in a certain sense standardized weighted average of the nearest-neighbor's labels and its absolute value should tend to be large whenever p is clearly larger or smaller than q within the random Euclidean ball with center X_j and radius $\|X_j - X_j^k\|_2$.

2.2 Adjustment for multiple testing

The idea is to build up a multiple test, combining all possible local statistics T_{jkn}. The typical way is to consider the distribution of the supremum $\sup_{j,k} T_{jkn}$, see, e.g.

Gijbels and Heckmann [15]. The problem is that the distribution is driven by small scales with a corresponding loss of power at larger scales, as there are many more small scales which contribute to the supremum. Here, we aim at a supremum type test statistic

$$T_n := \sup_{0<k\le n-1} \sup_{1\le j\le n} \left\{ |T_{jkn}| - C_{jkn} \right\},$$

where the constants C_{jkn} are appropriately chosen correction terms (independent of the label vector Λ) for adjustment of multiple testing within every "scale" k of k-nearest-neighbor statistics. These correction terms in the calibration aim to treat all the scales roughly equally. Although the distribution of T_n under the null hypothesis depends on the unknown underlying distribution $p = q$, the conditional distribution $\mathcal{L}_0(T_n|\mathcal{X}_n)$ of the above statistic is invariant under permutation of the components of the label vector $\underline{\Lambda}$. Here and subsequently, the index "0" indicates the null hypothesis, i.e. any couple (p, q) with $p = q$. Precisely, let the random variable Π be uniformly distributed on the symmetric group $\mathcal{S}_n$ of order n, independent of $\underline{X}$. Then $\mathcal{L}_0\left(T_n \,\middle|\, \mathcal{X}_n\right) = \mathcal{L}\left(T_n \circ \Pi \,\middle|\, \mathcal{X}_n\right)$, where $(T_n \circ \Pi)(\underline{\Lambda}) := T_n\left(\Lambda_{\Pi_1}, \ldots, \Lambda_{\Pi_n}\right)$. Elementary calculation entails that

$$\mathbb{E}\left(T_{jkn} \circ \Pi \,\middle|\, \mathcal{X}_n\right) = 0 \quad \text{and Var}\left(T_{jkn} \circ \Pi \,\middle|\, \mathcal{X}_n\right) = 1.$$

Thus the null hypothesis is satisfied if, and only if, the hypothesis of permutation invariance (or complete randomness) conditional on $\mathcal{X}_n$ is satisfied.

An adequate calibration of the randomized nearest-neighbor statistics, i.e. the choice of smallest possible constants C_{jkn}, requires both, an exact understanding of their tail behavior and their dependency structure. Note that the randomized nearest-neighbor statistics have a geometrically involved dependency structure. Even in case of the rectangular kernel ψ it depends explicitly on the "random design" $\mathcal{X}_n$ which incomplicates the sharp-optimal calibration for multiple testing compared to univariate problems, where the dependency of the single test statistics remains typically invariant under monotone transformation of the design points. Also, the optimal correction originally designed for Gaussian tails in Dümbgen and Spokoiny [10] does not carry over as only the subsequent Bernstein type exponential tail bound is available.

A coupling exponential inequality Based on an explicit coupling, the following proposition extends and tightens the exponential bounds derived in Serfling [34] for a combinatorial process in the present framework. If not stated otherwise, the random variable Π is uniformly distributed on $\mathcal{S}_n$, independent of $\underline{X}$.

Proposition 1 *Let T_{jkn} be as introduced above and define*

$$\delta(m, n) := \left(\mathbb{E} \min\left(\frac{S}{m}, \frac{n-S}{n-m} \right) \right)^{-1} \quad \textit{with} \quad S \sim \text{Bin}\,(n, m/n).$$

Then

$$\mathbb{P}\left(\left|\,T_{jkn}\circ\Pi\,\right| > \delta(m,n)\eta\,\middle|\,\mathcal{X}_n\right) \le 2\,\exp\left(-\frac{\eta^2/2}{1+\eta\,n^{-1/2}\gamma_{jkn}^{-1}\,R_\psi(m,n)}\right),$$

where

$$R_\psi(m,n) := \frac{2\|\psi\|_{\sup}}{3}\frac{\max(m,n-m)}{\sqrt{m(n-m)}}.$$

Remark The expression $\delta(m,n)$ is the payment for decoupling which appears by replacing the tail probability of a hypergeometric ensemble by that of the Binomial analogon. For details we refer to Sect. 6.1. In the typical case $0 < \liminf_n(m/n) \le \limsup_n(m/n) < 1$ we obtain $\delta(m,n) = 1 + O(n^{-1/2})$. Compared to results obtained for weighted averages of standardized, independent Bernoullis, the above Bernstein type appears to be nearly optimal, i.e. subgaussian tail behavior (with leading constant $1/2$) is actually not present.

Via inversion of the above exponential inequality, additive correction terms C_{jkn} for adjustment of multiple testing are constructed. The next theorem motivates our approach. The construction is designed for typical arrangements of the observation values which appear with probability close to one. To avoid technical expenditure, we restrict our attention to compactly supported densities. d_w denotes the dual bounded Lipschitz metric (see, e.g. [37]) which generates the topology of weak convergence. "$\to_{\mathbb{P}_n}$" refers to convergence in probability along the sequence of distributions $(\mathbb{P}_n)$.

Theorem 2 *Define the test statistic*

$$T_n := \sup_{\substack{1\le j\le n\\ 0<k\le n-1}} \left\{|T_{jkn}| - C_{jkn}\right\}$$

with

$$C_{jkn} := 3\,R_n\gamma_{jkn}^{-1}\delta(m,n)\Gamma_{jkn} + \delta(m,n)\sqrt{2\,\Gamma_{jkn}},$$

where $R_n = n^{-1/2}R_\psi(m,n)$ *and* $\Gamma_{jkn} := \log\left(1/\gamma_{jkn}{}^2\right)$. *Assume that the sequence of mixed densities* $h_n := (m/n)\,p_n + (1-m/n)q_n$ *on* $[0,1]^d$ *is equicontinuous and uniformly bounded away from zero, while* $0 < \liminf_n m/n \le \limsup_n m/n < 1$. *Let* $\mathbb{P}_n$ *and* $\mathbb{Q}_n$ *denote the probability measures corresponding to the densities* p_n *and* q_n, *respectively.Then the sequence* $\mathcal{L}\left(T_n\circ\Pi\,\middle|\,\mathcal{X}_n\right)$ *of conditional distributions is tight in* $\left(\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}\right)$*-probability. Additionally,*

$$d_w\left(\mathcal{L}\left(T_n\circ\Pi\,\middle|\,\mathcal{X}_n\right),\ \mathcal{L}(T_{\mathbb{H}_n})\right) \longrightarrow_{\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}} 0,$$

where

$$T_{\mathbb{H}_n} := \sup_{\substack{t\in[0,1]^d,\\ 0<r\le \max_{x\in[0,1]^d}\|x-t\|_2}} \left\{ \frac{\left| \int_{[0,1]^d} \phi_{rt,n}(x)\, dW(x) \right|}{\gamma_{rt,n}} - \sqrt{2\log\left(1/{\gamma_{rt,n}}^2\right)} \right\}$$

with W a standard Brownian sheet in $[0,1]^d$, $\gamma_{rt,n} := \left(\int_{[0,1]^d} \phi_{rt,n}(x)^2 dx\right)^{1/2}$ *and*

$$\phi_{rt,n}(x) := \left[\psi\left(\frac{\|x-t\|_2}{r}\right) - \int\limits_{[0,1]^d} \psi\left(\frac{\|z-t\|_2}{r}\right) h_n(z) dz \right] \sqrt{h_n(x)}.$$

The extra-term $3\, R_n \gamma_{jkn}^{-1} \delta(m,n) \Gamma_{jkn}$ in the constant C_{jkn} results from the exponential inequality in Proposition 1 and can be viewed as an additional penalty for non-subgaussianity. The theorem entails in particular that the sequence $\mathcal{L}(T_n \circ \Pi \,|\mathcal{X}_n)$ is weakly approximated in probability by a tight sequence of *non-degenerate* distributions $\mathcal{L}(T_{\mathbb{H}_n})$ which indicates that our corrections C_{jkn} are appropriately defined and cannot be chosen essentially smaller. Note that the approximation $\mathcal{L}(T_{\mathbb{H}_n})$ depends on the unknown mixed distribution even under the null hypothesis.

2.3 The multiple rerandomization test

Let $\kappa_\alpha(\underline{\mathrm{X}}) := \operatorname{argmin}_{C>0} \{\mathbb{P}(T_n \circ \Pi \le C \,|\, \mathcal{X}_n) \ge 1-\alpha\}$ denote the generalized $(1-\alpha)$-quantile of $\mathcal{L}\left(T_n \circ \Pi \,\middle|\, \mathcal{X}_n\right)$. Then we propose the conditional test

$$\phi_\alpha(\underline{\mathrm{X}}) := \begin{cases} 0 & \text{if } T_n \le \kappa_\alpha(\underline{\mathrm{X}}) \\ 1 & \text{if } T_n > \kappa_\alpha(\underline{\mathrm{X}}). \end{cases}$$

Our method can be viewed as a multiple testing procedure. For a given set of observations $\{X_1, \ldots, X_n\}$, the corresponding test statistic exceeds the $(1-\alpha)$-quantile if, and only if, the random set

$$\mathcal{D}_\alpha := \left\{ B_{X_j}\left(\left\| X_j^k - X_j \right\|_2 \right) \,\middle|\, 1 \le j \le n,\ 0 < k \le n-1;\ T_{jkn}(\underline{\mathrm{X}}) > C_{jkn}(\underline{\mathrm{X}}) + \kappa_\alpha(\underline{\mathrm{X}}) \right\}$$

is nonempty, where $B_t(r)$ denotes the Euclidean ball in $\mathbb{R}^d$ with center t and radius r. Since the test is valid conditional on the set of observations, we may conclude that p deviates from q at significance level α on *every* Euclidean ball $B_t(r) \in \mathcal{D}_\alpha$. In order to reduce the computational expenditure and to increase sensitivity on smaller scales, one may restrict one's attention to pairs (j,k) for $k \le m$ for some integer $m \in \{1, \ldots, n-1\}$. Note the validity of the test does not require any assumption about the densities.

Recently, Walther [38] proposed a multiple test for cluster analysis in two dimensions based on a suitable combination of local log-likelihood ratio statistics, evaluated on a fixed choice of axis-parallel rectangles. These statistics are not linear in $\widehat{\mathbb{P}}_n$ and $\widehat{\mathbb{Q}}_n$, respectively, but result in a subgaussian tail-behavior.

3 Asymptotic power

3.1 Minimax-efficiency and spatial adaptivity: local alternatives I

In this section, we show that the above introduced multiple testing procedure possesses optimality properties in a certain minimax sense. Nonparametric comparison of different samples was recently investigated in the minimax approach by Butucea and Tribouley [4], in a rate-adaptive way and of a different sense from our results here. We focus mainly on the considerably more involved problem of efficient adaptivity. Let us first introduce some notation. For any set $J \subset [0,1]^d$ and function f from $[0,1]^d \to \mathbb{R}$, $\|f\|_J := \sup_{x\in J} |f(x)|$. For any convex $I \subset \mathbb{R}^d$ let $\mathcal{H}_d(\beta, L; I)$ denote the isotropic Hölder smoothness class, which for $\beta \leq 1$ equals

$$\mathcal{H}_d(\beta, L; I) := \Big\{\phi : I \to \mathbb{R} : \ \big|\phi(x) - \phi(y)\big| \ \leq L\|x-y\|_2^\beta\Big\}.$$

Let $\lfloor\beta\rfloor$ denote the largest integer strictly smaller than β. For $\beta > 1$, $\mathcal{H}_d(\beta, L; I)$ consists of all functions $f : I \to \mathbb{R}$ that are $\lfloor\beta\rfloor$ times continuously differentiable such that the following property is satisfied: if $P_y^{(f)}$ denotes the Taylor polynomial of f at the point $y \in I$ up to the $\lfloor\beta\rfloor$'th order,

$$\Big|\, f(x) - P_y^{(f)}(x)\Big| \ \leq L\|x-y\|_2^\beta \quad \text{for all } x, y \in I.$$

In particular the definition entails that $f \in \mathcal{H}_d(\beta, L; \mathbb{R}^d)$ implies $f \circ U \in \mathcal{H}_d(\beta, L; \mathbb{R}^d)$ for every orthonormal transformation $U \in \mathbb{R}^{d\times d}$. For any pair of densities (p, q) on $[0,1]^d$, let $h(m, n, p, q)$ denote the corresponding mixed density $(m/n)p + (1 - m/n)q$. Fix a continuous density $h > 0$ and define $\mathcal{F}_h^{(m,n)}(\beta, L)$ to be the set of pairs of densities such that

$$\phi(m, n, p, q) := \frac{p-q}{\sqrt{h(m, n, p, q)}} \in \mathcal{H}_d\left(\beta, L; [0,1]^d\right) \quad \text{and} \quad h(m, n, p, q) = h.$$

Reparametrizing the composite hypothesis With the notation above,

$$p = h \cdot \left(1 + (1 - m/n)\,\phi/\sqrt{h}\right) \quad \text{and} \quad q = h \cdot \left(1 - (m/n)\,\phi/\sqrt{h}\right).$$

Consequently "$p = q$" is equivalent to "$\phi \equiv 0$", and if $(m/n)p + (1 - m/n)q = h$ is kept fixed, the composite hypothesis "$p = q$" reduces to the simple hypothesis "$\phi \equiv 0$". In order to develop a meaningful notion of minimax-efficiency for the two-sample problem we treat subsequently the mixed density $h = h(m, n, p, q)$ as fixed

but unknown infinite dimensional nuisance parameter for testing the hypothesis

$$H_0 : \phi = 0 \quad \text{versus} \quad H_A : \phi \neq 0.$$

Note that in case that h is uniformly bounded away from zero and p is close to q, ϕ coincides approximately with the difference $2\left(\sqrt{p} - \sqrt{q}\right)$, see also the explanation subsequent to Theorem 3.

Remark It is worth being noticed that the optimal statistic for testing H_0 against any fixed alternative ϕ equals the likelihood ratio statistic

$$\frac{d\mathbb{P}_{(m,n,p,q)}}{d\mathbb{P}_{(m,n,h,h)}}(\underline{X}) = \prod_{i=1}^{m}\left(1 + (1-m/n)\frac{\phi}{\sqrt{h}}(X_i)\right)\prod_{j=m+1}^{n}\left(1 - (m/n)\frac{\phi}{\sqrt{h}}(X_j)\right),$$

whose distribution still depends on h under the null. Here and subsequently, the subscript (m, n, p, q) indicates the distribution with density $\prod_{i=1}^{m} p \prod_{i=m+1}^{n} q$. The rational behind the reparametrization is to eliminate the dependency on the nuisance parameter h in the expectation under the null of the first and second order term of the log-likelihood expansion, resulting in asymptotic independence of h for its distribution under the hypothesis for any local sequence (ϕ_n).

The subsequent theorem is about the lower bound of hypothesis testing within the above defined classes of densities.

Theorem 3 (Minimax lower bound) *Let*

$$\rho_{m,n} := \left(\frac{n \log n}{m(n-m)}\right)^{\beta/(2\beta+d)} \quad \text{and define } c(\beta, L) := \left(\frac{2\,d\,L^{d/\beta}}{(2\beta+d)\|\gamma_\beta\|_2^2}\right)^{\beta/(2\beta+d)},$$

where γ_β defines the solution to the optimal recovery problem (2) below. Assume that the sequence of mixed densities (h_n) on $[0, 1]^d$ is equicontinuous and uniformly bounded away from zero. Then for any fixed $\delta > 0$ and every nondegenerate rectangle $J \subset [0, 1]^d$,

$$\limsup_{n\to\infty} \inf_{\substack{(p,q)\in\mathcal{F}_{h_n}^{(m,n)}(\beta,L):\\ \|\phi\|_J \geq (1-\delta)c(\beta,L)\rho_{m,n}}} \mathbb{E}_{(m,n,p,q)}\,\psi_n \;\leq\; \alpha$$

for arbitrary tests ψ_n at significance level $\leq \alpha$.

Note that ψ_n may depend on (β, L) and even on the nuisance parameter h_n as already does the Neyman–Pearson test for testing H_0 against any one-point alternative.

We now turn to the investigation of the test introduced in Sect. 2. To motivate the choice of an optimal kernel for our test statistics and its relation to the optimal recovery problem, let us restrict our consideration to the Gaussian white noise context, leading

in case of univariate Hölder continuous densities on $[0, 1]$ with $\beta > 1/2$ to locally asymptotically equivalent experiments

$$dX_{1n}(t) = p_n(t)\,dt + \frac{\sqrt{h_n(t)}}{\sqrt{m}}\,dW_1(t) \quad \text{and} \quad dX_{2n}(t) = q_n(t)\,dt + \frac{\sqrt{h_n(t)}}{\sqrt{(n-m)}}\,dW_2(t)$$

for two independent Brownian motions W_1 and W_2 on the unit interval ([28], Theorem 2.7 with $f_0 = h_n$ and Remark 2.8). Actually, h_n is identifiable using the semimartingal quadratic variation of the processes X_{1n} and X_{2n}, respectively. A multiscale statistic built from standardized differences of kernel estimates

$$\frac{\sqrt{(m/n)(1-m/n)}}{\|\psi\sqrt{h_n}\|_2} \int \psi(t)\,(dX_{1n}(t) - dX_{2n}(t))$$

then yields a distribution under the null close to ours in Theorem 2, up to the fact that our local integrals in dimension one are taken with respect to a Brownian bridge, reformulated to a Wiener process integrator by change of the integrand. Concerning the optimization of ψ, the quantity to be maximized within this Gaussian white noise context appears to be the expectation of the single test statistics under the least favorable alternatives as their variances do not depend on the mean. In case $h_n \equiv 1$ this expression equals

$$\inf_{\substack{\phi \in \mathcal{H}_1(\beta, L; [0,1]):\\ \|\phi\|_J \geq \delta}} \frac{\int \phi(t)\,\psi(t)\,dt}{\|\psi\|_2},$$

leading to the dual representation of the optimal recovery problem (see [5]).

The optimal recovery problem in higher dimension In the framework of isotropic Hölder balls, the optimal recovery problem leads to the solution $\gamma = \gamma_\beta$ of the optimization problem

$$\text{Minimize } \|\gamma\|_2 \quad \text{over all } \gamma \in \mathcal{H}_d\left(\beta, 1; \mathbb{R}^d\right) \quad \text{with } \gamma(0) \geq 1. \tag{2}$$

The closedness of $\mathcal{H}_d(\beta, 1; \mathbb{R}^d) \cap \left\{\gamma : \mathbb{R}^d \to \mathbb{R} \,\middle|\, \gamma(0) \geq 1\right\}$ in L_2 entails that the solution exists, its convexity implies furthermore uniqueness whence by isotropy of the functional class $\mathcal{H}_d(\beta, 1; \mathbb{R}^d)$ it must be radially symmetric. In case $\beta \leq 1$, one easily verifies that $\gamma_\beta(x) = \psi_\beta(\|x\|_2) = \left(1 - \|x\|_2^\beta\right)_+$. In its generality, the optimal recovery problem in higher dimension has not yet been investigated. Considering the partial derivatives of γ_β along the coordinate axes entails that ψ_β is necessarily contained in $\mathcal{H}_1(\beta, L; \mathbb{R})$. However, the transferred optimization problem

$$\text{minimize} \int \psi(r)^2 |r|^{d-1} dr \quad \text{over all } \psi \text{ with } \psi(\|.\|_2) \in \mathcal{H}_d(\beta, 1; \mathbb{R}) \quad \text{and } \psi(0) \geq 1 \tag{3}$$

does not coincide with the univariate optimal recovery problem due to the additional weighting by $|r|^{d-1}$ which comes into play by polar coordinate transformation. Whether the solution of (3) for $\beta > 1$ is compactly supported or not is still open. For the case of univariate densities, it is known that the solution of the optimal recovery problem has compact support for any $\beta > 0$ [23], but an explicit solution in case $\beta > 1$ is known for $\beta = 2$ only. Concerning details and advice on its construction, see [6] and [24]. For dimension $d > 1$, see [19].

The next Theorem is about the asymptotic power of the multiple test developed in Sect. 2. We restrict our attention to compact rectangles of $(0, 1)^d$ to avoid boundary effects. This restriction may be relaxed by the use of suitable boundary kernels, extending those of [25] for the univariate regression case to higher dimension.

Theorem 4 (Adaptivity and minimax efficiency) *Let $\phi^*_{n,\alpha}$ denote the multiple rerandomization test at significance level α, based on the kernel $\psi_\beta I\{\cdot \geq 0\}$ rescaled to* $[0, 1]$. *In case of unbounded support of ψ_β, we may use a truncated solution $\psi_{\beta,K} = \psi_\beta I\{0 \leq \cdot \leq K\}$. Let $0 < \liminf_n m/n \leq \limsup_n m/n < 1$. Assume that (h_n) is equicontinuous and uniformly bounded away from zero. Then for any fixed $\delta > 0$, there exists a $K > 0$ such that*

$$\liminf_{n\to\infty} \inf_{\substack{(p,q)\in\mathcal{F}^{(m,n)}_{h_n}(\beta,L):\\ \|\phi\|_J \geq (1+\delta)c(\beta,L)\rho_{m,n}}} \mathbb{P}_{(m,n,p,q)}\left(\phi^*_{n,\alpha} = 1\right) = 1$$

for any nondegenerate compact rectangle $J \subset (0, 1)^d$.

In particular, the test is sharp-optimal adaptive with respect to the second Hölder parameter L. While in view of the results in [17] the optimal rate of testing may be expected, some technical effort had to be done to propose a calibration achieving even sharp minimax-optimality.

Remark It is worth being noticed that the procedure achieves the upper bound uniformly over a large class of possible mixed densities. The intrinsic reason is that conditioning on $\mathcal{X}_n$ is actually equivalent to conditioning on $\widehat{\mathbb{H}}_n$, which indeed is a sufficient and complete statistic for the nuisance functional $\mathbb{H}_n$.

Remark (Sharp adaptivity with respect to β and L) Our construction, including the procedure especially designed for the one-dimensional situation, involves one kernel, shifted and rescaled depending on location and volume of the nearest-neighbor cluster under consideration. Due to the dependency of the optimal recovery solution γ_β on β, the corresponding test statistic $T_n = T_n(\beta)$ achieves sharp adaptivity with respect to the second Hölder parameter L only. Taking in addition the supremum $T^*_n := \sup_{\beta\in[\beta_0,\beta_1]} T_n(\beta)$ over all kernels γ_β within a compact range $[\beta_0, \beta_1] \subset (0, \infty)$, sharp adaptivity with respect to both Hölder parameters may be attained, provided that the above supremum statistic still defines a tight sequence (in probability), i.e. the corresponding sequence of $1 - \alpha$-quantiles $\kappa^*_\alpha(\underline{X})$ is stochastically bounded. Then the convergence

$$\mathbb{P}_{(m,n,p_n,q_n)}\left(T_n^* > \kappa_\alpha^*(\underline{X})\right) \geq \mathbb{P}_{(m,n,p_n,q_n)}\left(T_{\widehat{j}_n\widehat{k}_n n}(\beta) - C_{\widehat{j}_n\widehat{k}_n n}(\beta) > \kappa_\alpha^*(\underline{X})\right) \to 1 \quad \text{as } n \to \infty$$

for a suitably chosen random couple $(\widehat{j}_n, \widehat{k}_n)$ and any choice of β could be be extracted from the proof of Theorem 4. At least for $\beta \in [\beta_0, 1]$ this tightness may be deduced from the fact that the unimodal and symmetric $\psi_\beta(\|.\|_2)$ depends continuously on β in the sup-norm—in particular $\mathcal{L}\left(\left(\int \phi_{rt}^{(\beta)}(x)dW(x)\right)_{(t,r)}\right)$ as defined in Theorem 2 with $\psi = \psi_\beta$ depends continuously on β in the topology of weak convergence. A general investigation especially for $\beta > 1$ is beyond the scope of this article.

The next theorem shows however that our procedure simply based on the rectangular kernel is rate-adaptive with respect to both Hölder parameters (β, L). Due to the fact that it combines locally all nearest-neighbor scales at the same time, it even adapts to inhomogeneous smoothness of $p - q$, i.e. achieves *spatial adaptivity*.

Theorem 5 (Spatial rate-optimality) *Let $\phi_{n,\alpha}^*$ denote the multiple rerandomization test based on the rectangular kernel. Assume that $0 < \liminf_n m/n \leq \limsup_n m/n < 1$. Then for any fixed $k \in \mathbb{N}$ and parameters $(\beta_1, \ldots, \beta_k, L_1, \ldots, L_k)$, $K > 0$ and any collection of disjoint compact rectangles $J_i \subset [0,1]^d$, $i = 1, \ldots, k$, there exist constants $d_i = d(\beta_i, L_i, K)$ with*

$$\liminf_{n\to\infty} \inf_{\substack{(p,q):\\ (p-q)_{|J_i} \in \mathcal{H}_d(\beta_i, L_i; J_i)\\ \|p-q\|_{J_i} \geq d_i\, \rho_{m,n}(\beta_i),\\ h(m,n,p,q)_{|J_i} \geq K}} \mathbb{P}_{(m,n,p,q)}\left(J_i \cap \mathcal{D}_\alpha(\mathcal{X}_n) \neq \emptyset \; \forall\, i = 1, \ldots, k\right) = 1.$$

3.2 The stylized type of locally constant alternatives on small and large scales: local alternatives II

The results from the previous paragraph deal with small scales of different (arbitrary) order depending on the smoothness classes under consideration. In particular, the minimax lower bound is concerned with scales tending to zero as $m, n \to \infty$, and it is not yet clear that there is no substantial loss at rather large scales. The size of possible deviation $\|\phi\|_{\sup}$ and the scale (here $\sim (\|\phi\|_{\sup}/L)^{1/\beta}$) are linked in a specific way depending on the smoothness class under consideration, because the smoothness assumptions do not allow for arbitrarily fast decay to zero. The next theorem is different in spirit. We do not focus on smoothness classes but on stylized situations with ϕ being lower bounded by a "plateau" of absolute value $c_n/\sqrt{n\delta_n^d}$ within a ball $B_x(\delta_n)$. With λ denoting the Lebesgue measure on $[0,1]^d$, define

$$\mathcal{J}_+^{(m,n)}(c, x, \delta) := \Big\{ p, q \text{ } \lambda\text{-densities on } [0,1]^d : \phi(m,n,p,q)(z) \geq \frac{c}{\sqrt{n\delta^d}} \; \forall\, z \in B_x(\delta),\; 0 < c \leq \sqrt{n\delta^d} \Big\},$$

$$\mathcal{J}_-^{(m,n)}(c,x,\delta) := \Big\{ p,q \ \lambda\text{-densities on } [0,1]^d : \ \phi(m,n,p,q)(z) \\ \leq \ \frac{-c}{\sqrt{n\delta^d}} \ \forall z \in B_x(\delta), \ 0 < c \leq \sqrt{n\delta^d} \Big\}$$

and

$$\mathcal{G}^{(m,n)}(c,x,\delta) := \mathcal{J}_+^{(m,n)}(c,x,\delta) \cup \mathcal{J}_-^{(m,n)}(c,x,\delta).$$

Theorem 6 *Assume that* $0 < \liminf_n m/n \leq \limsup_n m/n < 1$.

(i) *If* ψ_n *is any sequence of tests at significance level* $\alpha \in (0,1)$*, then*

$$\inf_{(p,q)\in\mathcal{G}^{(m,n)}(c_n,x,\delta_n)} \mathbb{E}_{(m,n,p,q)} \, \psi_n \ \to \ 1$$

implies that $n\delta_n^d \to \infty$ *and* $c_n \to \infty$.

(ii) *If* $\psi_{n,\alpha}^*$ *describes the multiple rerandomization test based on the rectangular kernel at significance level* $\alpha \in (0,1)$*,*

$$\inf_{\substack{(p,q)\in\mathcal{G}^{(m,n)}(c_n,x,\delta_n) \\ h(m,n,p,q)\geq K>0}} \mathbb{E}_{(m,n,p,q)} \, \psi_{n,\alpha}^* \ \to \ 1,$$

provided that $n\delta_n^d \to \infty$ *and* $\sqrt{\log(1/\delta_n)}/c_n \to 0$.

In particular our test is also consistent against local alternatives of the type $\kappa_n\phi/\sqrt{n}$ for $\kappa_n \to \infty$, $\phi \not\equiv 0$. Comparing (i) and (ii) demonstrates that the adaptive search for the location of deviations costs an additional logarithm of its inverse scale. One may read out of the proof that the restriction for the sequence (c_n) in (ii) can be slightly refined.

4 Application to classification

Suppose we are given an iid sample (X_i, Y_i), $i = 1, \ldots, n$, where the marginal distribution of X_i is assumed to be Lebesgue-continuous with density h on $\mathbb{R}^d$, and Y_i takes values in $\{0,1\}$ with

$$\mathbb{P}\left(Y_i = 1 \,\middle|\, X_i = x\right) = \rho(x).$$

Then $M := \sum_{i=1}^n Y_i \ \sim \ \text{Bin}\,(n,\lambda)$ with $\lambda := \int \rho(x)h(x)dx$. Assuming $\lambda \in (0,1)$ to be known, the question of local classification is to identify simultaneously subregions in $\mathbb{R}^d$ where ρ deviates significantly from λ which results in local testing the hypothesis

$$H_0 : \ \rho = \lambda \quad \text{versus} \quad H_A : \ \rho \neq \lambda.$$

Imitating our procedure introduced in Sect. 2, we may combine suitably standardized local weighted averages of labels, but the standardization differs due to the fact that the sum of (strictly) positive labels is random and not fixed, in particular $Y_1, \dots, Y_n$ are stochastically independent. Consequently, we may then rely the procedure on the classical Bernstein exponential inequality for weighted averages of standardized Bernoullis. Of course, the optimal separation constant for testing "$\rho = \lambda$" within some Euclidean ball $B_t(r)$ and its complement depends on the amount of observations in $B_t(r)$, whence analogously to the consideration above for the two-sample problem we may use the reparametrization of (ρ, h) to (ϕ, h) with

$$\phi := \frac{\rho - \lambda}{\lambda(1-\lambda)}\sqrt{h}.$$

The power optimality results carry over to the classification context with similar arguments as used in the proof of Theorem 4. We omit its explicit formulation at this point.

5 Distribution-freeness via quantile transformation: the case $d = 1$

The one-dimensional situation allows for an alternative and more elegant approach based on order relations. For let $X_{(1)}, \dots, X_{(n)}$ denote the order statistic built from the pooled sample and define for any $0 \le j < k \le n$ the local test statistics

$$U_{jkn} := \frac{\sqrt{(m/n)(1-m/n)}}{\eta_{jkn}} \frac{1}{\sqrt{n}} \sum_{i=j+1}^{k} \psi\left(\frac{i}{k-j}\right) \Lambda(X_{(i)}),$$

where

$$\eta_{jkn}{}^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(\psi\left(\frac{i-j}{k-j}\right) - \frac{1}{n} \sum_{l=1}^{n} \psi\left(\frac{l-j}{k-j}\right) \right)^2.$$

Compared to the procedure described in the previous section, we omit the explicit dependence of the weights on the observed values. Note that in contrast to nearest-neighbor relations, the order remains invariant under quantile transformation, i.e. $\mathrm{rank}(H_n(X_i)) = \mathrm{rank}(X_i)$, resulting in distribution-freeness of the corresponding multiscale statistic under the null. Suppose the null hypothesis is satisfied for some Lebesgue continuous distribution on the real line. Then conditional on the order statistics as well as unconditional, the label vector is uniformly distributed on the set

$$\left\{ \Lambda \in \{n/m, -n/(n-m)\}^n : \sum_{i=1}^{n} \Lambda_i^{-1} = 0 \right\}.$$

The described test statistics are local versions of classical Wilcoxon rank sum statistics. We omit any further investigation as the calibration for multiple testing can be

done analogously to that proved in Theorem 2—but keep in mind that the approximating Gaussian multiscale statistic under the null hypothesis will be independent of the nuisance functional $\mathbb{H}_n$ due to the quantile transformation. Note that the use of typical mathematical tools for power investigation of rank statistics like Hoeffding's decomposition is getting involved because the kernel ψ_β for $\beta \leq 1$ is not differentiable.

6 Proofs and further probabilistic results

6.1 Decoupling inequality and coupling exponential bounds

This section contains the coupling exponential bounds, i.e. in this context for weighted averages from a hypergeometric ensemble. Using a different technique, namely an explicit coupling construction, the subsequent proposition extends results of Hoeffding [16] on decoupling of expectations of convex functions in the arithmetic mean of a sample without replacement. Whereas in the latter case decoupling with constant 1 is actually correct, a simple counterexample for an ensemble of two elements already shows that the result does not extend to arbitrary weighted averages, and some payment for decoupling appears to be necessary.

Proposition 7 (Decoupling inequality) *Let $Z_1, Z_2, \ldots, Z_n$ be iid with*

$$\mathbb{P}(Z_i = 1) = \frac{m}{n} \quad \textit{and} \quad \mathbb{P}(Z_i = 0) = 1 - \frac{m}{n}, \quad 0 < m < n.$$

Let $a \in \mathbb{R}^n$ with $\sum_{i=1}^n a_i = 0$ and $\Psi : \mathbb{R} \to \mathbb{R}$ be convex. Then

$$\mathbb{E}\left(\Psi\left(\sum_{i=1}^n a_i Z_i \right) \,\middle|\, \sum_{i=1}^n Z_i = m \right) \leq \mathbb{E}\, \Psi\left(\delta(m,n) \sum_{i=1}^n a_i Z_i \right),$$

with

$$\delta(m,n)^{-1} := \mathbb{E} \min\left(\frac{S}{m}, \frac{n-S}{n-m} \right), \quad S \sim \mathrm{Bin}\left(n, \frac{m}{n}\right).$$

In particular, $\delta(m,n)^{-1} = 1 + O(n^{-1/2})$ for $m/n \to \lambda \in (0,1)$.

Proof Let X be uniformly distributed on the set

$$\left\{ x \in \{0,1\}^n : \sum_{i=1}^n x_i = m \right\}$$

and let $S \sim \mathrm{Bin}(n, m/n)$ such that X and S are independent. Define

$$M := \{i : X_i = 1\}.$$

Conditional on X and S, the random vector $Z \in \{0,1\}^n$ is constructed as follows:

If $S > m$, let $Z_i = 1$ for all $i \in M$ and let $(Z_i)_{i \in M^c}$ be uniformly distributed on the set

$$\left\{ z \in \{0,1\}^{M^c} : \sum_{i \in M^c} z_i = S - m \right\}.$$

For $S \le m$, let $Z_i = 0$ for all $i \in M^c$ and let $(Z_i)_{i \in M}$ be uniformly distributed on

$$\left\{ z \in \{0,1\}^{M} : \sum_{i \in M} z_i = S \right\}.$$

Note that $Z_1, \dots, Z_n$ are iid $\mathrm{Bin}(1, m/n)$. Then

$$\begin{aligned}
\mathbb{E}\,\Psi\left(\sum_{i=1}^n a_i Z_i\right) &= \mathbb{E}\mathbb{E}\left(\Psi\left(\sum_{i=1}^n a_i Z_i\right) \Big| X, S\right) \\
&\ge \mathbb{E}\,\Psi\left(\mathbb{E}\left(\sum_{i=1}^n a_i Z_i \Big| X, S\right)\right) \quad \text{(Jensen inequality)} \\
&= \mathbb{E}\,\Psi\left(I\{S \le m\}\frac{S}{m}\sum_{i \in M} a_i + I\{S > m\}\left(\sum_{i \in M} a_i + \frac{S-m}{n-m}\sum_{i \in M^c} a_i\right)\right) \\
&= \mathbb{E}\,\Psi\left(I\{S \le m\}\frac{S}{m}\sum_{i \in M} a_i \;+\; I\{S > m\}\frac{n-S}{n-m}\sum_{i \in M} a_i\right) \quad \text{(since } \textstyle\sum_{i=1}^n a_i = 0\text{)} \\
&= \mathbb{E}\,\Psi\left(\min\left(\frac{S}{m}, \frac{n-S}{n-m}\right)\sum_{i=1}^n a_i X_i\right) \\
&= \mathbb{E}\mathbb{E}\left[\Psi\left(\min\left(\frac{S}{m}, \frac{n-S}{n-m}\right)\sum_{i=1}^n a_i X_i\right) \Big| X\right] \\
&\ge \mathbb{E}\,\Psi\left(\mathbb{E}\left\{\min\left(\frac{S}{m}, \frac{n-S}{n-m}\right)\right\}\sum_{i=1}^n a_i X_i\right) \quad \text{(Jensen inequality).}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\,\min\left(\frac{S}{m}, \frac{n-S}{n-m}\right) &= 1 - \mathbb{E}\left(\frac{(S-m)_-}{m} + \frac{(S-m)_+}{n-m}\right) \\
&\ge 1 - \mathbb{E}\left(\frac{|S-m|}{\min(m, n-m)}\right) \\
&\ge 1 - \frac{\lambda(m,n)}{\sqrt{n}}
\end{aligned}$$

with $\lambda(m,n) := \sqrt{m(n-m)}\big/\min(m, n-m)$, which is uniformly bounded for $m/n \to \lambda \in (0,1)$. □

Using the decoupling above, the next proposition presents the exponential bounds for the combinatorial process which are essential for our construction. It implies Proposition 1 and improves in particular exponential tail bounds for the hypergeometric distribution of Serfling [34] in the coefficient in front of η^2 for m/n close to zero or one, moderate η and large n. Note that this coefficient is crucial for the efficiency of the testing procedure. The results may also be compared with the decoupling based exponential tail bounds in de la Peña [29,30].

Proposition 8 (Coupling exponential inequalities) *Let $Z_1, \ldots, Z_n$ be iid with*

$$\mathbb{P}(Z_i = 1) = \frac{m}{n} \quad \textit{and} \quad \mathbb{P}(Z_i = 0) = 1 - \frac{m}{n}, \quad 0 < m < n.$$

Let $\psi_1, \ldots, \psi_n$ real valued numbers with $\bar{\psi}$ its arithmetic mean and denote

$$\gamma_{m,n}{}^2 := \mathrm{Var}\left(\sum_{i=1}^n \psi_i Z_i \,\middle|\, \sum_{i=1}^n Z_i = m \right) = \frac{m(n-m)}{n(n-1)} \sum_{i=1}^n \left(\psi_i - \bar{\psi}\right)^2.$$

Then in case of $\gamma_{m,n} \neq 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left| \frac{1}{\gamma_{m,n}} \sum_{i=1}^n \psi_i \left(Z_i - \frac{m}{n} \right) \right| > \delta(m,n)\eta \,\middle|\, \sum_{i=1}^n Z_i = m \right) \\
&\leq 2 \exp\left(- \frac{\eta^2/2}{1 + \eta\, R(\psi, m, n)} \right) \\
&\leq 2 \exp\left(- \frac{3\eta}{2c(m,n)} + \frac{9}{2c(m,n)^2} \right),
\end{aligned}$$

where

$$\begin{aligned}
R(\psi, m, n) &:= \frac{\max_i |\psi_i - \bar{\psi}|}{3\,\gamma_{m,n}} \max\left(\frac{m}{n}, 1 - \frac{m}{n} \right) \quad \textit{and} \\
c(m,n) &:= \frac{\max(m, n-m)}{\sqrt{m(n-m)}}.
\end{aligned}$$

Proof With

$$M := \frac{\max_i |\psi_i - \bar{\psi}|}{\gamma_{m,n}} \max\left(\frac{m}{n}, 1 - \frac{m}{n} \right)$$

we obtain for any $t > 0$

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{\gamma_{m,n}}\sum_{i=1}^{n}\psi_i\left(Z_i-\frac{m}{n}\right) > \delta(m,n)\eta \,\middle|\, \sum_{i=1}^{n} Z_i = m\right)\\
&= \mathbb{P}\left(\frac{1}{\gamma_{m,n}}\sum_{i=1}^{n}(\psi_i-\bar{\psi})\left(Z_i-\frac{m}{n}\right) > \delta(m,n)\eta \,\middle|\, \sum_{i=1}^{n} Z_i = m\right)\\
&\le \exp\left(-t\,\frac{\eta}{M}\right)\mathbb{E}\left\{\exp\left(\frac{t\,\delta(m,n)^{-1}}{M\,\gamma_{m,n}}\sum_{i=1}^{n}(\psi_i-\bar{\psi})\left(Z_i-\frac{m}{n}\right)\right)\,\middle|\, \sum_{i=1}^{n} Z_i = m\right\}\\
&\le \exp\left(-t\,\frac{\eta}{M}\right)\mathbb{E}\exp\left(\frac{t}{M\,\gamma_{m,n}}\sum_{i=1}^{n}(\psi_i-\bar{\psi})\left(Z_i-\frac{m}{n}\right)\right) \quad \text{(Proposition 7)}\\
&\le \exp\left(\frac{1}{M^2}\left(e^t-1-t\right)-t\,\frac{\eta}{M}\right), \qquad (4)
\end{aligned}$$

whereby the last inequality follows from the fact that for any random variable Y with $|Y| \le 1$, $\mathbb{E}Y = 0$ and $\mathrm{Var}(Y) = \sigma^2$,

$$\mathbb{E}\exp(tY) \le 1+\sigma^2(e^t-1-t) \le \exp\left(\sigma^2(e^t-1-t)\right).$$

Elementary algebra shows that (4) is minimized with the choice $t := \log(1+\eta M)$, which yields first a Bennett exponential bound (see [3]) and because of $(1+x)\log(1+x)-x \ge (x^2/2)/(1+x/3)$ consequently the Bernstein type

$$\mathbb{P}\left(\frac{1}{\gamma_{m,n}}\sum_{i=1}^{n}\psi_i\left(Z_i-\frac{m}{n}\right) > \delta(m,n)\eta \,\middle|\, \sum_{i=1}^{n} Z_i = m\right) \le \exp\left(-\frac{\eta^2/2}{1+\eta\,M/3}\right).$$

A symmetry argument provides the same bound for ψ_i replaced by $-\psi_i$, which completes the proof of the first inequality. Using that $\gamma_{m,n} \ge \sqrt{(m/n)(1-m/n)}\,\max_i|\psi_i - \bar{\psi}|$, we obtain the second asserted inequality from

$$\begin{aligned}
\frac{\eta^2/2}{1+\eta\,M/3} &\ge \frac{\eta^2/2}{1+\eta\,c(m,n)/3}\\
&= \frac{\eta}{2c(m,n)/3} - \frac{\eta}{2c(m,n)/3(1+\eta\,c(m,n)/3)}\\
&\ge \frac{\eta}{2c(m,n)/3} - \frac{1}{2c(m,n)^2/9}.
\end{aligned}$$

□

6.2 Auxiliary results about empirical processes

This section collects results in the context of empirical processes which are essential for the next section. For any totally-bounded pseudo-metric space $(\mathcal{T}, \rho)$, we define the covering number

$$N(\varepsilon, \mathcal{T}, \rho) := \min\left\{\sharp\mathcal{T}_0 : \mathcal{T}_0 \subset \mathcal{T}, \inf_{t_0 \in \mathcal{T}_0} \rho(t, t_0) \leq \varepsilon \text{ for all } t \in \mathcal{T}\right\}.$$

Let $\mathcal{B}(\mathcal{T})$ denote the Borel-σ-field on $\mathcal{T}$ induced by the pseudo-metric ρ (which induces a topology in the usual sense, although without the Hausdorff-property if it is not a metric) and let $\mathcal{F} \subset [0,1]^{\mathcal{T}}$ be a family of measurable functions. For any probability measure P on $\mathcal{B}(\mathcal{T})$, consider the pseudo-distance $d_P(f,g) := \int |f-g|\, dP$ for $f, g \in \mathcal{F}$. Then for any $u > 0$, the uniform covering numbers of $\mathcal{F}$ are defined as $\mathcal{N}(u, \mathcal{F}) := \sup_P N(u, \mathcal{F}, d_P)$, where the supremum is running over all probability measures P on $\mathcal{B}(\mathcal{T})$.

Theorem 9 ([11, technical report]) *Let $Z = (Z(t))_{t\in\mathcal{T}}$ be a stochastic process on a totally bounded pseudo-metric space $(\mathcal{T}, \rho)$. Let K be some positive constant, and for $\delta > 0$ let $G(\cdot, \delta)$ a nondecreasing function on $[0, \infty)$ such that for all $\eta \geq 0$ and $s, t \in \mathcal{T}$,*

$$\mathbb{P}\left\{\frac{|Z(s) - Z(t)|}{\rho(s,t)} > G(\eta, \delta)\right\} \leq K \exp(-\eta) \quad \text{if } \rho(s,t) \geq \delta. \tag{5}$$

Then for arbitrary $\delta > 0$ and $a \geq 1$,

$$\mathbb{P}\{|Z(s) - Z(t)| \geq 12 J(\rho(s,t), a) \text{ for some } s, t \in \mathcal{T}_* \text{ with } \rho(s,t) \leq \delta\} \leq \frac{K\delta}{2a},$$

where $\mathcal{T}_*$ is a dense subset of $\mathcal{T}$, and

$$J(\epsilon, a) := \int_0^{\epsilon} G(\log(a D(u)^2/u), u)\, du,$$

$$D(u) = D(u, \mathcal{T}, \rho) := \max\{\#\mathcal{T}_o : \mathcal{T}_o \subset \mathcal{T}, \rho(s,t) > u \text{ for different } s, t \in \mathcal{T}_o\}.$$

Remark Suppose that $G(\eta, \delta) = \widetilde{q}\, \eta^q$ for some constants $\widetilde{q}, q > 0$. In addition let $D(u) \leq A u^{-B}$ for $0 < u \leq 1$ with constants $A \geq 1$ and $B > 0$. Then elementary calculations show that for $0 < \epsilon \leq 1$ and $a \geq 1$, $J(\epsilon, a) \leq C\, \epsilon\, (\log(e/\epsilon))^q$ with $C = \widetilde{q}\, \max\left(1 + 2B, \log(aA^2)\right)^q \int_0^1 (\log(e/z))^q\, dz$.

For the proof of Theorem 2 the subsequent extension of the Chaining Lemma VII.9 in Pollard [31] and Theorem 8 in the technical report to Dmbgen and Walther [11] will be used. It complements in particular the existing multiscale theory by a uniform tightness result and to a situation where only a sufficiently sharp *uniform stochastic*

bound on local covering numbers is available, which typically involves additional logarithmic terms. The situation arises for example in the multivariate random design case where a non-stochastic bound obtained via uniform covering numbers and VC-theory may be too rough.

Theorem 10 (Chaining) *Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of random variables such that Y_n takes values in some Polish space $\mathcal{Y}_n$. For any $y_n \in \mathcal{Y}_n$, let $(Z_n(t; y_n))_{t\in\mathcal{T}_{y_n}}$ be a stochastic process on some countable, metric space $\big(\mathcal{T}_{y_n}, \rho_n(.,.; y_n)\big)$, where $\rho_n(.,.; y_n) \le 1$. Suppose that the following conditions are satisfied*:

(i) *There are measurable functions $\sigma_n(.; Y_n) : \mathcal{T}_{Y_n} \to (0, 1]$ and $G_n(., \delta) : [0, \infty) \to [0, \infty)$ such that for arbitrary $s, t \in \mathcal{T}_{Y_n}$, $\eta \ge 0$ and $\delta > 0$,*

$$\mathbb{P}\left(|Z_n(t, Y_n)| \ge \sigma_n(t; Y_n) G_n(\eta, \delta) \,\middle|\, Y_n\right) \le 2\exp(-\eta) \quad \text{if } \sigma_n(t; Y_n) \ge \delta,$$

$$\sup_{s,t\in\mathcal{T}_{Y_n}} \frac{|\sigma_n(t; Y_n) - \sigma_n(s; Y_n)|}{\rho_n(s, t; Y_n)} \le C < \infty \text{ for some constant } C > 0,$$

$$\big\{t \in \mathcal{T}_{Y_n} : \sigma_n(t; Y_n) \ge \delta\big\} \text{ is compact, and } G_o := \sup_{n\in\mathbb{N}} \sup_{\eta\ge 0, 0<\delta\le 1} \frac{G_n(\eta, \delta)}{1+\eta} < \infty.$$

(ii) *There exists a sequence $(\mathcal{C}_n)_{n\in\mathbb{N}}$ of measurable sets and positive constants A, B, W, α such that*

$$N\big(u\delta, \{t \in \mathcal{T}_{Y_n} : \sigma_n(t; Y_n) \le \delta\}, \rho_n(.,.; Y_n)\big) \le A u^{-B} \delta^{-W} \left[\log\left(e/(u\delta)\right)\right]^{\alpha} \quad \text{for } u,\ \delta \in (0, 1]$$

whenever $Y_n \in \mathcal{C}_n$.

For constants $q, Q > 0$ define

$$\mathcal{A}_n(\delta, q, Q; Y_n) := \left\{ \sup_{s,t\in\mathcal{T}_{Y_n}: \rho_n(s,t;Y_n)\le\delta} \frac{|Z_n(s; Y_n) - Z_n(t; Y_n)|}{\rho_n(s, t; Y_n)[\log(e/\rho_n(s, t; Y_n))]^q} \le Q \right\}.$$

Then there exists a constant $C = C(G_o, A, B, W, \alpha, q, Q) > 0$ such that for $0 < \delta \le 1$

$$\mathbb{P}\Bigg(\frac{|Z_n(t; Y_n)|}{\sigma_n(t; Y_n)} \le G_n\left(W \log\left(1/\sigma_n(t; Y_n)\right) + C \log\log\left(e/\sigma_n(t; Y_n)\right), \sigma_n(t; Y_n)\right) + C \Big/ \log(e/\sigma_n(t; Y_n)) \text{ on } \{t : \sigma_n(t; Y_n) \le \delta\} \,\Bigg|\, Y_n\Bigg)$$

is at least $\mathbb{P}\left(\mathcal{A}_n(2\delta, q, Q; Y_n) \,\middle|\, Y_n\right) - C \Big/ \log(e/\delta)$ whenever $Y_n \in \mathcal{C}_n$.

If in particular $\mathbb{P}^{Y_n}(\mathcal{C}_n) \to 1$ *and* $\lim_{\delta \searrow 0} \inf_n \mathbb{P}\left(\mathcal{A}_n(\delta, q, Q; Y_n) \,\middle|\, Y_n\right) = 1$ *a.s., then the sequence*

$$\mathcal{L}\Bigg(\sup_{t \in \mathcal{T}_{Y_n}} \Bigg\{ \frac{|Z_n(t; Y_n)|}{\sigma_n(t; Y_n)} - G_n\Big(W \log\big(1/\sigma_n(t; Y_n)\big) + C \log\log\big(e/\sigma_n(t; Y_n)\big),\ \sigma_n(t; Y_n)\Big)\Bigg\} \,\Bigg|\, Y_n\Bigg)$$

is tight in $\left(\mathbb{P}^{Y_n}\right)$*-probability, provided that* $\inf_n \sup_{t \in \mathcal{T}_{Y_n}} \sigma_n(t; Y_n) > 0$ *a.s.*

Remark Note that in case of $G(\eta, \delta) = (\kappa \eta)^{1/\kappa}$ with $\kappa > 1$, we obtain by the series expansion of $(1+z)^\alpha$ for $0 < \alpha < 1$ and $|z| < 1$

$$\begin{aligned} &G\left(W \log(1/\delta) + C \log\log(e/\delta), \delta\right) + C \Big/ \log(e/\delta) \\ &= (\kappa W \log(1/\delta))^{1/\kappa} + O\left(\log\log(e/\delta)[\log(e/\delta)]^{1/\kappa - 1}\right) \\ &= (\kappa W \log(1/\delta))^{1/\kappa} + o(1) \quad \text{as } \delta \searrow 0. \end{aligned}$$

Proof The proof of the first part follows in spirit that of Dümbgen and Walther [11], technical report, and is sketched in the extended version of this article.

Concerning the tightness in probability as stated in the second part of the theorem, notice that the result does not follow by an immediate continuity argument because the metric (and the metric space) change with both, Y_n and n, hence some additional uniformity is required. For $0 \le \delta < \delta' \le 1$ let $U_n(\delta, \delta'; Y_n)$ be defined by

$$\sup_{\substack{\sigma_n(t; Y_n) \in (\delta, \delta'] \\ t \in \mathcal{T}_n}} \Bigg\{ \frac{|Z_n(t; Y_n)|}{\sigma_n(t; Y_n)} - G_n(W \log(1/\sigma_n(t; Y_n)) + C \log\log(e/\sigma_n(t; Y_n)),\ \sigma_n(t; Y_n))\Bigg\}.$$

First observe that for any fixed $K > 0$,

$$\begin{aligned} \mathbb{P}\left(U_n(0, 1; Y_n) > K \,\middle|\, Y_n\right) \le\ & \mathbb{P}\left(U_n(0, \delta; Y_n) > K/2 \,\middle|\, Y_n\right) \\ & + \mathbb{P}\left(U_n(\delta, 1; Y_n) > K/2 \,\middle|\, Y_n\right). \end{aligned} \tag{6}$$

The first part of Theorem 10 implies that the first term on the right-hand-side in (6) is bounded by $1 - \mathbb{P}\left(\mathcal{A}_n(2\delta, q, Q; Y_n) \,\middle|\, Y_n\right) + C/\log(e/\delta)$ for $K > 2C/\log(e/\delta)$ whenever $Y_n \in \mathcal{C}_n$. Concerning the second term in (6), note that

$$U_n(\delta, 1; Y_n) \le -\inf_{\delta' \in [\delta, 1]} H_n(\delta'; Y_n) + \frac{1}{\delta} \sup_{\substack{t \in \mathcal{T}_{Y_n}: \\ \sigma_n(t; Y_n) \ge \delta}} \Big| Z_n(t; Y_n) \Big|.$$

Then the conclusion follows if we establish that

$$\lim_{K\to\infty}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{t\in\mathcal{T}_{Y_n}}\left|\,Z_n(t;Y_n)\,\right|>K;\ Y_n\in\mathcal{C}_n\,\middle|\,Y_n\right)=0\ \ a.s.$$

For $\varepsilon>0$ and $y_n\in\mathcal{C}_n$, let $t_1(y_n),\ldots,t_{m(y_n)}(y_n)$ be a maximal subset of $\mathcal{T}_{y_n}$ with $\rho_n(t_i,t_j;y_n)>\varepsilon$ for arbitrary different indices $i,j\in\{1,\ldots,m(y_n)\}$. Note that $m(y_n)\le A\varepsilon^{-B}[\log(e/\varepsilon)]^{\alpha}$ by assumption (ii). Then condition (i) implies that

$$\lim_{K\to\infty}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{i=1,\ldots,m(y_n)}\left|\,Z_n\left(t_i(y_n);Y_n\right)\,\right|>K\,\middle|\,Y_n=y_n\right)=0\ \ a.s. \tag{7}$$

On the other hand, we have on the set $\mathcal{A}_n(\varepsilon,q,Q;Y_n)$ the bound

$$\sup_{t\in\mathcal{T}_{Y_n}}|Z_n(t;Y_n)|\le Q\varepsilon[\log(e/\varepsilon)]^q+\sup_{i=1,\ldots,m(Y_n)}\left|\,Z_n(t_i(Y_n);Y_n)\,\right|. \tag{8}$$

With ε tending to zero sufficiently slowly, (7) and (8) show together with the stochastic equicontinuity condition $\lim_{\delta\searrow 0}\inf_n\mathbb{P}\left(\mathcal{A}_n(\delta,q,Q;Y_n)\,\middle|\,Y_n\right)=1$ a.s.

$$\lim_{K\to\infty}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{t\in\mathcal{T}_{y_n}}\left|\,Z_n(t;Y_n)\,\right|>K\,\middle|\,Y_n=y_n\right)=0\ \ a.s.$$

Since the assumption $\inf_n\sup_{t\in\mathcal{T}_{Y_n}}\sigma_n(t;Y_n)>0$ a.s. guarantees

$$\lim_{K\to\infty}\sup_n\mathbb{P}\left(U_n(Y_n)<-K\,\middle|\,Y_n\right)=0\ \ a.s.,$$

the tightness in $(\mathbb{P}^{Y_n})$-probability is proved. □

6.3 Proofs of the main results

Proof of Theorem 2 Let $\lambda_n:=m/n$. In view of the T_{jkn}'s, the behavior of the process

$$\left(\frac{\sqrt{\lambda_n(1-\lambda_n)}}{\sqrt{n}}\sum_{i=0}^{k}\psi\left(\frac{\left\|X_j-X_j^i\right\|_2}{\left\|X_j-X_j^k\right\|_2}\right)(\Lambda\circ\Pi)\left(X_j^i\right)\right)_{1\le j\le n,\,0<k\le n-1}$$

conditional on $\mathcal{X}_n$ needs to be investigated, where $\Lambda\circ\Pi|\mathcal{X}_n$ is uniformly distributed on the set

$$\left\{\lambda:\mathcal{X}_n\to\{1/\lambda_n,-1/(1-\lambda_n)\}:\sum_{x\in\mathcal{X}_n}\lambda(x)=0\right\}.$$

For notational convenience it seems useful to redefine the process on the random index set

$$\widehat{\mathcal{T}}_n := \left\{ \left(X_j, \left\| X_j - X_j^k \right\|_2 \right) : 1 \le j \le n,\ 0 < k \le n-1 \right\}$$

via the map $(j,k) \mapsto \left(X_j, \left\| X_j - X_j^k \right\|_2 \right)$ and extend it to a process $(Y_n(t,r))_{(t,r)\in\mathcal{T}}$ with $\mathcal{T} := \left\{ (t,r) : t \in [0,1]^d,\ 0 < r \le \max_{x\in[0,1]^d} \|x-t\|_2 \right\}$ by the definition

$$Y_n(t,r) := \sqrt{n}\sqrt{\lambda_n(1-\lambda_n)} \int \psi\left(\frac{\|t-x\|_2}{r} \right) \left(d\widehat{\mathbb{P}}_n^{\Pi}(x) - d\widehat{\mathbb{Q}}_n^{\Pi}(x) \right),$$

where $\widehat{\mathbb{P}}_n^{\Pi}$ and $\widehat{\mathbb{Q}}_n^{\Pi}$ denote the empirical measures based on the permutated variables $X_{\Pi(1)}, \ldots, X_{\Pi(m)}$ and $X_{\Pi(m+1)}, \ldots, X_{\Pi(n)}$, respectively. Let

$$\begin{aligned} \widehat{\gamma}_n(t,r)^2 &:= \operatorname{Var}\left(Y_n(t,r) \,\middle|\, \mathcal{X}_n \right) \\ &= \frac{n}{n-1} \int \left[\psi\left(\frac{\|t-x\|_2}{r} \right) - \int \psi\left(\frac{\|t-z\|_2}{r} \right) d\widehat{\mathbb{H}}_n(z) \right]^2 d\widehat{\mathbb{H}}_n(x), \end{aligned}$$

with $\widehat{\mathbb{H}}_n$ the empirical measure of the observations $X_1, \ldots, X_n$.

In the sequel we make use of the results in the previous section twice—in order to prove the tightness and weak approximation in probability of the sequence of conditional test statistics and within the "loop" we use the chaining arguments again to establish a sufficiently tightened uniform stochastic bound for the covering numbers below.

I. (Subexponential increments and Bernstein type tail behavior) The inversion of the conditional Bernstein type exponential inequality in Proposition 8 shows that for any $\eta > 0$,

$$\mathbb{P}\left(\left| \frac{Y_n(t,r)}{\widehat{\gamma}_n(t,r)} \right| > G_n\left(\eta, \widehat{\gamma}_n(t,r)\right) \,\middle|\, \mathcal{X}_n \right) \le 2\exp(-\eta),$$

where

$$G_n\left(\eta, \widehat{\gamma}_n(t,r)\right) := R_n\left(\widehat{\gamma}_n(t,r)\right)\eta + \left(\left(R_n\left(\widehat{\gamma}_n(t,r)\right)\eta\right)^2 + 2\delta(m,n)^2\eta \right)^{1/2}$$

with

$$R_n(\tau) := \delta(m,n) \frac{2\|\psi\|_{\sup}\sqrt{\lambda_n(1-\lambda_n)}}{3\,\min(\lambda_n, 1-\lambda_n)\sqrt{n}\,\tau}.$$

Let the random pseudo-metric $\widehat{\rho}_n$ on $\mathcal{T}$ be defined by

$$\begin{aligned}\widehat{\rho}_n\left((t,r),(t',r')\right)^2 &:= \operatorname{Var}\left(Y_n(t,r) - Y_n(t',r') \,\middle|\, \mathcal{X}_n\right)\\ &= \frac{n}{n-1}\Bigg[\int (\psi_{tr}(x) - \psi_{t'r'}(x))^2\, d\widehat{\mathbb{H}}_n(x)\\ &\quad - \left(\int (\psi_{tr}(x) - \psi_{t'r'}(x))\, d\widehat{\mathbb{H}}_n(x)\right)^2\Bigg],\end{aligned}$$

with $\psi_{tr}(x) := \psi\left(\frac{\|t-x\|_2}{r}\right)$. Then the application of the second exponential inequality of Proposition 8 implies for any fixed $(t,r), (t',r') \in \mathcal{T}$ that

$$\mathbb{P}\left(\left|\, Y_n(t,r) - Y_n(t',r')\,\right| > \widehat{\rho}_n\left((t,r),(t',r')\right)\, q\,\eta \,\middle|\, \mathcal{X}_n\right) \le 2\,\exp(-\eta),$$

where

$$q := 2\left(1 + \frac{9\lambda_n(1-\lambda_n)}{2\max(\lambda_n, 1-\lambda_n)^2}\,(\log 2)^{-1}\right).$$

II. (Random local covering numbers) We need a bound for the local random covering numbers $N\left((u\delta)^{1/2}, \left\{(t,r) \in \widehat{\mathcal{T}}_n : \widehat{\gamma}_n(t,r)^2 \le \delta\right\}, \widehat{\rho}_n\right)$. This is the most involved part of the proof. In contrast to previous work we aim at a uniform stochastic bound. In order to establish a sufficiently sharp upper bound, the following two claims are established:
(i) Let

$$\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2 := \int (\psi_{tr}(x) - \psi_{t'r'}(x))^2\, d\widehat{\mathbb{H}}_n(x)$$

and define d_n for arbitrary different points in $\widehat{\mathcal{T}}_n$ via

$$d_n{}^2 := \max\left[\mathbb{E}\,\widehat{\rho}_{2,n}^2, 4/n\right]\left(1 + C\log\left(4\,e \Big/ \max\left[\mathbb{E}\,\widehat{\rho}_{2,n}^2, 4/n\right]\right)\right),$$

with C a positive constant to be chosen later. Note that the map $x \mapsto x\sqrt{1 + 2C\log(\sqrt{e}/x)}$ is subadditive for $x \in (0,1]$, hence d_n defines a metric. Furthermore let $\gamma_n{}^2 := \mathbb{E}\,\widehat{\gamma}_{2,n}^2 - \left(\mathbb{E}\widehat{\gamma}_{1,n}\right)^2$, where

$$\widehat{\gamma}_{1,n}(t,r)^2 := \left(\int \psi_{tr}(x) d\widehat{\mathbb{H}}_n(x)\right)^2 \quad \text{and} \quad \widehat{\gamma}_{2,n}(t,r)^2 := \int \psi_{tr}(x)^2 d\widehat{\mathbb{H}}_n(x).$$

Then there exist a constant $C' > 0$ and a sequence $(\mathcal{C}_n)_{n\in\mathbb{N}}$ of measurable sets with $\mathbb{P}_n^{\otimes m} \otimes \mathbb{Q}_n^{\otimes(n-m)}(\mathcal{C}_n) \to 1$, such that for any $\delta > 0, u \in (0,1]$ with $u\delta \ge 4/n$ and any

realization $(X_1, \ldots, X_n) \in \mathcal{C}_n$

$$N\left((u\delta)^{1/2}, \left\{(t,r) \in \widehat{\mathcal{T}}_n : \widehat{\gamma}_n(t,r)^2 \le \delta\right\}, \widehat{\rho}_n\right)$$
$$\le N\left((u\delta)^{1/2}, \left\{(t,r) \in \widehat{\mathcal{T}}_n : \gamma_{2,n}(t,r)^2 \le C'\delta\,(\log(e/\delta))^4\right\}, d_n\right),$$

if ψ is not rectangular. In case of the rectangular kernel, the set

$$\left\{(t,r) \in \widehat{\mathcal{T}}_n : \gamma_{2,n}(t,r)^2 \le C'\delta\,(\log(e/\delta))^4\right\}$$

in the covering number has to be replaced by

$$\left\{(t,r) \in \widehat{\mathcal{T}}_n : \gamma_{2,n}(t,r)^2 \le C'\delta\,(\log(e/\delta))^4\right\}$$
$$\cup \left\{(t,r) \in \widehat{\mathcal{T}}_n : \gamma_{2,n}(t,r)^2 \ge 1 - C'\delta\,(\log(e/\delta))^4\right\}.$$

(ii) There exists a constant $A > 0$, independent of u, δ and n, such that whenever $u\delta \ge 4/n$, the upper bound given in (i) is again bounded from above by $Au^{-(d+1)}\delta^{-1}\,(\log[\,e/(u\delta)])^{5(d+1)}$. Moreover, the latter bound remains valid with $\mathcal{T}$ in place of $\widehat{\mathcal{T}}_n$.

Note that we do not rely our bound directly on uniform covering numbers and Vapnik–Cervonenkis (VC) theory as the envelope $I\{X \in \mathcal{X}_n\}$ only allows for a bound of order $u^{-2}\delta^{-2}$, which would result in the loss of efficiency of the procedure, and a pre-partitioning of $\widehat{\mathcal{T}}_n$ as used in the proof of (ii) seems to be rather involved.
Proof of (i): we first derive a uniform stochastic bound for the random metric $\widehat{\rho}_{2,n}$. Recall that every function ψ of bounded total variation is representable as a difference of isotonic functions $\psi^{(1)}$ and $\psi^{(2)}$. With the definition of the subgraphs

$$\mathrm{sgr}\left(\psi_{tr}^{(i)}\right) := \left\{(x,y) \in [0,1]^d \times \mathbb{R} : y \le \psi_{tr}^{(i)}(x)\right\}, \quad i = 1, 2,$$

the set $\left\{\mathrm{sgr}\left(\psi_{tr}^{(i)}\right) : (t,r) \in \mathcal{T}\right\}$ has a VC-dimension bounded by $d + 3$ [37] with envelope $TV(\psi)$. Consequently, the uniform covering numbers $N(\varepsilon, \mathcal{F})$ with

$$\mathcal{F} := \left\{(\psi_{tr} - \psi_{t'r'})^2 : (t,r), (t',r') \in \mathcal{T}\right\}$$

are bounded by $C\varepsilon^{-\alpha}$ for some real-valued $\alpha > 0$ and some constant $C > 0$. The boundedness of ψ shows that $\mathcal{F}$ is uniform Glivenko–Cantelli in particular (see [8], for instance). As an immediate consequence,

$$\lim_{n\to\infty} \mathbb{P}\left(\left\|\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2 - \mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\right\|_{\mathcal{T}\times\mathcal{T}} > \delta\right) = 0, \qquad (9)$$

for any $\delta > 0$. However such a bound is not sufficient for our purposes. Because of $\|\psi\|_{\sup} \le 1$, the squared random metric $\widehat{\rho}_{2,n}^2$ is $1/n$ times the sum of n independent

random variables with absolute values ≤ 4, hence

$$\begin{aligned}\operatorname{Var}\left(\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\right) &\leq \frac{4}{n}\,\mathbb{E}\left(\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\right)\\ &\leq \max\left\{\frac{4}{n},\,\mathbb{E}\left(\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\right)\right\}^2.\end{aligned}$$

Now the application of Bernstein's exponential inequality (see [35]) entails

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2-\mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2}{\max[4/n,\,\mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2]}\right|>\eta\right) &\leq 2\exp\left(-\frac{\eta^2/2}{1+\eta/3}\right)\\ &\leq 2\exp\left(-\frac{3}{2}\eta+\frac{9}{2}\right)\end{aligned}$$

for arbitrary points $(t,r),(t',r')\in\mathcal{T}$. I.e. $\widehat{\rho}_{2,n}^2-\mathbb{E}\widehat{\rho}_{2,n}^2$, standardized by $\max\left\{4/n,\,\mathbb{E}\widehat{\rho}_{2,n}^2\right\}$, has (uniformly) subexponential tails. Analogously, the process $\widehat{\rho}_{2,n}^2-\mathbb{E}\widehat{\rho}_{2,n}^2$ has subexponential increments with respect to the metric $\tilde{D}_n$ given by

$$\tilde{D}_n(a,b):=\max\left[1/n,\,\mathbb{E}\left(\widehat{\rho}_{2,n}^2(a)-\widehat{\rho}_{2,n}^2(b)\right)^2\right]I\{a\neq b\},\quad a,b\in\mathcal{T}\times\mathcal{T}.$$

Note that $\max[4/n,\,\mathbb{E}\widehat{\rho}_{2,n}^2]$ is Lipschitz continuous with respect to $\tilde{D}_n$. Theorem 9 shows that the above ingredients imply that $\lim_{\delta\searrow 0}\inf_n\mathbb{P}\left(\mathcal{A}_n(\delta,1,Q;\mathcal{X}_n)\,\middle|\,\mathcal{X}_n\right)=1$ for some adequately chosen $Q>0$, where we use the definition of $\mathcal{A}_n$ from Theorem 10 with $Y_n=\mathcal{X}_n$ and $Z_n=\widehat{\rho}_{2,n}^2-\mathbb{E}\widehat{\rho}_{2,n}^2$. Now we may apply the latter to conclude that there exists some universal constant $C>0$ such that the probability of the event

$$\begin{aligned}\Big\{&\left|\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2-\mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\right|\\ &>C\,\max\left[4/n,\,\mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\right]\\ &\quad\times\log\left(4\,e\Big/\max\left[4/n,\,\mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\right]\right)\\ &\quad\text{for some }(t,r),(t',r')\text{ with }\mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2\leq\delta'\Big\}\end{aligned}\tag{10}$$

is bounded by some function $\varepsilon(\delta')$ independent of n with $\lim_{\delta'\searrow 0}\varepsilon(\delta')=0$. Since the probability in (9) is antitonic in δ for any fixed n with limes 0 as $n\to\infty$ for any fixed δ, there exists a sequence $\delta_n\searrow 0$ along which the result of (9) still holds true. Thus, combining (9) and (10) for a sequence $\delta'=\delta'_n\searrow 0$ sufficiently slowly implies the existence of a sequence of sets $(\mathcal{A}_n)_{n\in\mathbb{N}}$ with $\mathbb{P}^{\otimes m}\otimes\mathbb{Q}^{\otimes(n-m)}(\mathcal{A}_n)\to 1$ such that

$$\begin{aligned}\widehat{\rho}_{2,n}\leq\max\left[4/n,\,\mathbb{E}\widehat{\rho}_{2,n}^2\right]^{1/2}&\left(1+C\log\left(4\,e\Big/\max\left[4/n,\,\mathbb{E}\widehat{\rho}_{2,n}^2\right]\right)\right)^{1/2}\\ &\text{whenever }\underline{\mathrm{X}}\in\mathcal{A}_n.\end{aligned}$$

The treatment of the random set

$$\widehat{\mathcal{B}}_\delta := \left\{(t,r) \in \widehat{\mathcal{T}}_n : \widehat{\gamma}_n(t,r)^2 \le \delta\right\}$$

is similar in spirit but more involved because the random quantity $\widehat{\gamma}_n^2$ is not representable as a sum of independent variables. However we can use the decomposition $[(n-1)/n]\widehat{\gamma}_n^2 = \widehat{\gamma}_{2,n}^2 - \widehat{\gamma}_{1,n}^2$. Before deriving a stochastic bound, we notice the following: if ψ describes the rectangular kernel, we have $\widehat{\gamma}_{2,n}^2 = \widehat{\gamma}_{1,n}$, i.e.

$$\widehat{\gamma}_{2,n}^2 - \widehat{\gamma}_{1,n}^2 = \widehat{\gamma}_{2,n}^2\left(1 - \widehat{\gamma}_{2,n}^2\right).$$

In this case, the random set $\widehat{\mathcal{B}}_\delta$ is consequently contained in the union

$$\left\{\widehat{\gamma}_{2,n}^2 \le 2\delta\right\} \cup \left\{\widehat{\gamma}_{2,n}^2 \ge 1 - 2\delta\right\}. \tag{11}$$

Consider the general case. Using that

$$\mathrm{Var}\left(\widehat{\gamma}_{1,n}(t,r)\right) = \frac{1}{n^2}\sum_{i=1}^n \left(\mathbb{E}\psi_{tr}(X_i)^2 - (\mathbb{E}\psi_{tr}(X_i))^2\right) \le \frac{1}{n}\mathbb{E}\left(\widehat{\gamma}_{2,n}(t,r)^2\right) \tag{12}$$

and

$$\mathrm{Var}\left(\widehat{\gamma}_{2,n}(t,r)^2\right) = \frac{1}{n^2}\sum_{i=1}^n \left(\mathbb{E}\psi_{tr}(X_i)^4 - \left(\mathbb{E}\psi_{tr}(X_i)^2\right)^2\right) \le \frac{1}{n}\mathbb{E}\left(\widehat{\gamma}_{2,n}(t,r)^2\right), \tag{13}$$

we may apply the above chain of arguments for $\widehat{\rho}_{2,n}^2$ to $\widehat{\gamma}_{1,n}$ and $\widehat{\gamma}_{2,n}^2$ together with the upper bounds in (12) and (13) for the standardization respectively and obtain the existence of a constant $C_1 > 0$ such that

$$\begin{aligned}
&\gamma_{1,n} - \frac{C_1 \max\left[1/n, \gamma_{2,n}^2\right]^{1/2}}{\sqrt{n}} \log\left(e\sqrt{n}\Big/ \max\left[1/n, \gamma_{2,n}^2\right]^{1/2}\right) \\
&\le \widehat{\gamma}_{1,n} \le \gamma_{1,n} + \frac{C_1 \max\left[1/n, \gamma_{2,n}^2\right]^{1/2}}{\sqrt{n}} \log\left(e\sqrt{n}\Big/ \max\left[1/n, \gamma_{2,n}^2\right]^{1/2}\right)
\end{aligned}$$

whenever $\underline{X} \in \mathcal{D}_n$ for some sequence $(\mathcal{D}_n)_{n\in\mathbb{N}}$ with asymptotic probability 1, uniformly evaluated at $(t,r) \in \widehat{\mathcal{T}}_n$. Note that $\widehat{\gamma}_{1,n} \ge 1/n$, $\widehat{\gamma}_{2,n}^2 \ge 1/n$ for all $(t,r) \in \widehat{\mathcal{T}}_n$. The same holds true with a constant $C_2 > 0$ and a sequence $(\mathcal{D}'_n)_{n\in\mathbb{N}}$ with asymptotic probability 1 and $\widehat{\gamma}_{1,n}$ and $\gamma_{1,n}$ replaced by $\widehat{\gamma}_{2,n}^2$ and $\gamma_{2,n}^2$. Using the lower bound for

$\widehat{\gamma}_{2,n}^2$ and the upper bound for $\widehat{\gamma}_{1,n}$, a bit of algebra yields

$$\widehat{\mathcal{B}}_\delta \subset \left\{\gamma_{2,n}^2 - \gamma_{1,n}^2 \le \delta + \max\left[1/n, \gamma_{2,n}^2\right]^{1/2} \frac{K}{\sqrt{n}} \left[\log\left(e\sqrt{n}\Big/ \max\left[1/n, \gamma_{2,n}^2\right]^{1/2}\right)\right]^2\right\}$$

whenever $\underline{X} \in \mathcal{D}_n \cap \mathcal{D}_n'$, $\delta \ge 1/n$. Here and from now on, K denotes some universal constant, not dependent on n and (t, r). Its value may be different in different expressions. Now we first consider the case

$$\sup_{n\in\mathbb{N}} \sup_{(t,r)\in\mathcal{T}} \left(\gamma_{1,n}^2 \Big/ \gamma_{2,n}^2\right) \le C' < 1.$$

Then the above condition shows that

$$\begin{aligned}\gamma_{2,n}^2(1 - C') &\le \delta + \max\left[1/n, \gamma_{2,n}^2\right]^{1/2} \frac{K}{\sqrt{n}} \left[\log\left(e\sqrt{n}\Big/ \max\left[1/n, \gamma_{2,n}^2\right]^{1/2}\right)\right]^2 \\ &\le 2\max\left\{\delta,\ \max\left[1/n, \gamma_{2,n}^2\right]^{1/2} \frac{K}{\sqrt{n}} \left[\log\left(e\sqrt{n}\Big/ \max\left[1/n, \gamma_{2,n}^2\right]^{1/2}\right)\right]^2\right\},\end{aligned}$$

which entails that $\gamma_{2,n}^2 \le K\,\delta[\log(e/\delta)]^4$ for $\delta \ge 1/n$ by the isotonicity of $x \mapsto x[\log(e/x)]^4$ on $(0, 1]$. On the other hand, the case

$$\sup_{n\in\mathbb{N}} \sup_{(t,r)\in\mathcal{T}} \left(\gamma_{1,n}^2 \Big/ \gamma_{2,n}^2\right) = 1 \tag{14}$$

implies already that ψ is equal to the rectangular kernel: if the sup is attained it is obvious. The equicontinuity of $(h_n)_{n\in\mathbb{N}}$ and its uniformly bounded L_1-norm $\|h_n\|_1 = 1$ imply its uniform boundedness, hence relative compactness in the topology of uniform convergence by the Arzelà–Ascoli-Theorem. There therefore exists at least a uniformly convergent subsequence $(h_{m(n)})$ with (uniformly) continuous limit, say h, along this result holds true as well, because $\max_{(t,r)\in\mathcal{T}} \left(\gamma_{1,n}^2 \big/ \gamma_{2,n}^2\right)$ depends continuously on the mixed density. This however implies that ψ describes the rectangular kernel, because the uniform limit h of that subsequence is bounded away from zero. Hence in case of (14), we consequently obtain by (11)

$$\begin{aligned}\widehat{\mathcal{B}}_\delta \subset \left\{\gamma_{2,n}^2 \le K\delta\,(\log(e/\delta))^4\right\} \cup \left\{\gamma_{2,n}^2 \ge 1 - K\delta\,(\log(e/\delta))^4\right\} \\ \text{whenever } \underline{X} \in \mathcal{D}_n \cap \mathcal{D}_n',\ \delta \ge 1/n.\end{aligned}$$

Proof of(ii): Since ψ is of bounded total variation, there exists some finite measure μ such that for any $0 \le z_1 < z_2 \le 1$, $|\psi(z_1) - \psi(z_2)| \le \mu[z_1, z_2]$. With

$$M_x(t, t', r, r') := \left[0, \frac{\|t - x\|_2}{r}\right] \Delta \left[0, \frac{\|t' - x\|_2}{r'}\right]$$

we obtain

$$\begin{aligned}\mathbb{E}\widehat{\rho}_{2,n}\left((t,r),(t',r')\right)^2 &\le \int (\psi_{tr}(x)-\psi_{t'r'}(x))^2\, d\mathbb{H}_n(x)\\ &\le K\int \mu\left(M_x(t,t',r,r')\right)\, d\mathbb{H}_n(x)\\ &= K\int\int I\{y\in M_x(t,t',r,r')\}\, d\mathbb{H}_n(x)d\mu(y) \quad \text{(Fubini)}\\ &\le K\sup_{y\in[0,1]}\int I\left\{y\in M_x(t,t',r,r')\right\}\, d\mathbb{H}_n(x). \qquad (15)\end{aligned}$$

Then $y\in M_x(t,t',r,r')$ implies that $x\in B_t(ry)\,\Delta B_{t'}\left(r'y\right)$. Since h_n is uniformly bounded from above, we obtain that (15) is not greater than $K\lambda\left(B_t(r)\Delta B_{t'}(r')\right)$. Consequently, $d_n\le K\,d$ if $d_n\ge 4/n$ with the metric d defined in (16), due to the isotonicity of $x\mapsto x(1+C\log(e/x))$ for $x\in(0,1]$, $C>0$. ψ attains its maximum 1 at 0, hence there exists some $r^*>0$ such that $\psi(\|x\|_2)\ge 1/2$ whenever $\|x\|_2\le r^*$. Using in addition the uniform boundedness of h_n away from zero we obtain $\gamma_{2,n}(t,r)^2\ge K\cdot r^d\ \ (t,r)\in\mathcal{T}$. We now start bounding the covering numbers

$$N\left((u\delta)^{1/2},\ \left\{(t,r)\in\mathcal{T}:\gamma_{2,n}(t,r)^2\le K\delta\left(\log(e/\delta)\right)^4\right\},\ d\right),$$

where the metric d on $\mathcal{T}\times\mathcal{T}$ is pointwise defined by

$$d\left((t,r),(t',r')\right)^2:=\lambda\left(B_t(r)\Delta B_{t'}(r')\right)\left(1+C\log\left[V\,e\Big/\lambda\left(B_t(r)\Delta B_{t'}(r')\right)\right]\right) \qquad (16)$$

with $V=\lambda(B_0(\sqrt{d}))$ the volume of the d-dimensional Euclidean ball with radius $\sqrt{d}$. Again by the isotonicity of $x\mapsto x\log(e/x)$ for $x\in(0,1]$, the inequality $\tilde{d}\left((t,r),(t',r')\right):=\lambda\left(B_t(r)\Delta B_{t'}(r')\right)^{1/2}\le\varepsilon/\sqrt{\log(V\,e/\varepsilon^2)}$ implies that $d\left((t,r),(t',r')\right)$ is not greater than $(2C+1)^{1/2}\varepsilon$. Thus in order to finish claim (ii), it is sufficient to bound

$$N\left(\left(\frac{u\delta}{\log(e/(u\delta))}\right)^{1/2},\ \left\{(t,r)\in\mathcal{T}:r^d\le\delta\left(\log(e/\delta)\right)^4\right\},\ \tilde{d}\right). \qquad (17)$$

First note that there exists a finite collection of at most $m\le K/(\delta[\log(e/\delta)]^4)$ points $t_1,\ldots,t_m$ such that the set $\left\{(t,r)\in\mathcal{T}:r^d\le\delta(\log(e/\delta))^4\right\}$ is contained in the union $\cup_{i=1}^m\mathcal{A}_i$ with

$$\mathcal{A}_i:=\left\{(t,r)\in\mathcal{T}:B_t(r)\subset B_{t_i}\left([K'\delta(\log(e/\delta))^4]^{1/d}\right)\right\}$$

for some universal $K' > 0$. The rotation and translation invariance of the Lebesgue measure leads to the rescaling invariance for the covering numbers

$$N\left(\varepsilon^{1/2},\, \{(t,r) : B_t(r) \subset B_0(R)\},\, \tilde{d}\,\right) = N\left((\varepsilon/R^d)^{1/2},\, \{(t,r) : B_t(r) \subset B_0(1)\},\, \tilde{d}\right). \tag{18}$$

But a minimal $\tilde{d}$- $(\varepsilon/R^d)^{1/2}$-net of the set $\left\{(t,r) \in \mathcal{T} : B_t(r) \subset B_0(1), r = r'\right\}$ for some *fixed* $r' > \varepsilon^{1/d}/R$ contains not more than $M = K[R^d/\varepsilon]^d$ elements $(t_1, r'), \ldots, (t_M, r')$ with K uniformly in $r' \in (\varepsilon^{1/d}/R, \sqrt{d}]$, noticing that $\lambda(B_t(r)\Delta B_{t'}(r)) \le K\|t - t'\|_2 r^{d-1}$ and $r \le \sqrt{d}$. Now fix a $K(\varepsilon/R^d)$-net $t_1, \ldots, t_M$ with respect to $\|\cdot\|_2$ and observe that $\lambda(B_t(r)\Delta B_t(r')) \le Kr^{d-1}(r - r')$ for $r > r', r \le \sqrt{d}$, which shows that the quantity (18) is bounded by $K(R^d/\varepsilon)^{d+1}$ (with K uniformly in ε and R). Correspondingly, this holds true for $N\left((u\delta/\log[e/(u\delta)])^{1/2}, \mathcal{A}_i, \tilde{d}\right)$, hence the covering number (17) is bounded by $A\delta^{-1}u^{-(d+1)}(\log(e/u\delta))^{5(d+1)}$ for some universal constant $A > 0$. An analogous bound holds for $\widehat{\mathcal{T}}_n$ in place of $\mathcal{T}$ (and $u\delta \ge 4/n$): If $(t_1, r_1), \ldots, (t_k, r_k)$ denotes an ε-net with respect to d in $B \subset \mathcal{T}$, we may define a $2\,\varepsilon$-net $(\widehat{t}_1, \widehat{r}_1), \ldots, (\widehat{t}_k, \widehat{r}_k)$ in $\widehat{\mathcal{T}}_n \cap B$ via the definition $(\widehat{t}_i, \widehat{r}_i) := \operatorname{argmin}_{(t,r)\in\widehat{\mathcal{T}}_n\cap B} d\left((t,r),(t_i,r_i)\right)$. The corresponding covering numbers in case of the rectangular kernel for the sets $\left\{\gamma_{2,n}^2 \ge 1 - K\delta\log\left(e/\delta\right)^4\right\}$ can be treated with similar arguments, which concludes the proof of (ii).

In order to line up with the requirements of Theorem 10, let us remark that the proof of that chaining requires only the special choice $u = u(\delta) = (\log(e/\delta))^\gamma$ for some exponent $\gamma < 0$, which entails that $\delta \le n^{-1}(\log n)^\alpha$ for some $\alpha > 0$ in case $u\delta \le 4/n$. But for any $\alpha' > 0$, $\sharp\left\{(t,r) \in \widehat{\mathcal{T}}_n : r^d \le Kn^{-1}(\log n)^{\alpha'}\right\} = \sum_{i=1}^n \sharp\left\{(X_i, r) \in \widehat{\mathcal{T}}_n : r^d \le Kn^{-1}(\log n)^{\alpha'}\right\}$, and with the same arguments as used in (i) we obtain for $r_n^d = n^{-1}(\log n)^{\alpha'}$ that the inequality $\widehat{\mathbb{H}}_n\left(B_t(r_n)\right) \le K\lambda(B_t(r_n))\log n$ holds, uniformly in $t \in [0,1]^d$, with asymptotic probability 1, which entails $\sharp\left\{(t,r) \in \widehat{\mathcal{T}}_n : r^d \le Kn^{-1}(\log n)^{\alpha'}\right\} = O_p\left(n(\log n)^{\alpha''}\right)$ for some $\alpha'' > 0$.

III. (TIGHTNESS AND WEAK APPROXIMATION IN PROBABILITY) As a consequence of the above exponential inequalities in step I and the bound for the uniform covering numbers $N(\delta, \mathcal{T})$, Theorem 9 shows

$$\lim_{\delta\searrow 0}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\widehat{\rho}_n((t,r),(t',r'))\le\delta}\frac{|Y_n(t,r) - Y_n(t',r')|}{\widehat{\rho}_n((t,r),(t',r'))\log\left(e/\widehat{\rho}_n((t,r),(t',r'))\right)} > \varepsilon \,\middle|\, \mathcal{X}_n\right) = 0, \tag{19}$$

where the sup within the brackets is even running over elements of $\mathcal{T} \times \mathcal{T}$. Now the application of Theorem 10 entails that $\mathcal{L}\left(T_n \circ \Pi \,|\, \mathcal{X}_n\right)$ is tight in $\left(\mathbb{P}_n^{\otimes m} \otimes \mathbb{Q}_n^{\otimes(n-m)}\right)$-probability. What remains being proved is the weak approximation. Starting from (19), the uniform convergence (9) implies in particular the asymptotic stochastic

equicontinuity

$$\lim_{\delta \searrow 0} \limsup_{n\to\infty} \mathbb{E}^*_{(p_n,q_n,\lambda_n)} \mathbb{P}^* \left(\sup_{\rho_n((t,r),(t',r'))\le\delta} \left| Y_n(t,r) - Y_n(t',r') \right| > \varepsilon \,\middle|\, \mathcal{X}_n \right) = 0$$

for all $\varepsilon > 0$.

Since to any subsequence of the metric ρ_n there exists some uniformly convergent subsubsequence as a consequence of the relative compactness of $(h_n)_{n\in\mathbb{N}}$ in the uniform topology, it suffices (via proof of contradiction) for the weak approximation in probability

$$d_w \left\{ \mathcal{L}\left((Y_n(t,r))_{(t,r)\in\mathcal{T}} \,\middle|\, \mathcal{X}_n \right),\ \mathcal{L}\left((Z_n(t,r))_{(t,r)\in\mathcal{T}} \right) \right\} \longrightarrow_{\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}} 0$$

to establish the convergence of finite dimensional distributions. Here, d_w is defined via the outer expectations $\mathbb{E}^*$. For let $\{(t_1,r_1),\ldots,(t_k,r_k)\}$ be a collection of points from $\mathcal{T}$. Denote furthermore $a_{rt}(X_i) := n^{-1/2}\sqrt{\lambda_n(1-\lambda_n)}\,\psi_{tr}(X_i)$. Then

$$(Y_n(t,r))_{(t,r)\in\mathcal{T}} = \left(\sum_{i=1}^n a_{rt}(X_i)\Lambda(t^i) \right)_{(t,r)\in\mathcal{T}},$$

with t^i the i'th nearest-neighbor of t within $\mathcal{X}_n$. Let $(Z_n(t,r))_{(t,r)\in\mathcal{T}}$ be pointwise be defined by $Z_n(t,r) := \sqrt{\lambda_n(1-\lambda_n)} \int \phi_{rt}^{(n)}(x)\, dW(x)$. Using that $2\,\mathrm{cov}\,(X_1,X_2)$ equals $\mathrm{Var}(X_1)+\mathrm{Var}(X_2)-\mathrm{Var}(X_1-X_2)$ for two random variables X_1 and X_2, one finds that $[(n-1)/n]\mathrm{cov}\left(Y_n(t,r), Y_n(t',r') \,\middle|\, \mathcal{X}_n\right)$ equals

$$\begin{aligned}
&-\frac{1}{2}\int (\psi_{tr}(x)-\psi_{t'r'}(x))^2\, d\widehat{\mathbb{H}}_n(x) + \frac{1}{2}\left(\int (\psi_{tr}(x)-\psi_{t'r'}(x))\, d\widehat{\mathbb{H}}_n(x) \right)^2 \\
&+\frac{1}{2}\int \psi_{tr}(x)^2 d\widehat{\mathbb{H}}_n(x) \\
&-\frac{1}{2}\left(\int \psi_{tr}(x) d\widehat{\mathbb{H}}_n(x) \right)^2 + \frac{1}{2}\int \psi_{t'r'}(x)^2 d\widehat{\mathbb{H}}_n(x) - \frac{1}{2}\left(\int \psi_{t'r'}(x) d\widehat{\mathbb{H}}_n(x) \right)^2 .
\end{aligned} \tag{20}$$

Replacing the empirical measure $\widehat{\mathbb{H}}_n$ by its expectation $\mathbb{H}_n$, the above six expressions in (20) coincide with the covariance $\mathrm{cov}\left(Z_n(t,r), Z_n(t',r')\right)$ of the limiting process Z_n. Define $\bar{a}_{r_jt_j}^{(n)} := n^{-1}\sum_{i=1}^n a_{r_jt_j}^{(n)}(X_i)$, $j=1,\ldots,k$. Since

$$\sum_{j=1}^k \frac{\max_i \left(a_{r_jt_j}^{(n)}(X_i) - \bar{a}_{r_jt_j}^{(n)} \right)^2}{\sum_{i=1}^n \left(a_{r_jt_j}^{(n)}(X_i) - \bar{a}_{r_jt_j}^{(n)} \right)^2} \longrightarrow_{\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}} 0 \quad (n\to\infty)$$

and $\left|\operatorname{cov}\left(Y_n(t,r), Y_n(t',r') \,\middle|\, \mathcal{X}_n\right) - \operatorname{cov}\left(Z_n(t,r), Z_n(t',r')\right)\right| \longrightarrow_{\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}} 0$ by an application of the weak law of large numbers for triangular arrays to each of the expressions in (20) separately, Hájek's Central Limit Theorem for permutation statistics extended for the multivariate setting yields the desired weak convergence in probability of the finite dimensional distributions. For notational convenience, define

$$T_n^{\Pi}(\delta,\delta') := \sup_{\substack{(j,k):\\ \delta<\gamma_n(j,k)\le\delta'}} \left\{\left|T_{jkn}\circ\Pi\right| - C_{jkn}\right\}$$

and

$$S_n(\delta,\delta') := \sup_{\substack{(t,r):\\ \delta<\gamma_n(t,r)\le\delta'}} \left\{\frac{\left|\int \phi_{rt}^{(n)}(x)\,dW(x)\right|}{\gamma_n(t,r)} - \sqrt{2\log\left(1/\gamma_n(t,r)^2\right)}\right\}.$$

Since $\sup_{t\in\mathcal{T}\setminus\widehat{\mathcal{T}}_n} d(t,\widehat{\mathcal{T}}_n) \to_{\mathbb{P}^{\otimes m}\otimes\mathbb{Q}^{\otimes(n-m)}} 0$ and $\sup_{(j,k):\,\gamma_n(j,k)\ge\delta}\left|C_{jkn}-(2\,\Gamma_{jkn})^{1/2}\right| \to_{\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}} 0$ as $n\to\infty$, it follows from the above established results that

$$d_w\left(\mathcal{L}\left(T_n^{\Pi}(\delta,1)\,\middle|\,\mathcal{X}_n\right),\ \mathcal{L}\left(S_n(\delta,1)\right)\right) \longrightarrow_{\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}} 0$$

for any fixed $\delta\in(0,1]$. An application of Theorem 10 as well as its subsequent Remark imply that

$$\lim_{\delta\searrow 0}\limsup_{n\to\infty}\mathbb{E}\,\mathbb{P}\left(T_n^{\Pi}(0,\delta)\ge\varepsilon\,\middle|\,\mathcal{X}_n\right)=0 \quad\text{and}\quad \lim_{\delta\searrow 0}\limsup_{n\to\infty}\mathbb{P}\left(S_n(0,\delta)\ge\varepsilon\right)=0$$

for any $\varepsilon>0$. Thus, because obviously $\lim_{\delta\searrow 0}\limsup_{n\to\infty}\mathbb{P}\left(S_n(\delta,1)\le-\varepsilon\right)=0$, we obtain

$$d_w\left(\mathcal{L}\left(T_n^{\Pi}(0,1)\,\middle|\,\mathcal{X}_n\right),\ \mathcal{L}\left(S_n(0,1)\right)\right) \longrightarrow_{\mathbb{P}_n^{\otimes m}\otimes\mathbb{Q}_n^{\otimes(n-m)}} 0.$$

□

Proof of Theorem 3 Let $\mathcal{C}$ be some compact rectangle of J. Fix $\beta>0$. For any integer $k>1$ let $\mathcal{C}_{n,k}\subset\mathcal{C}$ be some maximal subset of points such that $\|x-y\|_2\ge 2k\delta_n$ and $B_x(k\delta_n)\subset\mathcal{C}$ for arbitrary different points $x,y\in\mathcal{C}_{n,k}$. Then $\sharp\mathcal{C}_{n,k}\sim(k\delta_n)^{-d}$. Now let $\phi_{x,n}$ be the solution of the subsequent optimization problem:
(∗) Minimize $\|g\|_2$ under the constraints

$$g\in\mathcal{H}_d(\beta,L;\mathbb{R}^d),\quad \operatorname{supp}(g)\subseteq B_x(k\delta_n),\quad g(x)=L\delta_n^{\beta},\quad \int g(z)\sqrt{h_n(z)}dz=0.$$

These constraints define a closed and convex set in $L_2\left([0,1]^d\right)$ which is non-empty for k sufficiently large [and uniformly in n due to the equicontinuity of (h_n) and

the rescaling property, see subsequently to (24)]. Consequently in the latter case, the argmin $\phi_{x,n}$ exists and is unique. The resulting density candidates

$$p_{x,n} = h_n \cdot \left(1 + (1-(m/n))\,\phi_{x,n}/\sqrt{h_n}\right) \quad \text{and} \quad q_{x,n} = h_n \cdot \left(1 - (m/n)\phi_{x,n}/\sqrt{h_n}\right)$$

are non-negative and thus contained in $\mathcal{F}_{h_n}^{(m,n)}$ as soon as additionally

$$-\frac{\sqrt{h_n(.)}}{1-m/n} \le \phi_{x,n}(.) \le \frac{\sqrt{h_n(.)}}{m/n} \quad \text{for all } x \in \mathcal{C}_n.$$

This is guaranteed for sufficiently large n when sequence $(\delta_n)_{n\in\mathbb{N}}$ tends to zero. For any statistical level-α-test $\psi = \psi(\beta, L, h_n) : \mathbb{R}^{d\times n} \to [0,1]$ for testing the hypothesis "$\phi = 0$" it holds true that

$$\begin{aligned}
\min_{x\in\mathcal{C}_{n,k}} \mathbb{E}_{(m,n,p_{x,n},q_{x,n})}\psi - \alpha &\le \min_{x\in\mathcal{C}_{n,k}} \mathbb{E}_{(m,n,p_{x,n},q_{x,n})}\psi - \mathbb{E}_{(m,n,h_n,h_n)}\psi \\
&\le \frac{1}{\sharp\mathcal{C}_{n,k}} \sum_{x\in\mathcal{C}_{n,k}} \mathbb{E}_{(m,n,p_{x,n},q_{x,n})}\psi - \mathbb{E}_{(m,n,h_n,h_n)}\psi \\
&\le \mathbb{E}_{(m,n,h_n,h_n)} \left| \frac{1}{\sharp\mathcal{C}_{n,k}} \sum_{x\in\mathcal{C}_{n,k}} \frac{d\mathbb{P}_{(m,n,p_{x,n},q_{x,n})}}{d\mathbb{P}_{(m,n,h_n,h_n)}}(\underline{\mathrm{X}}) - 1 \right|.
\end{aligned} \tag{21}$$

For short we write $\mathbb{E}_0$ for $\mathbb{E}_{(m,n,h_n,h_n)}$ in the sequel. Note that the test is allowed to depend on the nuisance functional h_n (in fact the log-likelihood and its distribution do). Now we aim at determining δ_n such that the right-hand-side tends to zero as n goes to infinity. Although $\lambda\left(\mathrm{supp}(\phi_{x,n}) \cap \mathrm{supp}(\phi_{y,n})\right) = 0$ for any different $x, y \in \mathcal{C}_{n,k}$, the likelihood-ratios

$$\begin{aligned}
L_{x,n} &:= \frac{d\mathbb{P}_{(m,n,p_{x,n},q_{x,n})}}{d\mathbb{P}_{(m,n,h_n,h_n)}}(\underline{\mathrm{X}}) \\
&= \prod_{i=1}^{m} \left(1 + (1-(m/n))\,\frac{\phi_{x,n}}{\sqrt{h_n}}(X_i)\right) \prod_{i=m+1}^{n} \left(1 - (m/n)\frac{\phi_{x,n}}{\sqrt{h_n}}(X_i)\right),
\end{aligned}$$

are not independent. However, they are independent conditional on the random vector $\Delta_n = (\Delta_{x,n})_{x\in\mathcal{C}_{k,n}}$ with entries

$$\Delta_{x,n} := (\sharp\{i \le m : \|X_i - x\|_2 \le k\delta_n\},\ \sharp\{i > m : \|X_i - x\|_2 \le k\delta_n\}).$$

Note that $\mathbb{E}_0(L_{x,n}|\Delta_n) = \mathbb{E}_0 L_{x,n} = 1$. Following at this point standard truncation arguments as, for instance, in Dümbgen and Walther [11], proof of Lemma 7.4, it turns out to be sufficient for the convergence to zero of (21) to find δ_n and $\gamma = \gamma_n \in (0,1]$

such that the ratio

$$\max_{x\in\mathcal{C}_{n,k}} \frac{1}{(\sharp\,\mathcal{C}_{n,k})^{\gamma}} \mathbb{E}_0\, L_{x,n}^{1+\gamma} \tag{22}$$

tends to zero as n goes to infinity. But

$$\begin{aligned}
\mathbb{E}_0 L_{x,n}^{1+\gamma} &= \left\{ \int h_n(z) \left(1 + (1-m/n)\frac{\phi_{x,n}(z)}{\sqrt{h_n(z)}}\right)^{1+\gamma} dz \right\}^m \\
&\quad \times \left\{ \int h_n(z) \left(1 - (m/n)\frac{\phi_{x,n}(z)}{\sqrt{h_n(z)}}\right)^{1+\gamma} dz \right\}^{n-m} \\
&= \left\{ 1 + \frac{1}{2}\gamma(1+\gamma)\left(1 + O\left(\delta_n^{\beta}\right)\right)(1-(m/n))^2 \int_0^1 \phi_{x,n}(z)^2 dz \right\}^m \\
&\quad \times \left\{ 1 + \frac{1}{2}\gamma(1+\gamma)\left(1 + O\left(\delta_n^{\beta}\right)\right)(m/n)^2 \int_0^1 \phi_{x,n}(z)^2 dz \right\}^{n-m}, \qquad (23)
\end{aligned}$$

using the bound $(1+\Delta)^{1+\gamma} \le 1 + (1+\gamma)\Delta + 2^{-1}\gamma(1+\gamma)\Delta^2 + 3\gamma\Delta^2|\Delta|$ for $|\Delta| \le 1$. Now let $\tilde{\phi}_k$ be the solution to the following optimization problem
$(**)$ Minimize $\|g\|_2$ subject to

$$g \in \mathcal{H}_d(\beta, L; \mathbb{R}^d), \quad \mathrm{supp}(g) \subseteq B_0(k), \quad g(0) = 1, \quad \int g(x)dx = 0. \tag{24}$$

Notice the rescaling property $L\delta_n^{\beta} g(./\delta_n) \in \mathcal{H}_d(\beta, L; \mathbb{R}^d)$ with $\mathrm{supp}\left(L\delta_n^{\beta} g(./\delta_n)\right) = B_0(\delta_n k)$ and $L\delta_n^{\beta} g(0) = L\delta_n^{\beta} \Leftrightarrow g \in \mathcal{H}_d(\beta, L; \mathbb{R}^d)$ with $\mathrm{supp}(g) = B_0(k)$ and $g(0) = 1$. Due to the equicontinuity of $(h_n)_{n\in\mathbb{N}}$,

$$\lim_{\delta \searrow 0} \sup_{x \in B_z(\delta)} \sup_n \left| h_n(x) - h_n(z) \right| = 0,$$

whence

$$\int \phi_{x,n}(z)^2\, dz = (1 + o(1))\, L^2 \delta_n^{2\beta+d} \|\tilde{\phi}_k\|_2^2 \tag{25}$$

because the minimum in $(*)$ depends continuously on the mixed density h_n as can be seen using a Lagrange multiplier for the centering constraint. Note that the $o(1)$-term is uniformly in $x \in \mathcal{C}_{k,n}$. Now the combination of (23) and (25) shows that for δ_n sufficiently small, (22) is bounded by

$$\exp\left(n(m/n)(1-m/n)\frac{1}{2}\gamma(1+\gamma)L^2\delta_n^{2\beta+d}\|\tilde{\phi}_k\|_2^2\,(1+o(1)) - \gamma\log(\sharp\,\mathcal{C}_{k,n}) \right).$$

By construction, $\sharp\, \mathcal{C}_{k,n} \geq d_k \cdot \delta_n^{-d}$ for some constant $d_k > 0$. Now fix $\delta > 0$ and define

$$c_k(\beta, L) := \left(\frac{2\, d\, L^{d/\beta}}{(2\beta + d)\|\tilde{\phi}_k\|_2^2} \right)^{\beta/(2\beta+d)}.$$

Observe that the sequence $c_k(\beta, L)$ is increasing in k. We need to check that $\lim_{k\to\infty} \|\tilde{\phi}_k\|_2 = \|\gamma_\beta\|_2$. Note that in contrast to (24), the solution of (2) does not integrate to zero in general and it remains still open if γ_β is compactly supported for $d \geq 2$ and $\beta > 1$. Starting from γ_β, it is sufficient to construct a sequence $\tilde{\gamma}_{\beta,k}$ satisfying the constraints of the optimization problem $(**)$ such that $\lim_{k\to\infty} \|\tilde{\gamma}_{\beta,k}\|_2 = \|\gamma_\beta\|_2$. Then the equality $\lim_{k\to\infty} \|\tilde{\phi}_k\|_2 = \|\gamma_\beta\|_2$ follows from $\|\tilde{\gamma}_{\beta,k}\|_2 \geq \|\tilde{\phi}_k\|_2$. The existence is sketched in the appendix of the extended version of this article. As a consequence there exists some $k' \in \mathbb{N}$ such that $c(\beta, L)(1-\delta) < c_{k'}(\beta, L)(1-\delta/2)$. Now one verifies that the lower bound is established with the choice

$$\delta_n := \left(\frac{c_{k'}(\beta, L)(1-\delta/2)\rho_n}{L} \right)^{1/\beta}$$

and some sequence $\gamma = \gamma_n \to 0$ with $\lim_n \gamma_n (\log n)^{1/2} = \infty$. □

Proof of Theorem 4 By virtue of Theorem 2, the sequence $\mathcal{L}\left(T_n \circ \Pi \,\middle|\, \mathcal{X}_n\right)$ is tight in $\left(\mathbb{P}_n^{\otimes m} \otimes \mathbb{Q}_n^{\otimes(n-m)}\right)$-probability, resulting in stochastic boundedness of the sequence of random quantiles $\left(\kappa_\alpha(\underline{X})\right)_{n\in\mathbb{N}}$. The bounded total variation of the kernel for $\beta \leq 1$ is a consequence of its monotonicity, for $\beta > 1$ it results from the continuous differentiability of $\psi_{\beta,K}$ and its compact support. For notational convenience the dependency on β and K is suppressed. They are arbitrary but fixed unless stated otherwise. First note that for any random couple $(\widehat{j}_n, \widehat{k}_n)$ it holds true that

$$\mathbb{P}_{(m,n,p_n,q_n)}\left(T_n > \kappa_\alpha(\underline{X})\right) \geq \mathbb{P}_{(m,n,p_n,q_n)}\left(T_{\widehat{j}_n\widehat{k}_n n} - C_{\widehat{j}_n\widehat{k}_n n} > \kappa_\alpha(\underline{X})\right).$$

Hence it is sufficient to prove that for any sequence $(\phi_n)_{n\in\mathbb{N}}$ of admissible alternatives there exists a random sequence of $(\widehat{j}_n, \widehat{k}_n)_{n\in\mathbb{N}}$ with $T_{\widehat{j}_n\widehat{k}_n n} - C_{\widehat{j}_n\widehat{k}_n n} \longrightarrow_{\mathbb{P}^{\otimes m}\otimes\mathbb{Q}^{\otimes(n-m)}} \infty$. As in the proof of Theorem 2 define $\gamma_n(t,r)^2 := \mathbb{E}\widehat{\gamma}_{2,n}(t,r)^2 - \left(\mathbb{E}\widehat{\gamma}_{1,n}(t,r)\right)^2$, $(t,r) \in \mathcal{T}$. Let $t_n := \operatorname{argmax}_{x\in J} |\phi_n(x)|$ and $r_n := \left(\|\phi_n\|_{\sup}/L\right)^{1/\beta}$. Define $\left(\widehat{t}_n, \widehat{r}_n\right) := \left(X_{\widehat{j}_n}, \left\|X_{\widehat{j}_n} - X_{\widehat{k}_n}\right\|_2\right)$ with

$$\left(\widehat{j}_n, \widehat{k}_n\right) := \operatorname*{argmin}_{j,k=1,\ldots,n} \lambda\left(B_{t_n}(r_n)\, \Delta\, B_{X_j}\left(\|X_j - X_k\|_2\right)\right).$$

Now let the process S_n on $\mathcal{T}$ pointwise be defined by

$$S_n(t,r) := \frac{\sqrt{\lambda_n(1-\lambda_n)}}{\sqrt{n}} \sum_{i=1}^{n} \psi\left(\frac{\|X_i - t\|_2}{r}\right) \Lambda(X_i).$$

Furthermore, let us introduce the random variables $(\widehat{t}_{ni}, \widehat{r}_{ni})$, based on the indices $(\widehat{j}_{ni}, \widehat{k}_{ni})$ which are defined analogously to $(\widehat{j}_n, \widehat{k}_n)$ but with the minimum running over the set $j, k \in \{1, \dots, n\}\backslash\{i\}$ only. Then, recalling the definition $\psi_{tr}(x) := \psi\left(\frac{\|t-x\|_2}{r}\right)$,

$$
\begin{aligned}
&\frac{1}{\gamma_n(t_n, r_n)} \left| \mathbb{E}\left(S_n(\widehat{t}_n, \widehat{r}_n) - S_n(t_n, r_n)\right) \right| \\
&= \frac{\sqrt{\lambda_n(1-\lambda_n)}}{\gamma_n(t_n, r_n)} \frac{1}{\sqrt{n}} \left| \frac{n}{m} \sum_{i=1}^{m} \mathbb{E}\left(\psi_{\widehat{t}_n\widehat{r}_n}(X_i) - \psi_{t_n r_n}(X_i)\right) \right. \\
&\quad \left. - \frac{n}{n-m} \sum_{i=m+1}^{n} \mathbb{E}\left(\psi_{\widehat{t}_n\widehat{r}_n}(X_i) - \psi_{t_n r_n}(X_i)\right) \right| \\
&\le \frac{\sqrt{\lambda_n(1-\lambda_n)}}{\gamma_n(t_n, r_n)} \frac{1}{\sqrt{n}} \left| \frac{n}{m} \sum_{i=1}^{m} \mathbb{E}\left(\psi_{\widehat{t}_n\widehat{r}_n}(X_i) - \psi_{\widehat{t}_{ni}\widehat{r}_{ni}}(X_i)\right) \right. \\
&\quad \left. - \frac{n}{n-m} \sum_{i=m+1}^{n} \mathbb{E}\left(\psi_{\widehat{t}_n\widehat{r}_n}(X_i) - \psi_{\widehat{t}_{ni}\widehat{r}_{ni}}(X_i)\right) \right| \\
&\quad + \frac{\sqrt{\lambda_n(1-\lambda_n)}}{\gamma_n(t_n, r_n)} \frac{1}{\sqrt{n}} \left| \frac{n}{m} \sum_{i=1}^{m} \mathbb{E}\left(\psi_{\widehat{t}_{ni}\widehat{r}_{ni}}(X_i) - \psi_{t_n r_n}(X_i)\right) \right. \\
&\quad \left. - \frac{n}{n-m} \sum_{i=m+1}^{n} \mathbb{E}\left(\psi_{\widehat{t}_{ni}\widehat{r}_{ni}}(X_i) - \psi_{t_n r_n}(X_i)\right) \right| \\
&\le \frac{\sqrt{\lambda_n(1-\lambda_n)}}{\gamma_n(t_n r_n)} \frac{4}{\sqrt{n}} \|\psi\|_{\sup} \max\left(\frac{n}{m}, \frac{n}{n-m}\right) \\
&\quad + \frac{\sqrt{\lambda_n(1-\lambda_n)}}{\gamma_n(t_n, r_n)} \frac{1}{\sqrt{n}} \left| \mathbb{E}\left\{ \frac{n}{m} \sum_{i=1}^{m} \int \left(\psi_{\widehat{t}_{ni}\widehat{r}_{ni}}(x) - \psi_{t_n r_n}(x)\right) p_n(x) dx \right. \right. \\
&\quad \left. \left. - \frac{n}{n-m} \sum_{i=m+1}^{n} \int \left(\psi_{\widehat{t}_{ni}\widehat{r}_{ni}}(x) - \psi_{t_n r_n}(x)\right) q_n(x) dx \right\} \right|, \qquad (26)
\end{aligned}
$$

whereby we used for the first term in the last inequality that $(\widehat{t}_{ni}, \widehat{r}_{ni})$ differs from $(\widehat{t}_n, \widehat{r}_n)$ for at most two indices $i, j \in \{1, \dots, n\}$; the second term follows by including and evaluating the conditional expectation given $(\widehat{t}_{ni}, \widehat{r}_{ni})$ as X_i is independent of $(\widehat{t}_{ni}, \widehat{r}_{ni})$. Replacing again $(\widehat{t}_{ni}, \widehat{r}_{ni})$ by $(\widehat{t}_n, \widehat{r}_n)$, the second expression behind the inequality in formula (26) is bounded by

$$
\begin{aligned}
&\frac{\sqrt{\lambda_n(1-\lambda_n)}}{\gamma_n(t_n, r_n)} \frac{4}{\sqrt{n}} \|\psi\|_{\sup} \max\left(\frac{n}{m}, \frac{n}{n-m}\right) \\
&\quad + \frac{\sqrt{n}\sqrt{\lambda_n(1-\lambda_n)}}{\gamma_n(t_n, r_n)} \left| \mathbb{E}\left[\int \left(\psi_{\widehat{t}_n\widehat{r}_n}(x) - \psi_{t_n r_n}(x)\right) \left(p_n(x) - q_n(x)\right) dx \right] \right|. \qquad (27)
\end{aligned}
$$

Now we can make use of the fact that $\left| p_n(x) - q_n(x) \right| = \left| \phi_n(x)\sqrt{h_n(x)} \right| \leq C\|\phi_n\|_{\sup}$ with $C := \sup_n \sup_x \left| \sqrt{h_n(x)} \right|$. Recall that $\|h_n\|_{\sup}$ is uniformly bounded due to the equicontinuity assumption on $(h_n)_{n\in\mathbb{N}}$ and the constraint on the L_1-norm $\|h_n\|_1 = 1$, whence the term in (27) is not greater than

$$C\frac{\sqrt{n}\|\phi_n\|_{\sup}}{\gamma_n(t_n, r_n)}\,\mathbb{E}\left(\int \left| \psi_{\widehat{t}_n\widehat{r}_n}(x) - \psi_{t_n r_n}(x) \right| dx\right). \tag{28}$$

Using the bounded total variation $TV(\psi)$ of ψ and M_x and μ as defined in the proof of Theorem 2, the integral which appears in (28) can be bounded by

$$\begin{aligned}
&\mathbb{E}\left(\int \left| \psi_{\widehat{t}_n\widehat{r}_n}(x) - \psi_{t_n r_n}(x) \right| dx\right)\\
&\leq \mathbb{E}\left(\int \mu\left(M_x(t_n, r_n, \widehat{t}_n, \widehat{r}_n)\right) dx\right)\\
&= \mathbb{E}\left(\int\int I\left\{y \in M_x(t_n, r_n, \widehat{t}_n, \widehat{r}_n)\right\} dx d\mu(y)\right) \qquad \text{(Fubini)}\\
&\leq TV(\psi)\,\mathbb{E}\sup_{y\in[0,1]}\left(\int I\left\{y \in M_x(t_n, r_n, \widehat{t}_n, \widehat{r}_n)\right\} dx\right)\\
&= TV(\psi)\,\mathbb{E}\,\lambda\left(B_{t_n}(r_n)\Delta B_{\widehat{t}_n}(\widehat{r}_n)\right)\\
&= O\left(r_n^{d-1} n^{-1/d}\right),
\end{aligned} \tag{29}$$

using in the last bound besides the stochastic convergence rate $n^{-1/d}$ the uniform integrability of the sequences $\left(n^{1/d}\|\widehat{t}_n - t_n\|_2\right)$, $\left(n^{1/d}|\widehat{r}_n - r_n|\right)$ which result from $\mathbb{P}\left(\|\widehat{t}_n - t_n\|_2 > x\right) = \prod_{i=1}^n \mathbb{P}\left(X_i \notin B_{t_n}(x)\right) \sim \left(1 - \lambda(B_{t_n}(x) \cap [0,1]^d)\right)^n$ $(= (1 - Vx^d)^n$ if $B_{t_n}(x) \subset [0,1]^d)$ and $\mathbb{P}\left(|\widehat{r}_n - r_n| > x\right) \leq 2\,\mathbb{P}\left(\|\widehat{t}_n - t_n\|_2 > x/2\right)$. Here, V denotes the volume of the d-dimensional Euclidean unit ball, i.e. $V = \pi^{d/2}\Gamma(d/2+1)$. Together with (26–28) this shows that for any sequence of admissible alternatives $(\phi_n)_{n\in\mathbb{N}}$

$$\frac{\left|\mathbb{E}\left(S_n(\widehat{t}_n, \widehat{r}_n) - S_n(t_n, r_n)\right)\right|}{\gamma_n(t_n, r_n)} = O\left(r_n^{d/2-1+\beta} n^{-1/d+1/2}\right). \tag{30}$$

If in particular $\|\phi_n\|_{\sup} = O\left(((\log n)/n)^{\beta/(2\beta+d)}\right)$, the term in (30) is of order

$$O\left((\log n)^{(\beta+d/2-1)/(2\beta+d)} n^{-(2\beta/d)/(2\beta+d)}\right).$$

We need to check that

$$\frac{\gamma_n(t_n, r_n)}{\widehat{\gamma}_n(\widehat{t}_n, \widehat{r}_n)} \longrightarrow_{\mathbb{P}^{\otimes m}\otimes Q^{\otimes(n-m)}} 1. \tag{31}$$

For this we use the decomposition $[(n-1)/n]\widehat{\gamma}_n(t,r)^2 = \widehat{\gamma}_{2,n}(t,r)^2 - \widehat{\gamma}_{1,n}(t,r)^2$. To this end note first that

$$\begin{aligned}
&\left|\widehat{\gamma}_{n,1}(\widehat{t}_n,\widehat{r}_n) - \widehat{\gamma}_{n,1}(t_n,r_n)\right| \\
&\le \left\|\psi_{\widehat{t}_n\widehat{r}_n} - \psi_{t_nr_n}\right\|_{\sup}\frac{1}{n}\sum_{i=1}^n I\left\{X_i \in B_{\widehat{t}_n}(\widehat{r}_n)\cap B_{t_n}(r_n)\right\} \\
&\quad + 2\|\psi\|_{\sup}\frac{1}{n}\sum_{i=1}^n I\left\{X_i \in B_{\widehat{t}_n}(\widehat{r}_n)\,\Delta\, B_{t_n}(r_n)\right\} \\
&\le \left\|\psi_{\widehat{t}_n\widehat{r}_n} - \psi_{t_nr_n}\right\|_{\sup}\frac{1}{n}\sum_{i=1}^n I\left\{X_i \in B_{t_n}(r_n)\right\} \\
&\quad + 2\|\psi\|_{\sup}\frac{1}{n}\sum_{i=1}^n I\left\{X_i \in B_{\widehat{t}_n}(\widehat{r}_n)\,\Delta\, B_{t_n}(r_n)\right\} \\
&= o_p(1)O_p(r_n^d) + O_p\left(r_n^{d-1}n^{-1/d}\right) = o_p\left(\gamma_{n,1}(t_n,r_n)\right).
\end{aligned}$$

The "$o_p(1)$"-term results from the Hölder continuity of ψ (for $\beta > 1$ the first derivative of ψ is uniformly bounded on $[-K,K]$), $\operatorname{supp}\left(\psi_{t_nr_n} - \psi_{\widehat{t}_n\widehat{r}_n}\right) = B_{t_n}(r_n)\cup B_{\widehat{t}_n}(\widehat{r}_n)$ and the fact that $r_n > \left(c(\beta,L)\rho_{m,n}/L\right)^{1/\beta}$ while $\widehat{t}_n - t_n \sim n^{-1/d}$, $\widehat{r}_n - r_n \sim n^{-1/d}$. The case $i=2$ is done analogously (taking the square). To verify (31) it remains to be shown that $\widehat{\gamma}_n(t_n,r_n)/\gamma_n(t_n,r_n) - 1 = o_p(1)$ which however is a simple consequence of Chebychef's inequality since for any $\beta > 0$ and any sequence of admissible alternatives $(\phi_n)_{n\in\mathbb{N}}$, the sequence $\gamma_n(t_n,r_n) \sim r_n^{d/2}$ or some subsequence decreases (if it decreases) at a slower rate than $n^{-1/2}$. The above considerations show in particular that

$$\begin{aligned}
C_{\widehat{j}_n\widehat{k}_nn} &= \frac{3\,R_\psi(m,n)}{\sqrt{n}\,\widehat{\gamma}_n(\widehat{t}_n,\widehat{r}_n)}\delta(m,n)\log\left(\widehat{\gamma}_n(\widehat{t}_n,\widehat{r}_n)^{-2}\right) + \delta(m,n)\sqrt{2\log\left(\widehat{\gamma}_n(\widehat{t}_n,\widehat{r}_n)^{-2}\right)} \\
&= \sqrt{2\log\left(\gamma_n(t_n,r_n)^{-2}\right)} + o_p(1),
\end{aligned}$$

using in addition that $\delta(m,n) = 1 + O(n^{-1/2})$. Consequently,

$$T_{\widehat{j}_n\widehat{k}_nn} - C_{\widehat{j}_n\widehat{k}_nn} = O_p(1) + \frac{\mathbb{E}S_n(t_n,r_n)}{\gamma_n(t_n,r_n)}\left(1+o_p(1)\right) - \sqrt{2\log\left(\gamma_n(t_n,r_n)^{-2}\right)}, \tag{32}$$

and it has to be verified that the latter quantity goes to infinity. Recall that

$$\gamma_n(t_n, r_n)^2 = \int_{[0,1]^d} \psi_{t_n r_n}(x)^2 h_n(x) dx - \left(\int_{[0,1]^d} \psi_{t_n r_n}(x) h_n(x) dx \right)^2$$
$$= \left(1 + O(r_n^d)\right) \int_{[0,1]^d} \psi_{t_n r_n}(x)^2 h_n(x) dx. \tag{33}$$

We first assume that $r_n = o(1)$, i.e. $\|\phi_n\|_{\sup} = o(1)$. Using that

$$\lim_{\delta \searrow 0} \sup_n \sup_{t \in [0,1]^d} \sup_{x \in B_t(\delta)} \left| h_n(x) - h_n(t) \right| = 0,$$

which follows by the same argument as used in Theorem 3 and the fact that any sequence of centers $(t_n)_{n \in \mathbb{N}}$ has a convergent subsequence by the compactness of $[0, 1]^d$,

$$\frac{\mathbb{E} S_n(t_n, r_n)}{\gamma_n(t_n, r_n)} = \sqrt{n} \sqrt{\lambda_n(1 - \lambda_n)} \frac{\int_{[0,1]^d} \psi_{t_n r_n}(x) \phi_n(x) dx}{\left[\int_{[0,1]^d} \psi_{t_n r_n}(x)^2 dx \right]^{1/2}} (1 + o(1)). \tag{34}$$

Using the approximation in (33) we obtain analogously

$$\sqrt{2 \log \left(\gamma_n(t_n, r_n)^{-2} \right)} = \left[2 \log \left(1 \Big/ O(1) \int_{[0,1]^d} \psi_{t_n r_n}(x)^2 dx \right) \right]^{1/2}. \tag{35}$$

Recall that $\psi = \psi_{\beta, K}$ with K the bound of the support. Standard calculation shows that the bounded L_2-norm of γ_β implies

$$\frac{\left| \int \psi_{t_n r_n; \beta, K}(x) \phi_n(x) dx \right|}{\left[\int \psi_{t_n r_n; \beta, K}(x)^2 dx \right]^{1/2}} = \frac{\left| \int \psi_{t_n r_n; \beta}(x) \phi_n(x) dx \right|}{\left[\int \psi_{t_n r_n; \beta}(x)^2 dx \right]^{1/2}} (1 + c_K)$$
$$\text{with } c_K \to 0 \text{ as } K \to \infty,$$

but note that the total variation $TV(\psi_{\beta, K})$ is increasing in K. Define now $\delta_n := (1 + \delta) c(\beta, L) \rho_{m,n}$. Then by its construction, $\delta_n \psi_{t_n r_n; \beta} \in \mathcal{H}_d\left(\beta, L; \mathbb{R}^d\right)$. Moreover, by the closedness in L_2 and the convexity of the sets $\left\{\phi \in \mathcal{H}_d(\beta, L; \mathbb{R}^d) : \phi(t_n) \geq \delta_n\right\}$ and $\left\{\phi \in \mathcal{H}_d(\beta, L; \mathbb{R}^d) : \phi(t_n) \leq -\delta_n\right\}$, it results finally from convex analysis and the definition of γ_β that

$$\frac{\left| \int \psi_{t_n r_n; \beta}(x) \phi_n(x) dx \right|}{\left[\int \psi_{t_n r_n; \beta}(x)^2 dx \right]^{1/2}} \geq \frac{\delta_n^{-1} \|\delta_n \psi_{t_n r_n; \beta}\|_2^2}{\|\psi_{t_n r_n; \beta}\|_2} = \delta_n r_n^{d/2} \|\gamma_\beta\|_2.$$

Combining (33–35), one verifies for the expression of the right hand side in (32) that it possesses the approximation

$$\begin{aligned}(32) &= O_p(1) + \sqrt{n}\sqrt{\lambda_n(1-\lambda_n)}\delta_n r_n^{d/2}\|\gamma_\beta\|_2\,(1+c_K) \\ &\quad - \left(\frac{2d}{2\beta+d}\right)^{1/2}\sqrt{\log\left(n/\log n\right)} \\ &= O_p(1) + \sqrt{\log n}\left(\frac{2dL^{d/\beta}}{(2\beta+d)\|\gamma_\beta\|_2^2}\right)^{1/2} L^{-d/(2\beta)}\|\gamma_\beta\|_2(1+c_K)(1+\delta)^{d/(2\beta)+1} \\ &\quad - \left(\frac{2d}{2\beta+d}\right)^{1/2}\sqrt{\log\left(n/\log n\right)},\end{aligned}$$

which goes to infinity for K sufficiently large. If there exists a sequence $(\phi_n)_{n\in\mathbb{N}}$ of admissible alternatives such that $\limsup_{n\to\infty}\mathbb{P}_{(m,n,p_n,q_n)}\left(T_n > \kappa_\alpha(\underline{X})\right) < 1$, there exists by the considerations above a subsequence (for simplicity also denoted by (n)) along which $\|\phi_n\|_{\sup}$ stays uniformly bounded away from zero. But the bounds (30) and (31) show that

$$\frac{\mathbb{E}S_n(\widehat{t}_n,\widehat{r}_n) - \mathbb{E}S_n(t_n,r_n)}{\gamma_n(\widehat{t}_n,\widehat{r}_n)} = O\left(n^{-1/d+1/2}\right)\left(1+o_p(1)\right),$$

as well as the logarithmic correction term $C_{\widehat{j}_n\widehat{k}_n n}$ are in this case of smaller order than $|\mathbb{E}S_n(t_n,r_n)|$, which concludes the proof by contradiction. □

Proof of Theorem 5 Following the considerations of the proof of Theorem 4, it has to be established that there exist random sequences $\left(\widehat{j}_{ni},\widehat{k}_{ni}\right)_{n\in\mathbb{N}}$ with $B_{X_{\widehat{j}_{ni}}}\left(\left\|X_{\widehat{j}_{ni}} - X_{\widehat{k}_{ni}}\right\|_2\right) \subset J_i$, $i=1,\ldots,k$, such that for any sequence of alternatives as formulated in Theorem 5 and any fixed $K>0$

$$\liminf_{n\to\infty}\ \mathbb{P}_{(m,n,p_n,q_n)}\left(T_{\widehat{j}_{ni}\widehat{k}_{ni}} - C_{\widehat{j}_{ni}\widehat{k}_{ni}} > \kappa_\alpha(\underline{X})\right) = 1, \quad i=1,\ldots,k.$$

Then the result follows because the finite intersection of sets with asymptotic probability equal to 1 has asymptotically mass 1 as well. Inspired by the arguments in [32] for the univariate regression context, we first establish the following:

For $\phi_n \in \mathcal{H}_d\left(\beta, L; [0,1]^d\right)$ with $\|\phi_n\|_{\sup} \le 1$ and $x^* = \operatorname{argmax}_{x\in[0,1]^d}|\phi_n(x)|$, there exists some constant $c = c(\beta, L) > 0$ and a compact ball $B = B(\phi_n) \subset \mathbb{R}^d$ with center x^* such that

$$\lambda\left(B\cap[0,1]^d\right) \ge c|\phi_n(x^*)|^{d/\beta} \quad \text{and } \left|\phi_n(x)\right| \ge \frac{1}{2}\left|\phi_n(x^*)\right| \quad \text{for all } x\in B\cap[0,1]^d. \tag{36}$$

Assume that $\beta > 1$ (the above inequality is trivial in case $\beta \le 1$). With $j=(j_1,\ldots,j_d)$ we denote subsequently some multi-index, where $|j| = j_1+\cdots+j_d$ defines its length,

$x^j := \prod_{i=1}^d x_i^{j_i}$ and $D^j := \partial^{|j|}/[\partial x_1^{j_1}, \ldots, \partial x_m^{j_m}]$ the partial differential operator. Let $\phi \in \mathcal{H}_d\left(\beta, L; [0,1]^d\right)$ with $\|\phi\|_{\sup} = D > 0$. By the definition of the isotropic Hölder class we have $|\phi(x) - T_y^{(f)}(x)| \le L\|x-y\|_2^\beta$ $(\le L\sqrt{d}^\beta)$, which entails that $\sup_y \|T_y^{(f)}\|_{[0,1]^d} \le D + L\sqrt{d}^\beta$. In order to establish (36), note that for any polynomial $P = \sum_{|j|\le\lfloor\beta\rfloor} a_j x^j$, the topology induced by the metrics corresponding to the two norms $\|P\|_{(1)} = \sup_{x\in[0,1]^d}|P(x)|$ and $\|P\|_{(2)} := \max_j |a_j|$ respectively on the ring of polynomials of total degree at most $\lfloor\beta\rfloor$ on $[0,1]^d$ is the topology of uniform convergence, hence these two norms are equivalent. Consequently, the boundedness of the polynomial $T_y^{(f)}$ by $D + L\sqrt{d}^{\,\beta}$ uniformly in y implies that there exists some constant $C = C(\beta)$ such that $\|D^j\phi\|_{\sup} \le C\big(D+L\big)$ for all multi-indices j with $|j| \le \lfloor\beta\rfloor$. Now the Mean Value Theorem implies for some intermediate point $z \in \{x + t(x^* - x);\, 0 \le t \le 1\}$

$$\begin{aligned}\left|\phi(x) - \phi(x^*)\right| &= \left|\left(\nabla\phi(z)\right)^T\left(x - x^*\right)\right| \\ &\le \sqrt{d}\sup_{j:\,|j|=1}\left\|D^j\phi\right\|_{\sup}\|x - x^*\|_2 \\ &\le \sqrt{d}\,C\;(D+L)\,\|x-x^*\|_2.\end{aligned}$$

Thus,

$$|\phi(x)| \ge \frac{1}{2}|\phi(x^*)| \;\text{ for all } x \text{ in } B_{x^*}\left(\frac{D}{2\sqrt{d}C(D+L)}\right)\cap[0,1]^d.$$

If $\phi \in \mathcal{H}_d\left(\beta, L; [0,1]^d\right)$ with $\|\phi\|_{\sup} = \delta \le 1$, then the function g_δ, for $x \in [0,1]^d$ pointwise defined by $g_\delta(x) := \delta^{-1}\phi\left(\delta^{1/\beta}x + x^*\right)\cdot I\left\{\delta^{1/\beta}x + x^* \in [0,1]^d\right\}$ is element of $\mathcal{H}_d(\beta, L; \mathrm{supp}(g_\delta))$ with $\|g_\delta\|_{\sup} = 1$. Note that $\mathrm{supp}(g_\delta)$ is a convex set. Therefore, the above considerations imply that $|\phi(x)| \ge \delta/2$ on

$$B_{x^*}\left(\frac{\delta^{1/\beta}}{2\sqrt{d}C(1+L)}\right)\cap[0,1]^d.$$

But then its Lebesgue measure is always greater than $c|\delta|^{d/\beta}$ for some constant $c = c(\beta, L)$, independent of δ and x^*, hence (36) is established.

Let now $\beta_i, L_i \in (0,\infty)$ fixed but arbitrary, $J_i \subset [0,1]^d$ some nondegenerate rectangle, $\tilde{\phi}_n = p_n - q_n$ a sequence of functions with $\tilde{\phi}_{n|J_i} \in \mathcal{H}_d\left(\beta_i, L_i; J_i\right)$. It has to be shown that there exists a universal constant $k_i = k_i(\beta_i, L_i, c)$ such that $T_{\widehat{j}_n\widehat{k}_n n} - C_{\widehat{j}_n\widehat{k}_n n} \to_{\mathbb{P}^{\otimes m}\otimes\mathbb{Q}^{\otimes(n-m)}} \infty$ whenever $\|\tilde{\phi}_n\|_{J_i} \ge k_i\rho_{m,n}$. First, we choose a compact ball $B_i(\tilde{\phi}_n)$ with center $x_i^* := \mathrm{argmax}_{t\in J_i}|\tilde{\phi}_n(t)|$ satisfying $\lambda(B_i(\tilde{\phi}_n)\cap J_i) \ge c|\tilde{\phi}_n(x_i^*)|^{d/\beta}$ and (36). Let the couple $(\widehat{t}_n, \widehat{r}_n) := \left(X_{\widehat{j}_n}, \|X_{\widehat{j}_n} - X_{\widehat{k}_n}\|_2\right)$ be defined by

$$\left(\widehat{j}_n, \widehat{k}_n\right) := \operatorname*{argmin}_{j,k\in\{1,\ldots,n\}}\lambda\left(B_{X_j}\left(\|X_j - X_k\|_2\right)\,\Delta\, B_i(\tilde{\phi}_n)\right).$$

Consulting the proof of Theorem 4, this definition of $(\widehat{t}_n, \widehat{r}_n)$ allows for an approximation as in (32). Since $|\tilde{\phi}_n(x)| \geq 2^{-1}\|\tilde{\phi}_n\|_{J_i}$ for all $x \in B_i(\tilde{\phi}_n) \cap B_{\widehat{t}_n}(\widehat{r}_n) \cap J_i$,

$$\begin{aligned}\frac{\mathbb{E}S_n(t_n, r_n)}{\gamma_n(t_n, r_n)} &\geq \frac{1}{2}\|\tilde{\phi}_n\|_{J_i}\sqrt{n}\frac{\sqrt{\lambda_n(1-\lambda_n)}}{\sqrt{\max_x h_n(x)}}\left[\mathbb{E}\lambda\left(B_i(\tilde{\phi}_n) \cap B_{\widehat{t}_n}(\widehat{r}_n) \cap [0,1]^d\right)\right]^{1/2} \\ &\geq C\|\tilde{\phi}_n\|_{J_i}^{(\beta+d/2)/\beta}\sqrt{n}\,(1+o(1))\end{aligned}$$

for some universal constant $C = C(K, (\lambda_n)_{n\in\mathbb{N}}) > 0$. Now the asserted result is easily deduced for a sufficiently large constant k_i. □

Proof of Theorem 6 (i) Let (p, q, m, n) be such that $h = h_n = I_{[0,1]^d}$ and ϕ_n the sequence of piecewise constant functions on $[0, 1]^d$ with $\phi_n(z) = c_n/\sqrt{n\delta_n^d}$ for $z \in B_x(\delta_n)$, $\phi_n(z) = -c_n/\sqrt{n\delta_n^d}$ for $z \in B_x(\kappa\delta_n) \setminus B_x(\delta_n)$ and equals zero otherwise, where $\kappa = \kappa(d) > 1$ is such that $\lambda(B_x(\kappa\delta_n) \setminus B_x(\delta_n)) = \lambda(B_x(\delta_n))$ and $0 < c_n \leq \sqrt{n\delta_n^d}$. Then

$$\begin{aligned}\log\left(\frac{d\mathbb{P}_{(m,n,p_n,q_n)}}{d\mathbb{P}_{(m,n,h,h)}}(\underline{X})\right) &= \sum_{i=1}^{m}\log\left(1+(1-m/n)\phi_n(X_i)\right) \\ &\quad + \sum_{j=m+1}^{n}\log\left(1-(m/n)\phi_n(X_j)\right)\end{aligned}$$

with $(X_k)_{k\in\mathbb{N}}$ iid uniformly distributed on $[0, 1]^d$ under the hypothesis $\phi_n \equiv 0$. Note that

$$\begin{aligned}&\mathcal{L}_{(m,n,h,h)}\left[\log\left(\frac{d\mathbb{P}_{(m,n,p_n,q_n)}}{d\mathbb{P}_{(m,n,h,h)}}(\underline{X})\right)\right] \\ &= \mathcal{L}\left[\sum_{i=1}^{N_m}\log\left(1+(1-m/n)\frac{c_n}{\sqrt{n\delta_n^d}}R_i\right) + \sum_{j=1}^{N_{n-m}}\log\left(1-(m/n)\frac{c_n}{\sqrt{n\delta_n^d}}R'_j\right)\right]\end{aligned}$$

with $(R_k)_{k\in\mathbb{N}}$, $(R'_k)_{k\in\mathbb{N}}$, N_m and N_{n-m} all independent, where $(R_k)_{k\in\mathbb{N}}$ and $(R'_k)_{k\in\mathbb{N}}$ are sequences of iid Rademacher variables and

$$N_m \sim \mathrm{Bin}\left(m, V\kappa^d\delta_n^d\right), \quad N_{n-m} \sim \mathrm{Bin}\left(n-m, V\kappa^d\delta_n^d\right) \quad \text{and} \quad V = \pi^{d/2}\Gamma(d/2+1).$$

Suppose first that $n\delta_n^d \not\to \infty$. By extracting a subsequence if necessary we may assume that $m/n \to \lambda \in (0, 1)$, $c_n/\sqrt{n\delta_n^d} \to c \in [0, 1]$ and $n\delta_n^d \to V^{-1}\kappa^{-d}\gamma$. Then, with $\to_w$ denoting weak convergence,

$$\mathcal{L}_{(m,n,h,h)}\left[\log\left(\frac{d\mathbb{P}_{(m,n,p_n,q_n)}}{d\mathbb{P}_{(m,n,h,h)}}(\underline{X})\right)\right] \longrightarrow_w \mathbb{Q} \tag{37}$$

with the convolution

$$\mathbb{Q} := \left(\sum_{k=0}^{\infty} p_{\gamma(1-\lambda)}(k)\mathcal{L}\left(\sum_{i=1}^{k}\log\left(1-\lambda c R_i\right)\right)\right) \star \left(\sum_{k'=0}^{\infty} p_{\gamma\lambda}(k')\mathcal{L}\left(\sum_{j=1}^{k'}\log\left(1+(1-\lambda)c R'_j\right)\right)\right)$$

and the Poisson weights $p_\mu(k) := e^{-\mu}\mu^k/k!$. Since $\int e^z d\,\mathbb{Q}(z) = 1$, we can apply Le Cam's notion of contiguity ([20], Chapter 3) to conclude that

$$\limsup_{n\to\infty} \mathbb{E}_{(m,n,p_n,q_n)}\psi_n(\underline{X}) < 1.$$

Consequently $n\delta_n^d \to \infty$. Now assume that $n\delta_n^d \to \infty$ but $c_n \not\to \infty$. Without loss of generality we may assume that $c_n \to c'/\sqrt{V\kappa^d} \in [0,\infty)$. Then Lindeberg's CLT entails that (37) holds true with

$$\mathbb{Q} := \mathcal{N}\left(-\frac{(1-\lambda)\lambda^2 c'^2}{2},\ (1-\lambda)\lambda^2 c'^2\right) \star \mathcal{N}\left(-\frac{\lambda(1-\lambda)^2 c'^2}{2},\ \lambda(1-\lambda)^2 c'^2\right).$$

Again, the limiting distribution satisfies $\int e^z d\,\mathbb{Q}(z) = 1$, whence $c_n \to \infty$.

(ii) We begin as in the proof of Theorem 4, but with $t_n := x$, $r_n := \delta_n$ and $c_n/\sqrt{n\delta_n^d}$ the size of the lower-bounding plateau. Without loss of generality we may assume that $\phi_n(z) \geq c_n/\sqrt{n\delta_n^d}$ for all $z \in B_x(\delta_n)$, i.e. $\phi_n \in \mathcal{J}_+^{(m,n)}(c_n, x, \delta_n)$. Adjusting (26–30) yields

$$\frac{\left|\mathbb{E}\left(S_n(\widehat{x}_n,\widehat{\delta}_n) - S_n(x,\delta_n)\right)\right|}{\widehat{\gamma}_n(\widehat{x}_n,\widehat{\delta}_n)} = O\left(\delta_n^{-1}n^{-1/d}c_n\right)\left(1+o_p(1)\right).$$

The arguments of the proof of Theorem 4 apply again and lead to the expansion

$$T_{\widehat{j}_n\widehat{k}_n n} - C_{\widehat{j}_n\widehat{k}_n n} = O_p(1) + O\left(\delta_n^{-1}n^{-1/d}c_n\right)\left(1+o_p(1)\right) + \frac{\mathbb{E}S_n(x,\delta_n)}{\gamma_n(x,\delta_n)}\left(1+o_p(1)\right) - \sqrt{2\log\left(\gamma_n(x,\delta_n)^{-2}\right)}, \tag{38}$$

while with the same reasoning as in the proof of Theorem 5

$$\frac{\mathbb{E}S_n(x,\delta_n)}{\gamma_n(x,\delta_n)} \geq C(1+o(1))\sqrt{n}\frac{c_n}{\sqrt{n\delta_n^d}}\delta_n^{d/2} = C(1+o(1))c_n$$

for some constant $C = C\ (d, (\lambda_n)_{\in\mathbb{N}}) > 0$ and $\sqrt{2\log\left(\gamma_n(x,\delta_n)^{-2}\right)} = O(1)\sqrt{\log(1/\delta_n)}$. Thus, if $\sqrt{\log(1/\delta_n)}/c_n \to 0$ and $n\delta_n^d \to \infty$, (38) goes to infinity and the result follows. □

Acknowledgments Lutz Dümbgen's contribution to the decoupling subsequent to an extended discussion in Bern is gratefully acknowledged. Furthermore, I want to thank three unknown referees and an associate editor for their valuable comments and careful reading of the manuscript.

References

1. Behnen, K., Neuhaus, G., Ruymgaart, F.: Two sample rank estimators of optimal nonparametric score-functions and corresponding adaptive rank statistics. Ann. Stat. **11**, 588–599 (1983)
2. Belomestny, D., Spokoiny, V.: Spatial aggregation of local likelihood estimates with application to classification. Ann. Stat. **35**, 2287–2311 (2007)
3. Bennett, G.: Probability inequalities for sums of independent random variables. J. Am. Stat. Assoc. **57**, 33–45 (1962)
4. Butucea, C., Tribouley, K.: Nonparametric homogeneity tests. J. Stat. Plann. Inference **136**, 597–639 (2006)
5. Donoho, D.: Statical estimation and optimal recovery. Ann. Stat. **22**, 238–270 (1994a)
6. Donoho, D.: Asymptotic minimax risk for sup-norm loss—solution via optimal recovery. Probab. Theory Relat. Fields **99**, 145–170 (1994b)
7. Ducharme, G.R., Ledwina, T.: Efficient and adaptive nonparametric test for the two-sample problem. Ann. Stat. **31**, 2036–2058 (2003)
8. Dudley, R.M., Giné, E., Zinn, J.: Uniform and universal Glivenko–Cantelli classes. J. Theoret. Probab. **4**, 485–510 (1991)
9. Dümbgen, L.: Application of local rank tests to nonparametric regression. J. Nonparametric Stat. **14**, 511–537 (2002)
10. Dümbgen, L., Spokoiny, V.G.: Multiscale testing of qualitative hypotheses. Ann. Stat. **29**, 124–152 (2001)
11. Dümbgen, L., Walther, G.: Multiscale inference about a density. Ann. Stat. **36**, 1758–1785; accompanying technical report, version 2. Available at http://arxiv.org/abs/0706.3968 (2008)
12. Eubank, R.L., Hart, J.D.: Testing goodness-of-fit in regression via order selection criteria. Ann. Stat. **20**, 1412–1425 (1992)
13. Fan, J.: Test of significance based on wavelet thresholding and Neyman's truncation. J. Am. Stat. Assoc. **91**, 674–688 (1996)
14. Hájek, J., Šidak, Z.: Theory of Rank Tests. Academic press, New York (1967)
15. Gijbels, I., Heckmann, N.: Nonparametric testing for a monotone hazard function via normalized spacings. J. Nonparametric Stat. **16**, 463–477 (2004)
16. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
17. Ingster, Y.: Asymptotically minimax testing of nonparametric hypotheses. Prob. Theory Math. Stat. **I**, 553–574 (1987)
18. Janic-Wróblewska, A., Ledwina, T.: Data driven rank test for two-sample problem. Scand. J. Stat. **27**, 281–297 (2000)
19. Klemelä, J., Tsybakov, A.: Sharp adaptive estimation of linear functionals. Ann. Stat. **29**, 1567–1600 (2001)
20. Le Cam, L., Yang, G.: Asymptotics in Statistics: Some Basic Concepts. Springer, New York (2000)
21. Ledwina, T., Kallenberg, W.C.M.: Consistency and Monte Carlo simulation of a data-driven version of smooth goodness-of-fit tests. Ann. Stat. **23**, 1594–1608 (1995)
22. Ledwina, T.: Data-driven version of Neyman's smooth test of fit. J. Am. Stat. Assoc. **89**, 1000–1005 (1994)
23. Leonov, S.L.: On the solution of an optimal recovery problem and its applications in nonparametric Statics. Math. Methods Stat. **4**, 476–490 (1997)
24. Leonov, S.L.: Remarks on extremal problems in nonparametric curve estimation. Stat. Probab. Lett. **43**, 169–178 (1999)
25. Lepski, O., Tsybakov, A.: Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. Probab. Theory Relat. Fields **117**, 17–48 (2000)
26. Neuhaus, G.: H_0-contiguity in nonparametric testing problems and sample Pitman efficiency. Ann. Stat. **10**, 575–582 (1982)

27. Neuhaus, G.: Local asymptotics for linear rank Statics with estimated score functions. Ann. Stat. **15**, 491–512 (1987)
28. Nussbaum, M.: Asymptotic equivalence of density estimation and Gaussian white noise. Ann. Stat. **24**, 2399–2430 (1996)
29. de la Peña, V.H.: A bound on the moment generating function of a sum of dependent variables with an application to simple sampling without replacement. Ann. Inst. H. Poincaré Probab. Stat. **30**, 197–211 (1994)
30. de la Peña, V.H.: A general class of exponential inequalities for martingales and ratios. Ann. Prob. **27**, 537–564 (1999)
31. Pollard, D.: Convergence of Stochastic Processes. Springer, Heidelberg (1984)
32. Rohde, A.: Adaptive goodness-of-fit tests based on signed ranks. Ann. Stat. **36**, 1346–1374 (2008)
33. Rufibach, K., Walther, G.: A block criterion for multiscale inference about a density, with applications to other multiscale problems. J. Comp. Graph. Stat. (2009, to appear)
34. Serfling, R.J.: Probability inequalities for the sum of sampling without replacement. Ann. Stat. **2**, 39–48 (1974)
35. Shorack, G.R., Wellner, J.A.: Empirical Processes with Applications to Statistics. Wiley, New York (1986)
36. Spokoiny, V.: Adaptive hypothesis testing using wavelets. Ann. Stat. **24**, 2477–2498 (1996)
37. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, Heidelberg (1996)
38. Walther, G.: Optimal and fast detection of spatial clusters with scan statistics. Ann. Stat. **38** (2010, to appear)