The Annals of Statistics
2021, Vol. 49, No. 5, 2457–2490
https://doi.org/10.1214/20-AOS2041

OPTIMAL RATES FOR INDEPENDENCE TESTING VIA *U*-STATISTIC PERMUTATION TESTS

BY THOMAS B. BERRETT[1], IOANNIS KONTOYIANNIS[2,*] AND RICHARD J. SAMWORTH[2,†]

[1]*Department of Statistics, University of Warwick, tom.berrett@warwick.ac.uk*
[2]*Statistical Laboratory, Centre for Mathematical Sciences, *yiannis@maths.cam.ac.uk; †r.samworth@statslab.cam.ac.uk*

We study the problem of independence testing given independent and identically distributed pairs taking values in a σ-finite, separable measure space. Defining a natural measure of dependence $D(f)$ as the squared L^2-distance between a joint density f and the product of its marginals, we first show that there is no valid test of independence that is uniformly consistent against alternatives of the form $\{f : D(f) \geq \rho^2\}$. We therefore restrict attention to alternatives that impose additional Sobolev-type smoothness constraints, and define a permutation test based on a basis expansion and a U-statistic estimator of $D(f)$ that we prove is minimax optimal in terms of its separation rates in many instances. Finally, for the case of a Fourier basis on $[0, 1]^2$, we provide an approximation to the power function that offers several additional insights. Our methodology is implemented in the `R` package `USP`.

1. Introduction. Independence is a fundamental concept in both probability and statistics; it distinguishes the former from a mere branch of measure theory, and underpins both statistical theory and the way practitioners think about modelling. For statisticians, it is frequently important to ascertain whether or not assumptions of independence are realistic, both to determine whether certain theoretical properties of procedures can be expected to hold, and to assess the goodness-of-fit of a statistical model.

Classical approaches to independence testing have focused on the simple setting of univariate Euclidean data, and have often only had power against restricted classes of alternatives. These include tests based on Pearson's correlation (e.g., Pearson (1920)), Spearman's rank correlation coefficient (Spearman (1904)), Kendall's tau (Kendall (1938)) and Hoeffding's D (Hoeffding (1948)). However, motivated by a desire to handle the more general data types that are ubiquitous in modern-day practice, as well as to have power against broader classes of alternatives, the topic of independence testing has undergone a renaissance in recent years. Since, in settings of interest, no uniformly most powerful test exists (see Theorem 1 below and the surrounding discussion), several different perspectives and new tests have emerged, such as those based on the Hilbert–Schmidt independence criterion (Gretton et al. (2005), Li and Yuan (2019), Meynaoui et al. (2019), Pfister et al. (2018)), distance covariance (Sejdinovic et al. (2013), Székely, Rizzo and Bakirov (2007)), optimal transport and multivariate ranks (Deb and Sen (2019), Shi, Drton and Han (2020)), copula transforms (Kojadinovic and Holmes (2009)), sample space partitioning (Heller et al. (2016)) and nearest neighbour methods (Berrett and Samworth (2019)). For practical studies with discrete data, Pearson's chi-squared independence test remains ubiquitous in the scientific literature, despite the drawback that its size guarantees rely on pointwise asymptotic arguments that may fail to control the Type I error in finite samples; see Section 7 below. Independence tests

Received January 2020; revised November 2020.
MSC2020 subject classifications. 62G09, 62G10.
Key words and phrases. Independence testing, permutation tests, minimax separation rates, U-statistics, Stein's method.

for continuous data are also common in applications such as linguistics (Nguyen and Eisenstein (2017)), genetics (Steuer et al. (2002)) and public health (Reshef et al. (2011)), and have also been applied to functional data arising from credit card activity and geomagnetic records (Gabrys and Kokoszka (2007)).

This plethora of approaches gives rise to natural theoretical questions about the fundamental statistical difficulty of independence testing. In the setting where the marginal distributions are both univariate, early asymptotic results on minimax separation rates over certain classes of alternatives are given in Ingster (1989), Ermakov (1990) and Ingster (1996). There has been recent work on multivariate settings (Li and Yuan (2019), Meynaoui et al. (2019)), but many open questions remain.

Another issue with several of the tests mentioned above is that the asymptotic distribution of the test statistic under the null hypothesis of independence depends on unknown features of the relevant marginal distributions, so it is difficult to obtain an appropriate critical value. An attractive approach, therefore, is to use a permutation test, which uses permutations to mimic the null behaviour of the test statistic. Though the principle has been known for many decades (e.g., Pitman (1938); Fisher ((1935), Chapter 21)), permutation tests are becoming increasingly popular in modern statistics and machine learning (e.g., A/B testing), due to their ease of use and their guaranteed finite-sample Type I error control across the entire null hypothesis parameter space, assuming only that the data are exchangeable under the null. Besides (unconditional) independence testing, they have also been studied in problems such as conditional independence testing (Berrett et al. (2020)), two-sample testing (Chung and Romano (2013)) and changepoint analysis (Antoch and Hušková (2001)). We also highlight the work of Chung and Romano (2016), who show how a permutation test based on a U-statistic can extend the scope of the two-sample Wilcoxon test to null hypotheses of the form $\theta(P, Q) = \theta_0$ (where P and Q are the two underlying distributions), providing pointwise asymptotic size guarantees in general, and exact size guarantees when $P = Q$. For an overview of the study of permutation tests see, for example, Lehmann and Romano (2005) and Pesarin and Salmaso (2010).

In the context of permutation tests for independence, Romano (1989) considered a class of plug-in test statistics of the form $T_n = n^{1/2}\delta(\hat{P}_n, \hat{P}_n^X \hat{P}_n^Y)$, where $\delta(P, Q) = \sup_{V \in \mathcal{V}} |P(V) - Q(V)|$ for a Vapnik–Chervonenkis class of sets $\mathcal{V}$, and where $\hat{P}_n$, $\hat{P}_n^X$ and $\hat{P}_n^Y$ are the empirical distributions of the data pairs and their marginals, respectively. Fixing a sequence of alternatives (P_n), he showed that, under the condition that $\mathbb{P}_{P_n}(T_n \leq t) \to H(t)$ for some continuous function H, the asymptotic power of his permutation test coincides with that of the test that uses the true critical value. In the case of univariate marginals, Albert ((2015), Chapter 4) provides upper bounds on the minimax separation over Besov spaces using a test based on aggregrating many permutation tests. See also Albert et al. (2015) and Berrett and Samworth (2019). Despite these aforementioned works, however, there remains great interest in understanding better the power properties of permutation tests in the context of nonparametric independence testing. Indeed, shortly after an earlier version of this paper was made publicly available, Kim, Balakrishnan and Wasserman (2020) posted a complementary study of the power properties of permutation tests, with a greater focus on concentration inequalities for the test statistics as opposed to distributional results.

In this paper, we study the problem of independence testing in a general framework, where our data consist of independent copies of a pair (X, Y) taking values in a separable measure space $\mathcal{X} \times \mathcal{Y}$, equipped with a σ-finite measure μ. Assuming that the joint distribution of (X, Y) has a density f with respect to μ, we may define a measure of dependence $D(f)$, given by the squared $L^2(\mu)$ distance between the joint density and the product of its marginal densities. This satisfies the natural requirement that $D(f) = 0$ if and only if X and Y are independent. In fact, however, our hardness result in Theorem 1 reveals that it is not

possible to construct a valid independence test with nontrivial power against all alternatives satisfying a lower bound on $D(f)$. This motivates us to introduce classes satisfying an additional Sobolev-type smoothness condition as well as boundedness conditions on the joint and marginal densities.

The first main goal of this work is to determine the minimax separation rate of independence testing over these classes, and to this end, we define a new permutation test of independence based on a U-statistic estimator of $D(f)$. We refer to this test hereafter as the USP test, short for U-Statistic permutation test. Theorem 2 in Section 3 provides a very general upper bound on the separation rate of independence testing; the framework is broad enough to include both discrete and absolutely continuous data, as well as data that may take values in infinite-dimensional spaces, for instance. We show how the bound can be simplified in many special cases of interest, and, in Section 4, how to construct adaptive versions of our tests that incur only a small loss in effective sample size. Moreover, in Section 5, we go on to provide matching lower bounds in several instances, allowing us to conclude that our USP test attains the minimax optimal separation rate for independence testing in such settings. In Section 6, we elucidate an approximation to the power function of our test at local alternatives, thereby providing a very detailed description of its properties. Numerical properties of our procedure are studied in Section 7: we first show how an alternative representation of our test statistic dramatically reduces the computational complexity of our procedure, and then present a simulation study that reveals the strong empirical performance of our test in different settings. Section 8 provides further discussion. Proofs of some of our main results are given in Section 9; for other results, designated with (BKS(2020)), the proofs appear in the supplementary material (Berrett, Kontoyiannis and Samworth (2021)), where auxiliary results (labelled with an 'S' prefix) are also given. Our methodology is implemented in the `R` package `USP` (Berrett, Kontoyiannis and Samworth (2020)).

Further contributions of this paper are to introduce new sets of tools for studying both permutation tests and U-statistics; we believe both will find application beyond the scope of this work, in particular because many popular measures of dependence, such as distance covariance and the Hilbert–Schmidt independence criterion, can be estimated using U-statistics. Specifically, in the proof of Theorem 2, we develop moment bounds for U-statistics computed on permuted data sets. Moreover, Proposition 18 provides normal approximation error bounds in Wasserstein distance for degenerate U-statistics computed on permuted data sets (using Stein's method, and extending earlier results for unpermuted data, e.g., de Jong (1990), Döbler and Peccati (2019), Rinott and Rotar (1997)), and is the basis for our local power function result (Theorem 16). Finally, our minimax lower bound (Lemma 11) may also be of independent interest, in that it provides a general approach to constructing priors over the alternative hypothesis class whose distance from the null can be explicitly bounded.

Notation. We write $\mathbb{N} = \{1, 2, 3, \ldots\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and, for $n \in \mathbb{N}$, let $[n] := \{1, \ldots, n\}$. We also write $[\infty] := \mathbb{N}$. We write $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \leq Cb$, and write, for example, $a \lesssim_x b$ if there exists $C > 0$, depending only on x, such that $a \leq Cb$. We similarly define $a \gtrsim b$ and $a \gtrsim_x b$, and write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$, as well as $a \asymp_x b$ if $a \lesssim_x b$ and $a \gtrsim_x b$.

Let $\mathcal{S}_n$ denote the set of permutations of $[n]$. For a measure space $(\mathcal{Z}, \mathcal{C}, \nu)$ define $L^2(\nu) := \{f : \mathcal{Z} \to \mathbb{R} : \int_{\mathcal{Z}} f^2 \, d\nu < \infty\}$, with corresponding inner product $\langle f, g \rangle_{L^2(\nu)} := \int_{\mathcal{Z}} fg \, d\nu$ and norm $\|f\|_{L^2(\nu)} := \langle f, f \rangle_{L^2(\nu)}^{1/2}$. For a function $f : \mathcal{Z} \to \mathbb{R}$, we write $\|f\|_\infty := \sup_{z \in \mathcal{Z}} |f(z)| \in [0, \infty]$; if it is also $\mathcal{C}$-measurable, we write $\operatorname{ess\,inf}_{z \in \mathcal{Z}} f(z) := \sup\{y \in \mathbb{R} : \nu(f^{-1}(-\infty, y)) = 0\}$.

Let Φ denote the standard normal distribution function and let $\bar{\Phi} := 1 - \Phi$. Given a sample of independent and identically distributed random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ and a

$\sigma(X_1, Y_1, \ldots, X_n, Y_n)$-measurable random variable W, we write $\mathbb{E}_P(W)$ or $\mathbb{E}_f(W)$ for the expectation of W when (X_1, Y_1) has distribution P or density function f. Given probability measures μ and ν on $\mathcal{Z}$, we write $d_{\mathrm{TV}}(\mu, \nu) := \sup_{C \in \mathcal{C}} |\mu(C) - \nu(C)|$ for their total variation distance and, if both μ and ν are absolutely continuous with respect to another measure λ, then we write $d_{\chi^2}(\mu, \nu) = \{\int_{\mathcal{Z}} \frac{(d\mu/d\lambda)^2}{d\nu/d\lambda} d\lambda - 1\}^{1/2}$ for the square root of their χ^2-divergence. If $\mathcal{Z} = \mathbb{R}$, then we write

$$d_{\mathrm{W}}(\mu, \nu) := \inf_{(X,Y)\sim(\mu,\nu)} \mathbb{E}|X - Y|$$

for the Wasserstein distance between μ and ν, where the infimum is taken over all pairs (X, Y) defined on the same probability space with $X \sim \mu$ and $Y \sim \nu$. When $\mathcal{Z} = \mathbb{R}$, we will also write

$$d_{\mathrm{K}}(\mu, \nu) := \sup_{z\in\mathbb{R}} \big|\mu((-\infty, z]) - \nu((-\infty, z])\big|$$

for the Kolmogorov distance between μ and ν. If $X \sim \mu$ and $Y \sim \nu$, we sometimes write $d_{\mathrm{W}}(X, Y)$ and $d_{\mathrm{K}}(X, Y)$ as shorthand for $d_{\mathrm{W}}(\mu, \nu)$ and $d_{\mathrm{K}}(\mu, \nu)$ respectively. We use $\triangle$ to denote the symmetric difference operation on sets, so that $A \triangle B := (A \cap B^c) \cup (A^c \cap B)$.

Finally, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $q \in [1, \infty)$, we let $\|x\|_q := (\sum_{j=1}^d |x_j|^q)^{1/q}$, with the shorthand $\|x\| := \|x\|_2$, and for a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we let $\|A\|_{\mathrm{op}} := \sup_{x:\|x\|\le 1} \|Ax\|$ and $\|A\|_{\mathrm{F}} := \{\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} A_{jk}^2\}^{1/2}$ denote its operator and Frobenius norms, respectively.

2. Problem set-up and preliminaries. Let $(\mathcal{X}, \mathcal{A}, \mu_X)$ and $(\mathcal{Y}, \mathcal{B}, \mu_Y)$ be separable,[1] σ-finite measure spaces. In discrete settings, that is, when $\mathcal{X}$ is countable, μ_X would typically be counting measure on $\mathcal{X}$; more generally, it may be the relevant Lebesgue measure when $\mathcal{X}$ is a Euclidean space, or an appropriate measure on basis coefficients in infinite-dimensional examples such as Example 8 below. Both $L^2(\mu_X)$ and $L^2(\mu_Y)$ are then separable Hilbert spaces,[2] so there exist orthonormal bases $(p_j^X)_{j\in\mathcal{J}}$ and $(p_k^Y)_{k\in\mathcal{K}}$ of $L^2(\mu_X)$ and $L^2(\mu_Y)$ respectively, where $\mathcal{J}$ and $\mathcal{K}$ are countable indexing sets. Writing $\mu := \mu_X \otimes \mu_Y$ for the product measure on $\mathcal{X} \times \mathcal{Y}$, the product space $L^2(\mu)$ is also a separable Hilbert space,[3] and has an orthonormal basis given by $(p_{jk})_{j\in\mathcal{J},k\in\mathcal{K}}$, where $p_{jk}(\cdot, *) := p_j^X(\cdot) p_k^Y(*)$.

We may now define the subset $\mathcal{F}$ of $L^2(\mu)$ that consists of all density functions, that is,

$$\mathcal{F} := \Big\{ f \in L^2(\mu) : \operatorname*{ess\,inf}_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x, y) \ge 0, \int_{\mathcal{X}\times\mathcal{Y}} f \, d\mu = 1 \Big\}.$$

Given $f \in \mathcal{F}$, we may define the marginal density f_X by

$$f_X(x) := \int_{\mathcal{Y}} f(x, y) \, d\mu_Y(y),$$

and we may analogously define f_Y. From now on, we will work over the restricted space $\mathcal{F}^* := \{f \in \mathcal{F} : f_X \in L^2(\mu_X), f_Y \in L^2(\mu_Y)\}$, though we note that when μ_X and μ_Y are finite measures, we have $\mathcal{F}^* = \mathcal{F}$. For $f \in \mathcal{F}^*$, $j \in \mathcal{J}$ and $k \in \mathcal{K}$, we may define the coefficients

$$a_{jk}(f) := \int_{\mathcal{X}\times\mathcal{Y}} f p_{jk} \, d\mu, \qquad a_{j\bullet}(f) := \int_{\mathcal{X}} f_X p_j^X \, d\mu_X, \qquad a_{\bullet k}(f) := \int_{\mathcal{Y}} f_Y p_k^Y \, d\mu_Y.$$

[1] Recall that we say a measure space $(\mathcal{Z}, \mathcal{C}, \nu)$ is *separable* if, when equipped with the pseudo-metric $d(A, B) := \nu(A \triangle B)$, it has a countable dense subset.

[2] Since we were unable to find this precise statement in the literature, we provide a proof in Lemma S2.

[3] Likewise, we prove this statement in Lemma S3.

Then

$$f = \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} a_{jk}(f)p_{jk}, \qquad f_X = \sum_{j\in\mathcal{J}} a_{j\bullet}(f)p_j^X, \qquad f_Y = \sum_{k\in\mathcal{K}} a_{\bullet k}(f)p_k^Y.$$

We may therefore define the measure of dependence

$$\begin{aligned} D(f) &:= \int_{\mathcal{X}\times\mathcal{Y}} \{f(x,y) - f_X(x)f_Y(y)\}^2 \, d\mu(x,y) \\ &= \sum_{j\in\mathcal{J},k\in\mathcal{K}} \{a_{jk}(f) - a_{j\bullet}(f)a_{\bullet k}(f)\}^2, \end{aligned}$$

which, for $(X,Y)\sim f$, has the property that $D(f)=0$ if and only if $X \perp\!\!\!\perp Y$.

Given a sample $(X_1,Y_1),\dots,(X_n,Y_n)$ of independent and identically distributed copies of the pair (X,Y), we wish to test the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ of independence. A randomised independence test is measurable function $\psi : (\mathcal{X}\times\mathcal{Y})^n \to [0,1]$, with the interpretation that, after observing $(X_1,Y_1,\dots,X_n,Y_n) = (x_1,y_1,\dots,x_n,y_n)$, we reject H_0 with probability $\psi(x_1,y_1,\dots,x_n,y_n)$. We write Ψ for the set of all such randomised independence tests. Further, define the null space $\mathcal{P}_0$ as the set of all distributions on $\mathcal{X}\times\mathcal{Y}$ of pairs (X,Y) such that $X \perp\!\!\!\perp Y$, and, for a given $\alpha\in(0,1)$, define the set of valid size-α independence tests

$$\Psi(\alpha) := \Big\{\psi\in\Psi : \sup_{P\in\mathcal{P}_0} \mathbb{E}_P(\psi) \le \alpha\Big\}. \tag{1}$$

The first part of Theorem 1 below provides a preliminary result on the hardness of the independence testing problem when the alternative hypothesis H_1 consists of all densities $f\in\mathcal{F}^*$ of (X,Y) that satisfy a lower bound constraint on $D(f)$. In fact, the result can be stated more generally, allowing in addition for the possibility of a constraint on the smoothness of the alternatives that we consider. To this end, for an array $\theta = (\theta_{jk})_{j\in\mathcal{J},k\in\mathcal{K}} \in [0,\infty]^{\mathcal{J}\times\mathcal{K}}$, we define

$$S_\theta(f) := \sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \theta_{jk}^2\{a_{jk}(f) - a_{j\bullet}(f)a_{\bullet k}(f)\}^2.$$

Observe that when $\theta = 0_{\mathcal{J}\times\mathcal{K}}$, any nonnegative upper bound on $S_\theta(f)$ becomes vacuous, so that no smoothness constraint is imposed. This definition of smoothness is motivated by the nonparametric statistics literature (e.g., Laurent (1996)). An attractive feature is that, in contrast to some prior literature, smoothness is only imposed on the difference between the joint density and the product of the marginals, rather than on the individual densities themselves; Meynaoui et al. (2019) also adopt a similar approach to ours in this respect. At a high level, the first part of Theorem 1 is inspired by the work of Janssen (2000) and Shah and Peters (2020) on the hardness of goodness-of-fit testing and conditional independence testing respectively, though the proofs are completely different. The second part complements the first, as discussed below. Note that when μ_X is a probability measure, the constant function 1 belongs to $L^2(\mu_X)$, so can be included in our basis (as below).

THEOREM 1. *Suppose that μ_X and μ_Y are probability measures and that there exist $j_0\in\mathcal{J}$ and $k_0\in\mathcal{K}$ such that $p_{j_0}^X(x) = p_{k_0}^Y(y) = 1$ for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$. Let $n\in\mathbb{N}$ and $\alpha\in[0,1]$, and let $\psi\in\Psi$ be such that $\mathbb{E}_{p_{j_0k_0}}(\psi)\le\alpha$. Let $\theta = (\theta_{jk})_{j\in\mathcal{J},k\in\mathcal{K}} \in [0,\infty)^{\mathcal{J}\times\mathcal{K}}$ be given and, for $T\in[0,\infty)$, define*

$$\mathcal{M}_\theta(T) := \big\{(j,k)\in(\mathcal{J}\setminus\{j_0\})\times(\mathcal{K}\setminus\{k_0\}) : \theta_{jk}\le T\big\}.$$

Let $\underline{\theta} := \inf_{j\in\mathcal{J},k\in\mathcal{K}}\theta_{jk}$. *Then, for any* $\epsilon > 0$, *any* $\rho \in (0, 1/\sup_{j,k}\|p_{jk}\|_\infty]$ *and any* $r \in (\underline{\theta}\rho, \infty)$, *there exists* $f^* \in \mathcal{F}$ *with* $S_\theta(f^*) \le r^2$ *and* $D(f^*) = \rho^2$ *such that*

$$\mathbb{E}_{f^*}(\psi) \le \alpha + \epsilon + \left[\frac{\{(1+\rho^2)^n - 1\}\alpha}{|\mathcal{M}_\theta(r/\rho)|}\right]^{1/2}.$$

Moreover, there exists a permutation test $\psi_{f^*} \in \Psi(\alpha)$ *such that given any* $\beta \in (0, 1-\alpha)$, *we can find* $C = C(\alpha, \beta) > 0$ *with the property that* $\mathbb{E}_{f^*}(\psi_{f^*}) \ge 1 - \beta$ *whenever* $n > C/\rho^2$.

As a first conclusion, we can draw from Theorem 1, consider taking $\theta = 0_{\mathcal{J}\times\mathcal{K}}$, so that $|\mathcal{M}_\theta(r/\rho)| = (|\mathcal{J}| - 1)(|\mathcal{K}| - 1)$. In this case, Theorem 1 shows that in infinite-dimensional problems (where $|\mathcal{J} \times \mathcal{K}| = \infty$) with probability measures as base measures, there are no valid tests of independence that have uniformly nontrivial power against alternatives of the form $\{f \in \mathcal{F} : D(f) \ge \rho^2\}$, at least for $\rho > 0$ sufficiently small. The second part of the theorem then implies that in this setting there is no uniformly most powerful test. Thus, to develop a theory of minimax separation rates for independence testing, it is necessary to make additional assumptions about the structure of the alternative hypothesis. More generally, under the conditions of Theorem 1, whenever the set $\mathcal{M}_\theta(r/\rho)$ is infinite, there are no valid uniformly nontrivial independence tests against alternatives $f \in \mathcal{F}$ with $S_\theta(f) \le r^2$ and $D(f) \ge \rho^2$. We will therefore assume the following in much of our subsequent work:

(A1) The sets $\{(j,k) \in \mathcal{J} \times \mathcal{K} : \theta_{jk} \le T\}$ are finite for each $T \in (0, \infty)$.

Motivated by Theorem 1 above, for $\Xi := [0,\infty]^{\mathcal{J}\times\mathcal{K}} \times (0,\infty) \times [1,\infty)$, for $\xi = (\theta, r, A) \in \Xi$ and for $\rho > 0$, we will consider the space of alternatives given by

$$\mathcal{F}_\xi(\rho) := \big\{f \in \mathcal{F} : D(f) \ge \rho^2, S_\theta(f) \le r^2, \max\big(\|f\|_\infty, \|f_X\|_\infty, \|f_Y\|_\infty\big) \le A\big\}.$$

Although we make assumptions about the smoothness of our alternatives, we will not make any assumptions about the null distributions, and the fact that we are using a permutation test will guarantee uniform, nonasymptotic control of the probability of Type I error. In other words, we will prove that our test ψ belongs to $\Psi(\alpha)$ in (1).

Given $n \in \mathbb{N}$, $\alpha \in (0,1)$, $\xi = (\theta, r, A) \in \Xi$ and $\rho > 0$ we define the minimax risk with respect to $\mathcal{F}_\xi(\rho)$ as

$$\mathcal{R}(n, \alpha, \xi, \rho) := \alpha + \inf_{\psi\in\Psi(\alpha)} \sup_{f\in\mathcal{F}_\xi(\rho)} \mathbb{E}_f(1-\psi),$$

with the convention that $\mathcal{R}(n,\alpha,\xi,\rho) := \alpha$ if $\mathcal{F}_\xi(\rho) = \varnothing$. If we are also given a desired probability of Type II error $\beta \in (0, 1-\alpha)$, then we can consider the minimax separation radius

$$\rho^*(n, \alpha, \beta, \xi) := \inf\big\{\rho > 0 : \mathcal{R}(n,\alpha,\xi,\rho) \le \alpha + \beta\big\}.$$

3. Upper bounds. We now introduce our USP test that will allow us to establish upper bounds on the minimax separation ρ^*. This is based on a U-statistic estimator of $D(f)$ with kernel

$$\begin{aligned} &h\big((x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\big) \\ &\quad := \sum_{(j,k)\in\mathcal{M}} \big\{p_{jk}(x_1,y_1)p_{jk}(x_2,y_2) - 2p_{jk}(x_1,y_1)p_{jk}(x_2,y_3) \\ &\qquad + p_{jk}(x_1,y_2)p_{jk}(x_3,y_4)\big\}, \end{aligned} \tag{2}$$

where $\mathcal{M} \subseteq \mathcal{J} \times \mathcal{K}$ is a truncation set to be chosen later. The motivation for this definition comes from the observation that for any $f \in \mathcal{F}^*$ and when $\mathcal{M} = \mathcal{J} \times \mathcal{K}$, we have

$$\mathbb{E}_f\big\{h\big((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)\big)\big\} = D(f); \tag{3}$$

moreover, as we will see in the proof of Theorem 2 below, whenever Π is a uniformly random element of $\mathcal{S}_n$ that is independent of the data, we have

$$\mathbb{E}_f\{h((X_1, Y_{\Pi(1)}), (X_2, Y_{\Pi(2)}), (X_3, Y_{\Pi(3)}), (X_4, Y_{\Pi(4)}))\} = 0.$$

To reduce the effects of noise accumulation in the estimation of the summands, it will typically be necessary to choose $\mathcal{M}$ in (2) to be a proper subset of $\mathcal{J} \times \mathcal{K}$. The equality in (3) then no longer holds exactly for every $f \in \mathcal{F}^*$, but an appropriate choice of $\mathcal{M}$ allows us to control the bias-variance trade-off.

For $m \geq 2$, let $\mathcal{I}_m := \{(i_1, \ldots, i_m) \in [n]^m : i_1, \ldots, i_m \text{ all distinct}\}$. For $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n) \in \mathcal{Y}^n$, it is convenient to define

$$\mathcal{T}_{x,y} := \{(x_i, y_i) : i \in [n]\},$$

and for $\sigma \in \mathcal{S}_n$, set $\mathcal{T}_{x,y}^{(\sigma)} := \{(x_i, y_{\sigma(i)}) : i \in [n]\}$. Given independent pairs $\mathcal{T}_{X,Y} := \{(X_i, Y_i) : i = 1, \ldots, n\}$ with $n \geq 4$, we consider the test statistic

$$\hat{D}_n = \hat{D}_n^{\mathcal{M}}(\mathcal{T}_{X,Y}) := \frac{1}{4!\binom{n}{4}} \sum_{(i_1,\ldots,i_4)\in\mathcal{I}_4} h((X_{i_1}, Y_{i_1}), \ldots, (X_{i_4}, Y_{i_4})).$$

To define the critical value for our test, let $B \in \mathbb{N}$ and generate an independent sequence of uniform random permutations $\Pi_1, \ldots, \Pi_B$ taking values in $\mathcal{S}_n$, independently of $\mathcal{T}_{X,Y}$. It is important to note that we can typically choose B to be much smaller than $n!$ (the number of distinct permutations in $\mathcal{S}_n$); indeed, the choice $B = 99$ is common for permutation tests. For each $b \in [B]$, we construct the null statistics

$$\hat{D}_n^{(b)} := \hat{D}_n^{\mathcal{M}}(\mathcal{T}_{X,Y}^{(\Pi_b)}). \tag{4}$$

Finally, we can define the p-value

$$P := \frac{1 + \sum_{b=1}^{B} \mathbb{1}_{\{\hat{D}_n \leq \hat{D}_n^{(b)}\}}}{1 + B}, \tag{5}$$

and reject the null hypothesis if $P \leq \alpha$. Formally, this corresponds to the randomised test $\psi_\alpha \in \Psi$, given by

$$\psi_\alpha(\mathcal{T}_{x,y}) := \mathbb{P}\left(1 + \sum_{b=1}^{B} \mathbb{1}_{\{\hat{D}_n^{\mathcal{M}}(\mathcal{T}_{x,y}) \leq \hat{D}_n^{\mathcal{M}}(\mathcal{T}_{x,y}^{(\Pi_b)})\}} \leq (1 + B)\alpha\right),$$

where the only randomness here is in the permutations $\Pi_1, \ldots, \Pi_B$. Then, on observing $\mathcal{T}_{X,Y}$, we do indeed reject H_0 with probability $\psi_\alpha(\mathcal{T}_{X,Y})$. Under the null hypothesis, the sequence of data sets $\mathcal{T}_{X,Y}, \mathcal{T}_{X,Y}^{(\Pi_1)}, \ldots, \mathcal{T}_{X,Y}^{(\Pi_B)}$ is exchangeable, so every ordering of the components of $(\hat{D}_n^{\mathcal{M}}(\mathcal{T}_{X,Y}), \hat{D}_n^{\mathcal{M}}(\mathcal{T}_{X,Y}^{(\Pi_1)}), \ldots, \hat{D}_n^{\mathcal{M}}(\mathcal{T}_{X,Y}^{(\Pi_B)}))$ is equally likely if we break ties uniformly at random. In particular, the rank of $\hat{D}_n^{\mathcal{M}}(\mathcal{T}_{X,Y})$ among these $B + 1$ observations, which is a lower bound on the numerator in (5), is uniformly distributed on $\{1, \ldots, B + 1\}$, so $\psi_\alpha \in \Psi(\alpha)$.

A naive implementation of the test has computational complexity $O(n^4 B|\mathcal{M}|)$, due to the need to calculate fourth-order U-statistics. However, using an alternative representation of our test statistic inspired by Song et al. (2012), we can reduce the complexity to $O(n^2 B(|\mathcal{J}_0| + |\mathcal{K}_0|)$ when $\mathcal{M} = \mathcal{J}_0 \times \mathcal{K}_0$. See Section 7.1 for further details.

The following theorem provides a general upper bound on the minimax separation rate, and is obtained using the above test.

THEOREM 2. *Fix $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$ and let $\xi = (\theta, r, A) \in \Xi$. Then there exists $C = C(\alpha, \beta, A) > 0$ such that when $n \geq 16$, we have*

$$\rho^*(n, \alpha, \beta, \xi) \leq C \inf_{\mathcal{M} \subseteq \mathcal{J} \times \mathcal{K}} \max\left\{ \frac{r}{\inf\{\theta_{jk} : (j,k) \notin \mathcal{M}\}}, \frac{\min(\|h\|_\infty^{1/2}, |\mathcal{M}|^{1/4})}{n^{1/2}}, \frac{1}{n^{1/2}} \right\}.$$

An explicit upper bound showing the dependence of C on its arguments is given in (29) in the proof of Theorem 2. To give a heuristic explanation of the terms in the bound in Theorem 2, observe that in order for our test to have high power, we want ρ^2 to dominate the sum of the bias of the test statistic and its standard deviation under the null. The first term represents this bias, which is induced by truncating the sum in (2) to indices that lie in $\mathcal{M}$. The second term arises from bounding the variance of our U-statistic in terms of the symmetrised kernel $\bar{h}$, defined formally in (16) below. More precisely, under the null, our test statistic is a degenerate U-statistic, that is, $\mathbb{E}\{\bar{h}((x,y),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4))\} = 0$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, so its variance can be bounded above by a constant multiple of $n^{-2}\operatorname{Var}\{\bar{h}((X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4))\}$. This latter expression can in turn be bounded by $\min(\|h\|_\infty^2, |\mathcal{M}|)/n^2$. The final term in the maximum represents the parametric rate of convergence, and is generally unavoidable.

3.1. *Discrete case.* As a first application of Theorem 2, consider the relatively simple problem of testing independence with discrete data, where for some $J, K \in \mathbb{N} \cup \{\infty\}$ we have $\mathcal{X} = [J]$ and $\mathcal{Y} = [K]$ and we take μ_X and μ_Y to be the counting measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively. For $j, x \in [J]$ and $k, y \in [K]$, we can define the basis functions $p_j^X(x) := \mathbb{1}_{\{x=j\}}$ and $p_k^Y(y) := \mathbb{1}_{\{y=k\}}$. In this case, we have $\|h\|_\infty \leq 2$ independently of $\mathcal{M}$, and we may take $\mathcal{M} = [J] \times [K]$ so that there is in fact no truncation and our test statistic is an unbiased estimator of D. Note here that, since μ_X and μ_Y are not probability measures, Theorem 1 does not apply, and we will see that no structural assumptions are necessary on the alternative hypothesis. Indeed, we take $\xi = (0_{[J]\times[K]}, 1, 1) \in \Xi$, so that our alternative hypothesis class is simply

$$\mathcal{F}_\xi(\rho) = \left\{ f \in \mathcal{F} : \sum_{j \in [J], k \in [K]} \{f(j,k) - f_X(j) f_Y(k)\}^2 \geq \rho^2 \right\}.$$

The following result is a straightforward corollary of Theorem 2, noting that the cases where $n < 16$ can be handled using the fact that $\rho^*(n, \alpha, \beta, \xi) \leq 2^{1/2}$ for all n.

COROLLARY 3. *Fix $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$. Then there exists $C = C(\alpha, \beta) \in (0, \infty)$ such that*

$$\rho^*(n, \alpha, \beta, \xi) \leq C n^{-1/2}.$$

This behaviour should be contrasted with that found in Diakonikolas and Kane (2016), where the strength of the dependence is measured by the L_1 distance rather than the L_2 distance, and where the minimax optimal separation rates depend on the alphabet sizes; in fact, they are given by $\frac{J^{1/4}K^{1/4}}{n^{1/2}} \max(1, J^{1/4}/n^{1/4}, K^{1/4}/n^{1/4})$.

In fact, in this discrete setting, we can give a relatively simple, explicit form for the test. To this end, for $j \in [J]$, $k \in [K]$, let $N_{jk} := \sum_{i=1}^n \mathbb{1}_{\{X_i = j, Y_i = k\}}$, let $N_{j+} := \sum_{k=1}^K N_{jk}$ and let $N_{+k} := \sum_{j=1}^J N_{jk}$. Then, omitting terms that only depend on N_{j+} and N_{+k} (and hence remain fixed under permutation, so are irrelevant for the test), our test statistic becomes

$$\hat{T}_n := \frac{1}{n(n-3)} \sum_{j=1}^J \sum_{k=1}^K \left(N_{jk} - \frac{N_{j+}N_{+k}}{n} \right)^2 - \frac{4}{n^2(n-2)(n-3)} \sum_{j=1}^J \sum_{k=1}^K N_{jk} N_{j+} N_{+k}.$$

Thus, the test statistic can be computed using only the contingency table counts, as opposed to the original data. Moreover, the permutated data sets may also be generated using only these counts: indeed, writing $N^{(1)}_{jk}$ for the (j,k)th cell count under an independent, uniformly random permutation of the original data, we have

$$\mathbb{P}\big((N^{(1)}_{jk}) = (n_{jk}) | \mathcal{T}_{X,Y}\big) = \frac{(\prod_{j=1}^{J} N_{j+}!)(\prod_{k=1}^{K} N_{+k}!)}{n! \prod_{j=1}^{J}\prod_{k=1}^{K} n_{jk}!},$$

whenever (n_{jk}) is such that $\sum_{k=1}^{K} n_{jk} = N_{j+}$ for all $j \in [J]$ and $\sum_{j=1}^{J} n_{jk} = N_{+k}$ for all $k \in [K]$. This formula simplifies the computation of the permuted data sets, and one can sample from this distribution using Patefield's algorithm (Patefield (1981)), which is implemented in the `R` function `r2dtable`.

3.2. *Sobolev and infinite-dimensional examples.* To apply Theorem 2 in general, when a useful bound on $\|h\|_\infty$ is not available, we instead control the right-hand side by controlling $|\mathcal{M}|$. We remark that, when there exist $j_0 \in \mathcal{J}$ and $k_0 \in \mathcal{K}$ such that $p^X_{j_0}(x) = p^Y_{k_0}(y) = 1$ for all x, y, then $a_{j_0k} = a_{\bullet k}$, $a_{jk_0} = a_{j\bullet}$, $a_{j_0\bullet} = 1$, $a_{\bullet k_0} = 1$, so the $j = j_0$ and $k = k_0$ terms do not contribute to the value of $D(\cdot)$ and $S_\theta(\cdot)$ does not depend on $(\theta_{j_0k})_{k\in\mathcal{K}}$ or $(\theta_{jk_0})_{j\in\mathcal{J}}$. Thus the choice of $\mathcal{M}$ in the definition of $\hat{D}^{\mathcal{M}}_n$ does not need to contain any (j,k) with $j = j_0$ or $k = k_0$. For notational convenience, we will adopt the convention that, in such cases, $\theta_{jk} = \infty$ if either $j = j_0$ or $k = k_0$. When **(A1)** holds it is possible to arrange $\{\theta_{jk} : \theta_{jk} < \infty\}$ in increasing order, so that there exists a bijection $\omega : \mathbb{N} \to \{(j,k) : \theta_{jk} < \infty\}$ such that $\theta_{\omega(1)} \le \theta_{\omega(2)} \le \cdots$. Given $t \in (0,\infty)$, define[4]

$$m_0(t) := \min\big\{m \in \mathbb{N} : m^{1/2}\theta^2_{\omega(m)} > t\big\}.$$

We can now simplify the conclusion of Theorem 2 under **(A1)**.

COROLLARY 4. *Fix $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$ and let $\xi = (\theta, r, A) \in \Xi$. Assume* **(A1)**. *Then there exists $C = C(\alpha,\beta,A) > 0$ such that when $n \ge 16$, we have*

$$\rho^*(n,\alpha,\beta,\xi) \le C \inf_{m\in\mathbb{N}} \max\left\{\frac{r}{\theta_{\omega(m)}}, \frac{m^{1/4}}{n^{1/2}}\right\} \le \frac{C m_0^{1/4}(nr^2)}{n^{1/2}}. \tag{6}$$

We now further specialise our upper bound by making a specific choice of $\mathcal{J}$, $\mathcal{K}$ and weights $(\theta_{jk} : j \in \mathcal{J}, k \in \mathcal{K})$; such a choice yields a concrete upper bound on the minimax rate of independence testing for densities lying in a Sobolev space, as we illustrate in the example that follows. See Example 13 and Proposition 14 below for a discussion of optimality of this bound.

COROLLARY 5. *Fix $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$, fix $d_X, d_Y \in \mathbb{N}$ and s_X, s_Y, r, $A > 0$. Writing $\mathcal{J} = \mathbb{N}_0^{d_X}$, $\mathcal{K} = \mathbb{N}_0^{d_Y}$, set $\theta_{jk} = \|j\|_1^{s_X} \vee \|k\|_1^{s_Y}$ whenever $j \ne 0_{[d_X]}$ and $k \ne 0_{[d_Y]}$ and $\theta_{jk} = \infty$ otherwise. Then, with $\theta = \{\theta_{jk} : j \in \mathcal{J}, k \in \mathcal{K}\}$, there exists $C = C(d_X, d_Y, \alpha, \beta, A) > 0$ such that if $n \ge 16$ and $nr^2 \ge 1$, then*

$$\rho^*(n,\alpha,\beta,\xi) \le C\left(\frac{r^d}{n^{2s}}\right)^{1/(4s+d)},$$

where $d := d_X + d_Y$, $s := d/(d_X/s_X + d_Y/s_Y)$ and $\xi = (\theta, r, A)$.

[4]Here and throughout, if $\omega(m) = (j,k)$, we interpret $\theta_{\omega(m)}$ as θ_{jk} and $p_{\omega(m)}$ as p_{jk}.

The upper bound in Corollary 5 is obtained using our U-statistic permutation test. Here, $\theta_{\omega(m)} \asymp_{s_X,s_Y,d_X,d_Y} m^{s/d}$, so we can balance the two terms in the maximum in Corollary 4 by taking $\mathcal{M} = \{\omega(1), \ldots, \omega(m)\}$ with $m \asymp_{s_X,s_Y,d_X,d_Y} (nr^2)^{2d/(4s+d)}$. A natural application of (a minor variant of) this corollary is to absolutely continuous data, which for simplicity we restrict to lie in $[0,1]^{d_X} \times [0,1]^{d_Y}$. In this setting, the Fourier basis functions are an obvious choice.

EXAMPLE 6. Let $\mathcal{X} = [0,1]^{d_X}$ and $\mathcal{Y} = [0,1]^{d_Y}$, equipped with d_X-dimensional Lebesgue measure μ_X and d_Y-dimensional Lebesgue measure μ_Y, respectively. Taking $\mathcal{J} := \{(a,m) : a \in \{0,1\}, m \in \mathbb{N}_0^{d_X}\} \setminus \{(1, 0_{[d_X]})\}$ and $\mathcal{K} := \{(a,m) : a \in \{0,1\}, m \in \mathbb{N}_0^{d_Y}\} \setminus \{(1, 0_{[d_Y]})\}$, we can define the orthonormal Fourier basis functions[5] for $L^2([0,1]^{d_X})$ given by $p^X_{0,0} := 1$ and for $m = (m_1, \ldots, m_{d_X}) \neq 0_{[d_X]}$,

$$p^X_{a,m}(x_1, \ldots, x_{d_X}) := 2^{1/2} \operatorname{Re}\left(e^{-a\pi i/2} \prod_{\ell=1}^{d_X} e^{-2\pi i m_\ell x_\ell} \right). \tag{7}$$

The Fourier basis functions $\{p^Y_{a,m} : (a,m) \in \mathcal{K}\}$ for $L^2([0,1]^{d_Y})$ are defined similarly, but with d_Y replacing d_X. For $j = (a_X, m_X) \in \mathcal{J}$, $k = (a_Y, m_Y) \in \mathcal{K}$ and $s_X, s_Y > 0$, we can then take $\theta_{jk} = \|m_X\|_1^{s_X} \vee \|m_Y\|_1^{s_Y}$, $\theta = \{\theta_{jk} : j \in \mathcal{J}, k \in \mathcal{K}\}$ and $\xi = (\theta, r, A) \in \Xi$ to conclude from Corollary 4 that $\rho^*(n, \alpha, \beta, \xi) \leq C(r^d/n^{2s})^{1/(4s+d)}$ when $n \geq 16$ and $nr^2 \geq 1$, as in Corollary 5.

We mention here that Li and Yuan (2019) and Meynaoui et al. (2019) consider Gaussian kernel-based Hilbert–Schmidt Independence Criterion tests of independence in similar Sobolev settings to that in Example 6. Assuming the same level of Sobolev smoothness s for both the joint and marginal distributions, Li and Yuan (2019) show that the critical consistency level is of order $n^{-2s/(4s+d)}$ over tests that have asymptotically nominal size. Meynaoui et al. (2019) obtain the same rate in a nonasymptotic setting and only impose smoothness conditions on the difference between the joint and marginal distributions, at the expense of restricting the smoothness s to be at most 2, and having bounded null densities.

In fact, Corollary 4 also provides explicit upper bounds for certain infinite-dimensional models. Corollary 7 below illustrates this for a particular choice of $\mathcal{J}$, $\mathcal{K}$ and weights $(\theta_{jk} : j \in \mathcal{J}, k \in \mathcal{K})$.

COROLLARY 7 (BKS(2020)). *Let* $\mathbb{N}_0^{<\infty} := \{m = (m_1, m_2, \ldots) \in \mathbb{N}_0^{\mathbb{N}} : \sum_{\ell=1}^{\infty} \mathbb{1}_{\{m_\ell \neq 0\}} < \infty\}$, *and let* $\mathcal{J} = \mathcal{K} := \{(a,m) : a \in \{0,1\}, m \in \mathbb{N}_0^{<\infty}\} \setminus \{(1,0)\}$. *For* $m = (m_1, m_2, \ldots) \in \mathbb{N}_0^{<\infty}$, *write* $|m| := \max_{\ell \in \mathbb{N}} \ell^2 m_\ell$, *and if* $j = (a,m) \in \mathcal{J}$, *write* $|j| := |m|$. *For* $j \in \mathcal{J}$, $k \in \mathcal{K}$ *with* $|j| \wedge |k| > 0$, *and* $s_X, s_Y > 0$, *set*

$$\theta_{jk} = \exp(s_X |j|^{1/2}) \vee \exp(s_Y |k|^{1/2}),$$

and if either $|j| = 0$ *or* $|k| = 0$ *then set* $\theta_{jk} = \infty$. *Define the increasing function* $M : [0,\infty) \to [0,\infty)$ *by*

$$M(t) := \exp\left(\sum_{\ell=1}^{\infty} \log\left(1 + \left\lfloor \frac{t}{\ell^2} \right\rfloor\right)\right) - 1$$

[5]The fact that these functions form an orthonormal basis for $L^2([0,1]^{d_X})$ follows from a very similar (in fact, slightly simpler) argument to that given in Lemma S4, which relates to Example 8 below. The main difference is that in this example our functions are defined on finite-dimensional spaces.

and write

$$m_{0,s_X,s_Y}(t) := \min\left\{m \in \mathbb{N} : M\left(\frac{\log^2(t/m^{1/2})}{4s_X^2}\right)M\left(\frac{\log^2(t/m^{1/2})}{4s_Y^2}\right) < \frac{m}{4}\right\}.$$

(*i*) *Fix* $\alpha, \beta \in (0,1)$ *such that* $\alpha + \beta < 1$ *and fix* $r, s_X, s_Y, A > 0$. *Then, with* $\xi = (\theta, r, A) \in \Xi$ *there exists* $C = C(\alpha, \beta, s_X, s_Y, A) > 0$ *such that when* $n \geq 16$ *and* $nr^2 \geq C$ *we have*

$$\rho^*(n, \alpha, \beta, \xi) \leq \frac{C m_{0,s_X,s_Y}^{1/4}(nr^2)}{n^{1/2}}.$$

(*ii*) *Writing* $s := 2/(s_X^{-1} + s_Y^{-1})$ *and given* $\epsilon \in (0, 4s)$, *there exists* $C' = C'(s_X, s_Y, \epsilon) > 0$ *such that when* $t \geq C'$ *we have*

$$t^{\frac{2c_0-\epsilon}{2s+c_0}} \leq m_{0,s_X,s_Y}(t) \leq t^{\frac{2c_0+\epsilon}{2s+c_0}},$$

where $c_0 := \sum_{\ell=1}^{\infty}\{\ell^{-1/2} - (\ell+1)^{-1/2}\}\log(1+\ell) = 1.65\ldots$.

We will see in Proposition 15 below that the rate given in the first part of Corollary 7 is optimal in regimes of n and r of interest in the context of Example 8 below. The second part of the corollary shows that, if we ignore subpolynomial factors in nr^2, then we have $\rho^*(n, \alpha, \beta, \xi) \lesssim_{\alpha,\beta,s_X,s_Y,A} (r^{c_0}/n^s)^{1/(2s+c_0)}$. By comparison with Corollary 5, we can therefore interpret c_0 as the 'effective dimension' of each of $\mathcal{X}$ and $\mathcal{Y}$, when θ is selected in this way.

EXAMPLE 8. As an application of Corollary 7, consider the infinite-dimensional setting where $\mathcal{X} = \mathcal{Y} = [0,1]^{\mathbb{N}} := \{(x_1, x_2, \ldots) : x_\ell \in [0,1] \text{ for all } \ell \in \mathbb{N}\}$, equipped with the Borel σ-algebra in the product topology, and where $\mu_X = \mu_Y$ is the distribution of an infinite sequence $(U_1, U_2, \ldots)$ of Unif$[0,1]$ random variables. It follows from an application of the Stone–Weierstrass theorem (see Lemma S4) that an orthonormal basis for $L^2(\mu_X)$ is then given by $\{p_{a,m}^X(\cdot) : (a,m) \in \mathcal{J}\}$, where $p_{0,0}^X := 1$ and for $m \neq 0_{\mathbb{N}}$,

$$p_{a,m}^X(x_1, x_2, \ldots) := 2^{1/2}\,\mathrm{Re}\left(e^{-a\pi i/2}\prod_{\ell=1}^{\infty} e^{-2\pi i m_\ell x_\ell}\right).$$

We may take the same basis for $L^2(\mu_Y)$, so that $p_{a,m}^Y = p_{a,m}^X$ for all $a \in \{0,1\}$ and $m \in \mathbb{N}_0^{<\infty}$. Then Corollary 7 provides an upper bound on the minimax separation rate of independence testing in this example.

4. Adaptation. The practical implementation of our USP tests requires a choice of the truncation set $\mathcal{M}$. The optimal choice of $\mathcal{M}$, which yields the separation rates described in the previous section, typically depends on both θ and r, which may be unknown in practice. In this section, we therefore describe adaptive versions of our tests, that do not require knowledge of any unknown parameters and whose minimax risk can be shown in many cases to be only slightly inflated compared with the optimal tests. Our initial setting is rather general, but assumes that $\mathcal{J} \times \mathcal{K}$ has an ordering that is respected by every θ considered. Since this assumption does not hold in the setting of Corollary 5 unless $s_X = s_Y$ (as the relative magnitudes of s_X and s_Y affect the ordering of θ), we also illustrate the way in which this assumption can be relaxed, so that it remains possible to adapt to both of the unknown parameters separately in this Sobolev example.

To describe this initial setting, let $\omega : \mathbb{N} \to \mathcal{J} \times \mathcal{K}$ be injective, and, for a given $\theta_0 > 0$, let $\Theta(\omega, \theta_0) \subseteq [0,\infty]^{\mathcal{J}\times\mathcal{K}}$ denote the set of all $\theta = (\theta_{jk})_{j\in\mathcal{J},k\in\mathcal{K}}$ such that ω is a bijection from $\mathbb{N}$ to $\{(j,k) \in \mathcal{J} \times \mathcal{K} : \theta_{jk} < \infty\}$ and

$$\theta_0 \leq \theta_{\omega(1)} \leq \theta_{\omega(2)} \leq \cdots.$$

Here, ω denotes an ordering of $\mathcal{J} \times \mathcal{K}$ that ranks the importance of departures from independence in each direction. In our Sobolev example with $s_X = s_Y$, we could take ω to be any ordering of $(\mathbb{N}_0^{d_X} \setminus \{0_{[d_X]}\}) \times (\mathbb{N}_0^{d_Y} \setminus \{0_{[d_Y]}\})$ such that, writing $(j_m, k_m) := \omega(m)$, we have that $\max(\|j_1\|_1, \|k_1\|_1) \leq \max(\|j_2\|_1, \|k_2\|_1) \leq \cdots$. Taking $\gamma := \lceil 2 \log_2 n \rceil$, let $K_* := \{2^j : j \in [\gamma]\}$. Our adaptive procedure can now be described as follows. Given a desired Type II error probability $\beta \in (0, 1-\alpha)$, for each $m \in K_*$, carry out the permutation test from Section 3 with $\mathcal{M} = \{\omega(1), \ldots, \omega(m)\}$ and $B \geq 2(\frac{\gamma}{\alpha\beta} - 1)$ to yield p-values $p^{(1)}, \ldots, p^{(\gamma)}$. If $\min_{i \in [\gamma]} p^{(i)} < \alpha/\gamma$, then we reject H_0. As we have applied a standard Bonferroni correction, the Type I error of this omnibus test is controlled at the level α. The following result concerns its power.

PROPOSITION 9. *Let ω and $\theta_0 > 0$ be as above, and suppose that $\alpha \in (0,1)$, $\beta \in (0, 1-\alpha)$, $R_0 > 0$ and $A \geq 1$. Assume further that $f \in \mathcal{F}_\xi(\rho)$ for some $\xi = (\theta, r, A) \in \Xi$ with $\theta \in \Theta(\omega, \theta_0)$ and $r \in (0, R_0]$. Then there exists $C = C(\alpha, \beta, R_0, \theta_0, A) > 0$ such that we reject H_0 with probability at least $1-\beta$ whenever $n \geq C$ and*

$$\rho \geq C \max\left\{ \frac{\log^{1/4} n}{n^{1/2}} m_0^{1/4}\left(\frac{nr^2}{\log^{1/2} n}\right), \frac{\log^{1/2} n}{n^{1/2}} \right\}.$$

Comparing this result with the upper bound on the optimal separation in Corollary 4, we see that the price we pay for adaptation is that our effective sample size is reduced from n to $n/\log^{1/2} n$, at least provided that $m_0(nr^2/\log^{1/2} n) \gtrsim \log n$.

As mentioned above, in some applications, the set $\mathcal{J} \times \mathcal{K}$ will not be naturally ordered. Nevertheless, it may be the case that $\mathcal{J}$ and $\mathcal{K}$ are ordered separately, and in these cases it is still possible to adapt to unknown parameters. Consider the setting of Corollary 5, and define $\gamma_X := \lceil (2/d_X) \log_2 n \rceil$ and $K_X := \{2^j : j \in [\gamma_X]\}$ (with γ_Y and K_Y defined similarly). Similarly to before, given a desired Type II error probability $\beta \in (0, 1-\alpha)$, for each $(m_X, m_Y) \in K_X \times K_Y$, carry out the permutation test from Section 3 with $\mathcal{M} \equiv \mathcal{M}_{m_X, m_Y} = \{(j,k) \in \mathbb{N}_0^{d_X} \times \mathbb{N}_0^{d_Y} : 1 \leq \|j\|_1 \leq m_X, 1 \leq \|k\|_1 \leq m_Y\}$ and $B \geq 2(\frac{\gamma_X \gamma_Y}{\alpha\beta} - 1)$ to yield p-values $\{p^{(m_X m_Y)} : (m_X, m_Y) \in K_X \times K_Y\}$. This test again controls the Type I error at level α, and the following result shows that the critical separation radius is inflated by at most a logarithmic factor in n.

PROPOSITION 10. *Assume the setting of Corollary 5. Given $R_0 > 0$, suppose that $r \leq R_0$. Then there exists $C = C(\alpha, \beta, R_0, s_X, s_Y, d_X, d_Y, A) > 0$ such that we reject H_0 with probability at least $1-\beta$ whenever $n \geq C$ and*

$$\rho \geq C \left\{ \frac{r^d}{(n/\log n)^{2s}} \right\}^{1/(4s+d)}. \tag{8}$$

We note that a similar procedure could be applied in the setting of Corollary 7 to obtain an adaptive test there, too. Finally, in this section, we remark that in a more restricted setting it may be possible to improve the $\log n$ dependence to $\log\log n$ dependence using the very recent concentration results of Kim, Balakrishnan and Wasserman (2020).

5. Lower bounds. The goal of this section is to provide lower bounds to allow us to study the optimality of our USP test in different contexts. Slightly more precisely, we wish to determine the maximal departure from independence (measured in terms of our quantity $D(\cdot)$) that no valid independence test could reliably detect; equivalently, we seek the minimal separation level at which a valid independence test could have nontrivial power, uniformly

over the alternatives in our classes. To this end, we first prove a general lemma (Lemma 11 below), and then illustrate how it can be applied in different settings of interest.

Our lower bound results actually apply to a weaker notion of minimax risk, and will hold in settings where our base measures on $\mathcal{X}$ and $\mathcal{Y}$ are probability measures, and where our orthonormal bases contain the constant function 1, so that there exist $j_0 \in \mathcal{J}$ and $k_0 \in \mathcal{K}$ such that $p^X_{j_0}(x) = 1$ and $p_{k_0}(y) = 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Define

$$\tilde{\mathcal{R}}(n, \xi, \rho) := \inf_{\psi \in \Psi(1)} \Big\{ \mathbb{E}_{p_{j_0 k_0}}(\psi) + \sup_{f \in \mathcal{F}_\xi(\rho)} \mathbb{E}_f(1 - \psi) \Big\},$$

which only controls the sum of the error probabilities, and only considers a simple null, and further define

$$\tilde{\rho}^*(n, \gamma, \xi) := \inf\big\{ \rho > 0 : \tilde{\mathcal{R}}(n, \xi, \rho) \le \gamma \big\}.$$

Then, for any $n \in \mathbb{N}$, $\xi \in \Xi$, $\alpha, \beta \in (0, 1)$ with $\alpha + \beta < 1$, and $\rho \in (0, \infty)$, we have that $\tilde{\mathcal{R}}(n, \xi, \rho) \le \mathcal{R}(n, \alpha, \xi, \rho)$ and, therefore, also that $\tilde{\rho}^*(n, \alpha + \beta, \xi) \le \rho^*(n, \alpha, \beta, \xi)$. When our upper and lower bounds match, in terms of the separation rates, the problems of independence testing with simple and composite nulls are equivalent, and we have the same rates of convergence if we control the sum of error probabilities or if we control the error probabilities separately.

We are now in a position to state our main, general lower bound lemma. Recall that a Rademacher random variable ξ takes values 1 and -1, each with probability $1/2$.

LEMMA 11. *Suppose that μ_X and μ_Y are probability measures and that there exist $j_0 \in \mathcal{J}$ and $k_0 \in \mathcal{K}$ such that $p^X_{j_0}(x) = p^Y_{k_0}(y) = 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Let $(a_{jk})_{j \in \mathcal{J} \setminus \{j_0\}, k \in \mathcal{K} \setminus \{k_0\}}$ be a deterministic square-summable array of real numbers, let $(\xi_{jk})_{j \in \mathcal{J} \setminus \{j_0\}, k \in \mathcal{K} \setminus \{k_0\}}$ be an independent and identically distributed array of Rademacher random variables, and define a random element of $L^2(\mu)$ by*

$$p := p_{j_0 k_0} + \sum_{j \in \mathcal{J} \setminus \{j_0\}, k \in \mathcal{K} \setminus \{k_0\}} a_{jk} \xi_{jk} p_{jk}.$$

Assume $\{p \in \mathcal{F}\}$ is an event, and define f to be a random element of $\mathcal{F}$ that has the same distribution as $p | \{p \in \mathcal{F}\}$. Writing $\mathbb{E}\mathbb{P}_f^{\otimes n}$ for the resulting mixture distribution on $(\mathcal{X} \times \mathcal{Y})^n$ and $\mathbb{P}_{p_{j_0 k_0}}$ for the distribution on $\mathcal{X} \times \mathcal{Y}$ with density $p_{j_0 k_0}$, we have that

$$d^2_{\mathrm{TV}}\big(\mathbb{P}^{\otimes n}_{p_{j_0 k_0}}, \mathbb{E}\mathbb{P}^{\otimes n}_f\big) \le \frac{\exp\big(\frac{(n+1)^2}{2} \sum_{j \in \mathcal{J} \setminus \{j_0\}, k \in \mathcal{K} \setminus \{k_0\}} a^4_{jk}\big)}{4\mathbb{P}(p \in \mathcal{F})^2} - \frac{1}{4}.$$

Suppose that the f defined in Lemma 11 takes values in $\mathcal{F}_\xi(\rho)$ with probability one. Then we have that

$$\tilde{\mathcal{R}}(n, \xi, \rho) \ge \inf_{\psi \in \Psi(1)} \big\{ \mathbb{E}_{p_{j_0 k_0}}(\psi) + \mathbb{E}\mathbb{P}_f(1 - \psi) \big\} \ge 1 - d_{\mathrm{TV}}\big(\mathbb{P}^{\otimes n}_{p_{j_0 k_0}}, \mathbb{E}\mathbb{P}^{\otimes n}_f\big),$$

which reduces the problem of finding lower bounds for the minimax risk $\tilde{\mathcal{R}}(n, \xi, \rho)$ to the choice of an appropriate separation ρ and prior distribution over $\mathcal{F}_\xi(\rho)$.

The main challenge in applying Lemma 11 is in finding a suitable upper bound for $\mathbb{P}(p \notin \mathcal{F})$. Provided $\bar{p} := \sup_{j \in \mathcal{J}, k \in \mathcal{K}} \|p_{jk}\|_\infty < \infty$, we can ensure that $\mathbb{P}(p \notin \mathcal{F}) = 0$ by simply imposing the constraint that $\sum_{j \in \mathcal{J} \setminus \{j_0\}, k \in \mathcal{K} \setminus \{k_0\}} |a_{jk}| \le 1/\bar{p}$. If we do this then, we can prove the lower bound in Theorem 12 below.

THEOREM 12. *Suppose that μ_X and μ_Y are probability measures and that there exist $j_0 \in \mathcal{J}$ and $k_0 \in \mathcal{K}$ such that $p^X_{j_0}(x) = p^Y_{k_0}(y) = 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Assume that $\bar{p} < \infty$, and fix $\gamma \in (0,1)$ and $\xi = (\theta, r, A) \in \Xi$ such that* **(A1)** *holds. Then there exists $c = c(\gamma, A) \in (0, \infty)$ such that*

$$\tilde{\rho}^*(n, \gamma, \xi) \geq c \sup_{m \in \mathbb{N}} \min\biggl(\frac{r}{\theta_{\omega(m)}}, \frac{m^{1/4}}{n^{1/2}}, \frac{1}{m^{1/2}\bar{p}}\biggr).$$

Thinking of $\gamma = \alpha + \beta$, this lower bound matches the upper bound in Theorem 2 in certain cases, up to terms depending only on α, β and A, as we now explain. Suppose that $nr^2 \geq \underline{\theta}^2$, which means that $m_0(nr^2) \geq 2$, so we only rule out the case where the sample size is so small that the optimal truncation level is to include only one basis function. Suppose further that $m_0(nr^2) \leq Cn^{2/3}/\bar{p}^{4/3}$ for some $C = C(\alpha, \beta, A)$, which amounts to asking that the optimal truncation level does not grow too fast, or equivalently, that our alternatives are not too rough. Then

$$\begin{aligned}\sup_{m \in \mathbb{N}} \min\biggl(\frac{r}{\theta_{\omega(m)}}, \frac{m^{1/4}}{n^{1/4}}, \frac{1}{m^{1/2}\bar{p}}\biggr) &\geq \min\biggl(\frac{\{m_0(nr^2)-1\}^{1/4}}{n^{1/2}}, \frac{1}{\{m_0(nr^2)-1\}^{1/2}\bar{p}}\biggr) \\ &\geq \frac{m_0(nr^2)^{1/4}}{n^{1/2}} \min\bigl(2^{-1/4}, C^{-3/4}\bigr). \end{aligned} \tag{9}$$

A comparison of Corollary 4 and (9) allows us to conclude that our U-statistic permutation test attains the minimax optimal separation rate in wide generality (i.e., with few restrictions on the underlying spaces and the sequence θ), provided that nr^2 is sufficiently large and $m_0(nr^2) \leq Cn^{2/3}/\bar{p}^{4/3}$. The following example illustrates this latter condition in a specific case.

EXAMPLE 13. Write $\zeta = (s_X, s_Y, d_X, d_Y, \alpha, \beta, A)$. In our d-dimensional Sobolev setting of Example 6, when $t \geq 1$, we have $m_0(t) \asymp_\zeta t^{2d/(4s+d)}$, and hence when $n^{2s-d} \gtrsim_\zeta r^{3d}$ we have that $m_0(nr^2) \lesssim_\zeta n^{2/3}$. Since we may take $\bar{p} = 2^{1/2}$, it therefore follows that when $nr^2 \geq 1$ and $n^{2s-d} \gtrsim_\zeta r^{3d}$, the lower bound (9) holds and this matches the upper bound from Corollary 4.

Despite the attractive conclusions that can be drawn from Theorem 12, it remains desirable to weaken further the smoothness requirements on our alternatives. It turns out that in certain settings, we can use empirical process techniques to lower bound the $\mathbb{P}(p \in \mathcal{F})$ term in Lemma 11 without a bound on $\sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} |a_{jk}|$. This allows us to substantially widen the range of smoothnesses under which our upper and lower bounds match. We first illustrate this approach in our Sobolev example.

PROPOSITION 14. *In the context of Example 6, fix $\gamma \in (0,1)$. Then there exist $c_1, c_2 \in (0, \infty)$, each depending only on d_X, d_Y, γ, s_X, s_Y and A, such that if $nr^2 \geq 2$ and $(r^d/n^{2s})^{1/(4s+d)} \leq c_1/\log^{1/2}(nr^2)$, then*

$$\tilde{\rho}^*(n, \gamma, \xi) \geq c_2 \biggl(\frac{r^d}{n^{2s}}\biggr)^{1/(4s+d)}.$$

Thus, the lower bound of Proposition 14 matches the upper bound of Example 6 when $(r^d/n^{2s})^{1/(4s+d)} \leq c_1/\log^{1/2}(nr^2)$, or equivalently when $m_0(nr^2) \lesssim_\zeta n^2/\log^2(nr^2)$. This condition is rather weak, and holds whenever the minimax separation rate is polynomially decreasing in r^d/n^{2s}. Compared with Example 13, Proposition 14 extends the parameter regime over which the lower bound on the minimax separation rate for independence testing

matches the upper bound of Example 6, by also covering lower smoothness cases where $n^{2s-d} \ll r^{3d}$.

We remark that Proposition 14 generalises to more abstract settings. Assume that $\mathcal{X}$ and $\mathcal{Y}$ are equipped with metrics $\tau_{\mathcal{X}}$ and $\tau_{\mathcal{Y}}$, respectively, and write $H(\cdot, \mathcal{X})$ and $H(\cdot, \mathcal{Y})$ for the corresponding metric entropies. Suppose that there exist $\kappa_1, \kappa_2 \geq 0$ and functions $\ell_1, \ell_2 : (0, \infty) \to (0, \infty)$ that are slowly varying at infinity such that $H(u, \mathcal{X}) = u^{-2\kappa_1}\ell_1(1/u)$ and $H(u, \mathcal{Y}) = u^{-2\kappa_2}\ell_2(1/u)$; thus, if $\mathcal{X} = [0,1]^{d_X}$, then we may take $\kappa_1 = 0$ and $\ell_1(u) = d_X \log u$. Suppose further that there exist $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ such that

$$|p_{jk}(x, y) - p_{jk}(x', y')| \lesssim_\zeta \|j\|_1^{\alpha_1}\tau_{\mathcal{X}}(x, x')^{\beta_1} + \|k\|_1^{\alpha_2}\tau_{\mathcal{Y}}(y, y')^{\beta_2}$$

for all $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$, $j \in \mathcal{J}$, $k \in \mathcal{K}$, where ζ does not depend on n, r, x, x', y, y', j, k. In our Sobolev example, then we may take $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$. Finally, assume that $\bar{p} < \infty$. Then, taking $\xi = (\theta, r, A) \in \Xi$ and $\gamma \in (0, 1)$, writing $\gamma_1 := \frac{\kappa_1}{\beta_1((s_X/\alpha_1)\wedge 1)}$ and $\gamma_2 := \frac{\kappa_2}{\beta_2((s_Y/\alpha_2)\wedge 1)}$, and setting $s = d(d_X/s_X + d_Y/s_Y)^{-1}$, similar calculations to those in the proof of Proposition 14 reveal that

$$\tilde{\rho}^*(n, \gamma, \xi) \gtrsim_\zeta \left(\frac{r^d}{n^{2s}}\right)^{1/(4s+d)}$$

whenever $\max(\gamma_1, \gamma_2) < 1$ and $r \lesssim_{\zeta,\epsilon} \min(n^{\frac{2s(1-\gamma_1)}{d+4s\gamma_1}-\epsilon}, n^{\frac{2s(1-\gamma_2)}{d+4s\gamma_2}-\epsilon})$ for some $\epsilon > 0$. Thus, we match the upper bound of Corollary 5 even in this more general setting.

Our final lower bound applies similar empirical process techniques to show that the rate found by applying the first part of Corollary 7 to Example 8 for our infinite-dimensional example is optimal in certain regimes of (n, r).

PROPOSITION 15 (BKS(2020)). *Let $\mathcal{X}, \mathcal{Y}, \mu_X, \mu_Y, (p_{jk}), \mathcal{J}$ and $\mathcal{K}$ be as in Corollary 7 and Example 8. Fix $\alpha, \beta \in (0, 1)$ such that $\alpha + \beta < 1$ and $r, s_X, s_Y, A > 0$. For $j \in \mathcal{J}$, $k \in \mathcal{K}$, let $\theta_{jk} = \exp(s_X|j|^{1/2}) \vee \exp(s_Y|k|^{1/2})$, and let $\xi = (\theta, r, A) \in \Xi$. Recalling the definitions of s and c_0 from Corollary 7, suppose that $r^2 \leq n^{s/(s+c_0)-\epsilon}$ for some $\epsilon > 0$. Then there exist $C = C(\alpha, \beta, s_X, s_Y, A, \epsilon) > 0$ and $C' = C'(\alpha, \beta, s_X, s_Y, A, \epsilon) > 0$ such that when $\min(n, nr^2) \geq C'$ we have*

$$\rho^*(n, \alpha, \beta, \xi) \geq \frac{C m^{1/4}_{0,s_X,s_Y}(nr^2)}{n^{1/2}}.$$

6. Power function. In this section, we provide an approximation to the power function of our USP test from Section 3. For simplicity of exposition, we will restrict attention to the case where the $\mathcal{X} = \mathcal{Y} = [0, 1]$, and work with the Fourier basis (7) with respect to the respective Lebesgue base measures μ_X and μ_Y. Recall that in this case, $\mathcal{J} = \mathcal{K} = (\{0, 1\} \times \mathbb{N}_0) \setminus \{(1, 0)\}$. We will consider test statistics $\hat{D}_n$ with

$$\mathcal{M} = (\{0, 1\} \times [M]) \times (\{0, 1\} \times [M])$$

for a tuning parameter $M \in \mathbb{N}$ which will typically be large so that $\hat{D}_n$ is approximately normally distributed. When M is large and the dependence between X and Y is weak, we will see that the variance of $\hat{D}_n$ can be approximately expressed in terms of

$$\sigma^2_{M,X} \equiv \sigma^2_{M,X}(f) := 2M + 1 + \sum_{m=1}^{2M}(2M + 1 - m)\{a_{(0,m)\bullet}(f)^2 + a_{(1,m)\bullet}(f)^2\}$$

$$\asymp M\|f_X\|^2_{L^2(\mu_X)}$$

as $M \to \infty$, and the corresponding quantity $\sigma^2_{M,Y}$, in which f_X and μ_X above are replaced with f_Y and μ_Y, respectively, and $a_{j\bullet}(f)$ for $j \in \mathcal{J}$ is replaced with $a_{\bullet k}(f)$ for $k \in \mathcal{K}$.

Define $A_{M,X} \equiv A_{M,X}(f) := 1 + \sum_{m=1}^{2M} (|a_{(0,m)\bullet}(f)| + |a_{(1,m)\bullet}(f)|)$, with the corresponding definition of $A_{M,Y}$. We will see that the quantities $A_{M,X}$ and $A_{M,Y}$, which when $f \in \mathcal{F}$ are both $o(M^{1/2})$ as $M \to \infty$ by Lemma S5 in the supplement, will play a role in controlling the normal approximation error of our test statistic and the corresponding null statistics.

THEOREM 16 (BKS(2020)). *In the above setting, let $f \in \mathcal{F}$ with $\|f\|_\infty < \infty$, let $\alpha \in (0, 1)$ and let $B \in \mathbb{N}$. Write*

$$\Delta_f := \frac{\binom{n}{2}^{1/2} \sum_{(j,k)\in\mathcal{M}} \{a_{jk}(f) - a_{j\bullet}(f) a_{\bullet k}(f)\}^2}{\sigma_{M,X}\sigma_{M,Y}}$$

and, with $s = \lceil \alpha(B+1) \rceil - 1$, let $\mathrm{B}_{B-s,s+1} \sim \mathrm{Beta}(B-s, s+1)$. Let

$$\delta_* := \max\left\{ \frac{\Delta_f^{1/2}}{M^{1/2}}, \frac{1}{M^{1/2}}, D(f)^{1/4}, \left(\frac{M^2}{n}\right)^{1/2}, \frac{A_{M,X}A_{M,Y}}{M} \right\}^{1/3}.$$

Then there exists $C = C(\|f\|_\infty, \alpha) > 0$ such that the p-value P in (5) satisfies

$$\left|\mathbb{P}_f(P \le \alpha) - \mathbb{E}\bar{\Phi}\big(\Phi^{-1}(\mathrm{B}_{B-s,s+1}) - \Delta_f\big)\right| \le C \min\big\{B^{4/3}\delta_*, \big(B^{-1/3} \vee \delta_*^{1/3}\big)\big\}.$$

To understand the implications of this theorem, first consider the case where the null hypothesis holds, so that $\Delta_f = 0$, and further assume for simplicity that $\alpha(B+1)$ is an integer. Then the conclusion states that

$$\left|\mathbb{P}_f(P \le \alpha) - \alpha\right| \le C \min\big\{B^{4/3}\delta_*, \big(B^{-1/3} \vee \delta_*^{1/3}\big)\big\},$$

though in fact, we already know that $\mathbb{P}_f(P \le \alpha) = \alpha$ in this special case. More generally, Theorem 16 provides an approximation to the local power of our test when $D(f)$ is small and both n and M are large, with M^2/n small. It could be used by practitioners to guide the choice of B in cases where computation is expensive: given an anticipated effect size Δ_f, one can compare $\mathbb{E}\bar{\Phi}(\Phi^{-1}(\mathrm{B}_{B-s,s+1}) - \Delta_f)$ to $\bar{\Phi}(\Phi^{-1}(1-\alpha) - \Delta_f)$ to understand the trade-off between computation and power. Note also that $\bar{\Phi}(\Phi^{-1}(1-\alpha) - \Delta_f)$ is the limiting power of the oracle test that has access to the marginal distributions.

To illustrate Theorem 16, we conducted some simulations to verify the accuracy of the approximate power function. For a parameter $\rho \in [0, 1/2]$, we considered independent copies of pairs (X, Y) with density function

$$f_\rho(x, y) = 1 + 2\rho \sin(2\pi x) \sin(2\pi y) \tag{10}$$

for $x, y \in [0, 1]$, so that, marginally, $X, Y \sim U[0, 1]$. For these densities, we have $D(f_\rho) = \rho^2$ and $\sigma^2_{M,X} = \sigma^2_{M,Y} = 2M + 1$. In our simulations, we take $n = 300$, $M = 7$, $B = 99$ and $\alpha = 0.1$ so that

$$\Delta_{f_\rho} = \binom{300}{2}^{1/2} \rho^2/15 = \rho^2 \times 14.1\ldots.$$

Figure 1 plots the theoretical approximate power function, given by $\mathbb{E}\bar{\Phi}(\Phi^{-1}(\mathrm{B}_{B-s,s+1}) - \Delta_f)$, and the empirical power function, which was computed by averaging over 700 independent repetitions of the experiment for each value of ρ. The simulations reveal a good agreement between our approximations and empirical performance.

The proof of Theorem 16 uses careful bounds for the error in normal approximations to degenerate U-statistics, as well as corresponding bounds in the case where the U-statistic is computed on a permuted data set. In the unpermuted case, such bounds have been well

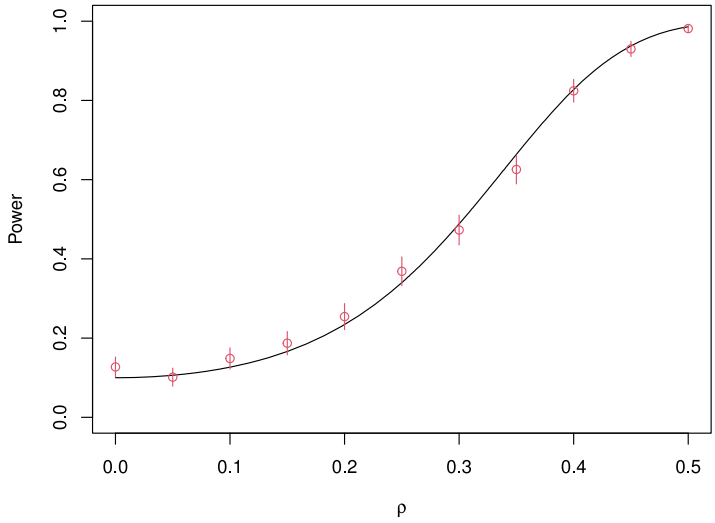

FIG. 1. *The theoretical approximate power function from Theorem 16 (black), and an empirical estimate of the true power function (red); error bars show two standard deviations. Here, the data were generated according to (10) with* $n = 300$, $B = 99$, $\alpha = 0.1$, $M = 7$.

studied, inspired by the work of Hall (1984) and de Jong (1990), who established asymptotic normality results for degenerate U-statistics. This is interesting because, in the classical theory, the asymptotic distribution of a degenerate U-statistic of order 2, for a fixed h, is given by a weighted infinite sum of independent chi-squared random variables (e.g., Serfling ((1980), page 194)). Indeed, from the form of the first term on the right-hand side of (11) below, it is not clear that a normal approximation error will be small. However, if we allow h to depend on the sample size n, then the weights in the infinite sum may become more diffuse, so that a normal approximation may be more appropriate. In our setting, the truncation set $\mathcal{M}$ will typically depend on n, in which case we are in a situation where the U-statistic kernel depends on the sample size. Rinott and Rotar (1997) derived error bounds in the normal approximation with respect to classes of probability integral metrics that include the Kolmogorov distance. Döbler and Peccati (2017, 2019) extended these results in two directions, first by working with multivariate U-statistics, and second by controlling the normal approximation error in the L_1-Wasserstein distance. We present a consequence of Döbler and Peccati ((2019), Theorem 3.3) below, because it it will help to contextualise our (new) error bound in the permuted case, which appears as Proposition 18.

PROPOSITION 17 (Döbler and Peccati ((2019), Theorem 3.3)). *For* $n \geq 2$, *let* $Z_1, \dots, Z_n$ *be independent and identically distributed random elements in a measurable space* $\mathcal{Z}$, *and let* $h : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ *be a symmetric measurable function that satisfies* $\mathbb{E}h(z, Z_1) = 0$ *for all* $z \in \mathcal{Z}$ *and* $\mathbb{E}\{h(Z_1, Z_2)^2\} = 1$. *Write* $g(x, y) := \mathbb{E}\{h(x, Z_1)h(y, Z_1)\}$ *and* $U := \frac{1}{2}\binom{n}{2}^{-1/2} \sum_{i \in \mathcal{I}_2} h(Z_{i_1}, Z_{i_2})$. *With* $W \sim N(0, 1)$, *there exists a universal constant* $C > 0$ *such that for* $n \geq 2$ *we have*

$$d_{\mathrm{W}}(U, W) \leq C \max\left[\frac{\mathbb{E}^{1/2}\{h^4(Z_1, Z_2)\}}{n^{1/2}}, \mathbb{E}^{1/2}\{g^2(Z_1, Z_2)\}\right]. \tag{11}$$

As mentioned above, Proposition 18 below extends Proposition 17 to the case of a permuted data set and, therefore, provides a useful stepping stone for analysing the power properties of permutation tests based on degenerate U-statistics.

PROPOSITION 18 (BKS(2020)). *For $n \geq 4$, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random elements in a product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let Π be a uniformly random element of $\mathcal{S}_n$, independent of $(X_i, Y_i)_{i=1}^n$. Let $h : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a symmetric measurable function that satisfies*

$$\mathbb{E}h\big((x, y), (x', Y_1)\big) = \mathbb{E}h\big((x, y), (X_1, y')\big) = 0$$

for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$, and also satisfies $\mathbb{E}h^2((X_1, Y_2), (X_3, Y_4)) = 1$. Write $g((x, y), (x', y')) := \mathbb{E}\{h((x, y), (X_1, Y_2))h((x', y'), (X_1, Y_2))\}$ and

$$U := \frac{1}{2}\binom{n}{2}^{-1/2} \sum_{(i_1,i_2)\in\mathcal{I}_2} h\big((X_{i_1}, Y_{\Pi(i_1)}), (X_{i_2}, Y_{\Pi(i_2)})\big).$$

Then, with $W \sim N(0, 1)$, there exists a universal constant $C > 0$ such that

$$d_{\mathrm{W}}(U, W) \leq C \max\Big[\frac{1}{n^{1/2}} \max_{\sigma\in\mathcal{S}_4} \mathbb{E}^{1/2}\big\{h^4\big((X_1, Y_{\sigma(1)}), (X_2, Y_{\sigma(2)})\big)\big\}, \tag{12}$$
$$\mathbb{E}^{1/2}\big\{g^2\big((X_1, Y_2), (X_3, Y_4)\big)\big\}, \mathbb{E}\big|\mathbb{E}\big\{h\big((X_1, Y_2), (X_3, Y_1)\big)\big|X_3, Y_2\big\}\big|\Big].$$

Comparing the bounds in Propositions 17 and 18, we see three differences caused by the permutation. The first term in (12) is slightly inflated by the maximum over the 24 permutations in $\mathcal{S}_4$; the second term involves distinct indices, which is to be expected since most permutations of $\mathcal{S}_n$ have only a small number of fixed points; and finally, there is an additional third term, which vanishes if X_1 and Y_1 are independent.

In fact, for a full description of the power properties of our permutation test, we require a multivariate normal approximation error bound for the random vector consisting of the original test statistic and the B test statistics computed on the permuted data sets. Since this statement is more complicated, we defer it to the online supplement (Lemma S1). Its main message for our purposes, however, is that these $B + 1$ statistics are approximately independent, which is what facilitates the power function approximation in Theorem 16.

7. Numerical results. In this section, we examine the empirical performance of our USP test, comparing it with alternative approaches where appropriate. We consider discrete, absolutely continuous and infinite-dimensional settings, following the main examples given earlier. First, however, we show how our test statistic can be computed much more efficiently than might initially appear to be the case.

7.1. *Computational trick.* Our test statistic $\hat{D}_n$ can be rewritten similarly to the test statistics in Song et al. (2012) to allow for quicker computation, in the case that $\mathcal{M} = \mathcal{J}_0 \times \mathcal{K}_0$ for some $\mathcal{J}_0 \subseteq \mathcal{J}$ and $\mathcal{K}_0 \subseteq \mathcal{K}$. Define matrices $J = (J_{i_1 i_2})_{i_1,i_2=1}^n$, $K = (K_{i_1 i_2})_{i_1,i_2=1}^n$ by

$$J_{i_1 i_2} := \sum_{j\in\mathcal{J}_0} p_j^X(X_{i_1}) p_j^X(X_{i_2}) \quad \text{and} \quad K_{i_1 i_2} := \sum_{k\in\mathcal{K}_0} p_k^Y(Y_{i_1}) p_k^Y(Y_{i_2}),$$

and let $\tilde{J}$ and $\tilde{K}$ be the corresponding matrices with the diagonal entries set to zero. Then, writing $\mathbf{1} \in \mathbb{R}^n$ for the all-ones vector, we have that

$$\hat{D}_n = \frac{1}{n(n-1)} \sum_{(i_1,i_2)\in\mathcal{I}_2} J_{i_1 i_2} K_{i_1 i_2} - \frac{2}{n(n-1)(n-2)} \sum_{(i_1,i_2,i_3)\in\mathcal{I}_3} J_{i_1 i_2} K_{i_1 i_3}$$
$$+ \frac{1}{n(n-1)(n-2)(n-3)} \sum_{(i_1,i_2,i_3,i_4)\in\mathcal{I}_4} J_{i_1 i_3} K_{i_2 i_4}$$

$$
\begin{aligned}
&= \frac{1}{n(n-1)} \sum_{i_1,i_2=1}^{n} \tilde{J}_{i_1 i_2} \tilde{K}_{i_1 i_2} - \frac{2}{n(n-1)(n-2)} \Bigg(\sum_{i_1,i_2,i_3=1}^{n} \tilde{J}_{i_1 i_2} \tilde{K}_{i_1 i_3} - \sum_{i_1,i_2=1}^{n} \tilde{J}_{i_1 i_2} \tilde{K}_{i_1 i_2} \Bigg) \\
&\quad + \frac{1}{n(n-1)(n-2)(n-3)} \\
&\quad \times \Bigg(\sum_{i_1,i_2,i_3,i_4=1}^{n} \tilde{J}_{i_1 i_3} \tilde{K}_{i_2 i_4} - 4 \sum_{i_1,i_2,i_3=1}^{n} \tilde{J}_{i_1 i_2} \tilde{K}_{i_1 i_3} + 2 \sum_{i_1,i_2=1}^{n} \tilde{J}_{i_1 i_2} \tilde{K}_{i_1 i_2} \Bigg) \\
&= \left\{ \frac{1}{n(n-1)} + \frac{2}{n(n-1)(n-2)} + \frac{2}{n(n-1)(n-2)(n-3)} \right\} \operatorname{tr}(\tilde{J}\tilde{K}) \\
&\quad - \left\{ \frac{2}{n(n-1)(n-2)} + \frac{4}{n(n-1)(n-2)(n-3)} \right\} \mathbf{1}^T \tilde{J} \tilde{K} \mathbf{1} + \frac{\mathbf{1}^T \tilde{J} \mathbf{1} \mathbf{1}^T \tilde{K} \mathbf{1}}{n(n-1)(n-2)(n-3)} \\
&= \frac{\operatorname{tr}(\tilde{J}\tilde{K})}{n(n-3)} - \frac{2 \mathbf{1}^T \tilde{J} \tilde{K} \mathbf{1}}{n(n-2)(n-3)} + \frac{\mathbf{1}^T \tilde{J} \mathbf{1} \mathbf{1}^T \tilde{K} \mathbf{1}}{n(n-1)(n-2)(n-3)}.
\end{aligned}
$$

From this final expression, we can see that $\hat{D}_n$ can be computed in $O(n^2(|\mathcal{J}_0| + |\mathcal{K}_0|))$ operations, with the most time-consuming part being the computation of the matrices $\tilde{J}$ and $\tilde{K}$.

7.2. *Discrete settings*. Here, we study two different examples, to illustrate the effects of sparse and dense dependence. The first is a 6×6 contingency table, so that $J = K = 6$, where the cell probabilities are of the form

$$
f(j,k) = \frac{2^{-(j+k)}}{(1-2^{-J})(1-2^{-K})} + \epsilon(\mathbb{1}_{\{j=k=1\}} + \mathbb{1}_{\{j=k=2\}}) - \epsilon(\mathbb{1}_{\{j=1,k=2\}} + \mathbb{1}_{\{j=2,k=1\}}),
$$

for $j, k \in [6]$. Here, $\epsilon \geq 0$ measures the strength of the dependence; in fact, $D(f) = 4\epsilon^2$. Our second example has $J = K = 8$ and cell probabilities of the form

$$
f(j,k) = \frac{1}{JK} + (-1)^{j+k-1}\epsilon,
$$

for which $D(f) = JK\epsilon^2$. Thus, the main difference between the examples is in the number of cells affected by the perturbation: in the first case, only the summands in $D(f)$ corresponding to $(j,k) \in \{1,2\} \times \{1,2\}$ are nonzero, whereas in the second example, all summands are nonzero.

Figure 2 plots estimates, computed as sample averages over 10,000 repetitions, of the power of our USP test as a function of ϵ in the two examples, with $n = 100$ in Figure 2(a) and $n = 50$ in Figure 2(b). In both cases, we set $\alpha = 0.05$ and $B = 99$. For comparison, we also plot corresponding power estimates for two versions of Pearson's chi-squared test. The first, corresponding to the more usual practice in applications, uses as a critical value for the test the $(1-\alpha)$th quantile of the chi-squared distribution with $(J-1)(K-1)$ degrees of freedom; the second computes the critical value using a permutation procedure similarly to that employed for our USP test. The advantage of the second approach is that it controls the Type I error at the nominal level. In both cases, our USP test has greater power than both versions of Pearson's test, particularly in the first example, which is especially striking given that the chi-squared quantile version of Pearson's test is anticonservative there.

7.3. *Sobolev example*. In this subsection, we consider a setting originally studied by Sejdinovic et al. (2013). For $\omega \in \mathbb{N}$ and $(x, y) \in [0,1]^2$, define the density function

$$
f_\omega(x,y) = 1 + \sin(2\pi\omega x)\sin(2\pi\omega y).
$$

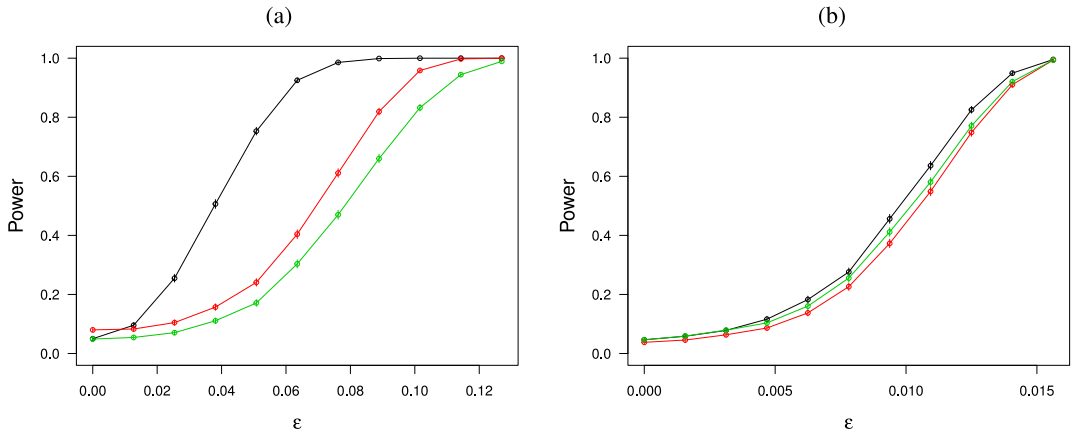

FIG. 2. *Estimated power functions in the two discrete settings for our U-statistic permutation test (black), as well as Pearson's chi-squared test with chi-squared quantile (red) and quantile obtained from permutations (green). Error bars show three standard errors; other parameters: $\alpha = 0.05$, $B = 99$, $n = 100$ (left), $n = 50$ (right).*

Berrett and Samworth (2019) also consider this family of densities, and explain why it becomes increasingly difficult to detect the dependence as ω increases, despite the fact that the mutual information does not depend on ω. In fact, we also have $D(f_\omega) = 1/4$ for every $\omega \in \mathbb{N}$, so this measure of dependence does not depend on ω either.

In Figure 3, we plot estimates of the power of our USP test, computed over 2000 repetitions with $n = 100, 200$. The choice of $\mathcal{M}$ is made as in Section 6, with $M = 2, 4$. As alternative approaches, we also study the HSIC test of Gretton et al. (2005), which is implemented in the `R` package `dHSIC` (Pfister and Peters (2017)), the MINTav test of Berrett and Samworth (2019), implemented in the R package `IndepTest` (Berrett, Grose and Samworth (2018)) with $k \in [5]$, a test based on the empirical copula process described by Kojadinovic and Holmes (2009) and implemented in the `R` package `copula` (Hofert et al. (2017)) and a test based on distance covariance implemented in the `R` package `energy` (Rizzo and Szekely (2017)). For these comparison methods, we used the default tuning parameter values recom-

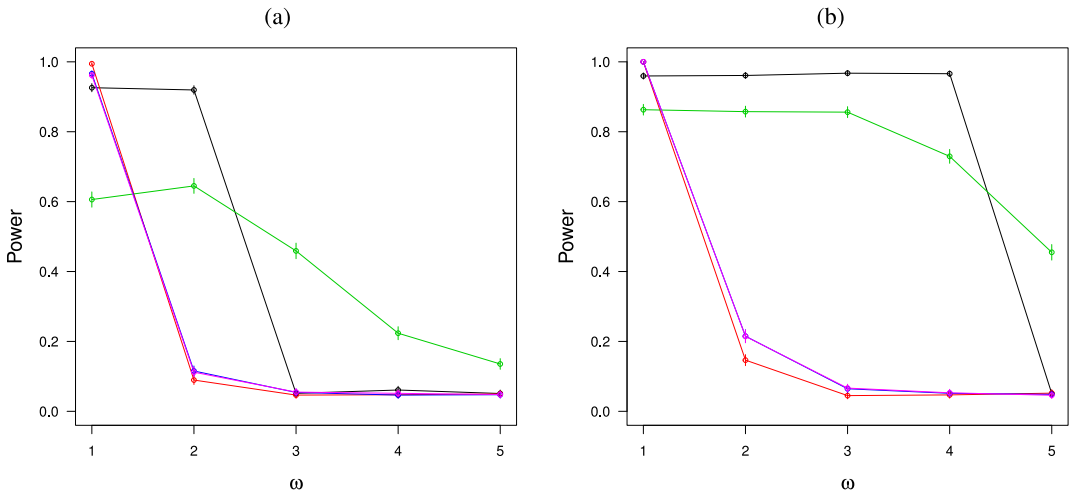

FIG. 3. *Estimated power functions in the Sobolev example for our U-statistic permutation test (black) with $M = 2$, $n = 100$ (left) and $M = 4$, $n = 200$ (right), HSIC (red), distance covariance (blue), copula (purple) and MINTav (green). Error bars show two standard errors; other parameters: $\alpha = 0.05$, $B = 99$.*

mended by the corresponding authors. The fact that the departures in this example are aligned with a single basis function for each choice of ω means that the power of our USP test is constant for $\omega \leq M$, and it performs extremely well in these cases. Once ω exceeds M, the test has no better than nominal power, as expected. Thus, M determines the number of directions of departure from independence that we can hope to detect with our USP test (we have $4M^2$ coefficients to estimate). Increasing the value of M would provide nontrivial power for larger values of ω, but would sacrifice some power for smaller values of ω.

7.4. *Infinite-dimensional example.* Our final example concerns potentially correlated Brownian motions on $[0, 1]$, as an illustration of our USP test applied to functional data. More precisely, our data come in the form of pairs (X, Y), where $X = (X_t)_{t\in[0,1]}$ is a standard Brownian motion, and where, for some $r \in [0, 1]$ and for another standard Brownian motion $Z = (Z_t)_{t\in[0,1]}$ that is independent of X, we have that $Y = (Y_t)_{t\in[0,1]}$ is given by

$$Y_t = rX_t + (1 - r^2)^{1/2} Z_t.$$

Thus, marginally, Y is also distributed as a standard Brownian motion.

By the Wiener representation of Brownian motion (e.g., Kahane (1997)), we can write

$$X_t = 2^{1/2} \sum_{\ell=1}^{\infty} \eta_\ell \frac{\sin((\ell - 1/2)\pi t)}{(\ell - 1/2)\pi},$$

where $(\eta_\ell)_{\ell=1}^{\infty}$ is a sequence of independent, standard normal random variables. For any $W = (W_t)_{t\in[0,1]} \in L^2[0, 1]$, we can compute the transformed coefficients

$$u_\ell(W) := \Phi\Big(2^{1/2}(\ell - 1/2)\pi \int_0^1 W_t \sin\big((\ell - 1/2)\pi t\big)\, dt\Big)$$

for $\ell \in \mathbb{N}$. We can therefore consider testing the independence of the random vectors $(u_1(X), \ldots, u_L(X))$ and $(u_1(Y), \ldots, u_L(Y))$, for some suitably chosen truncation level L. For $\ell, m \in \mathbb{N}$ and $x \in L^2[0, 1]$, let $p^X_{\ell m}(x) := 2^{1/2} \cos(2\pi m u_\ell(x))$, and define $p^Y_{\ell m}(\cdot)$ similarly. The U-statistic kernel in this example can be written as

$$\begin{aligned} h\big((x_1, y_1), \ldots, (x_4, y_4)\big) = \sum_{\ell_1,\ell_2=1}^{L} \sum_{m_1,m_2=1}^{M} \{ & p^X_{\ell_1 m_1}(x_1) p^Y_{\ell_2 m_2}(y_1) p^X_{\ell_1 m_1}(x_2) p^Y_{\ell_2 m_2}(y_2) \\ & - 2p^X_{\ell_1 m_1}(x_1) p^Y_{\ell_2 m_2}(y_1) p^X_{\ell_1 m_1}(x_2) p^Y_{\ell_2 m_2}(y_3) \\ & + p^X_{\ell_1 m_1}(x_1) p^Y_{\ell_2 m_2}(y_2) p^X_{\ell_1 m_1}(x_3) p^Y_{\ell_2 m_2}(y_4)\}, \end{aligned}$$

where $L, M \in \mathbb{N}$. In Figure 4, we plot the power functions of our USP test, estimated over 2000 repetitions, for three different sample sizes, namely $n \in \{50, 100, 200\}$, with $L = 2$ and $M = 1$. As expected, the power of our test increases with both r and n.

8. Discussion and outlook. In this paper, we have introduced a new permutation test of independence based on a U-statistic estimator of the squared L^2-distance between a joint distribution and the product of its marginals. Our methodology extends naturally to the problem of testing mutual independence of several random elements. We have further demonstrated its minimax optimality in various settings; to the best of our knowledge, this is the first time that minimax optimality results have been established for such permutation tests. We conclude by explaining how closely related ideas can be used to provide new goodness-of-fit tests and two-sample tests with desirable properties.

Consider $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P \in \mathcal{P}$, where $\mathcal{P}$ is a dominated class of distributions on a separable, σ-finite measure space $(\mathcal{Z}, \mathcal{C}, \nu)$. Suppose further that we wish to test $H_0 : P = P_0$

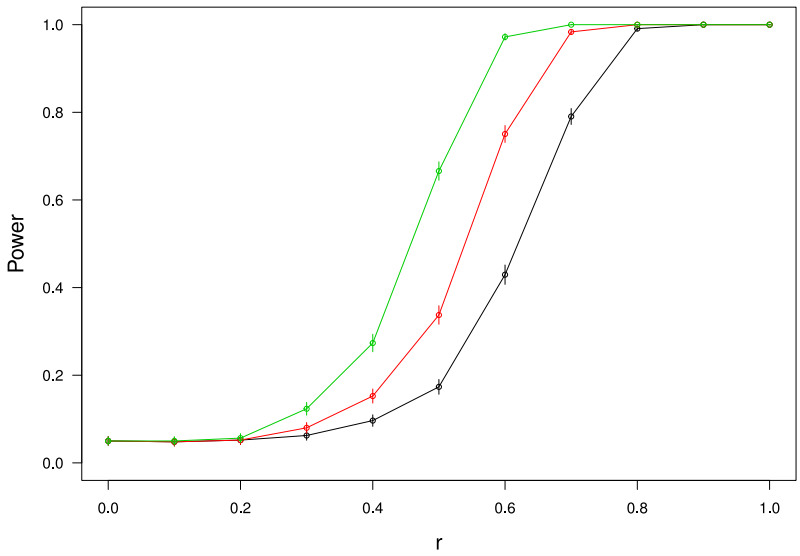

FIG. 4. *Estimated power functions for testing the independence of two Brownian motions with* $n = 50$ (*black*), $n = 100$ (*red*) *and* $n = 200$ (*green*). *Error bars show two standard errors*; *other parameters*: $\alpha = 0.05$, $B = 99$, $L = 2$, $M = 1$.

against $H_1 : P \neq P_0$, where $P_0 \in \mathcal{P}$. Then, writing f and f_0, respectively, for the Radon–Nikodym derivatives of P and P_0 with respect to ν, we can construct a U-statistic estimator of the squared $L^2(\nu)$ distance between f and f_0 in a very similar spirit to (4). Since the null hypothesis is simple, there is no need for permutations, and we can obtain a critical value for the test by sampling from P_0.

For two-sample tests, we can let $\mathcal{Y} = \{0, 1\}$, so that testing the independence of X and Y amounts to testing the equality of the distributions $X|\{Y = 0\}$ and $X|\{Y = 1\}$. A small observation here is that the sample sizes from each conditional distribution are random (having a binomial distribution), whereas these are often treated as fixed in the usual two-sample testing formulation. Our methodology and theory apply directly to this problem, therefore, further extending its scope.

9. Proofs of main results.

PROOF OF THEOREM 1. Since ψ is bounded, we have that $\psi \in L^2(\bigotimes_{i=1}^n \mu)$. Given $j \in \mathcal{J}$, $k \in \mathcal{K}$ and $I \subseteq [n]$ we write

$$b_{jk}^I := \left\langle \psi, \bigotimes_{i=1}^n \{\mathbb{1}_{\{i \in I\}} p_{jk} + \mathbb{1}_{\{i \notin I\}} p_{j_0 k_0}\} \right\rangle_{L^2(\bigotimes_{i=1}^n \mu)}.$$

Since $r > \underline{\theta}\rho$, we have that $\mathcal{M}_\theta(r/\rho) \neq \varnothing$. For $(j, k) \in \mathcal{M}_\theta(r/\rho)$ to be chosen later consider $f^* \equiv f_{jk}^* := p_{j_0 k_0} + \rho p_{jk} \in \mathcal{F}$, which satisfies $S_\theta(f^*) = \theta_{jk}^2 \rho^2 \leq r^2$ and $D(f^*) = \rho^2$. Then by Cauchy–Schwarz,

$$\begin{aligned} \mathbb{E}_{f^*}(\psi) &= \left\langle \psi, \bigotimes_{i=1}^n f^* \right\rangle_{L^2(\bigotimes_{i=1}^n \mu)} = \mathbb{E}_{p_{j_0 k_0}}(\psi) + \sum_{\varnothing \neq I \subseteq [n]} \rho^{|I|} b_{jk}^I \\ &\leq \alpha + \{(1 + \rho^2)^n - 1\}^{1/2} \Bigg\{ \sum_{\varnothing \neq I \subseteq [n]} (b_{jk}^I)^2 \Bigg\}^{1/2}. \end{aligned} \tag{13}$$

Now, observe that

$$\sum_{(j,k)\in\mathcal{M}_\theta(r/\rho)}\sum_{\varnothing\neq I\subseteq[n]}\bigl(b_{jk}^I\bigr)^2 \le \|\psi\|^2_{L^2(\bigotimes_{i=1}^n\mu)} = \mathbb{E}_{p_{j_0k_0}}\bigl(\psi^2\bigr) \le \alpha.$$

Hence, for any $\eta>0$ we may choose $(j,k)\in\mathcal{M}_\theta(r/\rho)$ such that

$$\sum_{\varnothing\neq I\subseteq[n]}\bigl(b_{jk}^I\bigr)^2 \le \frac{\alpha}{|\mathcal{M}_\theta(r/\rho)|}+\eta.$$

The first claim of Theorem 1 follows from this combined with (13).

For the second part, first note the definitions of Ξ, $\mathcal{F}_\xi(\rho)$ and $\rho^*(n,\alpha,\beta,\xi)$ immediately after **(A1)**. For the choice of j, k in the first part of the proof, let $\theta'=(\theta'_{j'k'})_{j'\in\mathcal{J},k'\in\mathcal{K}}$ be given by

$$\theta'_{j'k'} := \begin{cases} 0 & \text{if } j'=j \text{ and } k'=k,\\ \infty & \text{otherwise.}\end{cases}$$

Now $f^*\in\mathcal{F}_{(\theta',r',2)}(\rho)$ for any $r'>0$. Applying Theorem 2 with $\mathcal{M}=\{(j,k)\}$ then yields that there exists $C=C(\alpha,\beta)>0$ such that $\rho^*(n,\alpha,\beta,\xi)\le C^{1/2}/n^{1/2}$. In other words, there exists $\psi_{f^*}\in\Psi(\alpha)$ such that $\mathbb{E}_{f^*}(\psi_{f^*})\ge 1-\beta$ whenever $n>C/\rho^2$. Finally, the proof of Theorem 2 reveals that ψ_{f^*} may be taken to be a permutation test (in fact the permutation test described in Section 3 with $\mathcal{M}=\{(j,k)\}$), as required. □

PROOF OF THEOREM 2. Consider the test of Section 3. Choose $B\ge 2(\frac{1}{\alpha\beta}-1)$, and suppose $f\in\mathcal{F}^*$ were such that

$$D(f) \ge \max\Biggl[2\bigl|\mathbb{E}_f\bigl(\hat{D}_n-\hat{D}_n^{(1)}\bigr)-D(f)\bigr|, \Bigl\{\frac{8}{\alpha\beta}\operatorname{Var}_f\bigl(\hat{D}_n-\hat{D}_n^{(1)}\bigr)\Bigr\}^{1/2}\Biggr]. \tag{14}$$

Then, by two applications of Markov's inequality, we would have that

$$\begin{aligned}\mathbb{P}_f(P>\alpha) &= \mathbb{P}_f\Biggl(1+\sum_{b=1}^B \mathbb{1}_{\{\hat{D}_n\le\hat{D}_n^{(b)}\}} > (1+B)\alpha\Biggr) \le \frac{1+B\mathbb{P}_f(\hat{D}_n\le\hat{D}_n^{(1)})}{(1+B)\alpha}\\ &\le \frac{1}{(1+B)\alpha}\Biggl[1+\frac{B\operatorname{Var}_f(\hat{D}_n-\hat{D}_n^{(1)})}{\{\mathbb{E}_f(\hat{D}_n-\hat{D}_n^{(1)})\}^2}\Biggr] \le \frac{1}{(1+B)\alpha}\Bigl(1+\frac{B\alpha\beta}{2}\Bigr)\le\beta.\end{aligned}$$

We may think of $\hat{D}_n-\hat{D}_n^{(1)}$ as an estimator of $D(f)$, so that (14) ensures that the strength of the dependence $D(f)$ outweighs the bias and standard deviation of the estimator so that we can detect the dependence using our test, up to the given probabilities of error. The remainder of the proof is dedicated to bounding the bias and variance for a given $\xi\in\Xi$, which enables us to choose ρ so that (14) holds for all $f\in\mathcal{F}_\xi(\rho)$, and hence ensures that $\rho^*(n,\alpha,\beta,\xi)\le\rho$. Henceforth we will write Π as shorthand for Π_1; moreover, for some $\rho>0$ to be chosen later, we fix $f\in\mathcal{F}_\xi(\rho)$ and write D, a_{jk}, $a_{j\bullet}$, $a_{\bullet k}$ instead of $D(f)$, $a_{jk}(f)$, $a_{j\bullet}(f)$, $a_{\bullet k}(f)$, respectively.

Given $(i_1,i_2)\in\mathcal{I}_2$ write $\sigma_{i_1i_2}\in\mathcal{S}_n$ for the transposition of i_1 and i_2, and note that $\Pi\overset{d}{=}\Pi\circ\sigma_{i_1i_2}$. Thus $(\Pi(1),\Pi(2))\overset{d}{=}(\Pi(1),\Pi(3))$, so for every $(j,k)\in\mathcal{M}$ we have that

$$p_{jk}(X_1,Y_{\Pi(1)})p_{jk}(X_2,Y_{\Pi(2)}) \overset{d}{=} p_{jk}(X_1,Y_{\Pi(1)})p_{jk}(X_2,Y_{\Pi(3)}).$$

Similarly, $p_{jk}(X_1,Y_{\Pi(1)})p_{jk}(X_2,Y_{\Pi(3)})\overset{d}{=}p_{jk}(X_1,Y_{\Pi(2)})p_{jk}(X_3,Y_{\Pi(4)})$, so that

$$\begin{aligned}\mathbb{E}\bigl(\hat{D}_n^{(1)}\bigr) = \sum_{(j,k)\in\mathcal{M}} \mathbb{E}\bigl\{&p_{jk}(X_1,Y_{\Pi(1)})p_{jk}(X_2,Y_{\Pi(2)})\\ &-2p_{jk}(X_1,Y_{\Pi(1)})p_{jk}(X_2,Y_{\Pi(3)})+p_{jk}(X_1,Y_{\Pi(2)})p_{jk}(X_3,Y_{\Pi(4)})\bigr\}=0.\end{aligned}$$

Thus, using our Sobolev smoothness condition to bound the truncation error,

$$|\mathbb{E}(\hat{D}_n - \hat{D}_n^{(1)}) - D| = |\mathbb{E}(\hat{D}_n) - D| = \Big| \sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2 - D \Big|$$

$$= \sum_{(j,k)\in(\mathcal{J}\times\mathcal{K})\setminus\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})^2 \leq \frac{r^2}{\inf\{\theta_{jk}^2 : (j,k)\notin\mathcal{M}\}}. \tag{15}$$

We now turn to bounding $\mathrm{Var}(\hat{D}_n - \hat{D}_n^{(1)})$. First write $\bar{h}$ for the symmetrised version of h, given by

$$\bar{h}\big((x_1,y_1),\dots,(x_4,y_4)\big) := \frac{1}{4!}\sum_{\sigma\in\mathcal{S}_4} h\big((x_{\sigma(1)},y_{\sigma(1)}),\dots,(x_{\sigma(4)},y_{\sigma(4)})\big). \tag{16}$$

By, for example, Serfling ((1980), Lemma A, page 183), we have that

$$\mathrm{Var}(\hat{D}_n) = \mathrm{Var}\Big(\frac{1}{4!\binom{n}{4}} \sum_{(i_1,\dots,i_4)\in\mathcal{I}_4} \bar{h}\big((X_{i_1},Y_{i_1}),\dots,(X_{i_4},Y_{i_4})\big)\Big)$$

$$= \binom{n}{4}^{-1} \sum_{c=1}^{4} \binom{4}{c}\binom{n-4}{4-c}\zeta_c, \tag{17}$$

where $\zeta_c := \mathrm{Var}(\mathbb{E}\{\bar{h}((X_1,Y_1),\dots,(X_4,Y_4))|(X_1,Y_1),\dots,(X_c,Y_c)\})$, and moreover, $\zeta_1 \leq \zeta_2 \leq \zeta_3 \leq \zeta_4$. For each $j\in\mathcal{J}$ write $\mathcal{K}_j^{\mathcal{M}} := \{k\in\mathcal{K} : (j,k)\in\mathcal{M}\}$ and for each $k\in\mathcal{K}$ write $\mathcal{J}_k^{\mathcal{M}} := \{j\in\mathcal{J} : (j,k)\in\mathcal{M}\}$. Then, using Cauchy–Schwarz,

$$\begin{aligned}
\zeta_1 &= \mathrm{Var}\big(\mathbb{E}\{\bar{h}\big((X_1,Y_1),\dots,(X_4,Y_4)\big)|(X_1,Y_1)\}\big)\\
&= \frac{1}{4}\mathrm{Var}\Big(\sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})\{p_{jk}(X_1,Y_1) - p_j^X(X_1)a_{\bullet k} - a_{j\bullet}p_k^Y(Y_1)\}\Big)\\
&\leq \frac{3A}{4}\Big\{\Big\| \sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})p_{jk}\Big\|_{L^2(\mu)}^2 + \Big\| \sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})p_j^X a_{\bullet k}\Big\|_{L^2(\mu_X)}^2\\
&\quad + \Big\| \sum_{(j,k)\in\mathcal{M}} (a_{jk} - a_{j\bullet}a_{\bullet k})a_{j\bullet}p_k^Y\Big\|_{L^2(\mu_Y)}^2\Big\} \qquad (18)\\
&\leq \frac{3A}{4}\Big[D + \sum_{j\in\mathcal{J}}\Big\{\sum_{k\in\mathcal{K}_j^{\mathcal{M}}} (a_{jk} - a_{j\bullet}a_{\bullet k})a_{\bullet k}\Big\}^2 + \sum_{k\in\mathcal{K}}\Big\{\sum_{j\in\mathcal{J}_k^{\mathcal{M}}} (a_{jk} - a_{j\bullet}a_{\bullet k})a_{j\bullet}\Big\}^2\Big]\\
&\leq \frac{3AD}{4}\big(1 + \|f_Y\|_{L^2(\mu_Y)} + \|f_X\|_{L^2(\mu_X)}\big) \leq \frac{9A^2D}{4}.
\end{aligned}$$

Observe that we have

$$\zeta_4 = \mathrm{Var}\,\bar{h}\big((X_1,Y_1),\dots,(X_4,Y_4)\big) \leq \mathrm{Var}\,h\big((X_1,Y_1),\dots,(X_4,Y_4)\big).$$

One possibility, therefore, is to simply apply the bound $\zeta_4 \leq \|h\|_\infty^2$. On the other hand, by Cauchy–Schwarz, we can say that

$$\begin{aligned}
\zeta_4 &\leq A^4 \int_{\mathcal{X}\times\mathcal{Y}}\dots\int_{\mathcal{X}\times\mathcal{Y}} h^2\big((x_1,y_1),\dots,(x_4,y_4)\big)\,d\mu(x_1,y_1)\dots d\mu(x_4,y_4)\\
&\leq 18A^4 \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}}\Big\{\sum_{(j,k)\in\mathcal{M}} p_{jk}(x,y)p_{jk}(x',y')\Big\}^2 d\mu(x,y)\,d\mu(x',y') \qquad (19)\\
&\leq 18A^4|\mathcal{M}|.
\end{aligned}$$

We therefore have that

$$\operatorname{Var}(\hat{D}_n) \le \frac{16\zeta_1}{n} + \frac{72\zeta_4}{n(n-1)} \le \frac{36A^2D}{n} + \frac{72\min(\|h\|_\infty^2, 18A^4|\mathcal{M}|)}{n(n-1)}. \tag{20}$$

Next, with the same functions h and $\bar{h}$ as above, we may write

$$\hat{D}_n^{(1)} = \frac{1}{4!\binom{n}{4}} \sum_{(i_1,\dots,i_4)\in\mathcal{I}_4} \bar{h}\big((X_{i_1}, Y_{\Pi(i_1)}), \dots, (X_{i_4}, Y_{\Pi(i_4)})\big).$$

A simplifying property of $\bar{h}$ is that for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathbb{E}\big\{\bar{h}\big((x, y), (X_1, Y_2), (X_3, Y_4), (X_5, Y_6)\big)\big\} = 0. \tag{21}$$

Since we also have to deal with the uniformly random permutation Π, we cannot directly appeal to standard U-statistic theory for our bounds on $\operatorname{Var}(\hat{D}_n^{(1)})$. However, we can develop an analogue of (17) by writing

$$\begin{aligned}
\operatorname{Var}\big(\hat{D}_n^{(1)}\big) &= \frac{1}{4!\binom{n}{4}} \sum_{(i_1,\dots,i_4)\in\mathcal{I}_4} \operatorname{Cov}\big(\bar{h}\big((X_1, Y_{\Pi(1)}), \dots, (X_4, Y_{\Pi(4)})\big),\\
&\qquad \bar{h}\big((X_{i_1}, Y_{\Pi(i_1)}), \dots, (X_{i_4}, Y_{\Pi(i_4)})\big)\big)\\
&= \frac{1}{\binom{n}{4}} \sum_{c=0}^{4} \binom{4}{c}\binom{n-4}{4-c} \operatorname{Cov}\big(\bar{h}\big((X_1, Y_{\Pi(1)}), \dots, (X_4, Y_{\Pi(4)})\big),\\
&\qquad \bar{h}\big((X_1, Y_{\Pi(1)}), \dots, (X_c, Y_{\Pi(c)}), (X_5, Y_{\Pi(5)}), \dots, (X_{8-c}, Y_{\Pi(8-c)})\big)\big)\\
&=: \frac{1}{\binom{n}{4}} \sum_{c=0}^{4} \binom{4}{c}\binom{n-4}{4-c} \tilde{\zeta}_c.
\end{aligned} \tag{22}$$

For $c = 2, 3, 4$, we will use the crude bound

$$\begin{aligned}
\max(\tilde{\zeta}_2, \tilde{\zeta}_3, \tilde{\zeta}_4) &\le \max_{\sigma\in\mathcal{S}_n} \mathbb{E}\big\{h^2\big((X_1, Y_{\sigma(1)}), \dots, (X_4, Y_{\sigma(4)})\big)\big\}\\
&\le \min\big(\|h\|_\infty^2, 18A^8|\mathcal{M}|\big),
\end{aligned} \tag{23}$$

similar to (19). To bound $\tilde{\zeta}_0$ and $\tilde{\zeta}_1$, we must first bound two combinatorial probabilities. First,

$$\mathbb{P}\big(\big|[7] \cap \{\Pi(1), \dots, \Pi(7)\}\big| \ge 1\big) \le 7\mathbb{P}\big(\Pi(1) \in [7]\big) = \frac{49}{n}.$$

Now, similarly,

$$\begin{aligned}
\mathbb{P}\big(\big|[8] \cap \{\Pi(1), \dots, \Pi(8)\}\big| \ge 2\big) &\le \binom{8}{2}\mathbb{P}\big(\Pi(1), \Pi(2) \in [8]\big)\\
&= 2\binom{8}{2}^2 \mathbb{P}\big(\Pi(1) = 1, \Pi(2) = 2\big) = \frac{1568}{n(n-1)}.
\end{aligned} \tag{24}$$

The first of these allows us to use (21), Cauchy–Schwarz and (23) to write

$$\begin{aligned}\tilde{\zeta}_1 &= \operatorname{Cov}\big(\bar{h}\big((X_1, Y_{\Pi(1)}), \ldots, (X_4, Y_{\Pi(4)})\big),\\ &\qquad \bar{h}\big((X_1, Y_{\Pi(1)}), (X_5, Y_{\Pi(5)}), (X_6, Y_{\Pi(6)}), (X_7, Y_{\Pi(7)})\big)\big)\\ &\le \mathbb{P}\big([7] \cap \{\Pi(1), \ldots, \Pi(7)\} = \varnothing\big)\mathbb{E}\big\{\bar{h}\big((X_1, Y_8), (X_2, Y_9), (X_3, Y_{10}), (X_4, Y_{11})\big)\\ &\qquad \times \bar{h}\big((X_1, Y_8), (X_5, Y_{12}), (X_6, Y_{13}), (X_7, Y_{14})\big)\big\}\\ &\qquad + \frac{49}{n}\max_{\sigma\in\mathcal{S}_n}\mathbb{E}\big\{h^2\big((X_1, Y_{\sigma(1)}), \ldots, (X_4, Y_{\sigma(4)})\big)\big\}\\ &\le \frac{49}{n}\min\big(\|h\|_\infty^2, 18A^8|\mathcal{M}|\big).\end{aligned} \tag{25}$$

Finally, we may now use (21), Cauchy–Schwarz, (23) and (24) to similarly write

$$\begin{aligned}\tilde{\zeta}_0 &= \operatorname{Cov}\big(\bar{h}\big((X_1, Y_{\Pi(1)}), \ldots, (X_4, Y_{\Pi(4)})\big), \bar{h}\big((X_5, Y_{\Pi(5)}), \ldots, (X_8, Y_{\Pi(8)})\big)\big)\\ &\le \frac{1568}{n(n-1)}\min\big(\|h\|_\infty^2, 18A^8|\mathcal{M}|\big).\end{aligned} \tag{26}$$

From (22), (23), (25), (26), we have now established that

$$\begin{aligned}\operatorname{Var}\big(\hat{D}_n^{(1)}\big) &\le \tilde{\zeta}_0 + \frac{16}{n}\tilde{\zeta}_1 + \frac{72}{n(n-1)}\max(\tilde{\zeta}_2, \tilde{\zeta}_3, \tilde{\zeta}_4)\\ &\le \frac{2424\min(\|h\|_\infty^2, 18A^8|\mathcal{M}|)}{n(n-1)}.\end{aligned} \tag{27}$$

Thus, from (20) and (27) we deduce that

$$\operatorname{Var}\big(\hat{D}_n - \hat{D}_n^{(1)}\big) \le \frac{72A^2D}{n} + \frac{4992\min(\|h\|_\infty^2, 18A^8|\mathcal{M}|)}{n(n-1)}. \tag{28}$$

Now by substituting (15) and (28) into (14) we can see that if

$$D(f) \ge \max\left\{\frac{2r^2}{\inf\{\theta_{jk}^2 : (j,k)\notin\mathcal{M}\}}, \frac{1152A^2}{n\alpha\beta}, \frac{283\min(\|h\|_\infty, 5A^4|\mathcal{M}|^{1/2})}{\{n(n-1)\alpha\beta\}^{1/2}}\right\}, \tag{29}$$

then we have controlled the error probabilities as required. □

PROOF OF COROLLARY 5. There exists $C = C(d_X, d_Y) \in (1, \infty)$ such that for any $T > 0$ we have

$$\begin{aligned}\big|\{(j,k)\in\mathcal{J}\times\mathcal{K} : \theta_{jk}\le T\}\big| &= \big|\{j\in\mathcal{J} : \|j\|_1^{s_X}\le T\}\big|\big|\{k\in\mathcal{K} : \|k\|_1^{s_Y}\le T\}\big|\\ &\le \big(T^{1/s_X}+1\big)^{d_X}\big(T^{1/s_Y}+1\big)^{d_Y} < C(T\vee 1)^{d/s}.\end{aligned}$$

From this, we can infer that if $m > C$ then $\theta_{\omega(m)} > (m/C)^{s/d}$, and so

$$m_0\big(nr^2\big) \le \max\big\{C, \big(nr^2\big)^{2d/(4s+d)}C^{4s/(4s+d)}\big\} \le C\big\{\big(nr^2\big)\vee 1\big\}^{2d/(4s+d)}.$$

It now follows from (6) that there exists $C = C(d_X, d_Y, \alpha, \beta, A) > 0$ such that if $n \ge 16$ and $nr^2 \ge 1$ then

$$\rho^*(n, \alpha, \beta, \xi) \le C\left(\frac{r^d}{n^{2s}}\right)^{1/(4s+d)},$$

as required. □

PROOF OF PROPOSITION 9. By (29) in the proof of Theorem 2, we see that we reject H_0 with probability at least $1-\beta$, provided that $n \geq 16$ and

$$\rho^2 \geq \min_{m \in K_*} \max\biggl\{ \frac{2r^2}{\theta^2_{\omega(m+1)}}, \frac{1152A^2\gamma}{n\alpha\beta}, \frac{1415A^4 m^{1/2}\gamma^{1/2}}{\{n(n-1)\alpha\beta\}^{1/2}} \biggr\}.$$

Since $m_0(t) \leq t^2/\theta_0^4 + 1$, there exists $n_0 = n_0(R_0, \theta_0) \geq 16$ such that for all $n \geq n_0$ we have $m_0(nr^2/\log^{1/2} n) \leq 2^\gamma + 1$. But then, for $n \geq \max(n_0, e^3)$,

$$\begin{aligned}
&\min_{m \in K_*} \max\biggl\{ \frac{2r^2}{\theta^2_{\omega(m+1)}}, \frac{1152A^2\gamma}{n\alpha\beta}, \frac{1415A^4 m^{1/2}\gamma^{1/2}}{\{n(n-1)\alpha\beta\}^{1/2}} \biggr\} \\
&\leq 2^{1/2} \min_{m \in [2^\gamma]\setminus\{1\}} \max\biggl\{ \frac{2r^2}{\theta^2_{\omega(m+1)}}, \frac{1152A^2\gamma}{n\alpha\beta}, \frac{1415A^4 m^{1/2}\gamma^{1/2}}{\{n(n-1)\alpha\beta\}^{1/2}} \biggr\} \\
&\lesssim_{A,\alpha,\beta} \min_{m \in \{3,4,\ldots,2^\gamma+1\}} \max\biggl\{ \frac{r^2}{\theta^2_{\omega(m)}}, \frac{\log n}{n}, \frac{m^{1/2}\log^{1/2} n}{n} \biggr\} \\
&\leq \max\biggl\{ \frac{\log^{1/2} n}{n} m_0^{1/2}\biggl(\frac{nr^2}{\log^{1/2} n}\biggr), \frac{\log n}{n} \biggr\},
\end{aligned}$$

and the result follows. □

PROOF OF PROPOSITION 10. As in the proof of Proposition 9, by (29) in the proof of Theorem 2, we see that we reject H_0 with probability at least $1-\beta$ provided that $n \geq 16$ and

$$\rho^2 \geq \min_{\substack{m_X \in K_X \\ m_Y \in K_Y}} \max\biggl\{ \frac{2r^2}{m_X^{2s_X} \vee m_Y^{2s_Y}}, \frac{1152A^2\gamma_X\gamma_Y}{n\alpha\beta}, \frac{1415A^4 |\mathcal{M}_{m_X,m_Y}|^{1/2}(\gamma_X\gamma_Y)^{1/2}}{\{n(n-1)\alpha\beta\}^{1/2}} \biggr\}.$$

Since $|\mathcal{M}_{m_X,m_Y}| \asymp_{d_X,d_Y} m_X^{d_X} m_Y^{d_Y}$, if (m_X, m_Y) were not restricted to lie in $K_X \times K_Y$, then we would maximise the right-hand side here by taking $m_X^{s_X} \asymp m_Y^{s_Y} \asymp_{\alpha,\beta,d_X,d_Y,A} (nr^2/\log n)^{\frac{2s}{4s+d}}$. In fact, recalling that $d/s = d_X/s_X + d_Y/s_Y$, we have that

$$(nr^2/\log n)^{\frac{2s}{s_X(4s+d)}} \lesssim_{R_0,s_X,s_Y,d_X,d_Y} \biggl(\frac{n}{\log n}\biggr)^{\frac{2}{4s_X + d_X + d_Y s_X/s_Y}} \ll n^{2/d_X} \leq 2^{\gamma_X}.$$

As in the proof of Proposition 9, then we may choose $(m_X, m_Y) \in K_X \times K_Y$ so as to ensure that the separation in (8) suffices to guarantee power at least $1-\beta$. □

PROOF OF LEMMA 11. We will prove that

$$d_{\mathrm{TV}}^2\bigl(\mathbb{P}_{p_{j_0k_0}}^{\otimes n}, \mathbb{E}\mathbb{P}_f^{\otimes n}\bigr) \leq \frac{\exp(\frac{n^2}{2}\sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a_{jk}^4)}{4\mathbb{P}(p \in \mathcal{F})^2} - \frac{1}{4}$$

in the case that n is even. If, on the other hand, n is odd then we will use the fact that $d_{\mathrm{TV}}(\nu_1^{\otimes n}, \nu_2^{\otimes n}) \leq d_{\mathrm{TV}}(\nu_1^{\otimes(n+1)}, \nu_2^{\otimes(n+1)})$ for any probability measures ν_1, ν_2 to complete the proof.

Let $f^{(1)}$, $f^{(2)}$ be independent copies of f and let $p^{(1)}$, $p^{(2)}$ be independent copies of p. Then we have that

$$\begin{aligned}
\frac{1}{4} + d_{\mathrm{TV}}^2\bigl(\mathbb{P}_{p_{j_0k_0}}^{\otimes n}, \mathbb{E}\mathbb{P}_f^{\otimes n}\bigr) &\leq \frac{1}{4} + \frac{1}{4} d_{\chi^2}^2\bigl(\mathbb{P}_{p_{j_0k_0}}^{\otimes n}, \mathbb{E}\mathbb{P}_f^{\otimes n}\bigr) \\
&= \frac{1}{4}\int_{\mathcal{X}\times\mathcal{Y}} \cdots \int_{\mathcal{X}\times\mathcal{Y}} \bigl(\mathbb{E}\{f(x_1,y_1)\ldots f(x_n,y_n)\}\bigr)^2 \, d\mu(x_n,y_n)\ldots d\mu(x_1,y_1)
\end{aligned}$$

$$= \frac{\mathbb{E}\{\langle p^{(1)}, p^{(2)}\rangle^n_{L^2(\mu)} \mathbb{1}_{\{p^{(1)}, p^{(2)} \in \mathcal{F}\}}\}}{4\mathbb{P}(p^{(1)}, p^{(2)} \in \mathcal{F})} \leq \frac{\mathbb{E}\{\langle p^{(1)}, p^{(2)}\rangle^n_{L^2(\mu)}\}}{4\mathbb{P}(p \in \mathcal{F})^2},$$

and all that remains is to bound the numerator in this final expression. Let $(\xi^{(1)}_{jk})$, $(\xi^{(2)}_{jk})$ be independent copies of (ξ_{jk}) and write

$$Y := \sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a^2_{jk} \xi^{(1)}_{jk} \xi^{(2)}_{jk} \stackrel{d}{=} \sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a^2_{jk} \xi_{jk}.$$

The random variable Y has a distribution that is symmetric about the origin, so for odd m we have $\mathbb{E}(Y^m) = 0$. For $m, r \in \mathbb{N}$ with $r \leq m$, write $A_{m,r} := \{\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r : \alpha_1 + \cdots + \alpha_r = m\}$ and $(2m-1)!! = (2m-1)(2m-3)\ldots 3 = \frac{(2m)!}{m!2^m}$ for the double factorial. It is also convenient to define the multinomial coefficient: for $N \in \mathbb{N}$ and $m_1, \ldots, m_r \in \mathbb{N}_0$ with $m_1 + \cdots + m_r = N$, we set

$$\binom{N}{m_1, m_2, \ldots, m_r} := \frac{N!}{m_1! m_2! \ldots m_r!}.$$

Then, for every $m \in \{0, 1, \ldots, n/2\}$, we have

$$\begin{aligned}
\mathbb{E}(Y^{2m}) &= \sum_{\substack{j_1,\ldots,j_m \in \mathcal{J}\setminus\{j_0\} \\ k_1,\ldots,k_{2m} \in \mathcal{K}\setminus\{k_0\}}} a^2_{j_1k_1} \ldots a^2_{j_{2m}k_{2m}} \mathbb{E}(\xi_{j_1k_1} \ldots \xi_{j_{2m}k_{2m}}) \\
&= \sum_{r=1}^{m} \sum_{\alpha \in A_{m,r}} \sum_{\substack{(j_1,k_1),\ldots,(j_r,k_r) \\ \text{distinct}}} a^{4\alpha_1}_{j_1k_1} \ldots a^{4\alpha_r}_{j_rk_r} \times \frac{1}{r!} \binom{2m}{2\alpha_1, 2\alpha_2, \ldots, 2\alpha_r} \\
&= \sum_{r=1}^{m} \sum_{\alpha \in A_{m,r}} \sum_{\substack{(j_1,k_1),\ldots,(j_r,k_r) \\ \text{distinct}}} a^{4\alpha_1}_{j_1k_1} \ldots a^{4\alpha_r}_{j_rk_r} \times \frac{(2m-1)!! \binom{m}{\alpha_1,\ldots,\alpha_r}}{r!(2\alpha_1-1)!! \ldots (2\alpha_r-1)!!} \\
&\leq \sum_{r=1}^{m} \sum_{\alpha \in A_{m,r}} \sum_{\substack{(j_1,k_1),\ldots,(j_r,k_r) \\ \text{distinct}}} a^{4\alpha_1}_{j_1k_1} \ldots a^{4\alpha_r}_{j_rk_r} \times \frac{(2m-1)!! \binom{m}{\alpha_1,\ldots,\alpha_r}}{r!} \\
&= (2m-1)!! \Bigg(\sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a^4_{jk} \Bigg)^m.
\end{aligned}$$

It therefore follows that

$$\begin{aligned}
\mathbb{E}\{\langle p^{(1)}, p^{(2)}\rangle^n_{L^2(\mu)}\} &= \mathbb{E}\{(1+Y)^n\} = \sum_{m=0}^{n/2} \binom{n}{2m} \mathbb{E}(Y^{2m}) \\
&\leq \sum_{m=0}^{n/2} \frac{1}{m!} \Bigg(\frac{n^2}{2} \sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a^4_{jk} \Bigg)^m \leq \exp\Bigg(\frac{n^2}{2} \sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a^4_{jk} \Bigg),
\end{aligned}$$

as required. □

PROOF OF THEOREM 12. For $m \in \mathbb{N}$, set

$$c_m := \min\Bigg(\frac{r^2}{\theta^2_{\omega(m)}}, \frac{(2m)^{1/2}}{n+1} \log^{1/2}(1 + (1-\gamma)^2), \frac{(A-1)^2 \wedge 1}{m\bar{p}^2} \Bigg)$$

and

$$a_{\omega(\ell)} := \begin{cases} c_m^{1/2}/m^{1/2} & \text{for } \ell \in [m], \\ 0 & \text{otherwise.} \end{cases}$$

Then, with the convention that $\infty \cdot 0 = 0$, we have

$$\sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} \theta_{jk}^2 a_{jk}^2 = \frac{c_m}{m} \sum_{\ell=1}^{m} \theta_{\omega(\ell)}^2 \le c_m \theta_{\omega(m)}^2 \le r^2. \tag{30}$$

Moreover,

$$\sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a_{jk}^4 = \frac{c_m^2}{m} \le \frac{2}{(n+1)^2} \log\bigl(1 + (1-\gamma)^2\bigr),$$

and

$$\sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a_{jk} \|p_{jk}\|_\infty = m^{1/2} c_m^{1/2} \bar{p} \le (A-1) \wedge 1. \tag{31}$$

Now, writing $\rho = \{\sum_{\mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a_{jk}^2\}^{1/2}$, observe that the random element p of $L^2(\mu)$ defined in Lemma 11 has $D(p) = \rho^2$ with probability one. Furthermore, from (30) and (31), we have with probability one that $p \in \mathcal{F}_\xi(\rho)$. Since only finitely many elements of the set $\{a_{jk} : j \in \mathcal{J} \setminus \{j_0\}, k \in \mathcal{K} \setminus \{k_0\}\}$ are nonzero, $\{p \in \mathcal{F}\}$ is an event, so by Lemma 11 and the discussion immediately following it, we have

$$\tilde{\rho}(n, \gamma, \xi)^2 \ge \sup_{m \in \mathbb{N}} \sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a_{jk}^2 = \sup_{m \in \mathbb{N}} c_m,$$

and the result follows. □

PROOF OF PROPOSITION 14. For $m = \lceil nr^2 \rceil^{2d/(4s+d)}$, we set

$$d_m = \min\biggl(\frac{r^2}{\theta_{\omega(m)}^2}, \frac{(2m)^{1/2}}{n+1} \log^{1/2}\bigl(1 + (1-\gamma)^2\bigr) \biggr) \asymp_{s_X, s_Y, d_X, d_Y, \gamma} \biggl(\frac{r^d}{n^{2s}} \biggr)^{2/(4s+d)}$$

and

$$a_{\omega(\ell)} := \begin{cases} d_m^{1/2}/m^{1/2} & \text{for } \ell \in [m], \\ 0 & \text{otherwise.} \end{cases}$$

The rest of this proof is dedicated to showing that, for the p constructed in the statement of Lemma 11, we have

$$\mathbb{P}(p \notin \mathcal{F}) = \mathbb{P}\bigl(\operatorname*{ess\,inf}_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) < 0 \bigr) < 1 - \sqrt{\frac{1 + (1-\gamma)^2}{1 + 4(1-\gamma)^2}},$$

from which the result will follow from Lemma 11. We define the random function

$$F(x, y) := 1 - p(x, y) = - \sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a_{jk} \xi_{jk} p_{jk}(x, y)$$

and aim to bound $\mathbb{P}(\operatorname{ess\,sup}_{x \in \mathcal{X}, y \in \mathcal{Y}} F(x, y) > 1)$. The space $\mathcal{X} \times \mathcal{Y}$ can be equipped with the pseudo-metric

$$\tau\bigl((x, y), (x', y')\bigr) := \biggl[\sum_{j \in \mathcal{J}\setminus\{j_0\}, k \in \mathcal{K}\setminus\{k_0\}} a_{jk}^2 \bigl\{ p_{jk}(x, y) - p_{jk}(x', y') \bigr\}^2 \biggr]^{1/2},$$

which satisfies

$$\delta := \sup_{x\in\mathcal{X},y\in\mathcal{Y}} \tau\big((x,y),(x_0,y_0)\big) \le 4\Big\{\sum_{j\in\mathcal{J}\setminus\{j_0\},k\in\mathcal{K}\setminus\{k_0\}} a_{jk}^2\Big\}^{1/2} = 4d_m^{1/2}$$

for any $(x_0,y_0)\in\mathcal{X}\times\mathcal{Y}$. Now, for $m=(m_1,\dots,m_{d_X})\in\mathbb{N}_0^{d_X}$ and $x=(x_1,\dots,x_{d_X})\in\mathcal{X}$, we write $\langle m,x\rangle_{\mathcal{X}} := \sum_{\ell=1}^{d_X} m_\ell x_\ell$; similarly, for $m=(m_1,\dots,m_{d_Y})\in\mathbb{N}_0^{d_Y}$ and $y=(y_1,\dots,y_{d_Y})\in\mathcal{Y}$, we write $\langle m,y\rangle_{\mathcal{Y}} := \sum_{\ell=1}^{d_Y} m_\ell y_\ell$. Then, for any $x,x'\in\mathcal{X}$ and $y,y'\in\mathcal{Y}$,

$$\begin{aligned}
&\tau\big((x,y),(x',y')\big)^2\\
&\quad\le 4\sum_{\substack{(a_X,m_X)\in\mathcal{J}\setminus\{j_0\}\\(a_Y,m_Y)\in\mathcal{K}\setminus\{k_0\}}} a_{jk}^2\big\{\big|e^{-2\pi i\langle m_X,x-x'\rangle_{\mathcal{X}}}-1\big|+\big|e^{-2\pi i\langle m_Y,y-y'\rangle_{\mathcal{Y}}}-1\big|\big\}^2\\
&\quad\le 32\pi^2\sum_{\substack{(a_X,m_X)\in\mathcal{J}\setminus\{j_0\}\\(a_Y,m_Y)\in\mathcal{K}\setminus\{k_0\}}} a_{jk}^2\big(1\wedge\langle m_X,x-x'\rangle_{\mathcal{X}}^2+1\wedge\langle m_Y,y-y'\rangle_{\mathcal{Y}}^2\big) \qquad (32)\\
&\quad\le 32\pi^2\sum_{\substack{(a_X,m_X)\in\mathcal{J}\setminus\{j_0\}\\(a_Y,m_Y)\in\mathcal{K}\setminus\{k_0\}}} a_{jk}^2\big\{\big(\|m_X\|_1\|x-x'\|_\infty\big)^{2(s_X\wedge1)}+\big(\|m_Y\|_1\|y-y'\|_\infty\big)^{2(s_Y\wedge1)}\big\}\\
&\quad\le 64\pi^2r^2\max\big\{\|x-x'\|_\infty^{2(s_X\wedge1)},\|y-y'\|_\infty^{2(s_Y\wedge1)}\big\}.
\end{aligned}$$

For $u,v>0$, let $H_\infty(u,\mathcal{X})$ and $H_\infty(v,\mathcal{Y})$ be the u- and v-metric entropies of $\mathcal{X}$ and $\mathcal{Y}$, respectively, with respect to the appropriate supremum metric; thus, for example, there exists $\mathcal{X}_N := \{x_1,\dots,x_N\}$, where $\log N = H(u,\mathcal{X})$, such that given any $x\in\mathcal{X}$, there exists $x_{j^*}\in\mathcal{X}_N$ with $\|x-x_{j^*}\|_\infty\le u$. It follows from (32) that, if $H(w,\mathcal{X}\times\mathcal{Y})$ is the w-metric entropy of $(\mathcal{X}\times\mathcal{Y},\tau)$ in the metric τ, then

$$\begin{aligned}
H(w,\mathcal{X}\times\mathcal{Y}) &\le H_\infty\Big(\Big(\frac{w}{8\pi r}\Big)^{1/(s_X\wedge1)},\mathcal{X}\Big)+H_\infty\Big(\Big(\frac{w}{8\pi r}\Big)^{1/(s_Y\wedge1)},\mathcal{Y}\Big)\\
&\le d_X\log\Big(1+\Big(\frac{8\pi r}{w}\Big)^{1/(s_X\wedge1)}\Big)+d_Y\log\Big(1+\Big(\frac{8\pi r}{w}\Big)^{1/(s_Y\wedge1)}\Big)\\
&\le\Big(\frac{d_X}{s_X\wedge1}+\frac{d_Y}{s_Y\wedge1}\Big)\log(1+8\pi r/w).
\end{aligned}$$

This choice of metric allows us to write, for any $\lambda\in\mathbb{R}$, $x,x'\in\mathcal{X}$ and $y,y'\in\mathcal{Y}$,

$$\begin{aligned}
\log\mathbb{E}e^{\lambda\{F(x,y)-F(x',y')\}} &= \sum_{\substack{j\in\mathcal{J}\setminus\{j_0\}\\k\in\mathcal{K}\setminus\{k_0\}}}\log\cosh\big(\lambda a_{jk}\{p_{jk}(x,y)-p_{jk}(x',y')\}\big)\\
&\le\frac{\lambda^2}{2}\sum_{\substack{j\in\mathcal{J}\setminus\{j_0\}\\k\in\mathcal{K}\setminus\{k_0\}}} a_{jk}^2\{p_{jk}(x,y)-p_{jk}(x',y')\}^2=\frac{\lambda^2}{2}\tau\big((x,y),(x',y')\big)^2. \qquad (33)
\end{aligned}$$

We now apply a chaining argument. For each $t\in\mathbb{N}$, let $\delta_t := \delta 2^{-t}$, and let $\mathcal{Z}_t$ denote a δ_t-net of $\mathcal{X}\times\mathcal{Y}$ with respect to the pseudo-metric τ. Let $z_0=(x_0,y_0)$ be an arbitrary element of $\mathcal{X}\times\mathcal{Y}$ and $\mathcal{Z}_0 := \{z_0\}$. Then, for each $t\in\mathbb{N}_0$, we can define a map $\Pi_t:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}_t$ such that $\tau(z,\Pi_t(z))\le\delta_t$. Noting that $\mathbb{E}F(x_0,y_0)=0$ and writing $F_t := F\circ\Pi_t$, we have for every $T\in\mathbb{N}$ that

$$\mathbb{E}\big(\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} F(x,y)\big)\le\mathbb{E}\big(\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} F_T(x,y)+\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}}\big|F(x,y)-F_T(x,y)\big|\big)$$

$$\leq \sum_{t=1}^{T} \mathbb{E}\big[\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} \{F_t(x,y) - F_{t-1}(x,y)\}\big] + \sum_{j\in\mathcal{J}\setminus\{j_0\},k\in\mathcal{K}\setminus\{k_0\}} a_{jk}\big|p_{jk}(x,y) - p_{jk}\big(\Pi_T(x,y)\big)\big|.$$

Now $\tau(\Pi_t(x,y), \Pi_{t-1}(x,y)) \leq 3\delta_t$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $t \in \mathbb{N}$. Hence, by (33) and a standard sub-Gaussian maximal inequality (e.g., Boucheron, Lugosi and Massart ((2013), Theorem 2.5)),

$$\mathbb{E}\big(\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} F(x,y)\big) \leq 6\sum_{t=1}^{T} \delta_t H^{1/2}(\delta_t, \mathcal{X}\times\mathcal{Y}) + m\delta_T \leq 12\int_0^{\delta/2} H^{1/2}(u, \mathcal{X}\times\mathcal{Y})\,du + m\delta_T.$$

Since this bound holds for every $T \in \mathbb{N}$, we conclude that

$$\begin{aligned}\mathbb{E}\big(\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} F(x,y)\big) &\leq 12\int_0^{\delta/2} H^{1/2}(u, \mathcal{X}\times\mathcal{Y})\,du \\ &\leq 96\pi\bigg(\frac{d_X}{s_X\wedge 1} + \frac{d_Y}{s_Y\wedge 1}\bigg)^{1/2} r\int_0^{\frac{d_m^{1/2}}{2\pi r}} \log^{1/2}(1+1/v)\,dv \\ &\leq 24 d_m^{1/2}\bigg(\frac{d_X}{s_X\wedge 1} + \frac{d_Y}{s_Y\wedge 1}\bigg)^{1/2}\bigg\{\sqrt{\pi} + 2\sqrt{\log 2} + 2\sqrt{\log\bigg(\frac{4\pi r}{d_m^{1/2}}\bigg)}\bigg\}.\end{aligned}$$

Now with $\zeta = (s_X, s_Y, d_X, d_Y, \gamma)$ we have $d_m^{1/2} \asymp_\zeta (r^d/n^{2s})^{1/(4s+d)}$, so that $r/d_m^{1/2} \asymp_\zeta (nr^2)^{2s/(4s+d)}$, and hence there exists $c_1 = c_1(\zeta) \in (0,\infty)$ such that if $(r^d/n^{2s})^{1/(4s+d)} \leq c_1/\log^{1/2}(nr^2)$, then $\mathbb{E}\operatorname{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} F(x,y) \leq 1/2$.

Now by, for example, Boucheron, Lugosi and Massart ((2013), Theorem 12.1), the random variable $\sup_{x\in\mathcal{X},y\in\mathcal{Y}} F(x,y)$ is sub-Gaussian with variance proxy

$$\sum_{j\in\mathcal{J}\setminus\{j_0\},k\in\mathcal{K}\setminus\{k_0\}} a_{jk}^2 \|p_{jk}\|_\infty^2 \leq 2d_m.$$

By reducing $c_1 = c_1(\zeta) > 0$ if necessary, and since $nr^2 \geq 2$, we may assume that

$$d_m < -\frac{1}{16}\log\bigg(1 - \sqrt{\frac{1+(1-\gamma)^2}{1+4(1-\gamma)^2}}\bigg).$$

Hence, by a standard sub-Gaussian tail bound (e.g., Boucheron, Lugosi and Massart ((2013), page 25))

$$\begin{aligned}\mathbb{P}(p\notin\mathcal{F}) &\leq \mathbb{P}\big(\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} F(x,y) - \mathbb{E}\operatorname*{ess\,sup}_{x\in\mathcal{X},y\in\mathcal{Y}} F(x,y) > 1/2\big) \\ &\leq e^{-1/(16d_m)} < 1 - \sqrt{\frac{1+(1-\gamma)^2}{1+4(1-\gamma)^2}},\end{aligned}$$

as required. □

Acknowledgments. We are very grateful to the anonymous reviewers, whose constructive feedback helped to improve the paper. We would also like to thank Ilmun Kim for bringing the work of Song et al. (2012) to our attention; this inspired the computational improvements discussed in Section 7.1.

Funding. The first author was supported by the French National Research Agency (ANR) under the grants Labex Ecodec (ANR-11-LABEX-0047 and ANR-17-CE40-0003.

The third author was supported by Engineering and Physical Sciences Reseach Council (EPSRC) Programme grant EP/N031938/1 and EPSRC Fellowship EP/P031447/1.

SUPPLEMENTARY MATERIAL

Supplementary material: Optimal rates for independence testing via U-statistic permutation tests (DOI: 10.1214/20-AOS2041SUPP; .pdf). The supplement contains the remaining proofs of main results, as well as some auxiliary results.

REFERENCES

ALBERT, M. (2015). Tests of independence by bootstrap and permutation: An asymptotic and non-asymptotic study. Application to neurosciences. Ph.D. thesis. Available at https://tel.archives-ouvertes.fr/tel-01274647/file/2015NICE4079.pdf.

ALBERT, M., BOURET, Y., FROMONT, M. and REYNAUD-BOURET, P. (2015). Bootstrap and permutation tests of independence for point processes. *Ann. Statist.* **43** 2537–2564. MR3405603 https://doi.org/10.1214/15-AOS1351

ANTOCH, J. and HUŠKOVÁ, M. (2001). Permutation tests in change point analysis. *Statist. Probab. Lett.* **53** 37–46. MR1843339 https://doi.org/10.1016/S0167-7152(01)00009-8

BERRETT, T. B., GROSE, D. J. and SAMWORTH, R. J. (2018). IndepTest: Nonparametric independence tests based on entropy estimation. *R Package version 0.2.0*. Available at https://cran.r-project.org/web/packages/IndepTest/index.html.

BERRETT, T. B., KONTOYIANNIS, I. and SAMWORTH, R. J. (2020). USP: U-statistic permutation tests of independence for all data types, with improvement on Pearson's chi-squared test for discrete data. *R package version 0.1.0*. Available at https://cran.r-project.org/web/packages/USP/index.html.

BERRETT, T. B., KONTOYIANNIS, I. and SAMWORTH, R. J. (2021). Supplement to "Optimal rates for independence testing via U-statistic permutation tests." https://doi.org/10.1214/20-AOS2041SUPP

BERRETT, T. B. and SAMWORTH, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. MR3992389 https://doi.org/10.1093/biomet/asz024

BERRETT, T. B., WANG, Y., BARBER, R. F. and SAMWORTH, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 175–197. MR4060981

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. With a foreword by Michel Ledoux. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111 https://doi.org/10.1214/13-AOS1090

CHUNG, E. and ROMANO, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *J. Statist. Plann. Inference* **168** 97–105. MR3412224 https://doi.org/10.1016/j.jspi.2015.07.004

DE JONG, P. (1990). A central limit theorem for generalized multilinear forms. *J. Multivariate Anal.* **34** 275–289. MR1073110 https://doi.org/10.1016/0047-259X(90)90040-O

DEB, N. and SEN, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. Available at arXiv:1909.08733.

DIAKONIKOLAS, I. and KANE, D. M. (2016). A new approach for testing properties of discrete distributions. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 685–694. IEEE Comput. Soc., Los Alamitos, CA. MR3631031

DÖBLER, C. and PECCATI, G. (2017). Quantitative de Jong theorems in any dimension. *Electron. J. Probab.* **22** Paper No. 2. MR3613695 https://doi.org/10.1214/16-EJP19

DÖBLER, C. and PECCATI, G. (2019). Quantitative CLTs for symmetric U-statistics using contractions. *Electron. J. Probab.* **24** Paper No. 5. MR3916325 https://doi.org/10.1214/19-EJP264

ERMAKOV, M. S. (1990). Asymptotically minimax tests for nonparametric hypotheses concerning the distribution density. *J. Sov. Math.* **52** 2891–2898.

FISHER, R. A. (1935). *The Design of Experiments*, 1st ed. Oliver & Boyd, Edinburgh.

GABRYS, R. and KOKOSZKA, P. (2007). Portmanteau test of independence for functional observations. *J. Amer. Statist. Assoc.* **102** 1338–1348. MR2412554 https://doi.org/10.1198/016214507000001111

GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **3734** 63–77. Springer, Berlin. MR2255909 https://doi.org/10.1007/11564089_7

HALL, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14** 1–16. MR0734096 https://doi.org/10.1016/0047-259X(84)90044-7

HELLER, R., HELLER, Y., KAUFMAN, S., BRILL, B. and GORFINE, M. (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.* **17** Paper No. 29. MR3491123

HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19** 546–557. MR0029139 https://doi.org/10.1214/aoms/1177730150

HOFERT, M., KOJADINOVIC, I., MÄCHLER, M. and YAN, J. (2017). copula: Multivariate dependence with copulas. *R Package version 0.999-18*. Available at https://cran.r-project.org/web/packages/copula/index.html.

INGSTER, Y. I. (1989). Asymptotic minimax testing of independence hypothesis. *J. Sov. Math.* **44** 466–476.

INGSTER, Y. I. (1996). Minimax testing of the hypothesis of independence for ellipsoids in ℓ_p. *J. Math. Sci.* **81** 2406–2420.

JANSSEN, A. (2000). Global power functions of goodness of fit tests. *Ann. Statist.* **28** 239–253. MR1762910 https://doi.org/10.1214/aos/1016120371

KAHANE, J.-P. (1997). A century of interplay between Taylor series, Fourier series and Brownian motion. *Bull. Lond. Math. Soc.* **29** 257–279. MR1435557 https://doi.org/10.1112/S0024609396002913

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–89.

KIM, I., BALAKRISHNAN, S. and WASSERMAN, L. (2020). Minimax optimality of permutation tests. Available at https://arxiv.org/abs/2003.13208.

KOJADINOVIC, I. and HOLMES, M. (2009). Tests of independence among continuous random vectors based on Cramér–von Mises functionals of the empirical copula process. *J. Multivariate Anal.* **100** 1137–1154. MR2508377 https://doi.org/10.1016/j.jmva.2008.10.013

LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681. MR1394981 https://doi.org/10.1214/aos/1032894458

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

LI, T. and YUAN, M. (2019). On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. Available at arXiv:1909.03302.

MEYNAOUI, A., ALBERT, M., LAURENT, B. and MARREL, A. (2019). Adaptive test of independence based on HSIC measures. Available at arXiv:1902.06441.

NGUYEN, D. and EISENSTEIN, J. (2017). A kernel independence test for geographical language variation. *Comput. Linguist.* **43** 567–592. MR3708620 https://doi.org/10.1162/COLI_a_00293

PATEFIELD, W. M. (1981). Algorithm AS159. An efficient method of generating $r \times c$ tables with given row and column totals. *J. Roy. Statist. Soc. Ser. C* **30** 91–97.

PEARSON, K. (1920). Notes on the history of correlation. *Biometrika* **13** 25–45.

PESARIN, F. and SALMASO, L. (2010). *Permutation Tests for Complex Data. Theory, Applications and Software*. Wiley, Chichester, UK.

PFISTER, N. and PETERS, J. (2017). dHSIC: Independence testing via Hilbert Schmidt independence criterion. *R Package version 2.0*. https://cran.r-project.org/web/packages/dHSIC/index.html.

PFISTER, N., BÜHLMANN, P., SCHÖLKOPF, B. and PETERS, J. (2018). Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 5–31. MR3744710 https://doi.org/10.1111/rssb.12235

PITMAN, E. J. G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika* **29** 322–335.

RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., MCVEAN, G., TURNBAUGH, P. J., LANDER, E. S., MITZENMACHER, M. and SABETI, P. C. (2011). Detecting novel associations in large data sets. *Science* **334** 1518–1524.

RINOTT, Y. and ROTAR, V. (1997). On coupling constructions and rates in the CLT for dependent summands with applications to the antivoter model and weighted U-statistics. *Ann. Appl. Probab.* **7** 1080–1105. MR1484798 https://doi.org/10.1214/aoap/1043862425

RIZZO, M. L. and SZEKELY, G. J. (2017). energy: E-statistics: Multivariate inference via the energy of data. *R Package version 1.7-2*. Available at: https://cran.r-project.org/web/packages/energy/index.html.

ROMANO, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17** 141–159. MR0981441 https://doi.org/10.1214/aos/1176347007

SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. MR3127866 https://doi.org/10.1214/13-AOS1140

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0595165

SHAH, R. D. and PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann*. *Statist*. **48** 1514–1538. MR4124333 https://doi.org/10.1214/19-AOS1857

SHI, H., DRTON, M. and HAN, F. (2020). Distribution-free consistent independence tests via center-outward ranks and signs. *J*. *Amer*. *Statist*. *Assoc*. To appear. https://doi.org/10.1080/01621459.2020.1782223

SONG, L., SMOLA, A., GRETTON, A., BEDO, J. and BORGWARDT, K. (2012). Feature selection via dependence maximization. *J*. *Mach*. *Learn*. *Res*. **13** 1393–1434. MR2930643

SPEARMAN, C. (1904). The proof and measurement of association between two things. *Am*. *J*. *Psychol*. **15** 72–101.

STEUER, R., KURTHS, J., DAUB, C. O., WEISE, J. and SELBIG, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **18** 231–240.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann*. *Statist*. **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505