The Annals of Statistics
2020, Vol. 48, No. 5, 2879–2903
https://doi.org/10.1214/19-AOS1913

ASYMPTOTIC DISTRIBUTION AND DETECTION THRESHOLDS FOR TWO-SAMPLE TESTS BASED ON GEOMETRIC GRAPHS

BY BHASWAR B. BHATTACHARYA

Department of Statistics, The Wharton School, University of Pennsylvania, bhaswar@wharton.upenn.edu

In this paper, we consider the problem of testing the equality of two multivariate distributions based on geometric graphs constructed using the interpoint distances between the observations. These include the tests based on the minimum spanning tree and the K-nearest neighbor (NN) graphs, among others. These tests are asymptotically distribution-free, universally consistent and computationally efficient, making them particularly useful in modern applications. However, very little is known about the power properties of these tests. In this paper, using the theory of stabilizing geometric graphs, we derive the asymptotic distribution of these tests under general alternatives, in the Poissonized setting. Using this, the detection threshold and the limiting local power of the test based on the K-NN graph are obtained, where interesting exponents depending on dimension emerge. This provides a way to compare and justify the performance of these tests in different examples.

1. Introduction.

Given independent and identically distributed samples

$$\mathscr{X}_{N_1} = \{X_1, X_2, \ldots, X_{N_1}\} \quad \text{and} \quad \mathscr{Y}_{N_2} = \{Y_1, Y_2, \ldots, Y_{N_2}\}, \tag{1.1}$$

from two unknown densities f and g (with respect to the Lebesgue measure) in $\mathbb{R}^d$, respectively, the two-sample problem is to test the hypotheses

$$H_0 : f = g \quad \text{versus} \quad H_1 : f \neq g. \tag{1.2}$$

In this paper, we will derive asymptotic properties of two-sample tests based on geometric graphs in the usual limiting regime where the dimension d is fixed and the sample size $N_1 + N_2 := N \to \infty$, such that

$$\frac{N_1}{N_1 + N_2} \to p \in (0, 1), \qquad \frac{N_2}{N_1 + N_2} \to q := 1 - p. \tag{1.3}$$

For univariate data, there are several well-known nonparametric tests such as the two-sample Kolmogorov–Smirnoff maximum deviation test [29], the Wald–Wolfowitz runs test [31] and the Mann–Whitney rank test [22].

The nonparametric two-sample problem for multivariate data has been extensively studied, beginning with the work of Weiss [32] and Bickel [8]. Friedman and Rafsky [14] generalized the Wald–Wolfowitz runs test [31] to higher dimensions using the Euclidean minimal spanning tree (MST) of the pooled data. Thereafter, many other two-sample tests based on geometric graphs have been proposed. Schilling [28] and Henze [18] considered tests based on the K-nearest neighbor (K-NN) graph of the pooled sample. Later, Rosenbaum [26] developed a test based on matchings, and, more recently, Biswas et al. [10] proposed a test based on the Hamiltonian cycle, both of which are exactly distribution-free under the null. Recently, Chen and Friedman [12] proposed new modifications of these tests for high-dimensional and

Received March 2019; revised September 2019.
MSC2020 subject classifications. 62F07, 62G10, 60D05, 60F05, 60C05.
Key words and phrases. Efficiency, local power, geometric probability, nearest-neighbor graphs, nonparametric hypothesis testing.

object data. Maa et al. [21] provided certain theoretical motivations for using tests based on interpoint distances.

Another class of multivariate two-sample tests is the Liu–Singh rank sum statistics [20], which generalize the Mann–Whitney rank test using the notion of data depth [20, 30]. For other popular two-sample tests, refer to [3, 4, 15, 17, 27] and the references therein. The problem of testing the equality of two discrete distributions has also been extensively studied in recent years [5, 11].

1.1. *Graph-based two-sample tests*. Many of the tests mentioned above can be studied in the general framework of graph-based two-sample tests [6], which include the tests based on geometric graphs, as well as those based on data depth. To this end, we have the following definition: A *graph functional* $\mathscr{G}$ in $\mathbb{R}^d$ defines a graph for all finite subsets of $\mathbb{R}^d$, that is, given $S \subset \mathbb{R}^d$ finite, $\mathscr{G}(S)$ is a graph with vertex set S. A graph functional is said to be *undirected/directed* if the graph $\mathscr{G}(S)$ is an undirected/directed graph with vertex set S. We assume that $\mathscr{G}(S)$ has no self loops and multiple edges, that is, no edge is repeated more than once in the undirected case, and no edge in the same direction is repeated more than once in the directed case. The set of edges in the graph $\mathscr{G}(S)$ will be denoted by $E(\mathscr{G}(S))$.

DEFINITION 1.1 (Bhattacharya [6]). Let $\mathscr{X}_{N_1}$ and $\mathscr{Y}_{N_2}$ be i.i.d. samples of size N_1 and N_2 from densities f and g, respectively, as in (1.1). The *two-sample test statistic based on the graph functional* $\mathscr{G}$ is defined as

$$T\big(\mathscr{G}(\mathscr{X}_{N_1} \cup \mathscr{Y}_{N_2})\big) := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \mathbf{1}\big\{(X_i, Y_j) \in E\big(\mathscr{G}(\mathscr{X}_{N_1} \cup \mathscr{Y}_{N_2})\big)\big\}. \tag{1.4}$$

If $\mathscr{G}$ is an undirected graph functional, then the statistic (1.4) counts the number of edges in the graph $\mathscr{G}(\mathscr{X}_{N_1} \cup \mathscr{Y}_{N_2})$ with one end point in $\mathscr{X}_{N_1}$ and the other in $\mathscr{Y}_{N_2}$. If $\mathscr{G}$ is a directed graph functional, then (1.4) is the number of directed edges with the outward end in $\mathscr{X}_{N_1}$ and the inward end in $\mathscr{Y}_{N_2}$. The null hypothesis is generally rejected for "small" values of the statistic (1.4). This includes the Friedman–Rafsky (FR) test [14] (based on the MST), the test based on the K-NN graph [18, 28], the cross match test [26] (based on minimum non-bipartite matching), among others. These tests are asymptotically distribution-free, universally consistent and computationally efficient (both in sample size and in dimension), making them particularly attractive for modern statistical applications.

1.2. *Poissonization*. In the Poissonized setting, instead of taking N_1 samples from the density f and N_2 from the density g, we have $\mathrm{Pois}(N_1)$ from f and $\mathrm{Pois}(N_2)$ samples from g. To this end, suppose $\mathscr{X} = \{X_1, X_2, \ldots\}$ and $\mathscr{Y} = \{Y_1, Y_2, \ldots\}$ be i.i.d. samples from f and g, respectively, and

$$\mathscr{X}'_{N_1} = \{X_1, X_2, \ldots, X_{L_{N_1}}\} \quad \text{and} \quad \mathscr{Y}'_{N_2} = \{Y_1, Y_2, \ldots, Y_{L_{N_2}}\}, \tag{1.5}$$

where $L_{N_1} \sim \mathrm{Pois}(N_1)$ and $L_{N_2} \sim \mathrm{Pois}(N_2)$ are independent of each other, and of $\mathscr{X}$ and $\mathscr{Y}$. Poissonization is a common assumption in geometric probability, which simplifies calculations, due to the spatial independence of the Poisson process, and yields cleaner formulas for the asymptotic variances. One can expect to de-Poissionize the limit theorems derived below, using well-known de-Poissonization methods [23, 24]. However, de-Poissonization does not affect the rates of convergence, and the detection thresholds obtained below would remain unchanged (see Remark 3.3 for more on de-Poissonization).

Given a graph functional $\mathscr{G}$, the Poissonized two-sample statistic is defined as

$$T\big(\mathscr{G}\big(\mathscr{X}'_{N_1} \cup \mathscr{Y}'_{N_2}\big)\big) := \sum_{i=1}^{L_{N_1}} \sum_{j=1}^{L_{N_2}} \mathbf{1}\big\{(X_i, Y_j) \in E\big(\mathscr{G}\big(\mathscr{X}'_{N_1} \cup \mathscr{Y}'_{N_2}\big)\big)\big\}. \tag{1.6}$$

The distribution of this statistic can be described as follows: Let $\phi_N(x) := \frac{N_1}{N} f(x) + \frac{N_2}{N} g(x)$ and $Z_1, Z_2, \ldots,$ be independent random variables with common density $\phi_N(\cdot)$. Let L_N be an independent Poisson variable with mean $N_1 + N_2$. Then $\mathcal{Z}'_N = \{Z_1, Z_2, \ldots, Z_{L_N}\}$ is a nonhomogeneous Poisson process in $\mathbb{R}^d$ with rate function $N\phi_N = N_1 f + N_2 g$. Label each point of $z \in \mathcal{Z}'_N$ independently with

$$c_z = \begin{cases} 1 & \text{with probability } \dfrac{N_1 f(z)}{N_1 f(z) + N_2 g(z)}, \\ 2 & \text{with probability } \dfrac{N_2 g(z)}{N_1 f(z) + N_2 g(z)}. \end{cases} \tag{1.7}$$

Then the sets of points assigned labels 1 and 2 have the same distribution as $\mathscr{X}'_{N_1}$ and $\mathscr{Y}'_{N_2}$ (as in (1.5)), respectively. This implies that for a directed graph functional $\mathscr{G}$, the Poissonized two-sample test statistic (1.6) is equal in distribution to

$$T\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big) = \sum_{x,y \in \mathcal{Z}'_N} \psi(c_x, c_y) \mathbf{1}\big\{(x, y) \in E\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big)\big\}, \tag{1.8}$$

where $\psi(c_x, c_y) = \mathbf{1}\{c_x = 1, c_y = 2\}$. (Note that every undirected graph functional $\mathscr{G}$ can be modified to a directed graph functional $\mathscr{G}_+$ in a natural way: For $S \subset \mathbb{R}^d$ finite, $\mathscr{G}_+(S)$ is obtained by replacing every edge in $\mathscr{G}(S)$ with two directed edges, one in each direction. Thus, without loss of generality, it suffices to consider directed graph functionals.)

Denote by $\mathbb{E}_{H_0}$ and $\mathbb{E}_{H_1}$ the expectation under the null and the alternative, respectively. For a directed graph functional $\mathscr{G}$,

$$\mathbb{E}_{H_0}\big(T\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big)\big) = \frac{N_1 N_2}{(N_1 + N_2)^2} \mathbb{E}\big(\big|E\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big)\big|\big),$$

where $|E(\mathscr{G}(\mathcal{Z}'_N))|$ denotes the number of edges in the graph $\mathscr{G}(\mathcal{Z}'_N)$. For example, in the MST functional, $\mathbb{E}(|E(\mathcal{T}(\mathcal{Z}'_N))|) = N - 1$, and in the directed K-NN graph functional $\mathbb{E}(|E(\mathcal{N}_K(\mathcal{Z}'_N))|) = KN$, respectively. (Formal definitions of these graph-functionals are given in Section 1.3 below.) We will see later in Section 3 that for many geometric graphs, such as the MST and the K-NN graph, the statistic $T(\mathscr{G}(\mathcal{Z}'_N))$ is asymptotically normal and distribution-free under the null H_0, that is, $N^{-\frac{1}{2}}\{T(\mathscr{G}(\mathcal{Z}'_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}'_N)))\} \xrightarrow{D} N(0, \sigma^2_{\mathscr{G}})$, where $\sigma_{\mathscr{G}}$ depends on the graph functional $\mathscr{G}$, but not on the unknown null distribution. For such a graph functional $\mathscr{G}$, the asymptotically level α-test rejects H_0 when

$$\frac{1}{\sqrt{N}}\big\{T\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big) - \mathbb{E}_{H_0}\big(T\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big)\big)\big\} \le \sigma_{\mathscr{G}} z_\alpha, \tag{1.9}$$

where z_α is the standard normal quantile of level α.

1.3. *Stabilizing graphs.* Many geometric graphs such as the MST and the K-NN graph, have local dependence, that is, addition/deletion of a point only effects the edges incident on the neighborhood of that point. This was formalized by Penrose and Yukich [25], using the notion of stabilization. To describe this, a few definitions are needed: A subset $S \subset \mathbb{R}^d$ is said to be *locally finite*, if $S \cap C$ is finite, for all compact subsets $C \subset \mathbb{R}^d$. A locally finite set $S \subset \mathbb{R}^d$ is *nice* if all the interpoint distances among elements of S are distinct. If S is a set

of N i.i.d. points $W_1, W_2, \ldots, W_N$ from some continuous distribution function F, then the distribution of $\|W_1 - W_2\|$ does not have any point mass, and S is nice almost surely.

Let $\mathscr{G}$ be a graph functional defined for all locally finite subsets of $\mathbb{R}^d$. For $S \subset \mathbb{R}^d$ nice and $x \in \mathbb{R}^d$, let $E(x, \mathscr{G}(S))$ be the set edges incident on x in $\mathscr{G}(S \cup \{x\})$. Note that $|E(x, \mathscr{G}(S))| := d(x, \mathscr{G}(S))$, the (total) degree of the vertex x in $\mathscr{G}(S \cup \{x\})$. Finally, note that two graphs H_1, H_2 are said to be isomorphic if there is a bijection ϕ from the vertex set of H_1 to the vertex set of H_2 such that any two vertices u and v of H_1 are adjacent in H_1 if and only if $\phi(u)$ and $\phi(v)$ are adjacent in H_2.

DEFINITION 1.2. Given $S \subset \mathbb{R}^d$, $y \in \mathbb{R}^d$, and $a \in \mathbb{R}$, denote by $y + S = \{y + z : z \in S\}$ and $aS = \{az : z \in S\}$. A graph functional $\mathscr{G}$ is said to be *translation invariant* if the graphs $\mathscr{G}(x + S)$ and $\mathscr{G}(S)$ are isomorphic for all points $x \in \mathbb{R}^d$ and all locally finite $S \subset \mathbb{R}^d$. A graph functional $\mathscr{G}$ is *scale invariant* if $\mathscr{G}(aS)$ and $\mathscr{G}(S)$ are isomorphic for all points $a \in \mathbb{R}$ and and all locally finite $S \subset \mathbb{R}^d$.

For $\lambda \geq 0$, denote by $\mathcal{P}_\lambda$ the homogeneous Poisson process of intensity λ in $\mathbb{R}^d$, and $\mathcal{P}_\lambda^x := \mathcal{P}_\lambda \cup \{x\}$, for $x \in \mathbb{R}^d$. Penrose and Yukich [25] defined stabilization of graph functionals over homogeneous Poisson processes as follows.

DEFINITION 1.3 (Penrose and Yukich [25]). A translation and scale invariant graph functional $\mathscr{G}$ *stabilizes* on $\mathcal{P}_\lambda$ if, for almost all realizations $\mathcal{P}_\lambda$, there exists $R := R(\mathcal{P}_\lambda) < \infty$ such that

$$E\big(0, \mathscr{G}\big(\mathcal{P}_\lambda^0\big)\big) \stackrel{a.s.}{=} E\big(0, \mathscr{G}\big(\mathcal{P}_\lambda^0 \cap B(0, R) \cup \mathscr{A}\big)\big), \tag{1.10}$$

for all finite $\mathscr{A} \subset \mathbb{R}^d \setminus B(0, R)$, where $B(0, R)$ is the (Euclidean) ball of radius R centered at the origin $0 \in \mathbb{R}^d$.

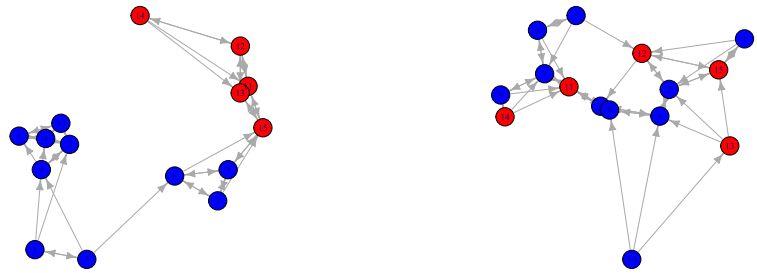
Informally, a graph functional is stabilizing if addition of finitely many points outside a ball of finite radius centered at the origin, does not effect the set of edges incident at the origin. The K-NN graph and the minimum spanning tree are known to be stabilizing ([25], Lemma 2.1). We discuss the two-sample tests associated with these graphs below.

1.3.1. *Friedman–Rafsky (FR) test.* Friedman and Rafsky [14] generalized the Wald and Wolfowitz runs test to higher dimensions by using the Euclidean minimal spanning tree of the pooled sample.

DEFINITION 1.4. Given a nice finite set $S \subset \mathbb{R}^d$, a *spanning tree* of S is a connected graph $\mathcal{T}$ with vertex-set S and no cycles. The *length* $w(\mathcal{T})$ of $\mathcal{T}$ is the sum of the Euclidean lengths of the edges of $\mathcal{T}$. A *minimum spanning tree* (MST) of S, denoted by $\mathcal{T}(S)$, is a spanning tree with the smallest length, that is, $w(\mathcal{T}(S)) \leq w(\mathcal{T})$ for all spanning trees $\mathcal{T}$ of S.

Thus, $\mathcal{T}$ defines a graph functional in $\mathbb{R}^d$, and given $\mathscr{X}'_{N_1}$ and $\mathscr{Y}'_{N_2}$ as in (1.5), the FR-test rejects H_0 for small values of

$$\begin{aligned} T\big(\mathcal{T}\big(\mathcal{Z}'_N\big)\big) &= \sum_{x,y \in \mathcal{Z}'_N} \mathbf{1}\{c_x \neq c_y\} \mathbf{1}\big\{(x, y) \in E\big(\mathcal{T}\big(\mathcal{Z}'_N\big)\big)\big\}, \\ &= \sum_{x,y \in \mathcal{Z}'_N} \psi(c_x, c_y) \mathbf{1}\big\{(x, y) \in E\big(\mathcal{T}_+\big(\mathcal{Z}'_N\big)\big)\big\}, \end{aligned} \tag{1.11}$$



(a) (b)

FIG. 1. *The directed* 3*-NN graph on a pooled sample of size* 15 *in* $\mathbb{R}^2$ *with* 10 *i.i.d. points from* $N(0, \mathrm{I}_2)$ (*colored blue*) *and* 5 *i.i.d. points from* $N(\Delta \cdot \mathbf{1}, \mathrm{I}_2)$ (*colored red*). *For* (*a*) $\Delta = 2$, *there are* 3 *directed edges starting from sample* 1 *and ending in sample* 2, *and for* (*b*) $\Delta = 0.05$, *there are* 8 *such edges*.

where $\mathcal{Z}'_N = \mathcal{X}'_{N_1} \cup \mathcal{Y}'_{N_2}$ and $\mathcal{T}_+(\mathcal{Z}'_N)$ is obtained by replacing every (undirected) edge in $\mathcal{T}(\mathcal{Z}'_N)$ with two directed edges, one in each direction. Note that this counts the number of edges in the MST of the pooled sample with one end-point in sample 1 and the other end-point in sample 2, which is expected to be small when the two distributions are different. Note that this reduces to the well-known Wald–Wolfowitz runs test when dimension $d = 1$, where the MST is the path through the data.

Friedman and Rafsky [14] calibrated (1.11) as a permutation test, and showed that it has good power in finite sample simulations. Later, Henze and Penrose [19] proved that the statistic $T(\mathcal{T}(\mathcal{Z}'_N))$ is asymptotically normal under H_0 and is consistent under all fixed alternatives.

1.3.2. *Test based on the K-nearest neighbor* (*K-NN*) *graph.* As in (1.11), a multivariate two-sample test can be constructed using the K-nearest neighbor graph of $\mathcal{Z}'_N$. This was originally suggested by Friedman and Rafsky [14], and later studied by Schilling [28] and Henze [18].

DEFINITION 1.5. Given a nice finite set $S \subset \mathbb{R}^d$, the (directed) K-nearest neighbor graph (K-NN) is a graph with vertex set S with a directed edge (a, b), for $a, b \in S$, if the Euclidean distance between a and b is among the K-th smallest distances from a to any other point in S. Denote the directed K-NN of S by $\mathcal{N}_K(S)$.

Given $\mathcal{Z}'_N = \mathcal{X}'_{N_1} \cup \mathcal{Y}'_{N_2}$ as in (1.5), the K-NN statistic is

$$T(\mathcal{N}_K(\mathcal{Z}'_N)) = \sum_{x,y \in \mathcal{Z}'_N} \psi(c_x, c_y) \mathbf{1}\{(x, y) \in E(\mathcal{N}_K(\mathcal{Z}'_N))\}. \tag{1.12}$$

As before, when the two distributions are different, the number of directed edges starting from sample 1 and ending in sample 2 will be small (see Figure 1), so the K-NN test rejects H_0 for small values of (1.12). This will be our main running example throughout the paper.

Another variant is the *symmetrized K-NN test statistic* [28]:

$$T_S(\mathcal{N}_K(\mathcal{Z}'_N)) = \sum_{x,y \in \mathcal{Z}'_N} \psi_S(c_x, c_y) \mathbf{1}\{(x, y) \in E(\mathcal{N}_K(\mathcal{Z}'_N))\}, \tag{1.13}$$

where $\psi_S(c_x, c_y) = \mathbf{1}\{c_x \neq c_y\}$, which counts the number of (directed) edges with the end-points in the different samples. This can be rewritten as a graph-based test (1.8) by considering the underlying undirected multigraph (which allows for multiple edges between two vertices).

1.4. *Summary of results.* The asymptotic null distribution and consistency of the tests described above are well known (see [19] for the FR test and [18, 28] for the K-NN test). However, a mathematical treatment of the power properties of these tests, which requires understanding the limiting distribution of the test statistics under the alternative, remained unavailable. In this paper, we address this problem by deriving the asymptotic distribution of (1.8), for stabilizing geometric graph functionals, under general alternatives, in the Poissonized setting described above. As a consequence, the exact detection threshold and the limiting local power of these tests can be derived.

We begin with a few notations: For a vector $x \in \mathbb{R}^p$, $\|x\|$ and $\|x\|_1$ will denote the L_2 and L_1 norms of x, respectively. For two nonnegative sequences, $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, $a_n = \Theta(b_n)$ means that there exist positive constants C_1, C_2, such that $C_1 b_n \leq a_n \leq C_2 b_n$, for all n large enough. Finally, for two positive sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we write $a_n \ll b_n$ or $a_n \gg b_n$, if $a_n/b_n \to 0$ or $a_n/b_n \to \infty$, respectively. The results obtained in this paper are summarized below.

1. The limiting distribution of graph-based two-sample tests under general alternatives is derived. The proof of this general result has two main steps: To begin with we show that for tests based on stabilizing geometric graphs, such as the Friedman–Rafsky test (1.11) and the test based on the K-nearest-neighbor (K-NN) graph (1.12), the statistic (1.8) has a limiting normal distribution, after centering by the conditional mean and scaling by $N^{-\frac{1}{2}}$ (Theorem 3.1). This result is of independent interest, as it leads to a new conditional test, and can be used for approximate power calculations (Remark 3.2). Next, under the stronger assumption of *exponential stabilization* [24], the conditional CLT can be strengthened to obtain the (unconditional) central limit theorem of (1.8) (Theorem 3.3).

2. The CLT proved above can be used to determine the detection threshold of the K-NN test, that is, the rate at which the alternatives shrink toward the null, such that the limiting power of the test transitions from 0 to 1. More precisely, suppose $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is a parametric family of distributions in $\mathbb{R}^d$, indexed by $\theta \in \Theta \subseteq \mathbb{R}^p$. Given samples $\mathscr{X}'_{N_1}$ and $\mathscr{Y}'_{N_2}$ from $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ as in (1.5), respectively, consider the testing problem

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N,$$

for a sequence $(\varepsilon_N)_{N\geq 1}$ in $\mathbb{R}^p$, such that $\|\varepsilon_N\| \to 0$. The detection threshold for the K-NN test is the magnitude of the sequence ε_N below which the test is powerless and above which the test has power going to 1. The parametric rate of detection is $O(N^{-\frac{1}{2}})$; however, results in [6] imply that tests based on geometric graphs, have no power in this scale, that is, they have zero Pitman efficiency, which makes the problem of determining the detection threshold of such tests particularly interesting. In Theorem 4.2, we determine the precise detection threshold of the K-NN test, which undergoes a remarkable phase transition at dimension $d \geq 9$, and compute the exact limiting power at the threshold. The result is pictorially represented in Figure 2 and summarized below:

- For dimension $d \leq 8$, the detection threshold of the test based on the K-NN graph (4.3) is at $\Theta(N^{-\frac{1}{4}})$, that is, the limiting power of the test undergoes a phase transition from the level α to 1, depending on whether $\|N^{\frac{1}{4}}\varepsilon_N\| \to 0$ or $\|N^{\frac{1}{4}}\varepsilon_N\| \to \infty$, respectively. Moreover, using the CLT above, we can derive the exact local power at the threshold $N^{\frac{1}{4}}\varepsilon_N \to h$.
- The detection threshold changes for dimension $d \geq 9$, where the situation becomes more delicate: Here, the K-NN test has power going to α or 1, depending on whether $\|N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N\| \to 0$ or $\|N^{\frac{2}{d}}\varepsilon_N\| \to \infty$, respectively. This shows that the detection threshold is somewhere between these two bounds, however, unlike in $d \leq 8$, the exact location of

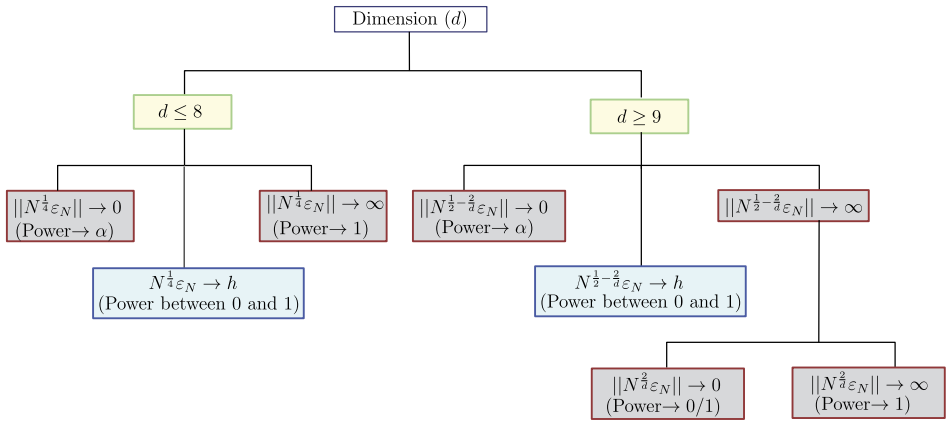


FIG. 2. *Detection threshold of the test based on the K-NN graph*: *Illustration of Theorem 4.2.*

the detection threshold has no universality: it depends on the distribution of the data under the null and the direction along which ε_N goes to zero. Note that the exponent in the lower bound $\frac{1}{2} - \frac{2}{d}$ increases to $\frac{1}{2}$ (the parametric detection threshold), and the exponent in the upper bound decreases to the 0 (which gives consistent fixed alternatives). We show that both these thresholds are tight in a truncated spherical normal problem, depending on the sign of the alternative. This is an example where the K-NN test exhibit a surprising *blessing of dimensionality*, that is, it becomes easier to detect local changes along certain directions as dimension increases (see Section 4.2.2 for details). The reason behind the phase transition of the detection threshold at dimension 9 is explained in Section 4.1.1, and the details of the proof are given in Appendix B.

1.5. *Organization.* The rest of the paper is organized as follows: The general consistency result is stated in Section 2. The central limit theorems for the statistic (1.8) are described in Section 3. The detection threshold and local power of the K-NN test are given in Section 4.1, and the performance of the different tests are compared in simulations in Section 4.2. The proofs of the results are given in the Supplementary Material [7].

2. Consistency. In this section, we prove consistency against all fixed alternatives of the test (1.9) for stabilizing graphs functionals. This unifies the proof of consistency of the test based on the K-NN graph [18, 28], and the FR test [19], generalizing the result to any stabilizing graph. We begin by recalling that $d(x, \mathscr{G}(S))$ is the total degree of the vertex x in $\mathscr{G}(S \cup \{x\})$, for $S \subset \mathbb{R}^d$ nice and $x \in \mathbb{R}^d$. Moreover, for a function $\psi : \mathbb{R}^d \to \mathbb{R}$, we denote by $\mathcal{P}_\psi$ the inhomogeneous Poisson process with intensity function ψ. (In particular, this means for any measurable set $A \subset \mathbb{R}^d$, the number of points in A is distributed as $\mathrm{Pois}(\int_A \psi(x)\,\mathrm{d}x)$.)

ASSUMPTION 2.1 (Degree moment condition). A translation and scale invariant graph functional $\mathscr{G}$ is said to satisfy the *β-degree moment condition* if it stabilizes on $\mathcal{P}_\lambda$, for all $\lambda \in (0, \infty)$, and

$$\sup_{N \in \mathbb{N}} \sup_{\substack{z \in \mathbb{R}^d, \\ \mathcal{A} \subset \mathbb{R}^d}} \mathbb{E}\{d^\beta(z, \mathscr{G}(\mathcal{P}_{N\phi_N} \cup \mathcal{A}))\} < \infty, \tag{2.1}$$

where $\mathcal{A}$ ranges over all finite subsets of $\mathbb{R}^d$, and $\phi_N = \frac{N_1}{N} f + \frac{N_2}{N} g$.

This condition ensures that the β-th moment of the degree function at a point z is uniformly bounded over z and over the addition of finitely many points to the data. Note that this is trivially satisfied for bounded degree graphs, such as the K-NN and the MST. Under this assumption, the weak limit of the statistic $\frac{1}{N}T(\mathscr{G}(\mathcal{Z}'_N))$ can be derived, which is given in terms of the Henze–Penrose dissimilarity measure between the two density functions.

DEFINITION 2.1. Given $p \in (0,1)$ and densities f and g in $\mathbb{R}^d$, the *Henze–Penrose dissimilarity* measure is defined as

$$\delta(f,g,p) = 1 - 2pq \int \frac{f(x)g(x)}{pf(x)+qg(x)}\,\mathrm{d}x. \tag{2.2}$$

This belongs to a general class of separation measures between probability distributions [16].

The following proposition gives the weak-limit of $\frac{1}{N}T(\mathscr{G}(\mathcal{Z}'_N))$ for stabilizing graph functionals satisfying the degree moment condition. The proof of the proposition closely mimics [19], Theorem 2, and is detailed in Section A.3.

PROPOSITION 2.1. *Let $\mathscr{G}$ be a translation and scale invariant directed graph functional which stabilizes on $\mathcal{P}_\lambda$ for all $\lambda \in (0,\infty)$. If $\mathscr{G}$ satisfies the β-degree moment condition for some $\beta > 4$, then*

$$\frac{1}{N}T\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big) \xrightarrow{P} \frac{\mathbb{E}\Delta_0^{\uparrow}}{2}\big(1-\delta(f,g,p)\big), \tag{2.3}$$

where $\Delta_0^{\uparrow} = d^{\uparrow}(0,\mathscr{G}(\mathcal{P}_1))$ is the out-degree of the origin in the graph $\mathscr{G}(\mathcal{P}_1 \cup \{0\})$.

Using the fact $\delta(f,g,p) \geq \delta(f,f,p) = p^2+q^2$ and that the inequality is strict for densities f and g differing on a set of positive measure (see [16], Theorem 1 and Corollary 1), it can be shown that various tests based on stabilizing graph functionals, which includes the MST and the K-NN graphs, are consistent for all fixed alternatives (1.2) (refer to Remark A.2 for details).

REMARK 2.1. Recently, Arias-Castro and Pelletier [2] showed that Rosenbaum's cross match test [26] based on non-bipartite matching (NBM), has the same limit as in (2.3), thus it is also consistent for general alternatives. Note that this does not follow from Proposition 2.1, because it is unknown whether the NBM graph functional is stabilizing. They show that the properties of stabilizing graphs required in the proof of consistency also hold for the NBM graph functional and, therefore, (2.3) holds for the cross match test as well.

3. Distribution under general alternatives. This section describes the central limit theorems of the Poissonized two-sample statistic $T(\mathscr{G}(\mathcal{Z}'_N))$ (recall (1.8)) for stabilizing graph functionals. Let $\mathscr{X}'_{N_1}$ and $\mathscr{Y}'_{N_2}$ be Poissonized samples from densities f and g in $\mathbb{R}^d$ as in (1.5). Define

$$\phi_N(x) = \frac{N_1}{N}f(x) + \frac{N_2}{N}g(x) \quad \text{and} \quad \phi(x) = pf(x)+qg(x). \tag{3.1}$$

Recall from Section 1.2 that the joint distribution of $\mathscr{X}'_{N_1}$ and $\mathscr{Y}'_{N_2}$ can be described as follows: Let $\mathcal{Z}' = \{Z_1, Z_2, \ldots\}$ be independent random variables with common density ϕ_N. Then $\mathcal{Z}'_N = \{Z_1, Z_2, \ldots, Z_{L_N}\}$, where $L_N \sim \mathrm{Pois}(N)$ is independent of $\mathcal{Z}'$, and each point of $\mathcal{Z}'_N$ is labeled 1 or 2 as in (1.7). Then the sets of points assigned labels 1 and 2 have the

same distribution as $\mathscr{X}'_{N_1}$ and $\mathscr{Y}'_{N_2}$. In this section, we derive the limiting distribution of the test statistic

$$\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N)) = \frac{1}{\sqrt{N}}\{T(\mathscr{G}(\mathcal{Z}'_N)) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}'_N)))\} \tag{3.2}$$

for stabilizing graph functionals. This involves the following two steps:

(1) The first step is to derive the CLT of the *test statistic centered by the conditional mean* $\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})$, where $\mathcal{F} := \sigma(\mathcal{Z}', L_N)$ is the sigma-algebra generated by $\mathcal{Z}'$ and the Poisson random variable L_N, that is,

$$\mathcal{R}_1(\mathscr{G}(\mathcal{Z}'_N)) = \frac{1}{\sqrt{N}}\{T(\mathscr{G}(\mathcal{Z}'_N)) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})\}, \tag{3.3}$$

for a stabilizing graph functional $\mathscr{G}$. Note that conditional on $\mathcal{F}$, the randomness comes from the labeling (1.7). As the labeling is independent across the vertices of the graph, the dependence in (3.3) is local, and the CLT can be proved using Stein's method based on dependency graphs (Theorem 3.1). This can be used to devise and calibrate a conditional test (see Remark 3.2), which might be of independent interest.

(2) The second step is to derive the CLT of the *conditional mean*

$$\mathcal{R}_2(\mathscr{G}(\mathcal{Z}'_N)) = \frac{1}{\sqrt{N}}\{\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F}) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}'_N)))\}. \tag{3.4}$$

This requires the additional assumption of exponential stabilization (Definition 3.2), and is proved in Proposition 3.2.

The above results can be combined to obtain the CLT of (3.2), since $\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N)) = \mathcal{R}_1(\mathscr{G}(\mathcal{Z}'_N)) + \mathcal{R}_2(\mathscr{G}(\mathcal{Z}'_N))$ (see Theorem 3.3 below for details).

3.1. *The conditional CLT.* For a directed graph functional $\mathscr{G}$, $S \subset \mathbb{R}^d$ finite and a point $x \in \mathbb{R}^d$, let $d^{\uparrow}(x, \mathscr{G}(S))$ be the out-degree of the vertex x in the graph $\mathscr{G}(S \cup \{x\})$, that is, the number of outgoing edges (x, y), where $y \in S \cup \{x\}$, in the graph $\mathscr{G}(S \cup \{x\})$. Similarly, let $d^{\downarrow}(x, \mathscr{G}(S))$ be the in-degree of the vertex x in the graph $\mathscr{G}(S \cup \{x\})$, that is, the number of incoming edges (y, x), where $y \in S \cup \{x\}$, in the graph $\mathscr{G}(S \cup \{x\})$. Note that $d(x, \mathscr{G}(S)) = d^{\downarrow}(x, \mathscr{G}(S)) + d^{\uparrow}(x, \mathscr{G}(S))$ is the total degree of the vertex x in the graph $\mathscr{G}(S \cup \{x\})$.

Moreover, let

$$T_2^{\uparrow}(x, \mathscr{G}(S)) = \binom{d^{\uparrow}(x, \mathscr{G}(S))}{2}, \qquad T_2^{\downarrow}(x, \mathscr{G}(S)) = \binom{d^{\downarrow}(x, \mathscr{G}(S))}{2} \tag{3.5}$$

be the number of outward 2-stars and inward 2-stars incident on x in $\mathscr{G}(S)$, respectively. Finally, let $T_2^{+}(x, \mathscr{G}(S))$ be the number of 2-stars incident on x in $\mathscr{G}(S)$ with different directions on the two edges. For notational brevity, denote

$$\Delta_0^{\uparrow} = d^{\uparrow}(0, \mathscr{G}(\mathcal{P}_1)), \qquad \Delta_0^{\downarrow} = d^{\downarrow}(0, \mathscr{G}(\mathcal{P}_1)), \tag{3.6}$$

and $\Delta_0^{+} := |\{z \in \mathcal{P}_1 : (0, z), (z, 0) \in E(\mathscr{G}(\mathcal{P}_1^0))\}|$. (Note that $\Delta_0^{\uparrow}$ was already defined in the statement of Proposition 2.1.) Similarly, let

$$T_2^{\uparrow} = T_2^{\uparrow}(0, \mathscr{G}(\mathcal{P}_1)), \qquad T_2^{\downarrow} = T_2^{\downarrow}(0, \mathscr{G}(\mathcal{P}_1)), \tag{3.7}$$

and $T_2^{+} := T_2^{+}(0, \mathscr{G}(\mathcal{P}_1))$.

To derive the CLT of (3.3), we need some control on the maximum degree of the graph functional $\mathscr{G}$. The natural assumption of bounded maximum degree includes most of the natural graphs, such as the MST and the K-NN graph. The slightly weaker polynomial upper bound given below includes other stabilizing geometric graphs, like the Delaunay graph [25].

ASSUMPTION 3.1 (Maximum degree condition). A graph functional $\mathscr{G}$ is said to satisfy the *maximum degree condition* if

$$\sup_{z\in\mathcal{P}_{N\phi_N}} d\big(z,\mathscr{G}(\mathcal{P}_{N\phi_N})\big) = o_P\big(N^{\frac{1}{40}}\big). \tag{3.8}$$

The following theorem gives the CLT of the test statistic centered by the conditional mean, as in (3.3), for stabilizing graph functionals. Recall $\phi(x) = pf(x) + qg(x)$.

THEOREM 3.1. *Let $\mathscr{G}$ be a translation and scale invariant directed graph functional which stabilizes on $\mathcal{P}_\lambda$, for all $\lambda \in (0,\infty)$. If $\mathscr{G}$ satisfies the β-degree moment condition for $\beta > 4$ and the maximum degree condition* (3.8), *then*

$$\mathcal{R}_1\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big) \xrightarrow{D} N\big(0,\kappa^2_{\mathscr{G}}\big),$$

where

$$\kappa^2_{\mathscr{G}} = \frac{r}{4}\int \frac{f(x)g(x)}{\phi^3(x)} L(x)\,\mathrm{d}x, \tag{3.9}$$

with $r := 2pq$, and

$$L(x) := 2\mathbb{E}\Delta_0^{\uparrow}\phi^2(x) + 4\phi(x)\big(q\mathbb{E}T_2^{\uparrow}g(x) + p\mathbb{E}T_2^{\downarrow}f(x)\big) - 4pq\mathbb{E}\Gamma_0 f(x)g(x),$$

where $\Gamma_0 := T_2^{\uparrow} + T_2^{\downarrow} + T_2^{+} + \frac{\Delta_0^{+}}{2} + \frac{\Delta_0^{\uparrow}}{2}$, with $\Delta_0^{\uparrow}$, $\Delta_0^{\downarrow}$, $T_2^{\uparrow}$, and $T_2^{\downarrow}$ as defined in (3.6) *and* (3.7).

The proof of theorem is given in Section A.4. The limit of the conditional variance (3.9) is derived using properties of stabilizing graphs, and the CLT is proved using Stein's method based on dependency graphs. In fact, our proof suggests that it is possible to extend the CLT in Theorem 3.1 to other distance functions in $\mathbb{R}^d$, whenever the maximum degree condition (Assumption 3.1) holds, and the conditional variance of $\mathcal{R}_1(\mathscr{G}(\mathcal{Z}'_N))$ has a limit in probability, in the graph $\mathscr{G}(\mathcal{Z}'_N)$ constructed using that metric. This is because our proof technique proceeds by conditioning on the randomness of the graph and, therefore, as long as the associated graph quantities that arise in $\mathrm{Var}(\mathcal{R}_1(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})$ converge in probability (as in Lemma A.5), and the dependence is local (which is ensured by Assumption 3.1), the Stein's method argument applies and the asymptotic normality in Theorem 3.1 would hold.

REMARK 3.1 (Null distribution). Given the graph functional $\mathscr{G}$, the limit of the conditional variance $\kappa_{\mathscr{G}}$ depends on the densities f and g and the limiting proportion p of the samples. Under the null ($f = g$) this simplifies to

$$\kappa^2_{\mathscr{G},H_0} = \frac{r}{4}\big\{2\mathbb{E}\Delta_0^{\uparrow} + 4\big(q\mathbb{E}T_2^{\uparrow} + p\mathbb{E}T_2^{\downarrow}\big) - 2r\Gamma_0\big\}. \tag{3.10}$$

- $\mathscr{G} = \mathcal{N}_K$ is the K-NN nearest neighbor graph functional: In this case, $\Delta_0^{\uparrow} = K$, $T_2^{\uparrow} = \frac{K(K-1)}{2}$, $T_2^{+} = \Delta_0^{\uparrow}\Delta_0^{\downarrow} - \Delta_0^{+} = K\Delta_0^{\downarrow} - \Delta_0^{+}$, $\mathbb{E}\Delta^{\downarrow} = K$, and (3.10) simplifies to

$$\kappa^2_{\mathcal{N}_K,H_0} = \frac{r}{2}\big\{Kpq + (p-q)^2K^2 + p^2\,\mathrm{Var}\big(\Delta_0^{\downarrow}\big) + pq\mathbb{E}\Delta_0^{+}\big\}. \tag{3.11}$$

- $\mathscr{G}$ is an undirected graph functional: In this case (3.10) simplifies to

$$\kappa^2_{\mathscr{G},H_0} = \frac{r}{2}\big\{r\mathbb{E}\Delta_0 + \mathbb{E}\big(\Delta_0^2\big)(1-2r)\big\}, \tag{3.12}$$

since $\mathbb{E}T_2^{\uparrow} = \mathbb{E}T_2^{\downarrow} = \mathbb{E}\frac{\Delta_0(\Delta_0-1)}{2}$, $\mathbb{E}T_2^{+} = \mathbb{E}\Delta_0(\Delta_0-1)$ and $\mathbb{E}\Delta_0^{+} = \mathbb{E}\Delta_0$. For example, when $\mathscr{G} = \mathcal{T}$ is the MST graph functional as in the Friedman–Rafsky test (1.11), $\Delta_0 = 2$ ([1], Lemma 7), and (3.12) becomes $\kappa^2_{\mathcal{T},H_0} = r\{r + \frac{1}{2}\mathbb{E}(\Delta_0^2)(1-2r)\}$.

The above discussion suggests that the CLT in Theorem 3.1 can be used to derive a conditional test for (1.2).

REMARK 3.2 (A conditional test and its power). For concreteness, suppose $\mathscr{G} = \mathcal{T}$ is the MST. Then under the null $\mathbb{E}_{H_0}(T(\mathcal{T}(\mathcal{Z}'_N))|\mathcal{F}) = \frac{N_1N_2}{(N_1+N_2)^2}|E(\mathcal{T}(\mathcal{Z}'_N))| = \frac{N_1N_2}{(N_1+N_2)^2}(L_N - 1)$, and given the data, we reject H_0 whenever

$$\frac{1}{\sqrt{N}}\left\{T(\mathcal{T}(\mathcal{Z}'_N)) - \frac{N_1N_2}{(N_1+N_2)^2}L_N\right\} \le \kappa_{\mathscr{G},H_0}z_\alpha.$$

By Theorem 3.1, this test has asymptotically level α. Moreover, it can be shown that (see Section A.3 for details) that

$$\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F}) = \sum_{1\le i\ne j\le L_N} \frac{N_1N_2f(Z_i)g(Z_j)\mathbf{1}\{(Z_i,Z_j)\in E(\mathscr{G}(\mathcal{Z}'_N))\}}{(N_1f(Z_i)+N_2g(Z_j))(N_1f(Z_i)+N_2g(Z_j))}.$$

The proof of Proposition 2.1 reveals that this test is consistent against all fixed alternatives, and using Theorem 3.1 we can compute the approximate power of this test as

$$\Phi\left(\frac{\kappa_{\mathscr{G},H_0}z_\alpha - \Xi(\mathcal{Z}'_N)}{\kappa_{\mathscr{G}}}\right), \tag{3.13}$$

where $\Xi(\mathcal{Z}'_N) = \frac{1}{\sqrt{N}}(\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F}) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F}))$ is the difference of the conditional means under the alternative and the null, which can be calculated from the data. The approximation in (3.13) can be justified because Stein's method gives uniform control on the corresponding distribution functions (see Proposition A.2 in Appendix A.4.2). (Note that the argument above holds for any stabilizing graph functional, as long as the number of edges $|E(\mathscr{G}(\mathcal{Z}'_N))|$ does not depend on the unknown null distribution, as is the case for the Friedman–Rafsky test and the test based on the K-NN graph.)

3.2. *CLT of the test statistic under general alternatives.* In this section, the (unconditional) CLT of the test statistic (3.2) is derived. This involves finding the CLT of the conditional mean (3.4), which requires the stronger notion of exponential stabilization [24]. For any locally finite point set $\mathcal{H} \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, define the *out-degree measure* of a graph functional $\mathscr{G}$ as follows: For all Borel sets $A \subset \mathbb{R}^d$,

$$d^{\uparrow}_{\mathscr{G}}(x,\mathcal{H},A) = \sum_{y\in\mathcal{H}^x\cap A} \mathbf{1}\{(x,y)\in E(\mathscr{G}(\mathcal{H}^x))\}, \tag{3.14}$$

where $\mathcal{H}^x = \mathcal{H}\cup\{x\}$. In other words, the out-degree measure of a set A, with respect to $\mathcal{H}$ and x is the number of edges incident on x with the other end point in $\mathcal{H}^x \cap A$ in the graph $\mathscr{G}(\mathcal{H}^x)$. The following definition formalizes the notion of "radius of stabilization" of a point, which is the smallest radius outside which addition of finitely many points does not affect the degree measure at the point.

DEFINITION 3.1. Fix a locally finite point set $\mathcal{H}$, a point $x \in \mathbb{R}^d$, and a Borel set $A \subseteq \mathbb{R}^d$. The *radius of stabilization* of the degree measure (3.14) at x with respect to $\mathcal{H}$ and A (to be denoted by $R(x,\mathcal{H},A)$) is the smallest $R \ge 0$ such that

$$d^{\uparrow}_{\mathscr{G}}(x, x+\{\mathcal{H}\cap B(0,R)\cup\mathcal{Y}\}, x+B) = d^{\uparrow}_{\mathscr{G}}(x, x+\{\mathcal{H}\cap B(0,R)\}, x+B), \tag{3.15}$$

for all finite $\mathcal{Y} \subseteq A\backslash B(0,R)$ and all Borel subsets $B \subseteq A$, where $B(0,R)$ is the (Euclidean) ball of radius R with center at the point $0 \in \mathbb{R}^d$. If no such R exists, then set $R(x,\mathcal{H},A) = \infty$.

Throughout this section, we will assume that f and g have a common support S, which is compact and convex, and $N \to \infty$ such that

$$\sqrt{N}\left(\frac{N_1}{N_1+N_2} - p\right) \to 0 \quad \text{and} \quad \sqrt{N}\left(\frac{N_1}{N_1+N_2} - q\right) \to 0. \tag{3.16}$$

DEFINITION 3.2. Let $R_N(x) := R(x, \mathcal{P}_{N\phi_N}, S)$ be the radius of stabilization of out-degree measure $d^{\uparrow}_{\mathscr{G}}$ at x with respect to the Poisson process $\mathcal{P}_{N\phi_N}$ and S. Define

$$\tau(s) := \sup_{N\in\mathbb{N}} \sup_{x\in\mathbb{R}^d} \mathbb{P}\big(R_{N\phi_N}(x) > N^{-\frac{1}{d}} s\big). \tag{3.17}$$

The out-degree measure $d^{\uparrow}_{\mathscr{G}}$ is said to be

- *power law stabilizing of order* q if $\sup_{s\geq 1} s^q \tau(s) < \infty$,
- *exponentially stabilizing* if $\limsup_{s\to\infty} \frac{1}{s} \log \tau(s) < 0$.

Conditions on the decay of the tail of the radius of stabilization, similar to (3.17) above, is a standard requirement for proving limit theorems of functionals of random geometric graphs [24, 33]. Using this machinery, we prove the following theorem, which gives the CLT of the conditional mean (3.4) for exponentially stabilizing random geometric graphs.

PROPOSITION 3.2. *Let $\mathscr{G}$ be a translation and scale invariant directed graph functional in $\mathbb{R}^d$ which satisfies the β-degree moment condition (2.1) for some $\beta > 2$. If the out-degree measure $d^{\uparrow}_{\mathscr{G}}$ is power law stabilizing of order $q > \frac{\beta}{\beta-2}$, then*

$$\lim_{N\to\infty} \operatorname{Var}\big(\mathcal{R}_2\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big)\big) = \tau^2_{\mathscr{G}}, \tag{3.18}$$

where

$$\tau^2_{\mathscr{G}} = \frac{r^2}{4} L_0 \int \frac{f^2(x) g^2(x)}{\phi^3(x)} \,\mathrm{d}x, \tag{3.19}$$

where $L_0 := \int (\mathbb{E}\{d^{\uparrow}(0, \mathscr{G}(\mathcal{P}^z_1)) d^{\uparrow}(z, \mathscr{G}(\mathcal{P}^0_1))\} - (\mathbb{E}\Delta^{\uparrow}_0)^2) \,\mathrm{d}z + \mathbb{E}(\Delta^{\uparrow}_0)^2$. *Moreover, if $d^{\uparrow}_{\mathscr{G}}$ is exponentially stabilizing then* $\mathcal{R}_2(\mathscr{G}(\mathcal{Z}'_N)) \xrightarrow{D} N(0, \tau^2_{\mathscr{G}})$.

The proof of theorem is given in Section A.6.1. Combining Theorem 3.1 and Proposition 3.2, the CLT of $\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N))$ (defined in (3.2)) can be obtained. The proof is in Section A.6.2.

THEOREM 3.3. *Let $\mathscr{G}$ be a translation and scale invariant directed graph functional which satisfies the β-degree moment condition for some $\beta > 4$ and the maximum out-degree condition (3.8). If the degree measure $d^{\uparrow}_{\mathscr{G}}$ is exponentially stabilizing, then*

$$\mathcal{R}\big(\mathscr{G}\big(\mathcal{Z}'_N\big)\big) \xrightarrow{D} N\big(0, \sigma^2_{\mathscr{G}}\big), \tag{3.20}$$

where $\sigma^2_{\mathscr{G}} = \kappa^2_{\mathscr{G}} + \tau^2_{\mathscr{G}}$, with $\kappa^2_{\mathscr{G}}$ and $\tau^2_{\mathscr{G}}$ as defined in (3.9) and (3.19), respectively.

Many random geometric graphs, such as the K-NN graph and the Delaunay graph [24, 25] are exponentially stabilizing. This theorem gives the asymptotic distribution of two-sample tests based on such graphs, under general alternatives. This can be used to the compute power of such tests as in Remark 3.2. Moreover, using this we can understand the asymptotic performances of the tests, by identifying testable local alternatives, as elaborated in the following

section for the test based on the K-NN graph. The techniques used in this section might also be useful in studying limiting distributions of multivariate goodness-of-fit tests based on nearest neighbors [9, 13].

To see why the asymptotic variance in (3.20) is the sum of two terms, note that

$$\operatorname{Var}(\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N))) = \mathbb{E}(\operatorname{Var}(\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})) + \operatorname{Var}(\mathbb{E}(\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})),$$

where $\mathcal{F} := \sigma(\mathcal{Z}', L_N)$ is the sigma-algebra generated by $\mathcal{Z}'$ and the Poisson random variable L_N (recall notation from Section 1.2). Now, recalling (3.3) shows $\operatorname{Var}(\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F}) = \operatorname{Var}(\mathcal{R}_1(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})$, and (3.4) gives $\operatorname{Var}(\mathbb{E}(\mathcal{R}(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})) = \operatorname{Var}(\mathcal{R}_2(\mathscr{G}(\mathcal{Z}'_N)))$. In the proof of Theorem 3.1, we show that $\operatorname{Var}(\mathcal{R}_1(\mathscr{G}(\mathcal{Z}'_N))|\mathcal{F})$ converges in L_2 to $\kappa^2_{\mathscr{G}}$ (Lemma A.5), while the proof of Proposition 3.2 shows that $\operatorname{Var}(\mathcal{R}_2(\mathscr{G}(\mathcal{Z}'_N))) \to \tau^2_{\mathscr{G}}$ (Section A.6.1), hence the asymptotic variance in (3.20) is the sum of these two terms.

REMARK 3.3. (Comments on de-Poissonization) Poissionization is a commonly used trick in geometric probability, where calculations become simpler because of the spatial independence of the Poisson process. In fact, when the sample sizes are large, one can pretend that the data comes from Poissonized samples with a slightly smaller mean, since a Poisson random variable is tightly concentrated around its expectation. De-Poissionization techniques are well known ([23], Section 2.5 and [24], Theorem 2.3), using which one can expect to de-Poissonize the CLT in Theorem 3.3 for the test based on the K-NN graph. The only thing that would change is the formula of the asymptotic variance, but its derivation is quite tedious for general alternatives. However, for the implementation of the test, we are more interested in the asymptotic null variance, where the calculations are much simpler, and the de-Poissonized null variance can be easily computed (see Section A.5). In fact, de-Poissonization would only change the asymptotic variance (not the order), and the constants in the limiting power (but, not the rates). Therefore, de-Poissonization would not affect (most of) the results of Section 4 as these mainly focus on detection thresholds. This is also validated by the simulations in Section 4.2.

4. Local power of the K-NN test. The test based on the K-NN graph is exponentially stabilizing and, therefore, the results obtained in the previous section apply. Recall that we assume f, g have a common support S which is compact and convex, and $N \to \infty$ such that (3.16) hold. Then we have the following corollary of Theorem 3.3.

COROLLARY 4.1. *For the K-NN graph functional $\mathcal{N}_K$ and f and g as above*:

$$\mathcal{R}(\mathcal{N}_K(\mathcal{Z}'_N)) \xrightarrow{D} N(0, \sigma^2_{\mathcal{N}_K}), \tag{4.1}$$

where $\sigma^2_{\mathcal{N}_K} = \kappa^2_{\mathcal{N}_K} + \frac{r^2K^2}{4}\int \frac{f^2(x)g^2(x)}{\phi^3(x)}\,\mathrm{d}x$, *with* $\kappa_{\mathcal{N}_K}$ *as defined in* (3.9).

PROOF. Note that the graph functional $\mathcal{N}_K$ is exponentially stabilizing [24] and satisfies the degree moment condition for $\beta > 4$. Therefore, by Theorem 3.3, (4.1) holds with $\sigma^2_{\mathcal{N}_K} = \kappa^2_{\mathcal{N}_K} + \tau^2_{\mathcal{N}_K}$. The result follows by noting that $\tau^2_{\mathcal{N}_K} = \frac{r^2K^2}{4}\int \frac{f^2(x)g^2(x)}{\phi^3(x)}\,\mathrm{d}x$ (recall (3.19)). □

REMARK 4.1. Under the null ($f = g$), $\tau^2_{\mathcal{N}_K,H_0} = \frac{r^2K^2}{4}$, and using (3.11), the asymptotic variance (4.1) simplifies to

$$\begin{aligned}\sigma^2_K &:= \sigma^2_{\mathcal{N}_K,H_0} = \kappa^2_{\mathcal{N}_K,H_0} + \tau^2_{\mathcal{N}_K,H_0} \\ &= \frac{r}{2}\{K(K+1)pq + (p-q)^2K^2 + p^2\operatorname{Var}(\Delta_0^{\downarrow})\}.\end{aligned} \tag{4.2}$$

Then recalling (1.9), the two-sample test based on $\mathcal{N}_K$ rejects when

$$\frac{1}{\sqrt{N}}\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_0}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\} \le \sigma_K z_\alpha. \tag{4.3}$$

4.1. *Power against local alternatives.* In this section, we determine the power of the K-NN test against local alternatives, that is, the power when the alternatives shrink (with increasing sample size) toward the null at a certain rate. To this end, let $\Theta \subseteq \mathbb{R}^p$ be the parameter space and $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ be a parametric family of distributions in $\mathbb{R}^d$ with density $f(\cdot|\theta)$. Let $\mathcal{X}'_{N_1}$ and $\mathcal{Y}'_{N_2}$ be samples from $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ as in (1.5), respectively, and consider the testing problem

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N, \tag{4.4}$$

for a sequence $(\varepsilon_N)_{N\ge1}$ in $\mathbb{R}^p$, such that $\|\varepsilon_N\| \to 0$. The *limiting power* of the two-sample test based on the K-NN graph $\mathcal{N}_K$ (4.3) is

$$\lim_{N\to\infty} \mathbb{P}_{\theta_2=\theta_1+\varepsilon_N}(N^{-\frac{1}{2}}\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_0}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\} \le \sigma_K z_\alpha),$$

where σ_K is the variance of the K-NN test under the null (recall (4.2)). Our goal is to find the threshold on ε_N where the K-NN test transitions from powerless to powerful. More precisely, we want to determine the sequence $a_N \to 0$, such that for $\|\varepsilon_N\| \ll a_N$, the limiting power is α, and for $\|\varepsilon_N\| \gg a_N$, the limiting power is 1. The sequence $(a_N)_{N\ge1}$ is often known as the *detection-threshold* of the test.

The parametric rate of detection is $O(N^{-\frac{1}{2}})$; however, results in [6] imply that the test based on the K-NN graph has no power in this scale. As a result, the asymptotic performance of these tests cannot be compared using their Pitman efficiencies (limiting local power when $\varepsilon_N = hN^{-\frac{1}{2}}$, which happens to be zero in this case), making the problem of determining the exact detection threshold particularly important. We answer this question in Theorem 4.2, where the exact detection threshold of the K-NN test is determined. Quite interestingly, the threshold depends on several things, such as the dimension d, the distribution of the data, and the direction of the alternative.

To state the assumptions required for computing the detection threshold, we need a few definitions: For a function $g(z_1, z_2) : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$, $\nabla_{z_1} g(z_1, z_2)$ denotes the $d \times 1$ gradient vector and $\mathrm{H}_{z_1} g(z_1, z_2)$ the $d \times d$ Hessian matrix of g, with respect to z_1 (with z_2 held fixed). Similarly, $\nabla_{z_2} g(z_1, z_2)$ and $\mathrm{H}_{z_2} g(z_1, z_2)$ is the $p \times 1$ gradient vector and the $p \times p$ Hessian matrix of g, with respect to z_2, respectively.

ASSUMPTION 4.1. Suppose the parameter space $\Theta \subseteq \mathbb{R}^p$ is convex, and the family of distributions $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ satisfy:

(a) For all $\theta \in \Theta$, the density $f(\cdot|\theta)$ has a compact and convex support $S \subset \mathbb{R}^d$, with a nonempty interior, not depending on θ.

(b) $\int_{\partial S} f(z|\theta)\,\mathrm{d}z = 0$, for all $\theta \in \Theta$, where ∂S denotes the boundary of S.

(c) For all $\theta \in \Theta$, the functions $f(\cdot|\theta)$ and $\nabla_\theta f(\cdot|\theta)$ are three times continuously differentiable in the interior of S, and the expected squared of the score function: $\mathbb{E}_{X\sim f(\cdot|\theta)}[\frac{h^\top \nabla_\theta f(X|\theta)}{f(X|\theta)}]^2 > 0$, for all $h \in \mathbb{R}^p\backslash\{0\}$.

(d) For all $x \in S$, $f(x|\cdot)$ is three times continuously differentiable in the interior of Θ.

The compactness of the support is required for establishing the CLT for exponentially stabilizing graph functionals (recall Corollary 4.1). However, we expect the CLT, and hence our results, to hold even when the support is not compact, as long as, the distributions have

"nice" tails (see simulations in Section 4.2 below). Under the above assumptions, the following theorem characterizes the detection threshold of the K-NN test and determines the exact limiting power at the threshold. To state the theorem, we need to introduce some notation: Recall that $\mathcal{P}_1^0$ denotes the Poisson process of rate 1 in $\mathbb{R}^d$ with the origin $0 \in \mathbb{R}^d$ added to it. Define

$$C_{K,s} := \mathbb{E}\Big\{ \sum_{x \in \mathcal{P}_1^0} \|x\|^s \mathbf{1}\{(0,x) \in E(\mathcal{N}_K(\mathcal{P}_1^0))\} \Big\}, \tag{4.5}$$

which is the expectation of the sum of the s-th power of the lengths of the outward edges incident at the origin 0 in the graph $\mathcal{N}_K(\mathcal{P}_1^0)$. This can be computed explicitly in terms of Gamma functions (see (B.11) in the Supplementary Material for details). Finally, define

$$a_{K,\theta_1}(h) := -\frac{rpC_{K,2}}{4d\sigma_K} \int_S h^\top \nabla_{\theta_1} \Big(\frac{\operatorname{tr}(\mathrm{H}_x f(x|\theta_1))}{f(x|\theta_1)} \Big) f^{\frac{d-2}{d}}(x|\theta_1)\,\mathrm{d}x, \tag{4.6}$$

where $C_{K,2}$ is defined above in (4.5), σ_K as in (4.2), and

$$b_{K,\theta_1}(h) := \frac{r^2 K}{2\sigma_K} \mathbb{E}\Big[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)} \Big]^2, \tag{4.7}$$

where the expectation is with respect to $X \sim f(\cdot|\theta_1)$.

THEOREM 4.2. *Let $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ be a family of distributions satisfying Assumption 4.1, and $\mathscr{X}'_{N_1}$ and $\mathscr{Y}'_{N_2}$ be samples from $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ as in (1.5), respectively. Consider the two-sample test based on the K-NN graph functional $\mathcal{N}_K$ with rejection region (4.3) for the testing problem (4.4).*

(i) *If the dimension $d \le 8$, then the following hold*:

- $\|N^{\frac14}\varepsilon_N\| \to 0$: *The limiting power of the test (4.3) is α.*
- $N^{\frac14}\varepsilon_N \to h$: *Then if dimension dimension $d \le 7$, limiting power of the test (4.3) is*

$$\Phi\big(z_\alpha + b_{K,\theta_1}(h)\big). \tag{4.8}$$

Otherwise, dimension $d = 8$ and the limiting power is

$$\Phi\big(z_\alpha + a_{K,\theta_1}(h) + b_{K,\theta_1}(h)\big), \tag{4.9}$$

where $a_{K,\theta_1}(h)$ and $b_{K,\theta_1}(h)$ are as defined above.

- $\|N^{\frac14}\varepsilon_N\| \to \infty$: *The limiting power of the test (4.3) is 1.*

(ii) *If the dimension $d \ge 9$, then the following hold*:

- $\|N^{\frac12-\frac2d}\varepsilon_N\| \to 0$: *The limiting power of the test (4.3) is α.*
- $N^{\frac12-\frac2d}\varepsilon_N \to h$: *The limiting power of the test (4.3) is*

$$\Phi\big(z_\alpha + a_{K,\theta_1}(h)\big). \tag{4.10}$$

- $\|N^{\frac12-\frac2d}\varepsilon_N\| \to \infty$ *such that* $\|N^{\frac2d}\varepsilon_N\| \to 0$: *Then depending on whether*

$$N^{\frac12-\frac2d} \int_S \varepsilon_N^\top \nabla_{\theta_1} \Big(\frac{\operatorname{tr}(\mathrm{H}_x f(x|\theta_1))}{f(x|\theta_1)} \Big) f^{\frac{d-2}{d}}(x|\theta_1)\,\mathrm{d}x \to \begin{cases} \infty, \\ -\infty, \end{cases}$$

the limiting power of the test (4.3) is 0 or 1, respectively.

- $N^{\frac2d}\varepsilon_N \to h$: *The limiting power of the test (4.3) is 0 or 1, depending on whether $a_{K,\theta_1}(h) + b_{K,\theta_1}(h)$ is negative or positive, respectively.*
- $\|N^{\frac2d}\varepsilon_N\| \to \infty$: *The limiting power of the test (4.3) is 1.*

The theorem is pictorially summarized in Figure 2, and the proof is given in the Appendix. We elaborate on the implications of this result, and its several interesting consequences below:

(a) Theorem 4.2 shows that for dimension $d \leq 7$, the detection threshold of the test is at $N^{-\frac{1}{4}}$. More precisely, if we fix an alternative direction $h \in \mathbb{R}^p$, and suppose $\theta_2 = \theta_1 + \delta_N h$, for some positive sequence $\delta_N \to 0$, then, by Theorem 4.2, the limiting power of the test (4.3) is

$$\begin{cases} \alpha & \text{if } N^{\frac{1}{4}}\delta_N \to 0, \\ \Phi\big(z_\alpha + \lambda^2 b_{K,\theta_1}(h)\big) > \alpha & \text{if } N^{\frac{1}{4}}\delta_N \to \lambda > 0, \\ 1 & \text{if } N^{\frac{1}{4}}\delta_N \to \infty. \end{cases}$$

Note that the power at the threshold $N^{\frac{1}{4}}\delta_N \to \lambda$ is always greater than α, because $b_{K,\theta_1}(h) > 0$ by Assumption 4.1(c). Here, the limiting power is obtained from the limit of the Hessian of the mean difference (defined below in (4.13)), which can be thought of as the *second-order efficiency* of the test (4.3), in comparison to the first-order (Pitman) efficiency, which is zero in this case (see Section 4.1.1 below for more on this analogy).

(b) For dimension $d = 8$, the behavior is similar to the case above, but there is a subtle difference when $N^{\frac{1}{4}}\delta_N \to \lambda$. Here, for an alternative direction $h \in \mathbb{R}^p$ and $\theta_2 = \theta_1 + \delta_N h$ as above, the limiting power of the test (4.3) is

$$\begin{cases} \alpha & \text{if } N^{\frac{1}{4}}\delta_N \to 0, \\ \Phi\big(z_\alpha + \lambda a_{K,\theta_1}(h) + \lambda^2 b_{K,\theta_1}(h)\big) & \text{if } N^{\frac{1}{4}}\delta_N \to \lambda > 0, \\ 1 & \text{if } N^{\frac{1}{4}}\delta_N \to \infty. \end{cases}$$

Note that at the threshold $N^{\frac{1}{4}}\delta_N \to \lambda$, the limiting power can be greater than α or less than α, depending on whether $\lambda a_{K,\theta_1}(h) + \lambda^2 b_{K,\theta_1}(h)$ is positive or negative. In particular, considering the power as a function of λ gives: if $\lambda < -\frac{a_{K,\theta_1}(h)}{b_{K,\theta_1}(h)}$, then the limiting power is less than α, and if $\lambda > -\frac{a_{K,\theta_1}(h)}{b_{K,\theta_1}(h)}$, then the limiting power is greater than α. Therefore, in dimension 8, the limiting power function is nonmonotone if $a_{K,\theta_1}(h) < 0$. The asymptotic power starts off at α, decreases for a while, going below α and making it *asymptotically biased* (i.e., the limiting power is less than the size of the test), then starts to increase, going past α and eventually becoming 1, as $\lambda \to \infty$. This also shows that for every direction $h \in \mathbb{R}^d$ such that $a_{K,\theta_1}(h) < 0$, there is a "special" point $\lambda = -\frac{a_{K,\theta_1}(h)}{b_{K,\theta_1}(h)} > 0$, where the limiting power is exactly α.

(c) A surprising phenomenon happens for dimension $d \geq 9$: Here, unlike for dimension 8 or smaller, the precise location of the detection threshold depends on the distribution of the data under the null $f(\cdot|\theta_1)$ and the direction of the alternative. As before, fix an alternative direction $h \in \mathbb{R}^p$, and suppose $\theta_2 = \theta_1 + \delta_N h$, for some positive sequence $\delta_N \to 0$. Then, depending on the sign of $a_{K,\theta_1}(h)$ (recall (4.6)), there are two cases:

– Suppose $a_{K,\theta_1}(h) > 0$. Then, by Theorem 4.2, the limiting power of the test (4.3) is

$$\begin{cases} \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\big(z_\alpha + \lambda a_{K,\theta_1}(h)\big) > \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 1 & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty. \end{cases} \tag{4.11}$$

Here, the detection threshold of the test (4.3) is at $\Theta(N^{-\frac{1}{2}+\frac{2}{d}})$, that is, the limiting power transitions from α to 1 at $\delta_N = \Theta(N^{-\frac{1}{2}+\frac{2}{d}})$. Note that the detection threshold $N^{-\frac{1}{2}+\frac{2}{d}}$

improves with dimension, moving closer to the parametric rate of $N^{-\frac{1}{2}}$ as the dimension d grows to infinity, exhibiting a *blessing of dimensionality*. An example where this is attained is the truncated spherical normal problem (see Section 4.2.2 below).

– Suppose $a_{K,\theta_1}(h) < 0$. By Theorem 4.2, the limiting power of the test (4.3) is

$$\begin{cases} \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\big(z_\alpha + \lambda a_{K,\theta_1}(h)\big) < \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 0 & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty \text{ and } N^{\frac{2}{d}}\delta_N \to 0, \\ 0 & \text{if } N^{\frac{2}{d}}\delta_N \to \kappa \text{ and} \\ & \kappa a_{K,\theta_1}(h) + \kappa^2 b_{K,\theta_1}(h) < 0, \\ 1 & \text{if } N^{\frac{2}{d}}\delta_N \to \kappa \text{ and} \\ & \kappa a_{K,\theta_1}(h) + \kappa^2 b_{K,\theta_1}(h) > 0, \\ 1 & \text{if } N^{\frac{2}{d}}\delta_N \to \infty. \end{cases}$$

Note that in this case the limiting power function is nonmonotone and asymptotically biased, it starts off at α, then goes below α, eventually drops to zero, and then transitions up to 1. This surprising phenomenon happens because the test (4.3) has a one-sided rejection region, and it is universally consistent. Therefore, the limiting power when $N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0$ is given by the normal lower tail, more precisely, $\Phi(z_\alpha + \lambda a_{K,\theta_1}(h))$. Therefore, for a direction chosen such that $a_{K,\theta_1}(h) < 0$, the power drops below α and then goes to zero when $\lambda \to \infty$, but it has to eventually go up to 1 because of consistency, hence the nonmonotonicity. In this case, the power transitions from 0 to 1, at $\delta_N = \Theta(N^{-\frac{2}{d}})$, which becomes worse with dimension (converging finally to fixed difference alternatives as $d \to \infty$), exhibiting a curse of dimensionality. Again, this is attained in the truncated spherical normal problem (see Section 4.2.2 below). Theorem 4.2 also gives the limiting power at the threshold $N^{\frac{2}{d}}\delta_N \to \kappa > 0$. Here, the limiting power of the test (4.3) converges to 0 or 1, depending on whether $\kappa a_{K,\theta_1}(h) + \kappa^2 b_{K,\theta_1}(h)$ is negative or positive, respectively. In other words, considering the limiting power as a function of κ gives: if $\kappa < -\frac{a_{K,\theta_1}(h)}{b_{K,\theta_1}(h)}$, then the limiting power is 0, and if $\kappa > -\frac{a_{K,\theta_1}(h)}{b_{K,\theta_1}(h)}$, then the limiting power is 1. This happens because at the threshold $N^{\frac{2}{d}}\delta_N \to \kappa > 0$, the gradient and Hessian of the mean difference (defined below in (4.13)) are of the same order, and the limiting power is 0 or 1 depending on whether the sum of the gradient and the Hessian diverges to ∞ or $-\infty$, which is in turn determined by the sign of $\kappa a_{K,\theta_1}(h) + \kappa^2 b_{K,\theta_1}(h)$. (Note that, similar to case (b) above, there is a "special" point $\kappa = -\frac{a_{K,\theta_1}(h)}{b_{K,\theta_1}(h)}$, where the theorem is unable to say anything about the limiting power, when $N^{\frac{2}{d}}\delta_N \to \kappa > 0$. If this happens, then the limiting power depends on the higher-order expansions of the gradient and the Hessian of the mean difference, which has to be calculated individually for specific examples.)

The discussion above shows that for dimension 9 and higher, given a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ and an alternative direction $h \in \mathbb{R}^p$, there are some "good directions" (where $a_{K,\theta_1}(h) > 0$) where the test (4.3) exhibits a blessing of dimensionality, but at the same time there are "bad directions" (where $a_{K,\theta_1}(h) < 0$) where one sees a curse of dimensionality. For simulations illustrating this phenomenon, refer to Section 4.2.2 below.

(d) Note that Theorem 4.2 does not tell us what the detection threshold is when $a_{K,\theta_1}(h) = 0$. These are the "degenerate directions," for which the precise location of the detection threshold has to be determined on a case by case basis: For example, this happens in the normal location problem (see Section 4.2.1 below), where a direct calculation shows that, irrespective of the dimension, the detection threshold is at $\Theta(N^{-\frac{1}{4}})$, for all directions.

(e) The rates obtained in Theorem 4.2 can be summarized in terms of the *critical exponents*,

$$\beta_d = \begin{cases} \frac{1}{4} & \text{if } d \le 8, \\ \frac{1}{2} - \frac{2}{d} & \text{if } d \ge 9, \end{cases} \qquad \gamma_d = \begin{cases} \frac{1}{4} & \text{if } d \le 8, \\ \frac{2}{d} & \text{if } d \ge 9. \end{cases} \tag{4.12}$$

Theorem 4.2 says that (irrespective of the distribution of the data) for the testing problem (4.4): (1) if $\|N^{\beta_d}\varepsilon_N\| \to 0$, the limiting power of the test (4.3) is α; and (2) if $\|N^{\gamma_d}\varepsilon_N\| \to \infty$, the limiting power of the test (4.3) is 1. Note that they are equal up to dimension $d = 8$, after which β_d increases with d to $\frac{1}{2}$ (recall the K-NN test has no power for $N^{-\frac{1}{2}}$ alternatives [6]), and γ_d decreases with d to zero (the K-NN test always has power against fixed alternatives).

4.1.1. *Proof outline*. The proof of Theorem 4.2 is given in Appendix B. Here, we give an outline of the proof. To find the limiting local power of the K-NN test (4.3), it suffices to derive the asymptotic distribution of

$$\frac{1}{\sqrt{N}}\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_0}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\} = T_1 + T_2,$$

where $T_1 := \frac{1}{\sqrt{N}}\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\}$ and the *mean difference*

$$T_2 := \frac{1}{\sqrt{N}}\mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}'_N))) - \mathbb{E}_{H_0}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\}, \tag{4.13}$$

when $\theta_2 = \theta_1 + \varepsilon_N$, where ε_N is as in (4.4). The proof of Corollary 4.1 shows that the first term converges in distribution to $N(0, \sigma_K^2)$. Therefore, determining the limiting power boils down to computing the limit of the mean difference T_2. In the parametric setup of (4.4), $\mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}'_N))) := \delta_N(\theta_1, \theta_2)$ for some function $\delta_N : \Theta^2 \to \mathbb{R}$. (The expression of δ_N is given in (B.1) in the Supplementary Material. Note that δ_N is related to the function μ_N in the Appendix as: $\delta_N(\theta_1, \theta_2) = \frac{N_1 N_2}{N^2}\mu_N(\theta_1, \theta_2)$.) Then by a Taylor series expansion in the second coordinate (and ignoring the error term) gives

$$\begin{aligned} T_2 &= \frac{1}{\sqrt{N}}\{\delta_N(\theta_1, \theta_1 + \varepsilon_N) - \delta_N(\theta_1, \theta_1)\} \\ &\approx \frac{\varepsilon_N^\top}{\sqrt{N}}\nabla\delta_N(\theta_1, \theta_1) + \frac{\varepsilon_N^\top \mathrm{H}[\delta_N(\theta_1, \theta_1)]\varepsilon_N}{2\sqrt{N}}, \end{aligned}$$

where $\nabla\delta_N(\theta_1, \theta_1) := \nabla_\theta\delta_N(\theta_1, \theta)|_{\theta=\theta_1} \in \mathbb{R}^p$ is the gradient vector (with respect to the second coordinate θ) of $\delta_N(\theta_1, \theta)$ evaluated at $\theta = \theta_1$, and $\mathrm{H}[\delta_N(\theta_1, \theta_1)] \in \mathbb{R}^{p\times p}$ is the Hessian matrix (with respect to θ) of $\delta_N(\theta_1, \theta)$. The proof of Theorem 4.2 involves showing the following steps:

- $\frac{1}{\sqrt{N}}\varepsilon_N^\top\nabla\delta_N(\theta_1, \theta_1)$ has finite limit when $\varepsilon_N = \frac{h}{N^{\frac{1}{2}-\frac{2}{d}}}$ (Lemma B.1 in Appendix B).
- $\frac{1}{\sqrt{N}}\varepsilon_N^\top\mathrm{H}[\delta_N(\theta_1, \theta_1)]\varepsilon_N$ has finite limit when $\varepsilon_N = \frac{h}{N^{\frac{1}{4}}}$ (Lemma B.2 in Appendix B).

Note that when $d \le 7$, $N^{-\frac{1}{2}+\frac{2}{d}} \ll N^{-\frac{1}{4}}$ and the Hessian term dominates the gradient term, giving the formula in (4.8). This can be thought of as the *second-order efficiency* of the test (4.3). (This is in analogy with the classical first-order (Pitman) efficiency, which is derived under local alternatives $\varepsilon_N = h/\sqrt{N}$. However, the K-NN test (4.3) has zero-Pitman efficiency, because the first-order term $\frac{h^\top}{N}\nabla\delta_N(\theta_1, \theta_1)$ is asymptotically zero in this scale, hence the local power is given by the second-order Hessian term.) On the other hand, when $d = 8$,

the rate of convergence of the gradient and the Hessian terms match, and, as a result the contributions from both the terms show up in (4.9). Finally, when $d \geq 9$, the gradient term dominates the Hessian term (since $N^{-\frac{1}{2}+\frac{2}{d}} \gg N^{-\frac{1}{4}}$), which explains the shift in the location of the detection threshold at dimension 8 and gives the expression in (4.10).

4.2. *Examples*. In this section, we discuss examples which attain the threshold obtained in Theorem 4.2. In order to meet compactness assumption in Theorem 4.2 (recall Assumption 4.1), we consider standard distributions truncated to a compact, convex set. However, as mentioned earlier, we expect the results to hold for the un-truncated family (with "nice" tails), as well.

4.2.1. *Example*: *Normal location*. Let $A \subset \mathbb{R}^d$ be a compact and convex set which is symmetric around the origin $\mathbf{0} \in \mathbb{R}^d$, that is, $A = -A$. For $\theta \in \mathbb{R}^d$, define a family of densities $\phi_A(x|\theta) = \frac{1}{Z_A(\theta)} e^{-\frac{1}{2}\|x-\theta\|^2}$, where $Z_A(\theta) := \int_A e^{-\frac{1}{2}\|x-\theta\|^2} \, \mathrm{d}x$, is the normalizing constant. This is the d-dimensional multivariate normal $N(\theta, \mathrm{I}_d)$ truncated to the set A.

Now, consider the problem of testing (4.4) based on (4.3), given i.i.d. samples $\mathscr{X}_{N_1}$ and $\mathscr{Y}_{N_2}$ from $\phi_A(\cdot|\theta_1)$ and $\phi_A(\cdot|\theta_2)$, respectively. There are two cases depending whether the true θ_1 is zero or nonzero. Here, we discuss the case $\theta_1 = \mathbf{0}$: When $\theta_1 = \mathbf{0}$, it is easy to check that

$$\int_A \nabla_{\theta=\mathbf{0}} \Big(\frac{\mathrm{tr}(\mathrm{H}_x \phi_A(x|\theta))}{\phi_A(x|\theta)} \Big) \phi_A^{\frac{d-2}{d}}(x|\theta) \, \mathrm{d}x = 0,$$

which implies Theorem 4.2 cannot be directly applied to the case $d \geq 9$. However, in this case a direct calculation shows that the gradient term is exactly zero across all dimensions, which implies the following (calculations are given in Lemma C.1 in Appendix C): For any $d \geq 1$:

- If $\|N^{\frac{1}{4}} \varepsilon_N\| \to 0$, the limiting power of the test is α.
- If $\|N^{\frac{1}{4}} \varepsilon_N\| \to \infty$, the limiting power of the test is 1.
- If $N^{\frac{1}{4}} \varepsilon_N \to h$, for some $h \in \mathbb{R}^p \backslash \{\mathbf{0}\}$, the limiting power of the test is given by $\Phi(z_\alpha + \frac{r^2 K}{2\sigma_K} \mathbb{E}_{X \sim \phi_A(\cdot|\mathbf{0})} (h^\top X)^2)$.

Details of the other case $\theta_1 \neq \mathbf{0}$ can be found in Appendix C. In this case, because of the asymmetry introduced by the truncation,

$$\int_A \nabla_{\theta_1} \Big(\frac{\mathrm{tr}(\mathrm{H}_x \phi_A(x|\theta_1))}{\phi_A(x|\theta_1)} \Big) \phi_A^{\frac{d-2}{d}}(x|\theta_1) \, \mathrm{d}x \neq 0,$$

and hence, the detection threshold undergoes a phase-transition at dimension 8 as in Theorem 4.2. However, in the untruncated normal family ($\{\mathbb{P}_\theta \sim N(\theta, \mathrm{I}_d) : \theta \in \mathbb{R}^d\}$)

$$\int_{\mathbb{R}^d} \nabla_{\theta_1} \Big(\frac{\mathrm{tr}(\mathrm{H}_x \phi_{\mathbb{R}^d}(x|\theta_1))}{\phi_{\mathbb{R}^d}(x|\theta_1)} \Big) \phi_{\mathbb{R}^d}^{\frac{d-2}{d}}(x|\theta_1) = 0,$$

for all $\theta_1 \in \mathbb{R}^d$, that is, for the untruncated normal location problem we expect the detection threshold to be at $N^{-\frac{1}{4}}$, for all dimensions, as seen in the simulations below.

To illustrate the results above, we consider the following simulation: Consider the parametric family $\mathbb{P}_\theta \sim N(\theta, \mathrm{I}_d)$, for $\theta \in \mathbb{R}^d$. Figure 3 shows the empirical power (out of 100 repetitions) of the tests based on the 2-NN and 6-NN graphs, the test based on the symmetrized 3-NN graph (see Appendix E for details on the limiting power of the symmetrized 3-NN test), and the Hotelling's T^2 test, with $N_1 = 2000$ samples from $N(2 \cdot \mathbf{1}, \mathrm{I}_d)$ and $N_2 = 1000$ samples from $N(2 \cdot \mathbf{1} + \delta N^{-\frac{1}{4}} \mathbf{1}, \mathrm{I}_d)$, over a grid of 40 values of δ in $[-3, 3]$ (smoothed

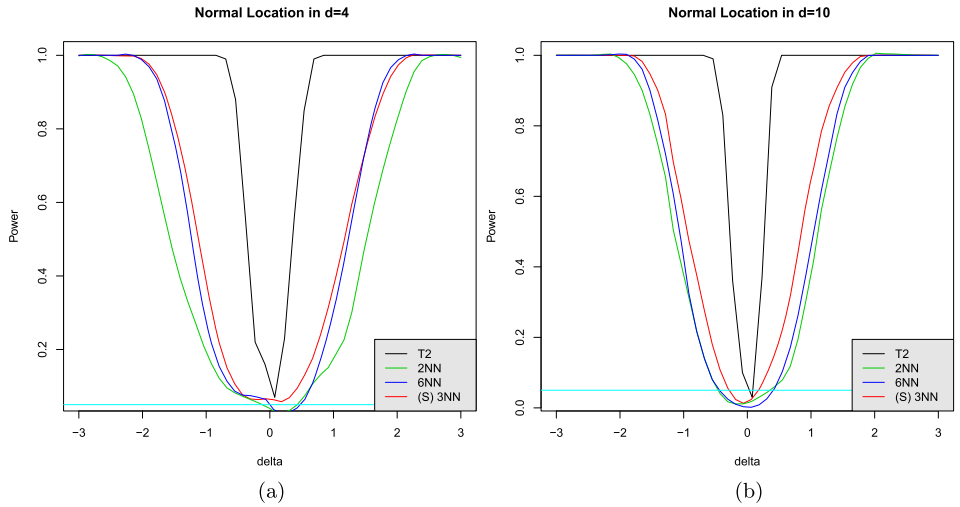


FIG. 3. *Empirical power for the normal location problem against* $\delta N^{-\frac{1}{4}}$ *alternatives, in dimension* (*a*) $d = 4$ *and* (*b*) $d = 10$.

out using the `loess` function in `R`), in (a) dimension 4 and (b) dimension 10. (Here, $N = N_1 + N_2 = 3000$.) The level of the tests are set to $\alpha = 0.05$. The plots show that the tests based on the NN graphs have nontrivial local power as a function of δ, as predicted by the calculations above. Note that, in this case, the most powerful test is the Hotelling's T^2-test, which has detection threshold at $N^{-\frac{1}{2}}$ and, therefore, has high power at the $N^{-\frac{1}{4}}$ scale, as seen in the plots.

Figure 4 shows the empirical power (out of 100 repetitions) of the different tests with $N_1 = 5000$ samples from $N(2 \cdot \mathbf{1}, \mathrm{I}_d)$ and $N_2 = 3000$ samples from $N(2 \cdot \mathbf{1} + N^{-b} \cdot \boldsymbol{h}, \mathrm{I}_d)$, where b varies over a grid of 100 values in $[0, 1]$, $\boldsymbol{h} = \mathbf{1}$ and dimension (a) $d = 4$ and (b) $d = 10$. Note that $b = 0$ corresponds to fixed alternatives where the power is expected to be near 1 because of consistency. The level of the tests are set to $\alpha = 0.25$. Note that the power of the tests based on the K-NN graphs transitions from α to 1 around $b = 0.25$, which

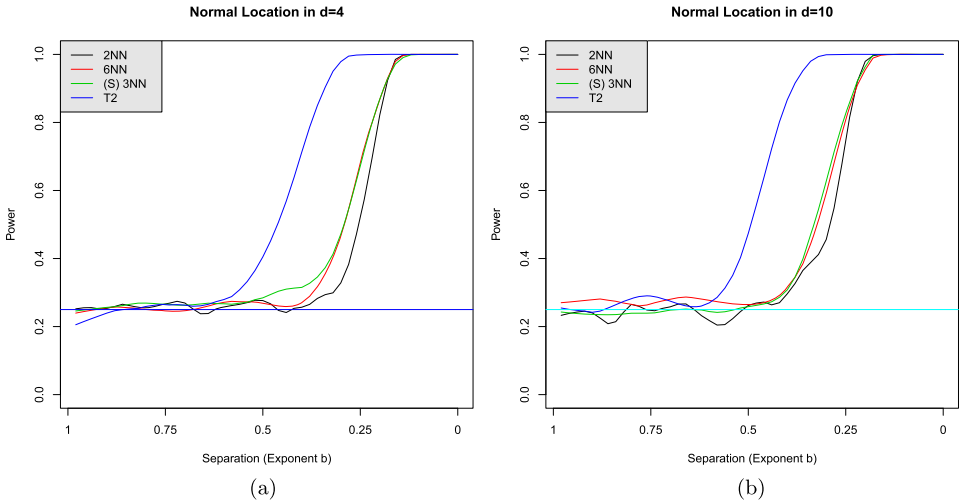


FIG. 4. *Empirical power in the normal location problem with* $N_1 = 5000$ *samples from* $N(2 \cdot \mathbf{1}, \mathrm{I}_d)$ *and* $N_2 = 3000$ *samples from* $N(2 \cdot \mathbf{1} + N^{-b} \cdot \boldsymbol{h}, \mathrm{I}_d)$, *where* b *varies over a grid of* 100 *values in* $[0, 1]$, $\boldsymbol{h} = \mathbf{1}$ *and dimension* (*a*) $d = 4$ *and* (*b*) $d = 10$.

corresponds to the rate $N^{-\frac{1}{4}}$, in both dimensions, as predicted by the calculations above. On the other hand, the power of the Hotelling's T^2 test transitions from α to 1 around $b = 0.5$, which corresponds to the parametric rate of $N^{-\frac{1}{2}}$. The corresponding plots for the negative direction $\boldsymbol{h} = -\mathbf{1}$ are given in Appendix F.1.

4.2.2. *Example*: *Spherical normal*. Let M be a convex, compact subset of $\mathbb{R}^d$. For $\lambda > 0$, define a family of densities $\phi_M(\cdot|\lambda^2)$:

$$\phi_M(x|\lambda^2) = \frac{1}{Z_M(\lambda^2)} e^{-\frac{1}{2\lambda^2}\|x\|^2} \quad \text{for } x \in M,$$

where $Z_M(\lambda^2) := \int_M e^{-\frac{1}{2\lambda^2}\|x\|^2} \mathrm{d}x$ is the normalizing constant. (Note that $Z_{\mathbb{R}^d}(\lambda^2) = (2\pi\lambda^2)^{\frac{d}{2}}$.) This is the d-dimensional spherical normal distribution $N(0, \lambda^2 \mathrm{I}_d)$ truncated to the set M. Now, consider the problem of testing (4.4) based on (4.3), given i.i.d. samples $\mathscr{X}_{N_1}$ and $\mathscr{Y}_{N_2}$ from $\phi_M(\cdot|\lambda_1^2)$ and $\phi_M(\cdot|\lambda_2^2)$, respectively. In this case, for $h \in \mathbb{R}$,

$$\begin{aligned} &\int_M h \cdot \nabla_{\lambda_1} \left(\frac{\operatorname{tr}(\mathrm{H}_x \phi_M(x|\lambda_1^2))}{\phi_M(x|\lambda_1^2)} \right) \phi_M^{\frac{d-2}{d}}(x|\lambda_1^2) \\ &\quad = -\frac{Z_M(\frac{d\lambda_1^2}{d-2})}{Z_M(\lambda_1^2)^{\frac{d-2}{d}}} \frac{2h}{\lambda_1^5} (2\mathbb{E}\|W\|^2 - \lambda_1^2 d), \end{aligned} \tag{4.14}$$

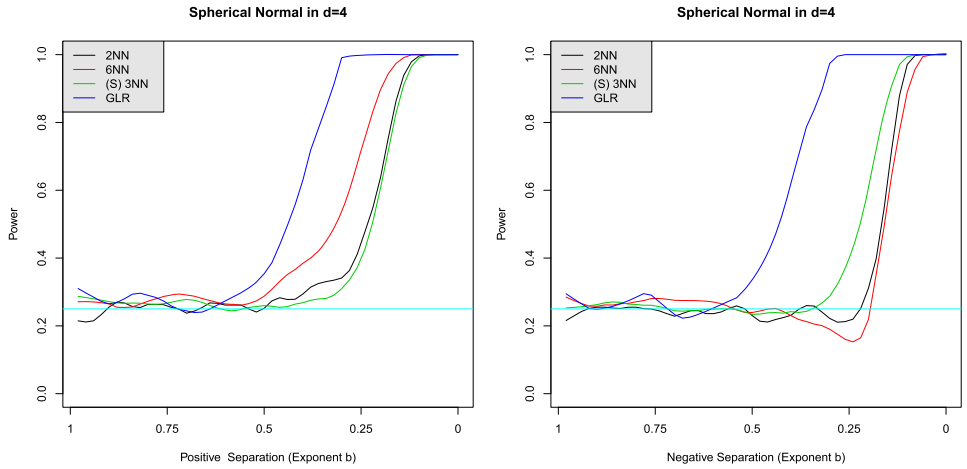
where $W = (W_1, W_2, \ldots, W_d)'$ are i.i.d. from the density $\phi_M(\cdot|\frac{d\lambda_1^2}{d-2})$ (see Section D in the Supplementary Material for details). Therefore (recall (4.6)),

$$a_{K,\lambda_1}(h) = \frac{rpC_{K,2}}{2d\sigma_K} \frac{Z_M(\frac{d\lambda_1^2}{d-2})}{Z_M(\lambda_1^2)^{\frac{d-2}{d}}} \frac{h}{\lambda_1^5} (2\mathbb{E}\|W\|^2 - \lambda_1^2 d), \tag{4.15}$$

which is positive or negative depending on whether h is positive or negative. Therefore, the limiting power of the test (4.3) for dimension $d \geq 9$, at the threshold $N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N \to h$, is $\Phi(z_\alpha + a_{K,\lambda_1}(h))$. (Note that in the simulations below we will consider the untruncated spherical normal family $\{\mathbb{P}_\lambda \sim N(0, \lambda^2 \mathrm{I}_d) : \lambda > 0\}$. The limiting power in this case can be obtained by choosing $M = [-L, L]^d$, and taking $L \to \infty$ in (4.15).)

Now, suppose we are given i.i.d. samples $\mathscr{X}_{N_1}$ from $\phi_M(\cdot|\lambda_1^2)$ and $\mathscr{Y}_{N_2}$ from $\phi_M(\cdot|\lambda_2^2)$, where $\lambda_2 = \lambda_1 + h\delta_N > 0$, for some h fixed and $\delta_N \to 0$, as $N \to \infty$. Then, by Theorem 4.2, depending on the dimension and the sign of h we have the following cases:

- For dimension $d \leq 8$, irrespective of the sign of h, the limiting power of the test (4.3) is 0 or 1, depending on whether $N^{\frac{1}{4}}\delta_N \to 0$ or $N^{\frac{1}{4}}\delta_N \to \infty$. At the threshold, $N^{\frac{1}{4}}\delta_N \to \kappa$, the limiting power is given by (4.8) or (4.9) (with h replaced by κh). This is illustrated in Figure 5, which shows the empirical power (out of 100 repetitions) of the tests based on the 2-NN and 6-NN graphs, the test based on the symmetrized 3-NN graph, and the generalized likelihood ratio test (GLR), in dimension $d = 4$, with $N_1 = 12{,}000$ samples from $N(0, 3^2 \cdot \mathrm{I}_d)$ and $N_2 = 6000$ samples from $N(0, (3 + hN^{-b})^2 \mathrm{I}_d)$, where b varies over a grid of 100 values in $[0, 1]$ and (a) $h = 2$ (b) $h = -2$. (Here, $N = N_1 + N_2 = 18{,}000$.) The level of the tests are set to $\alpha = 0.25$. Note that the power of the tests based on the K-NN graphs transitions from α to 1 around $b = 0.25$ (irrespective of the sign of h), which corresponds to the rate $N^{-\frac{1}{4}}$, as shown in the calculations above. On the other hand, the power of the GLR test transitions from α to 1 around $b = 0.5$, which corresponds to the parametric rate of $N^{-\frac{1}{2}}$.



(a) (b)

FIG. 5. *Empirical power in the spherical normal problem in dimension* $d = 4$ *with* $N_1 = 12{,}000$ *samples from* $N(0, 3^2 \cdot \mathrm{I}_d)$ *and* $N_2 = 6000$ *samples from* $N(0, (3 + hN^{-b})^2 \mathrm{I}_d)$, *where* b *varies over a grid of* 100 *values in* $[0, 1]$ *and* (*a*) $h = 2$ (*b*) $h = -2$.

- Next, suppose $d \geq 9$. Then depending on the sign of h the following cases arise:
  - Suppose $h > 0$ (then $a_{K,\lambda_1}(h) > 0$). By (4.11), the limiting power of the test (4.3) is

$$\begin{cases} \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\big(z_\alpha + \kappa a_{K,\lambda_1}(h)\big) > \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \kappa > 0, \\ 1 & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty, \end{cases}$$

    where $a_{K,\lambda_1}(h)$ is defined above in (4.15). Here, the detection threshold exhibits a blessing of dimensionality, improving with dimension to the parametric rate of $N^{-\frac{1}{2}}$ as the dimension d grows to infinity. This is illustrated in Figure 6(a), which shows the empirical power (out of 100 repetitions) of the different tests in dimension $d = 10$, with $N_1 = 300{,}000$ samples from $N(0, 3^2 \cdot \mathrm{I}_d)$ and $N_2 = 200{,}000$ samples from $N(0, (3 + hN^{-b})^2 \mathrm{I}_d)$, where b varies over a grid of 100 values in $[0, 1]$ and $h = 2$. As before, the level of the tests are set to $\alpha = 0.25$. Note that the power of the tests based on the K-NN graphs transitions from α to 1 around $b = \frac{1}{2} - \frac{2}{d} = 0.3$, which is the predicted rate of $N^{-\frac{1}{2}+\frac{2}{d}}$. As before, the power of the GLR test transitions from 0 to 1 around $b = 0.5$. To see the transitions more sharply and observe the local power of the different tests, we can zoom in at the thresholds (see Appendix F.2).
  - Suppose $h < 0$ (then $a_{K,\theta_1}(h) < 0$). By Theorem 4.2, the limiting power of the test (4.3) is

$$\begin{cases} \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\big(z_\alpha + \kappa a_{K,\theta_1}(h)\big) < \alpha & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \kappa > 0, \\ 0 & \text{if } N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty \text{ and } N^{\frac{2}{d}}\delta_N \to 0, \\ 1 & \text{if } N^{\frac{2}{d}}\delta_N \to \infty. \end{cases}$$

    This is illustrated in Figure 6(b), which shows the empirical power (out of 100 repetitions) of the different tests in dimension $d = 10$, with $N_1 = 200{,}000$ samples from $N(0, 3^2 \cdot \mathrm{I}_d)$ and $N_2 = 100{,}000$ samples from $N(0, (3 + hN^{-b})^2 \mathrm{I}_d)$, where b varies over a grid of 100 values in $[0, 1]$ and $h = -2$. Here, we observe the predicted non-monotonicity of the power of the K-NN tests. The asymptotic power starts of at the

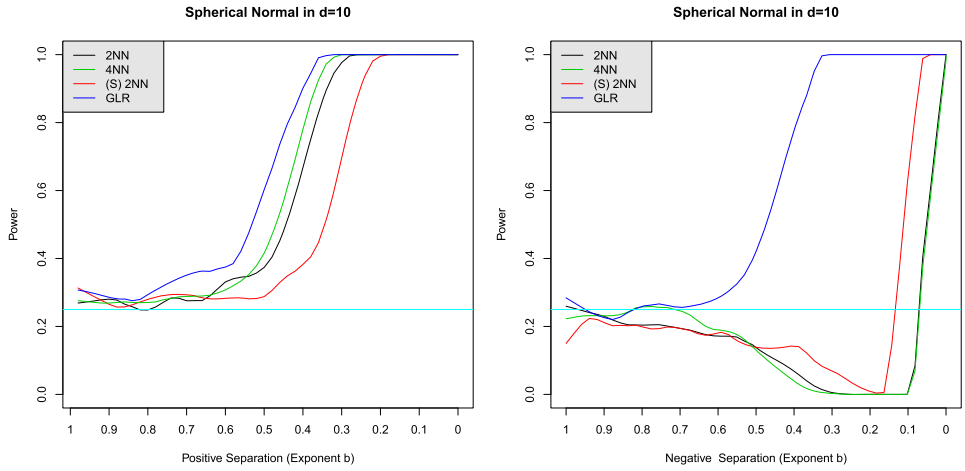


FIG. 6. *Empirical power in the spherical normal problem in dimension* $d = 10$ *with* N_1 *samples from* $N(0, 3^2 \cdot \mathrm{I}_d)$ *and* N_2 *samples from* $N(0, (3 + hN^{-b})^2 \mathrm{I}_d)$, *where* b *varies over a grid of* 100 *values in* $[0, 1]$ *and* (*a*) $h = 2$ (*b*) $h = -2$.

level $\alpha = 0.25$, goes down to zero (predicted by the theorem at $b = \frac{1}{2} - \frac{2}{d} = 0.3$), stays at zero for a while and jumps up to 1 (predicted by the theorem at $b = \frac{2}{d} = 0.2$). Additional simulations zooming in to the different thresholds are given in Appendix F.2.

Acknowledgments. The author is indebted to his advisor Persi Diaconis for introducing him to graph-based tests and for his constant encouragement and support. The author thanks Riddhipratim Basu, Sourav Chatterjee, Jerry Friedman, Shirshendu Ganguly and Susan Holmes for illuminating discussions and helpful comments. The author also thanks the Associate Editor and the anonymous referees for their detailed and thoughtful comments, which greatly improved the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs" (DOI: 10.1214/19-AOS1913SUPP; .pdf). The proofs of the results are given in the supplementary material, which is organized as follows: The consistency of tests based on stabilizing graphs (Proposition 2.1), and the central limit theorems of the test statistic under general alternatives (Theorem 3.1, Proposition 3.2 and Theorem 3.3) are proved in Appendix A. The detection thresholds for the test based on the K-NN graph (Theorem 4.2) is proved in Appendix B. Calculations for the normal location and the spherical normal examples are given in Appendix C and Appendix D, respectively. The symmetrized K-NN test is discussed in Appendix E. Additional simulations are given in Appendix F.

REFERENCES

[1] ALDOUS, D. and STEELE, J. M. (1992). Asymptotics for Euclidean minimal spanning trees on random points. *Probab. Theory Related Fields* **92** 247–258. MR1161188 https://doi.org/10.1007/BF01194923

[2] ARIAS-CASTRO, E. and PELLETIER, B. (2016). On the consistency of the crossmatch test. *J. Statist. Plann. Inference* **171** 184–190. MR3458077 https://doi.org/10.1016/j.jspi.2015.10.003

[3] ASLAN, B. and ZECH, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Stat. Comput. Simul.* **75** 109–119. MR2117010 https://doi.org/10.1080/00949650410001661440

[4] BARINGHAUS, L. and FRANZ, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.* **88** 190–206. MR2021870 https://doi.org/10.1016/S0047-259X(03)00079-4

[5] BATU, T., FORTNOW, L., RUBINFELD, R., SMITH, W. D. and WHITE, P. (2013). Testing closeness of discrete distributions. *J. ACM* **60** Art. 4, 25. MR3033221 https://doi.org/10.1145/2432622.2432626

[6] BHATTACHARYA, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 575–602. MR3961499

[7] BHATTACHARYA, B. B. (2020). Supplement to "Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs." https://doi.org/10.1214/19-AOS1913SUPP.

[8] BICKEL, P. J. (1968). A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Stat.* **40** 1–23. MR0256519 https://doi.org/10.1214/aoms/1177697800

[9] BICKEL, P. J. and BREIMAN, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Probab.* **11** 185–214. MR0682809

[10] BISWAS, M., MUKHOPADHYAY, M. and GHOSH, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika* **101** 913–926. MR3286925 https://doi.org/10.1093/biomet/asu045

[11] CHAN, S.-O., DIAKONIKOLAS, I., VALIANT, G. and VALIANT, P. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms* 1193–1203. ACM, New York. MR3376448 https://doi.org/10.1137/1.9781611973402.88

[12] CHEN, H. and FRIEDMAN, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112** 397–409. MR3646580 https://doi.org/10.1080/01621459.2016.1147356

[13] EBNER, B., HENZE, N. and YUKICH, J. E. (2018). Multivariate goodness-of-fit on flat and curved spaces via nearest neighbor distances. *J. Multivariate Anal.* **165** 231–242. MR3768763 https://doi.org/10.1016/j.jmva.2017.12.009

[14] FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. MR0532236

[15] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[16] GYŐRFI, L. and NEMETZ, T. (1977). f-dissimilarity: A general class of separation measures of several probability measures. In *Topics in Information Theory. Colloq. Math. Soc. János Bolyai*, **16** 309–321. MR0459923

[17] HALL, P. and TAJVIDI, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89** 359–374. MR1913964 https://doi.org/10.1093/biomet/89.2.359

[18] HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783. MR0947577 https://doi.org/10.1214/aos/1176350835

[19] HENZE, N. and PENROSE, M. D. (1999). On the multivariate runs test. *Ann. Statist.* **27** 290–298. MR1701112 https://doi.org/10.1214/aos/1018031112

[20] LIU, R. Y. and SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* **88** 252–260. MR1212489

[21] MAA, J.-F., PEARL, D. K. and BARTOSZYŃSKI, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann. Statist.* **24** 1069–1074. MR1401837 https://doi.org/10.1214/aos/1032526956

[22] MANN, H. B. and WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18** 50–60. MR0022058 https://doi.org/10.1214/aoms/1177730491

[23] PENROSE, M. (2003). *Random Geometric Graphs. Oxford Studies in Probability* **5**. Oxford Univ. Press, Oxford. MR1986198 https://doi.org/10.1093/acprof:oso/9780198506263.001.0001

[24] PENROSE, M. D. (2007). Gaussian limits for random geometric measures. *Electron. J. Probab.* **12** 989–1035. MR2336596 https://doi.org/10.1214/EJP.v12-429

[25] PENROSE, M. D. and YUKICH, J. E. (2003). Weak laws of large numbers in geometric probability. *Ann. Appl. Probab.* **13** 277–303. MR1952000 https://doi.org/10.1214/aoap/1042765669

[26] ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. MR2168202 https://doi.org/10.1111/j.1467-9868.2005.00513.x

[27] ROUSSON, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. *J. Multivariate Anal.* **80** 43–57. MR1889832 https://doi.org/10.1006/jmva.2000.1981

[28] SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

[29] SMIRNOFF, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow Univ. Math. Bull.* **2** 16. MR0002062

[30] TUKEY, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (*Vancouver, B. C.*, 1974), *Vol.* 2 523–531. MR0426989

[31] WALD, A. and WOLFOWITZ, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Stat.* **11** 147–162. MR0002083 https://doi.org/10.1214/aoms/1177731909

[32] WEISS, L. (1960). Two-sample tests for multivariate distributions. *Ann. Math. Stat.* **31** 159–164. MR0119305 https://doi.org/10.1214/aoms/1177705995

[33] YUKICH, J. (2013). Limit theorems in discrete stochastic geometry. In *Stochastic Geometry, Spatial Statistics and Random Fields. Lecture Notes in Math.* **2068** 239–275. Springer, Heidelberg. MR3059650 https://doi.org/10.1007/978-3-642-33305-7_8