Convergence Rates for Empirical Estimation of Binary Classification Bounds

Salimeh Yasaei Sekeh [1],*[ID]**, Morteza Noshad** [2]**, Kevin R. Moon** [3] **and Alfred O. Hero** [2][ID]

[1] School of Computing and Information Science, University of Maine, Orono, ME 04469, USA
[2] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA; noshad@umich.edu (M.N.); hero@umich.edu (A.O.H.)
[3] Department of Mathematics and Statistics, Utah State University, Logan, UT 84322, USA; kevin.moon@usu.edu
* Correspondence: salimeh.yasaei@maine.edu

check for updates

Abstract: Bounding the best achievable error probability for binary classification problems is relevant to many applications including machine learning, signal processing, and information theory. Many bounds on the Bayes binary classification error rate depend on information divergences between the pair of class distributions. Recently, the Henze–Penrose (HP) divergence has been proposed for bounding classification error probability. We consider the problem of empirically estimating the HP-divergence from random samples. We derive a bound on the convergence rate for the Friedman–Rafsky (FR) estimator of the HP-divergence, which is related to a multivariate runs statistic for testing between two distributions. The FR estimator is derived from a multicolored Euclidean minimal spanning tree (MST) that spans the merged samples. We obtain a concentration inequality for the Friedman–Rafsky estimator of the Henze–Penrose divergence. We validate our results experimentally and illustrate their application to real datasets.

Keywords: classification; Bayes error rate; Henze–Penrose divergence; Friedman–Rafsky test statistic; convergence rates; bias and variance trade-off; concentration bounds; minimal spanning trees

1. Introduction

Divergence measures between probability density functions are used in many signal processing applications including classification, segmentation, source separation, and clustering (see [1–3]). For more applications of divergence measures, we refer to [4].

In classification problems, the Bayes error rate is the expected risk for the Bayes classifier, which assigns a given feature vector $\mathbf{x}$ to the class with the highest posterior probability. The Bayes error rate is the lowest possible error rate of any classifier for a particular joint distribution. Mathematically, let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbb{R}^d$ be realizations of random vector $\mathbf{X}$ and class labels $S \in \{0, 1\}$, with prior probabilities $p = P(S = 0)$ and $q = P(S = 1)$, such that $p + q = 1$. Given conditional probability densities $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$, the Bayes error rate is given by

$$\epsilon = \int_{\mathbb{R}^d} \min \{ p f_0(\mathbf{x}), q f_1(\mathbf{x}) \} d\mathbf{x}. \tag{1}$$

The Bayes error rate provides a measure of classification difficulty. Thus, when known, the Bayes error rate can be used to guide the user in the choice of classifier and tuning parameter selection. In practice, the Bayes error is rarely known and must be estimated from data. Estimation of the Bayes error rate is difficult due to the nonsmooth min function within the integral in (1). Thus, research has focused on deriving tight bounds on the Bayes error rate based on smooth relaxations of the min function.

Many of these bounds can be expressed in terms of divergence measures such as the Bhattacharyya [5] and Jensen–Shannon [6]. Tighter bounds on the Bayes error rate can be obtained using an important divergence measure known as the Henze–Penrose (HP) divergence [7,8].

Many techniques have been developed for estimating divergence measures. These methods can be broadly classified into two categories: (i) plug-in estimators in which we estimate the probability densities and then plug them in the divergence function [9–12], (ii) entropic graph approaches, in which the relationship between the divergence function and a graph functional in Euclidean space is derived [8,13]. Examples of plug-in methods include k-nearest neighbor (K-NN) and Kernel density estimator (KDE) divergence estimators. Examples of entropic graph approaches include methods based on minimal spanning trees (MST), K-nearest neighbors graphs (K-NNG), minimal matching graphs (MMG), traveling salesman problem (TSP), and their power-weighted variants.

Disadvantages of plug-in estimators are that these methods often require assumptions on the support set boundary and are more computationally complex than direct graph-based approaches. Thus, for practical and computational reasons, the asymptotic behavior of entropic graph approaches has been of great interest. Asymptotic analysis has been used to justify graph based approaches. For instance, in [14], the authors showed that a cross match statistic based on optimal weighted matching converges to the the HP-divergence. In [15], a more complex approach based on the K-NNG was proposed that also converges to the HP-divergence.

The first contribution of our paper is that we obtain a bound on the convergence rates for the Friedman and Rafsky (FR) estimator of the HP-divergence, which is based on a multivariate extension of the non-parametric run length test of equality of distributions. This estimator is constructed using a multicolored MST on the labeled training set where MST edges connecting samples with dichotomous labels are colored differently from edges connecting identically labeled samples. While previous works have investigated the FR test statistic in the context of estimating the HP-divergence (see [8,16]), to the best of our knowledge, its minimax MSE convergence rate has not been previously derived. The bound on convergence rate is established by using the umbrella theorem of [17], for which we define a dual version of the multicolor MST. The proposed dual MST in this work is different than the standard dual MST introduced by Yukich in [17]. We show that the bias rate of the FR estimator is bounded by a function of N, η and d, as $O\big((N)^{-\eta^2/(d(\eta+1))}\big)$, where N is the total sample size, d is the dimension of the data samples $d \geq 2$, and η is the Hölder smoothness parameter $0 < \eta \leq 1$. We also obtain the variance rate bound as $O\big((N)^{-1}\big)$.

The second contribution of our paper is a new concentration bound for the FR test statistic. The bound is obtained by establishing a growth bound and a smoothness condition for the multicolored MST. Since the FR test statistic is not a Euclidean functional, we cannot use the standard subadditivity and superadditivity approaches of [17–19]. Our concentration inequality is derived using a different Hamming distance approach and a dual graph to the multicolored MST.

We experimentally validate our theoretic results. We compare the MSE theory and simulation in three experiments with various dimensions $d = 2, 4, 8$. We observe that, in all three experiments, as sample size increases, the MSE rate decreases and, for higher dimensions, the rate is slower. In all sets of experiments, our theory matches the experimental results. Furthermore, we illustrate the application of our results on estimation of the Bayes error rate on three real datasets.

1.1. Related Work

Much research on minimal graphs has focused on the use of Euclidean functionals for signal processing and statistics applications such as image registration [20,21], pattern matching [22], and non-parametric divergence estimation [23]. A K-NNG-based estimator of Rényi and f-divergence measures has been proposed in [13]. Additional examples of direct estimators of divergence measures include statistic based on the nonparametric two sample problem, the Smirnov maximum deviation test [24], and the Wald–Wolfowitz [25] runs test, which have been studied in [26].

Many entropic graph estimators such as MST, K-NNG, MMG, and TSP have been considered for multivariate data from a single probability density f. In particular, the normalized weight function of graph constructions all converge almost surely to the Rényi entropy of f [17,27]. For N uniformly distributed points, the MSE is $O(N^{-1/d})$ [28,29]. Later, Hero et al. [30,31] reported bounds on L_γ-norm bias convergence rates of power-weighted Euclidean weight functionals of order γ for densities f belonging to the space of Hölder continuous functions $\Sigma_d(\eta, K)$ as $O(N^{-\alpha\eta/(\alpha\eta+1)\,1/d})$, where $0 < \eta \leq 1$, $d \geq 1$, $\gamma \in (1, d)$, and $\alpha = (d - \gamma)/d$. In this work, we derive a bound on convergence rate of FR estimator for the HP-divergence when the density functions belong to the Hölder class, $\Sigma_d(\eta, K)$, for $0 < \eta \leq 1$, $d \geq 2$ [32]. Note that throughout the paper we assume the density functions are absolutely continuous and bounded with support on the unit cube $[0, 1]^d$.

In [28], Yukich introduced the general framework of continuous and quasi-additive Euclidean functionals. This has led to many convergence rate bounds of entropic graph divergence estimators.

The framework of [28] is as follows: Let F be finite subset of points in $[0, 1]^d$, $d \geq 2$, drawn from an underlying density. A real-valued function L_γ defined on F is called a Euclidean functional of order γ if it is of the form $L_\gamma(F) = \min_{E \in \mathcal{E}} \sum_{e \in E} |e(F)|^\gamma$, where $\mathcal{E}$ is a set of graphs, e is an edge in the graph E, $|e|$ is the Euclidean length of e, and γ is called the edge exponent or power-weighting constant. The MST, TSP, and MMG are some examples for which $\gamma = 1$.

Following this framework, we show that the FR test statistic satisfies the required continuity and quasi-additivity properties to obtain similar convergence rates to those predicted in [28]. What distinguishes our work from previous work is that the count of dichotomous edges in the multicolored MST is not Euclidean. Therefore, the results in [17,27,30,31] are not directly applicable.

Using the isoperimetric approach, Talagrand [33] showed that, when the Euclidean functional L_γ is based on the MST or TSP, then the functional L_γ for derived random vertices uniformly distributed in a hypercube $[0, 1]^d$ is concentrated around its mean. Namely, with high probability, the functional L_γ and its mean do not differ by more than $C(N \log N)^{(d-\gamma)/2d}$. In this paper, we establish concentration bounds for the FR statistic: with high probability $1 - \delta$, the FR statistic differs from its mean by not more than $O\left((N)^{(d-1)/d}\left(\log(C/\delta)\right)^{(d-1)/d}\right)$, where C is a function of N and d.

1.2. Organization

This paper is organized as follows. In Section 2, we first introduce the HP-divergence and the FR multivariate test statistic. We then present the bias and variance rates of the FR-based estimator of HP-divergence followed by the concentration bounds and the minimax MSE convergence rate. Section 3 provides simulations that validate the theory. All proofs and relevant lemmas are given in the Appendices A–E.

Throughout the paper, we denote expectation by $\mathbb{E}$ and variance by abbreviation Var. Bold face type indicates random variables. In this paper, when we say number of samples we mean number of observations.

2. The Henze–Penrose Divergence Measure

Consider parameters $p \in (0, 1)$ and $q = 1 - p$. We focus on estimating the HP-divergence measure between distributions f_0 and f_1 with domain $\mathbb{R}^d$ defined by

$$D_p(f_0, f_1) = \frac{1}{4pq}\left[\int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})}\,d\mathbf{x} - (p - q)^2\right]. \tag{2}$$

It can be verified that this measure is bounded between 0 and 1 and, if $f_0(\mathbf{x}) = f_1(\mathbf{x})$, then $D_p = 0$. In contrast with some other divergences such as the Kullback–Liebler [34] and Rényi divergences [35], the HP-divergence is symmetrical, i.e., $D_p(f_0, f_1) = D_q(f_1, f_0)$. By invoking relation (3) in [8],

$$\int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})} \, \mathrm{d}\mathbf{x} = 1 - 4pq A_p(f_0, f_1),$$

where

$$
\begin{aligned}
A_p(f_0, f_1) &= \int \frac{f_0(\mathbf{x})f_1(\mathbf{x})}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})} \, \mathrm{d}\mathbf{x} = \mathbb{E}_{f_0}\left[\left(p\, \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})} + q\right)^{-1}\right], \\
u_p(f_0, f_1) &= 1 - 4pq\, A_p(f_0, f_1),
\end{aligned}
$$

one can rewrite D_p in the alternative form:

$$D_p(f_0, f_1) = 1 - A_p(f_0, f_1) = \frac{u_p(f_0, f_1)}{4pq} - \frac{(p-q)^2}{4pq}.$$

Throughout the paper, we refer to $A_p(f_0, f_1)$ as the HP-integral. The HP-divergence measure belongs to the class of ϕ-divergences [36]. For the special case $p = 0.5$, the divergence (2) becomes the symmetric χ^2-divergence and is similar to the Rukhin f-divergence. See [37,38].

2.1. The Multivariate Runs Test Statistic

The MST is a graph of minimum weight among all graphs $\mathcal{E}$ that span n vertices. The MST has many applications including pattern recognition [39], clustering [40], nonparametric regression [41], and testing of randomness [42]. In this section, we focus on the FR multivariate two sample test statistic constructed from the MST.

Assume that sample realizations from f_0 and f_1, denoted by $\mathfrak{X}_m \in \mathbb{R}^{m \times d}$ and $\mathfrak{Y}_n \in \mathbb{R}^{n \times d}$, respectively, are available. Construct an MST spanning the samples from both f_0 and f_1 and color the edges in the MST that connect dichotomous samples green and color the remaining edges black. The FR test statistic $\mathfrak{R}_{m,n} := \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$ is the number of green edges in the MST. Note that the test assumes a unique MST, therefore all inter point distances between data points must be distinct. We recall the following theorem from [7,8]:

Theorem 1. *As $m \to \infty$ and $n \to \infty$ such that $\dfrac{m}{n+m} \to p$ and $\dfrac{n}{n+m} \to q$,*

$$1 - \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\, \frac{m+n}{2mn} \to D_p(f_0, f_1), \quad a.s. \tag{3}$$

In the next section, we obtain bounds on the MSE convergence rates of the FR approximation for HP-divergence between densities that belong to $\Sigma_d(\eta, K)$, the class of Hölder continuous functions with Lipschitz constant K and smoothness parameter $0 < \eta \le 1$ [32]:

Definition 1 (Hölder class). *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space. The Hölder class $\Sigma_d(\eta, K)$, with η-Hölder parameter, of functions with the L_d-norm, consists of the functions g that satisfy*

$$\left\{ g : \left\| g(\mathbf{z}) - p_{\mathbf{x}}^{\lfloor \eta \rfloor}(\mathbf{z}) \right\|_d \le K \left\| \mathbf{x} - \mathbf{z} \right\|_d^\eta, \ \mathbf{x}, \mathbf{z} \in \mathcal{X} \right\}, \tag{4}$$

where $p_{\mathbf{x}}^k(\mathbf{z})$ is the Taylor polynomial (multinomial) of g of order k expanded about the point $\mathbf{x}$ and $\lfloor \eta \rfloor$ is defined as the greatest integer strictly less than η.

In what follows, we will use both notations $\mathfrak{R}_{m,n}$ and $\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$ for the FR statistic over the combined samples.

2.2. Convergence Rates

In this subsection, we obtain the mean convergence rate bounds for general non-uniform Lebesgue densities f_0 and f_1 belonging to the Hölder class $\Sigma_d(\eta, K)$. Since the expectation of $\mathfrak{R}_{m,n}$ can be closely approximated by the sum of the expectation of the FR statistic constructed on a dense partition of $[0, 1]^d$, $\mathfrak{R}_{m,n}$ is a quasi-additive functional in mean. The family of bounds (A16) in Appendix B enables us to achieve the minimax convergence rate for the mean under the Hölder class assumption with smoothness parameter $0 < \eta \leq 1, d \geq 2$:

Theorem 2 (Convergence Rate of the Mean). *Let $d \geq 2$, and $\mathfrak{R}_{m,n}$ be the FR statistic for samples drawn from Hölder continuous and bounded density functions f_0 and f_1 in $\Sigma_d(\eta, K)$. Then, for $d \geq 2$,*

$$\left| \frac{\mathbb{E}\left[\mathfrak{R}_{m,n}\right]}{m+n} - 2pq \int \frac{f_0(\mathbf{x})f_1(\mathbf{x})}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})} \, d\mathbf{x} \right| \leq O\left((m+n)^{-\eta^2/(d(\eta+1))} \right). \tag{5}$$

This bound holds over the class of Lebesgue densities $f_0, f_1 \in \Sigma_d(\eta, K), 0 < \eta \leq 1$. Note that this assumption can be relaxed to $f_0 \in \Sigma_d^s(\eta, K_0)$ and $f_1 \in \Sigma_d^s(\eta, K_1)$ that is Lebesgue densities f_0 and f_1 belong to the Strong Hölder class with the same Hölder parameter η and different constants K_0 and K_1, respectively.

The following variance bound uses the Efron–Stein inequality [43]. Note that in Theorem 3 we do not impose any strict assumptions. We only assume that the density functions are absolutely continuous and bounded with support on the unit cube $[0, 1]^d$. Appendix C contains the proof.

Theorem 3. *The variance of the HP-integral estimator based on the FR statistic, $\mathfrak{R}_{m,n}/(m+n)$ is bounded by*

$$Var\left(\frac{\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)}{m+n} \right) \leq \frac{32 \, c_d^2 \, q}{(m+n)}, \tag{6}$$

where the constant c_d depends only on d.

By combining Theorems 2 and 3, we obtain the MSE rate of the form $O\left((m+n)^{-\eta^2/(d(\eta+1))} \right) + O\left((m+n)^{-1} \right)$. Figure 1 indicates a heat map showing the MSE rate as a function of d and $N = m = n$. The heat map shows that the MSE rate of the FR test statistic-based estimator given in (3) is small for large sample size N.
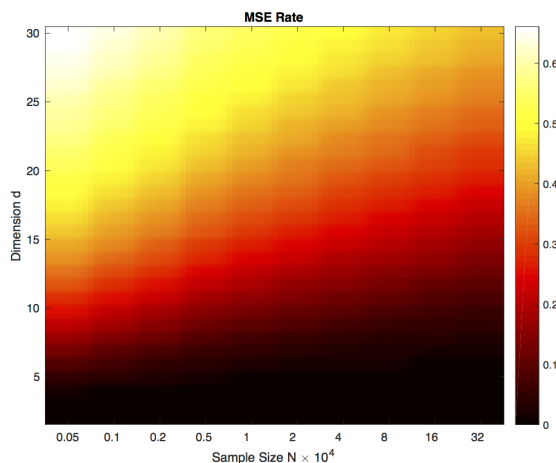

Figure 1. Heat map of the theoretical MSE rate of the FR estimator of the HP-divergence based on Theorems 2 and 3 as a function of dimension and sample size when $N = m = n$. Note the color transition (MSE) as sample size increases for high dimension. For fixed sample size N, the MSE rate degrades in higher dimensions.

2.3. Proof Sketch of Theorem 2

In this subsection, we first establish subadditivity and superadditivity properties of the FR statistic, which will be employed to derive the MSE convergence rate bound. This will establish that the mean of the FR test statistic is a quasi-additive functional:

Theorem 4. *Let $\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$ be the number of edges that link nodes from differently labeled samples $\mathfrak{X}_m = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and $\mathfrak{Y}_n = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ in $[0, 1]^d$. Partition $[0, 1]^d$ into l^d equal volume subcubes Q_i such that m_i and n_i are the number of samples from $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$, respectively, that fall into the partition Q_i. Then, there exists a constant c_1 such that*

$$\mathbb{E}\left[\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\right] \leq \sum_{i=1}^{l^d} \mathbb{E}\left[\mathfrak{R}_{m_i,n_i}\left((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i\right)\right] + 2\, c_1\, l^{d-1}\, (m+n)^{1/d}. \tag{7}$$

Here, $\mathfrak{R}_{m_i,n_i}$ is the number of dichotomous edges in partition Q_i. Conversely, for the same conditions as above on partitions Q_i, there exists a constant c_2 such that

$$\mathbb{E}\left[\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\right] \geq \sum_{i=1}^{l^d} \mathbb{E}\left[\mathfrak{R}_{m_i,n_i}\left((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i\right)\right] - 2\, c_2\, l^{d-1}\, (m+n)^{1/d}. \tag{8}$$

The inequalities (7) and (8) are inspired by corresponding inequalities in [30,31]. The full proof is given in Appendix A. The key result in the proof is the inequality:

$$\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \sum_{i=1}^{l^d} \mathfrak{R}_{m_i,n_i}\left((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i\right) + 2|D|,$$

where $|D|$ indicates the number of all edges of the MST which intersect two different partitions.

Furthermore, we adapt the theory developed in [17,30] to derive the MSE convergence rate of the FR statistic-based estimator by defining a dual MST and dual FR statistic, denoted by MST* and $\mathfrak{R}_{m,n}^*$ respectively (see Figure 2):
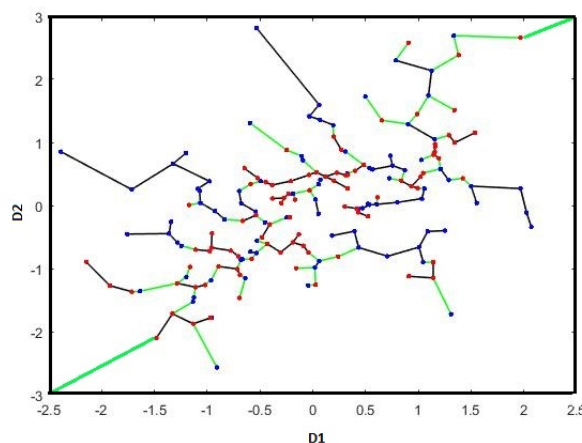

Figure 2. The dual MST spanning the merged set $\mathfrak{X}_m$ (blue points) and $\mathfrak{Y}_n$ (red points) drawn from two Gaussian distributions. The dual FR statistic ($\mathfrak{R}_{m,n}^*$) is the number of edges in the MST* (contains nodes in $\mathfrak{X}_m \cup \mathfrak{Y}_n \cup \{2 \text{ corner points}\}$) that connect samples from different color nodes and corners (denoted in green). Black edges are the non-dichotomous edges in the MST*.

Definition 2 (Dual MST, MST* and dual FR statistic $\mathfrak{R}_{m,n}^*$)**.** *Let $\mathbb{F}_i$ be the set of corner points of the subsection Q_i for $1 \leq i \leq l^d$. Then, we define $\text{MST}^*(\mathfrak{X}_m \cup \mathfrak{Y}_n \cap Q_i)$ as the boundary MST graph of partition Q_i [17], which contains $\mathfrak{X}_m$ and $\mathfrak{Y}_n$ points falling inside the section Q_i and those corner points in $\mathbb{F}_i$ which minimize total MST length. Notice it is allowed to connect the MSTs in Q_i and Q_j through points strictly*

contained in Q_i and Q_j and corner points are taken into account under condition of minimizing total MST length. Another word, the dual MST can connect the points in $Q_i \cup Q_j$ by direct edges to pair to another point in $Q_i \cup Q_j$ or the corner the corner points (we assume that all corner points are connected) in order to minimize the total length. To clarify this, assume that there are two points in $Q_i \cup Q_j$, then the dual MST consists of the two edges connecting these points to the corner if they are closed to a corner point; otherwise, dual MST consists of an edge connecting one to another. Furthermore, we define $\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i)$ as the number of edges in an MST* graph connecting nodes from different samples and number of edges connecting to the corner points. Note that the edges connected to the corner nodes (regardless of the type of points) are always counted in dual FR test statistic $\mathfrak{R}^*_{m,n}$.

In Appendix B, we show that the dual FR test statistic is a quasi-additive functional in mean and $\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \geq \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$. This property holds true since $\mathrm{MST}(\mathfrak{X}_m, \mathfrak{Y}_n)$ and $\mathrm{MST}^*(\mathfrak{X}_m, \mathfrak{Y}_n)$ graphs can only be different in the edges connected to the corner nodes, and in $\mathfrak{R}^*(\mathfrak{X}_m, \mathfrak{Y}_n)$ we take all of the edges between these nodes and corner nodes into account.

To prove Theorem 2, we partition $[0,1]^d$ into l^d subcubes. Then, by applying Theorem 4 and the dual MST, we derive the bias rate in terms of partition parameter l (see (A16) in Theorem A1). See Appendix B and Appendix E for details. According to (A16), for $d \geq 2$, and $l = 1, 2, \ldots$, the slowest rates as a function of l are $l^d(m+n)^{\eta/d}$ and $l^{-\eta d}$. Therefore, we obtain an l-independent bound by letting l be a function of $m + n$ that minimizes the maximum of these rates i.e.,

$$l(m+n) = arg \min_l \ \max \left\{ l^d(m+n)^{-\eta/d}, l^{-\eta d} \right\}.$$

The full proof of the bound in (2) is given in Appendix B.

2.4. Concentration Bounds

Another main contribution of our work in this part is to provide an exponential inequality convergence bound derived for the FR estimator of the HP-divergence. The error of this estimator can be decomposed into a bias term and a variance-like term via the triangle inequality:

$$\left| \mathfrak{R}_{m,n} - \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})} \mathrm{d}\mathbf{x} \right| \leq \underbrace{\left| \mathfrak{R}_{m,n} - \mathbb{E}\left[\mathfrak{R}_{m,n} \right] \right|}_{\text{variance-like term}}$$

$$+ \underbrace{\left| \mathbb{E}\left[\mathfrak{R}_{m,n} \right] - \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})} \mathrm{d}\mathbf{x} \right|}_{\text{bias term}}.$$

The bias bound was given in Theorem 2. Therefore, we focus on an exponential concentration bound for the variance-like term. One application of concentration bounds is to employ these bounds to compare confidence intervals on the HP-divergence measure in terms of the FR estimator. In [44,45], the authors provided an exponential inequality convergence bound for an estimator of Rény divergence for a smooth Hölder class of densities on the d-dimensional unite cube $[0,1]^d$. We show that if $\mathfrak{X}_m$ and $\mathfrak{Y}_n$ are the set of m and n points drawn from any two distributions f_0 and f_1, respectively, the FR criteria $\mathfrak{R}_{m,n}$ is tightly concentrated. Namely, we establish that, with high probability, $\mathfrak{R}_{m,n}$ is within

$$1 - O\left((m+n)^{-2/d} \epsilon^{*2} \right)$$

of its expected value, where ϵ^* is the solution of the following convex optimization problem:

$$
\begin{aligned}
\min_{\epsilon \geq 0} \quad & C'_{m,n}(\epsilon) \; \exp\left(\frac{-(t/(2\epsilon))^{d/(d-1)}}{(m+n)\tilde{C}}\right) \\
\text{subject to} \quad & \epsilon \geq O\big(7^{d+1}(m+n)^{1/d}\big),
\end{aligned}
\tag{9}
$$

where $\tilde{C} = 8(4)^{d/(d-1)}$ and

$$
C'_{m,n}(\epsilon) = 8\left(1 - O\big((m+n)^{-2/d}\epsilon^2\big)\right)^{-2}.
\tag{10}
$$

Note that, under the assumption $(m+n)^{1/d} \simeq 1$, $C'_{m,n}(\epsilon)$ becomes a constant depending only on ϵ by $8\left(1 - (c\,\epsilon^2)\right)^{-2}$, where c is a constant. This is inferred from Theorems 5 and 6 below as $(m+n)^{1/d} \simeq 1$. See Appendix D, specifically Lemmas A8–A12 for more detail. Indeed, we first show the concentration around the median. A median is by definition any real number M_e that satisfies the inequalities $P(X \leq M_e) \geq 1/2$ and $P(X \geq M_e) \geq 1/2$. To derive the concentration results, the properties of growth bounds and smoothness for $\mathfrak{R}_{m,n}$, given in Appendix D, are exploited.

Theorem 5 (Concentration around the median). *Let M_e be a median of $\mathfrak{R}_{m,n}$ which implies that $P\big(\mathfrak{R}_{m,n} \leq M_e\big) \geq 1/2$. Recall ϵ^* from (9) then we have*

$$
P\left(\left|\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - M_e\right| \geq t\right) \leq C'_{m,n}(\epsilon^*) \; \exp\left(\frac{-(t/\epsilon^*)^{d/(d-1)}}{(m+n)\tilde{C}}\right),
\tag{11}
$$

where $\tilde{C} = 8(4)^{d/(d-1)}$.

Theorem 6 (Concentration of $\mathfrak{R}_{m,n}$ around the mean). *Let $\mathfrak{R}_{m,n}$ be the FR statistic. Then,*

$$
P\left(\left|\mathfrak{R}_{m,n} - \mathbb{E}[\mathfrak{R}_{m,n}]\right| \geq t\right) \leq C'_{m,n}(\epsilon^*) \exp\left(\frac{-(t/(2\epsilon^*))^{d/(d-1)}}{(m+n)\,\tilde{C}}\right).
\tag{12}
$$

Here, $\tilde{C} = 8(4)^{d/(d-1)}$ and the explicit form for $C'_{m,n}(\epsilon^*)$ is given by (10) when $\epsilon = \epsilon^*$.

See Appendix D for full proofs of Theorems 5 and 6. Here, we sketch the proofs. The proof of the concentration inequality for $\mathfrak{R}_{m,n}$, Theorem 6, requires involving the median M_e, where $P(\mathfrak{R}_{m,n} \leq M_e) \geq 1/2$, inside the probability term by using

$$
\left|\mathfrak{R}_{m,n} - \mathbb{E}[\mathfrak{R}_{m,n}]\right| \leq \left|\mathfrak{R}_{m,n} - M_e\right| + \left|\mathbb{E}[\mathfrak{R}_{m,n}] - M_e\right|.
$$

To prove the expressions for the concentration around the median, Theorem 5, we first consider the h^d uniform partitions of $[0,1]^d$, with edges parallel to the coordinate axes having edge lengths h^{-1} and volumes h^{-d}. Then, by applying the Markov inequality, we show that with at least probability $1 - \big(\delta_{m,n}^h/\epsilon\big)$, where $\delta_{m,n}^h = O\big(h^{d-1}(m+n)^{1/d}\big)$, the FR statistic $\mathfrak{R}_{m,n}$ is subadditive with 2ϵ threshold. Afterward, owing to the induction method [17], the growth bound can be derived with at least probability $1 - \big(h\,\delta_{m,n}^h/\epsilon\big)$. The growth bound explains that with high probability there exists a constant depending on ϵ and h, $C_{\epsilon,h}$, such that $\mathfrak{R}_{m,n} \leq C_{\epsilon,h}(m\,n)^{1-1/d}$. Applying the law of total probability and semi-isoperimetric inequality (A108) in Lemma A11 gives us (A35). By considering the solution to convex optimization problem (9), i.e., ϵ^* and optimal $h = 7$ the claimed results (11) and (12) are derived. The only constraint here is that ϵ is lower bounded by a function of $\delta_{m,n}^h = O\big(h^{d-1}(m+n)^{1/d}\big)$.

Next, we provide a bound for the variance-like term with high probability at least $1 - \delta$. According to the previous results, we expect that this bound depends on ϵ^*, d, m and n. The proof is short and is given in Appendix D.

Theorem 7 (Variance-like bound for $\mathfrak{R}_{m,n}$). *Let $\mathfrak{R}_{m,n}$ be the FR statistic. With at least probability $1 - \delta$, we have*

$$\left| \mathfrak{R}_{m,n} - \mathbb{E}[\mathfrak{R}_{m,n}] \right| \leq O\left(\epsilon^* \, (m+n)^{(d-1)/d} \Big(\log \big(C'_{m,n}(\epsilon^*)/\delta \big) \Big)^{(d-1)/d} \right). \tag{13}$$

or, equivalently,

$$\left| \frac{\mathfrak{R}_{m,n}}{m+n} - \frac{\mathbb{E}[\mathfrak{R}_{m,n}]}{m+n} \right| \leq O\left(\epsilon^* \, (m+n)^{-1/d} \Big(\log \big(C'_{m,n}(\epsilon^*)/\delta \big) \Big)^{(d-1)/d} \right), \tag{14}$$

where $C'_{m,n}(\epsilon^*)$ depends on m, n, and d is given in (10) when $\epsilon = \epsilon^*$.

3. Numerical Experiments

3.1. Simulation Study

In this section, we apply the FR statistic estimate of the HP-divergence to both simulated and real data sets. We present results of a simulation study that evaluates the proposed bound on the MSE. We numerically validate the theory stated in Sections 2.2 and 2.4 using multiple simulations. In the first set of simulations, we consider two multivariate Normal random vectors **X**, **Y** and perform three experiments $d = 2, 4, 8$, to analyze the FR test statistic-based estimator performance as the sample sizes m, n increase. For the three dimensions $d = 2, 4, 8$, we generate samples from two normal distributions with identity covariance and shifted means: $\mu_1 = [0, 0]$, $\mu_2 = [1, 0]$ and $\mu_1 = [0, 0, 0, 0]$, $\mu_2 = [1, 0, 0, 0]$ and $\mu_1 = [0, 0, \ldots, 0]$, $\mu_2 = [1, 0, \ldots, 0]$ when $d = 2$, $d = 4$ and $d = 8$, respectively. For all of the following experiments, the sample sizes for each class are equal ($m = n$).

We vary $N = m = n$ up to 800. From Figure 3, we deduce that, when the sample size increases, the MSE decreases such that for higher dimensions the rate is slower. Furthermore, we compare the experiments with the theory in Figure 3. Our theory generally matches the experimental results. However, the MSE for the experiments tends to decrease to zero faster than the theoretical bound. Since the Gaussian distribution has a smooth density, this suggests that a tighter bound on the MSE may be possible by imposing stricter assumptions on the density smoothness as in [12].
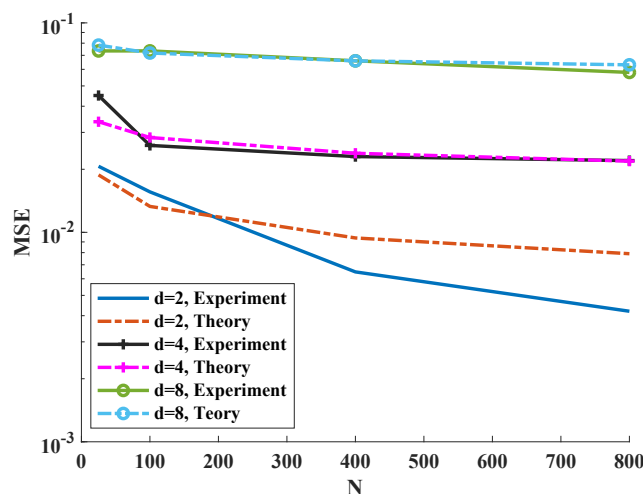

Figure 3. Comparison of the bound on the MSE theory and experiments for $d = 2, 4, 8$ standard Gaussian random vectors versus sample size from 100 trials.

In our next simulation, we compare three bivariate cases: first, we generate samples from a standard Normal distribution. Second, we consider a distinct smooth class of distributions i.e., binomial Gamma density with standard parameters and dependency coefficient $\rho = 0.5$. Third, we generate samples from Standard t-student distributions. Our goal in this experiment is to compare the MSE of the HP-divergence estimator between two identical distributions, $f_0 = f_1$, when f_0 is one of the Gamma, Normal, and t-student density function. In Figure 4, we observe that the MSE decreases as N increases for all three distributions.
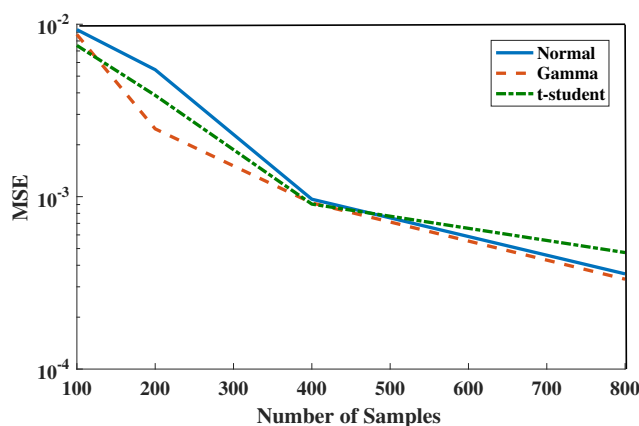

Figure 4. Comparison of experimentally predicted MSE of the FR-statistic as a function of sample size $m = n$ in various distributions Standard Normal, Gamma ($\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 1$, $\rho = 0.5$) and Standard t-Student.

3.2. Real Datasets

We now show the results of applying the FR test statistic to estimate the HP-divergence using three different real datasets [46]:

- Human Activity Recognition (HAR), Wearable Computing, Classification of Body Postures and Movements (PUC-Rio): This dataset contains five classes (sitting-down, standing-up, standing, walking, and sitting) collected on eight hours of activities of four healthy subjects.
- Skin Segmentation dataset (SKIN): The skin dataset is collected by randomly sampling B,G,R values from face images of various age groups (young, middle, and old), race groups (white, black, and asian), and genders obtained from the FERET and PAL databases [47].
- Sensorless Drive Diagnosis (ENGIN) dataset: In this dataset, features are extracted from electric current drive signals. The drive has intact and defective components. The dataset contains 11 different classes with different conditions. Each condition has been measured several times under 12 different operating conditions, e.g., different speeds, load moments, and load forces.

We focus on two classes from each of the HAR, SKIN, and ENGIN datasets, specifically, for HAR dataset two classes "sitting" and "standing" and for SKIN dataset the classes "Skin" and "Non-skin" are considered. In the ENGIN dataset, the drive has intact and defective components, which results in 11 different classes with different conditions. We choose conditions 1 and 2.

In the first experiment, we computed the HP-divergence using KDE plug-in estimator and then the MSE for the FR test statistic estimator is derived as the sample size $N = m = n$ increases. We used 95% confidence interval as the error bars. We observe in Figure 5 that the estimated HP-divergence ranges in $[0,1]$, which is one of the HP-divergence properties [8]. Interestingly, when N increases the HP-divergence tends to 1 for all HAR, SKIN, and ENGIN datasets. Note that in this set of experiments we have repeated the experiments on independent parts of the datasets to obtain the error bars. Figure 6 shows that the MSE expectedly decreases as the sample size grows for all three datasets. Here, we have used the KDE plug-in estimator [12], implemented on the all available samples, to determine the

true HP-divergence. Furthermore, according to Figure 6, the FR test statistic-based estimator suggests that the Bayes error rate is larger for the SKIN dataset compared to the HAR and ENGIN datasets.
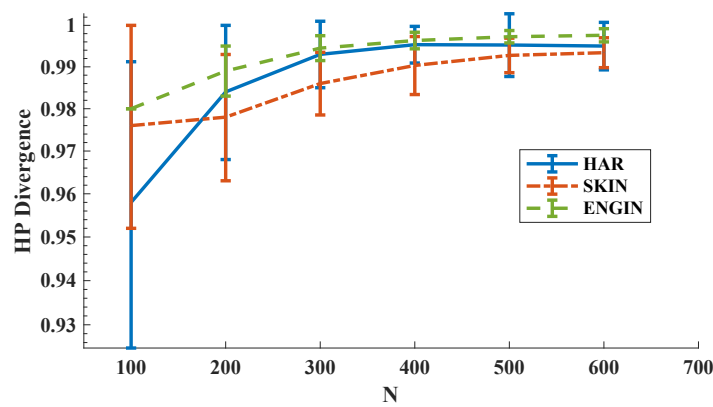

Figure 5. HP-divergence vs. sample size for three real datasets HAR, SKIN, and ENGIN.
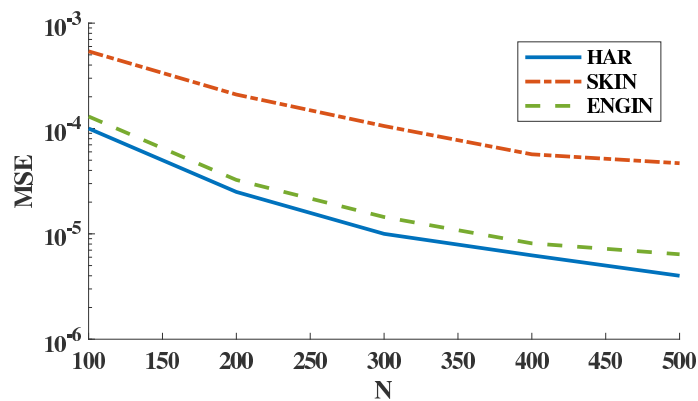

Figure 6. The empirical MSE vs. sample size. The empirical MSE of the FR estimator for all three datasets HAR, SKIN, and ENGIN decreases for larger sample size N.

In our next experiment, we add the first six features (dimensions) in order to our datasets and evaluate the FR test statistic's performance as the HP-divergence estimator. Surprisingly, the estimated HP-divergence doesn't change for the HAR sample; however, big changes are observed for the SKIN and ENGIN samples (see Figure 7).
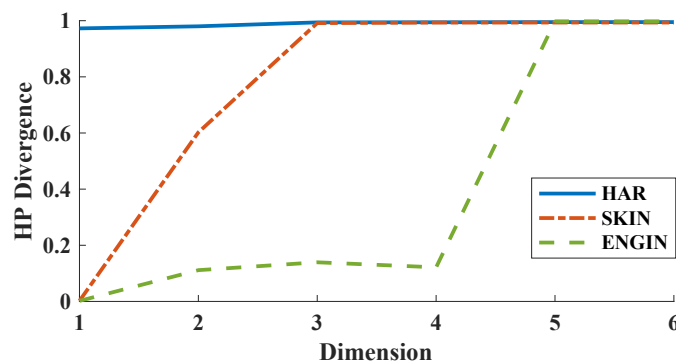

Figure 7. HP-divergence vs. dimension for three datasets HAR, SKIN, and ENGIN.

Finally, we apply the concentration bounds on the FR test statistic (i.e., Theorems 6 and 7) and compute theoretical implicit variance-like bound for the FR criteria with $\delta = 0.05$ error for the real

datasets ENGIN, HAR, and SKIN. Since datasets ENGIN, HAR, and SKIN have the equal total sample size $N = m + n = 1200$ and different dimensions $d = 14, 12, 4$, respectively; here, we first intend to compare the concentration bound (13) on the FR statistic in terms of dimension d when $\delta = 0.05$. For real datasets ENGIN, HAR, and SKIN, we obtain

$$P\left(|\Re_{m,n} - \mathbb{E}[\Re_{m,n}]| \leq \xi\right) \geq 0.95,$$

where $\xi = \xi'[0.257, 0.005, 0.6 \times 10^{-11}]$, respectively, and ξ' is a constant not dependent on d. One observes that as the dimension decreases the interval becomes significantly tighter. However, this could not be generally correct and computing bound (13) precisely requires the knowledge of distributions and unknown constants. In Table 1, we compute the standard variance-like bound by applying the percentiles technique and observe that the bound threshold is not monotonic in terms of dimension d. Table 1 shows the FR test statistic, HP-divergence estimate (denoted by $\Re_{m,n}$, $\widehat{D}_p$, respectively), and standard variance-like interval for the FR statistic using the three real datasets HAR, SKIN, and ENGIN.

Table 1. $\Re_{m,n}$, $\widehat{D}_p$, m, and n are the FR test statistic, HP-divergence estimates using $\Re_{m,n}$, and sample sizes for two classes, respectively.

	FR Test Statistic				
Dataset	$\mathbb{E}[\Re_{m,n}]$	$\widehat{D}_p$	m	n	**Variance-Like Interval**
HAR	3	0.995	600	600	(2.994,3.006)
SKIN	4.2	0.993	600	600	(4.196,4.204)
ENGIN	1.8	0.997	600	600	(1.798,1.802)

4. Conclusions

We derived a bound on the MSE convergence rate for the Friedman–Rafsky estimator of the Henze–Penrose divergence assuming the densities are sufficiently smooth. We employed a partitioning strategy to derive the bias rate which depends on the number of partitions, the sample size $m + n$, the Hölder smoothness parameter η, and the dimension d. However, by using the optimal partition number, we derived the MSE convergence rate only in terms of $m + n$, η, and d. We validated our proposed MSE convergence rate using simulations and illustrated the approach for the meta-learning problem of estimating the HP-divergence for three real-world data sets. We also provided concentration bounds around the median and mean of the estimator. These bounds explicitly provide the rate that the FR statistic approaches its median/mean with high probability, not only as a function of the number of samples, m, n, but also in terms of the dimension of the space d. By using these results, we explored the asymptotic behavior of a variance-like rate in terms of m, n, and d.

Abbreviations

HP	Henze-Penrose
BER	Bayes error rate
MST	Minimal Spanning Tree
FR	Friedman-Rafsky
MSE	Mean squared error

Appendix A. Proof of Theorem 4

In this section, we prove the subadditivity and superadditivity for the mean of FR test statistic. For this, first we need to illustrate the following lemma.

Lemma A1. *Let $\{Q_i\}_{i=1}^{l^d}$ be a uniform partition of $[0,1]^d$ into l^d subcubes Q_i with edges parallel to the coordinate axes having edge lengths l^{-1} and volumes l^{-d}. Let D_{ij} be the set of edges of MST graph between Q_i and Q_j with cardinality $|D_{ij}|$, then for $|D|$ defined as the sum of $|D_{ij}|$ for all $i,j = 1,\ldots,l^d$, $i \neq j$, we have $\mathbb{E}|D| = O(l^{d-1}\,n^{1/d})$, or more explicitly*

$$\mathbb{E}[|D|] \leq C' l^{d-1} n^{1/d} + O(l^{d-1} n^{(1/d)-s}), \tag{A1}$$

where $\eta > 0$ is the Hölder smoothness parameter and

$$s = \frac{(1-1/d)\eta}{d\left((1-1/d)\eta + 1\right)}.$$

Here, and in what follows, denote $\Xi_{MST}(\mathfrak{X}_n)$ the length of the shortest spanning tree on $\mathfrak{X}_n = \{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$, namely

$$\Xi_{MST}(\mathfrak{X}_n) := \min_T \sum_{e \in T} |e|,$$

where the minimum is over all spanning trees T of the vertex set $\mathfrak{X}_n$. Using the subadditivity relation for Ξ_{MST} in [17], with the uniform partition of $[0,1]^d$ into l^d subcubes Q_i with edges parallel to the coordinate axes having edge lengths l^{-1} and volumes l^{-d}, we have

$$\Xi_{MST}(\mathfrak{X}_n) \leq \sum_{i=1}^{l^d} \Xi_{MST}(\mathfrak{X}_n \cap Q_i) + C\, l^{d-1}, \tag{A2}$$

where C is constant. Denote D the set of all edges of $MST\left(\bigcup_{i=1}^{M} Q_i\right)$ that intersect two different subcubes Q_i and Q_j with cardinality $|D|$. Let $|e_i|$ be the length of i-th edge in set D. We can write

$$\sum_{i \in |D|} |e_i| \leq C l^{d-1} \quad \text{and} \quad \mathbb{E} \sum_{i \in |D|} |e_i| \leq C l^{d-1},$$

also we know that

$$\mathbb{E} \sum_{i \in |D|} |e_i| = \mathbb{E}_D \sum_{i \in |D|} \mathbb{E}[|e_i||D]. \tag{A3}$$

Note that using the result from ([31], Proposition 3), for some constants C_{i1} and C_{i2}, we have

$$\mathbb{E}|e_i| \leq C_{i1} n^{-1/d} + C_{i2} n^{-(1/d)-s}, \quad i \in |D|. \tag{A4}$$

Now, let $C_1 = \max_i \{C_{i1}\}$ and $C_2 = \max_i \{C_{i2}\}$, hence we can bound the expectation (A3) as

$$\mathbb{E}|D| \, (C_1 n^{-1/d} + C_2 (n^{-(1/d)-s})) \leq Cl^{d-1},$$

which implies

$$\mathbb{E}|D| \leq (C_1 n^{-1/d} + O(n^{-(1/d)-s}))$$

$$\leq C' l^{d-1} n^{1/d} + O(l^{d-1} n^{(1/d)-s}).$$

To aim toward the goal (7), we partition $[0,1]^d$ into $M := l^d$ subcubes Q_i of side $1/l$. Recalling Lemma 2.1 in [48], we therefore have the set inclusion:

$$MST\left(\bigcup_{i=1}^{M} Q_i\right) \subset \bigcup_{i=1}^{M} MST(Q_i) \cup D, \tag{A5}$$

where D is defined as in Lemma A1. Let m_i and n_i be the number of sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ respectively falling into the partition Q_i, such that $\sum_i m_i = m$ and $\sum_i n_i = n$. Introduce sets A and B as

$$A := MST\left(\bigcup_{i=1}^{M} Q_i\right), \quad B := \bigcup_{i=1}^{M} MST(Q_i).$$

Since set B has fewer edges than set A, thus (A5) implies that the difference set of B and A contains at most $2|D|$ edges, where $|D|$ is the number of edges in D. On the other word,

$$|A \Delta B| \leq |A - B| + |B - A| = |D| + |B - A|$$

$$= |D| + (|B| - |B \cap A| \leq |D| + (|A| - |B \cap A|) = 2|D|.$$

The number of edge linked nodes from different samples in set A is bounded by the number of edge linked nodes from different samples in set B plus $2|D|$:

$$\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \sum_{i=1}^{M} \mathfrak{R}_{m_i,n_i}\left((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i\right) + 2|D|. \tag{A6}$$

Here, $\mathfrak{R}_{m_i,n_i}$ stands with the number edge linked nodes from different samples in partition Q_i, M. Next, we address the reader to Lemma A1, where it has been shown that there is a constant c such that $\mathbb{E}|D| \leq c \, l^{d-1} \, (m+n)^{1/d}$. This concludes the claimed assertion (7). Now, to accomplish the proof, the lower bound term in (8) is obtained with similar methodology and the set inclusion:

$$\bigcup_{i=1}^{M} MST(Q_i) \subset MST\left(\bigcup_{i=1}^{M} Q_i\right) \cup D. \tag{A7}$$

This completes the proof.

Appendix B. Proof of Theorem 2

As many of continuous subadditive functionals on $[0,1]^d$, in the case of the FR statistic, there exists a dual superadditive functional $\mathfrak{R}^*_{m,n}$ based on dual MST, MST*, proposed in Definition 2. Note that, in the MST* graph, the degrees of the corner points are bounded by c_d, where it only depends on dimension d, and is the bound for degree of every node in MST graph. The following properties hold true for dual FR test statistic, $\mathfrak{R}^*_{m,n}$:

Lemma A2. *Given samples $\mathfrak{X}_m = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and $\mathfrak{Y}_n = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$, the following inequalities hold true:*

(i) *For constant c_d which depends on d:*

$$\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) + c_d \, 2^d,$$

$$\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n).$$

(A8)

(ii) *(Subadditivity on $\mathbb{E}[\mathfrak{R}^*_{m,n}]$ and Superadditivity) Partition $[0,1]^d$ into l^d subcubes Q_i such that m_i, n_i be the number of sample $\mathfrak{X}_m = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and $\mathfrak{Y}_n = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$ respectively falling into the partition Q_i with dual $\mathfrak{R}^*_{m_i,n_i}$. Then, we have*

$$\mathbb{E}\Big[\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\Big] \leq \sum_{i=1}^{l^d} \mathbb{E}\Big[\mathfrak{R}^*_{m_i,n_i}((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i)\Big] + c \, l^{d-1} \, (m+n)^{1/d},$$

$$\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \geq \sum_{i=1}^{l^d} \mathfrak{R}^*_{m_i,n_i}((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i) - 2^d c_d l^d,$$

(A9)

where c is a constant.

(i) Consider the nodes connected to the corner points. Since $\mathrm{MST}(\mathfrak{X}_m, \mathfrak{Y}_n)$ and $\mathrm{MST}^*(\mathfrak{X}_m, \mathfrak{Y}_n)$ can only be different in the edges connected to these nodes, and in $\mathfrak{R}^*(\mathfrak{X}_m, \mathfrak{Y}_n)$ we take all of the edges between these nodes and corner nodes into account, so we obviously have the second relation in (A8). In addition, for the first inequality in (A8), it is enough to say that the total number of edges connected to the corner nodes is upper bounded by $2^d \, c_d$.

(ii) Let $|D^*|$ be the set of edges of the MST^* graph which intersect two different partitions. Since MST and MST^* are only different in edges of points connected to the corners and edges crossing different partitions. Therefore, $|D^*| \leq |D|$. By eliminating one edge in set D in the worse scenario we would face two possibilities: either the corresponding node is connected to the corner which is counted anyways or any other point in MST graph which wouldn't change the FR test statistic. This implies the following subadditivity relation:

$$\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - |D| \leq \sum_{i=1}^{l^d} \mathfrak{R}^*_{m_i,n_i}((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i).$$

Further from Lemma A1, we know that there is a constant c such that $\mathbb{E}|D| \leq c \, l^{d-1} \, (m+n)^{1/d}$. Hence, the first inequality in (A9) is obtained. Next, consider $|D^*_c|$ which represents the total number of edges from both samples only connected to the all corners points in MST^* graph. Therefore, one can easily claim:

$$\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \geq \sum_{i=1}^{l^d} \mathfrak{R}^*_{m_i,n_i}((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i) - |D^*_c|.$$

In addition, we know that $|D^*_c| \leq 2^d l^d c_d$ where c_d stands with the largest possible degree of any vertex. One can write

$$\mathfrak{R}^*_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \geq \sum_{i=1}^{l^d} \mathfrak{R}^*_{m_i,n_i}((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i) - 2^d c_d l^d.$$

The following list of Lemmas A3, A4 and A6 are inspired from [49] and are required to prove Theorem A1. See Appendix E for their proofs.

Lemma A3. *Let $g(\mathbf{x})$ be a density function with support $[0,1]^d$ and belong to the Hölder class $\Sigma_d(\eta, L)$, $0 < \eta \leq 1$, stated in Definition 1. In addition, assume that $P(\mathbf{x})$ is a η-Hölder smooth function, such that its*

absolute value is bounded from above by a constant. Define the quantized density function with parameter l and constants ϕ_i as

$$\widehat{g}(\mathbf{x}) = \sum_{i=1}^{M} \phi_i \mathbf{1}\{\mathbf{x} \in Q_i\}, \quad \text{where } \phi_i = l^d \int_{Q_i} g(\mathbf{x}) \, d\mathbf{x}. \tag{A10}$$

Let $M = l^d$ and $Q_i = \{\mathbf{x}, \mathbf{x}_i : \|\mathbf{x} - \mathbf{x}_i\| < l^{-d}\}$. Then,

$$\int \left\| (g(\mathbf{x}) - \widehat{g}(\mathbf{x})) P(\mathbf{x}) \right\| \, d\mathbf{x} \le O(l^{-d\eta}). \tag{A11}$$

Lemma A4. *Denote $\Delta(\mathbf{x}, \mathcal{S})$ the degree of vertex $\mathbf{x} \in \mathcal{S}$ in the MST over set $\mathcal{S}$ with the n number of vertices. For given function $P(\mathbf{x}, \mathbf{x})$, one obtains*

$$\int P(\mathbf{x}, \mathbf{x}) g(\mathbf{x}) \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S})] \, d\mathbf{x} = 2 \int P(\mathbf{x}, \mathbf{x}) g(\mathbf{x}) \, d\mathbf{x} + \varsigma_\eta(l, n), \tag{A12}$$

where, for constant $\eta > 0$,

$$\varsigma_\eta(l, n) = \left(O(l/n) - 2 \, l^d/n \right) \int g(\mathbf{x}) P(\mathbf{x}, \mathbf{x}) \, d\mathbf{x} + O(l^{-d\eta}). \tag{A13}$$

Lemma A5. *Assume that, for given k, $g_k(\mathbf{x})$ is a bounded function belong to $\Sigma_d(\eta, L)$. Let $P : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0, 1]$ be a symmetric, smooth, jointly measurable function, such that, given k, for almost every $\mathbf{x} \in \mathbb{R}^d$, $P(\mathbf{x}, .)$ is measurable with $\mathbf{x}$ a Lebesgue point of the function $g_k(.)P(\mathbf{x}, .)$. Assume that the first derivative P is bounded. For each k, let $\mathbf{Z}_1^k, \mathbf{Z}_2^k, \dots, \mathbf{Z}_k^k$ be an independent d-dimensional variable with common density function g_k. Set $\mathfrak{Z}_k = \{\mathbf{Z}_1^k, \mathbf{Z}_2^k \dots, \mathbf{Z}_k^k\}$ and $\mathfrak{Z}_k^{\mathbf{x}} = \{\mathbf{x}, \mathbf{Z}_2^k, \mathbf{Z}_3^k \dots, \mathbf{Z}_k^k\}$. Then,*

$$\mathbb{E}\left[\sum_{j=2}^{k} P(\mathbf{x}, \mathbf{Z}_j^k) \mathbf{1}\{(\mathbf{x}, \mathbf{Z}_j^k) \in MST(\mathfrak{Z}_k^{\mathbf{x}})\} \right] = P(\mathbf{x}, \mathbf{x}) \, \mathbb{E}\left[\Delta(\mathbf{x}, \mathfrak{Z}_k^{\mathbf{x}}) \right] + \left\{ O(k^{-\eta/d}) + O(k^{-1/d}) \right\}. \tag{A14}$$

Lemma A6. *Consider the notations and assumptions in Lemma A5. Then,*

$$\left| k^{-1} \sum_{1 \le i < j \le k} \sum P(\mathbf{Z}_i^k, \mathbf{Z}_j^k) \mathbf{1}\{(\mathbf{Z}_i^k, \mathbf{Z}_j^k) \in MST(\mathfrak{Z}_k)\} - \int_{\mathbb{R}^d} P(\mathbf{x}, \mathbf{x}) g_k(\mathbf{x}) \, d\mathbf{x} \right|$$
$$\le \varsigma_\eta(l, k) + O(k^{-\eta/d}) + O(k^{-1/d}). \tag{A15}$$

Here, $MST(\mathcal{S})$ denotes the MST graph over nice and finite set $\mathcal{S} \subset \mathbb{R}^d$ and η is the smoothness Hölder parameter. Note that $\varsigma_\eta(l, k)$ is given as before in Lemma A4 (A13).

Theorem A1. *Assume $\mathfrak{R}_{m,n} := \mathfrak{R}(\mathfrak{X}_m, \mathfrak{Y}_n)$ denotes the FR test statistic and densities f_0 and f_1 belong to the Hölder class $\Sigma_d(\eta, L)$, $0 < \eta \le 1$. Then, the rate for the bias of the $\mathfrak{R}_{m,n}$ estimator for $d \ge 2$ is of the form:*

$$\left| \frac{\mathbb{E}\left[\mathfrak{R}_{m,n} \right]}{m + n} - 2pq \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})} \, d\mathbf{x} \right| \le O\left(l^d (m + n)^{-\eta/d}\right) + O(l^{-d\eta}). \tag{A16}$$

The proof and a more explicit form for the bound (A16) are given in Appendix E.

Now, we are at the position to prove the assertion in (5). Without loss of generality, assume that $(m + n)l^{-d} > 1$. In the range $d \ge 2$ and $0 < \eta \le 1$, we select l as a function of $m + n$ to be the sequence increasing in $m + n$ which minimizes the maximum of these rates:

$$l(m + n) = arg \, \min_l \max \left\{ l^d (m + n)^{-\eta/d}, \, l^{-\eta d} \right\}.$$

The solution $l = l(m + n)$ occurs when $l^d(m+n)^{-\eta/d} = l^{-\eta d}$, or equivalently $l = \lfloor (m+n)^{\eta/(d^2(\eta+1))} \rfloor$. Substitute this into l in the bound (A16), the RHS expression in (5) for $d \geq 2$ is established.

Appendix C. Proof of Theorems 3

To bound the variance, we will apply one of the first concentration inequalities which was proved by Efron and Stein [43] and further was improved by Steele [18].

Lemma A7 (The Efron–Stein Inequality). *Let* $\mathfrak{X}_m = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ *be a random vector on the space* $\mathcal{S}$. *Let* $\mathfrak{X}' = \{\mathbf{X}'_1, \ldots, \mathbf{X}'_m\}$ *be the copy of random vector* $\mathfrak{X}_m$. *Then, if* $f : \mathcal{S} \times \cdots \times \mathcal{S} \to \mathbb{R}$, *we have*

$$\mathbb{V}\big[f(\mathfrak{X}_m)\big] \leq \frac{1}{2} \sum_{i=1}^{m} \mathbb{E}\Big[\big(f(\mathbf{X}_1, \ldots, \mathbf{X}_m) - f(\mathbf{X}_1, \ldots, \mathbf{X}'_i, \ldots, \mathbf{X}_m)\big)^2\Big]. \tag{A17}$$

Consider two set of nodes $\mathbf{X}_i$, $1 \leq i \leq m$ and $\mathbf{Y}_j$ for $1 \leq j \leq n$. Without loss of generality, assume that $m < n$. Then, consider the $n - m$ virtual random points $\mathbf{X}_{m+1}, \ldots, \mathbf{X}_n$ with the same distribution as $\mathbf{X}_i$, and define $\mathbf{Z}_i := (\mathbf{X}_i, \mathbf{Y}_i)$. Now, for using the Efron–Stein inequality on set $\mathfrak{Z}_n = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, we involve another independent copy of $\mathfrak{Z}_n$ as $\mathfrak{Z}'_n = \{\mathbf{Z}'_1, \ldots, \mathbf{Z}'_n\}$, and define $\mathfrak{Z}_n^{(i)} := (\mathbf{Z}_1, \ldots, \mathbf{Z}_{i-1}, \mathbf{Z}'_i, \mathbf{Z}_{i+1}, \ldots, \mathbf{Z}_n)$, then $\mathfrak{Z}_n^{(1)}$ becomes $(\mathbf{Z}'_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n) = \big\{(\mathbf{X}'_1, \mathbf{Y}'_1), (\mathbf{X}_2, \mathbf{Y}_2), \ldots, (\mathbf{X}_m, \mathbf{Y}_n)\big\} =: (\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n^{(1)})$ where $(\mathbf{X}'_1, \mathbf{Y}'_1)$ is independent copy of $(\mathbf{X}_1, \mathbf{Y}_1)$. Next, define the function $r_{m,n}(\mathfrak{Z}_n) := \mathfrak{R}_{m,n}/(m+n)$, which means that we discard the random samples $\mathbf{X}_{m+1}, \ldots, \mathbf{X}_n$, and find the previously defined $\mathfrak{R}_{m,n}$ function on the nodes $\mathbf{X}_i$, $1 \leq i \leq m$ and $\mathbf{Y}_j$ for $1 \leq j \leq n$, and multiply by some coefficient to normalize it. Then, according to the Efron–Stein inequality, we have

$$Var(r_{m,n}(\mathfrak{Z}_n)) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[(r_{m,n}(\mathfrak{Z}_n) - r_{m,n}(\mathfrak{Z}_n^{(i)}))^2\right].$$

Now, we can divide the RHS as

$$\frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(r_{m,n}(\mathfrak{Z}_n) - r_{m,n}(\mathfrak{Z}_n^{(i)}))^2\right] = \frac{1}{2}\sum_{i=1}^{m} \mathbb{E}\left[(r_{m,n}(\mathfrak{Z}_n) - r_{m,n}(\mathfrak{Z}_n^{(i)}))^2\right]$$
$$+ \frac{1}{2}\sum_{i=m+1}^{n} \mathbb{E}\left[(r_{m,n}(\mathfrak{Z}_n) - r_{m,n}(\mathfrak{Z}_n^{(i)}))^2\right]. \tag{A18}$$

The first summand becomes

$$= \frac{1}{2}\sum_{i=1}^{m} \mathbb{E}\left[(r_{m,n}(\mathfrak{Z}_n) - r_{m,n}(\mathfrak{Z}_n^{(i)}))^2\right] = \frac{m}{2(m+n)^2} \mathbb{E}\left[(\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n^{(1)}))^2\right],$$

which can also be upper bounded as follows:

$$\left|\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n^{(1)})\right| \leq \left|\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n)\right|$$
$$+ \left|\mathfrak{R}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n^{(1)})\right|. \tag{A19}$$

For deriving an upper bound on the second line in (A19), we should observe how much changing a point's position modifies the amount of $\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$. We consider two steps of changing $\mathbf{X}_1$'s position: we first remove it from the graph, and then add it to the new position. Removing it would change $\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$ at most by $2c_d$ because X_1 has a degree of at most c_d, and c_d edges will be

removed from the MST graph, and c_d edges will be added to it. Similarly, adding $\mathbf{X}_1$ to the new position will affect $\mathfrak{R}_{m,n}(\mathfrak{X}_{m,n}, \mathfrak{Y}_{m,n})$ at most by $2c_d$. Thus, we have

$$\left| \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n) \right| \leq 4\, c_d,$$

and we can also similarly reason that

$$\left| \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n^{(1)}) \right| \leq 4\, c_d.$$

Therefore, totally we would have

$$\left| \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(1)}, \mathfrak{Y}_n^{(1)}) \right| \leq 8\, c_d.$$

Furthermore, the second summand in (A18) becomes

$$= \frac{1}{2} \sum_{i=m+1}^{n} \mathbb{E}\left[(r_{m,n}(\mathfrak{Z}_n) - r_{m,n}(\mathfrak{Z}_n^{(i)}))^2 \right] = K_{m,n} \mathbb{E}\left[(\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m^{(m+1)}, \mathfrak{Y}_n^{(m+1)}))^2 \right],$$

where $K_{m,n} = \frac{n-m}{2\,(m+n)^2}$. Since, in $(\mathfrak{X}_m^{(m+1)}, \mathfrak{Y}_n^{(m+1)})$, the point $\mathbf{X}_{m+1}'$ is a copy of virtual random point $\mathbf{X}_{m+1}$, therefore this point doesn't change the FR test statistic $\mathfrak{R}_{m,n}$. In addition, following the above arguments, we have

$$\left| \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n^{(m+1)}) \right| \leq 4\, c_d.$$

Hence, we can bound the variance as below:

$$Var(r_{m,n}(\mathfrak{Z}_n)) \leq \frac{8c_d^2(n-m)}{(m+n)^2} + \frac{32\, c_d^2\, m}{(m+n)^2}. \tag{A20}$$

Combining all results with the fact that $\dfrac{n}{m+n} \to q$ concludes the proof.

Appendix D. Proof of Theorems 5–7

We will need the following prominent results for the proofs.

Lemma A8. *For $h = 1, 2, \ldots$, let $\delta_{m,n}^h$ be the function $c\, h^{d-1}(m+n)^{1/d}$, where c is a constant. Then, for $\epsilon > 0$, we have*

$$P\left(\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_{m_i}, \mathfrak{Y}_{n_i}) + 2\epsilon \right) \geq \frac{\epsilon - \delta_{m,n}^h}{\epsilon}. \tag{A21}$$

Note that, in the case $\epsilon \leq \delta_{m,n}^h$, the above claimed inequality becomes trivial.

The subadditivity property for FR test statistic $\mathfrak{R}_{m,n}$ in Lemma A8, as well as Euclidean functionals, leads to several non-trivial consequences. The growth bound was first explored by Rhee (1993b) [50], and as is illustrated in [17,27] has a wide range of applications. In this paper, we investigate the probabilistic growth bound for $\mathfrak{R}_{m,n}$. This observation will lead us to our main goal in this appendix that is providing the proof of Theorem 6. For what follows, we will use $\delta_{m,n}^h$ notation for the expression $O\left(h^{d-1}(m+n)^{1/d}\right)$.

Lemma A9 (Growth bounds for $\mathfrak{R}_{m,n}$). *Let $\mathfrak{R}_{m,n}$ be the FR test statistic. Then, for given non-negative ϵ, such that $\epsilon \geq h^2 \, \delta_{m,n}^h$, with at least probability $g(\epsilon) := 1 - \dfrac{h \, \delta_{m,n}^h}{\epsilon}$, $h = 2, 3, \ldots$, we have*

$$\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq c_{\epsilon,h}'' \left(\#\mathfrak{X}_m \, \#\mathfrak{Y}_n \right)^{1-1/d}. \tag{A22}$$

Here, $c_{\epsilon,h}'' = O\left(\dfrac{\epsilon}{h^{d-1} - 1} \right)$ depending only on ϵ and h.

The complexity of $\mathfrak{R}_{m,n}$'s behavior and the need to pursue the proof encouraged us to explore the smoothness condition for $\mathfrak{R}_{m,n}$. In fact, this is where both subadditivity and superadditivity for the FR statistic are used together and become more important.

Lemma A10 (Smoothness for $\mathfrak{R}_{m,n}$). *Given observations of*

$$\mathfrak{X}_m := (\mathfrak{X}_{m'}, \mathfrak{X}_{m''}) = \{\mathbf{X}_1, \ldots, \mathbf{X}_{m'}, \mathbf{X}_{m'+1}, \ldots, \mathbf{X}_m\},$$

where $m' + m'' = m$ and $\mathfrak{Y}_n := (\mathfrak{Y}_{n'}, \mathfrak{Y}_{n''}) = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n'}, \mathbf{Y}_{n'+1}, \ldots, \mathbf{Y}_n\}$, where $n' + n'' = n$, denote $\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$ as before, the number of edges of $\mathrm{MST}(\mathfrak{X}_m, \mathfrak{Y}_n)$ which connect a point of $\mathfrak{X}_m$ to a point of $\mathfrak{Y}_n$. Then, for given integer $h \geq 2$, for all $(\mathfrak{X}_n, \mathfrak{Y}_m) \in [0,1]^d$, $\epsilon \geq h^2 \, \delta_{m,n}^h$ where $\delta_{m,n}^h = O\big(h^{d-1}(m+n)^{1/d} \big)$, we have

$$P\left(\left| \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'}) \right| \leq \tilde{c}_{\epsilon,h} \left(\#\mathfrak{X}_{m''} \, \#\mathfrak{Y}_{n''} \right)^{1-1/d} \right)$$
$$\geq 1 - \frac{2h \, \delta_{m,n}^h}{\epsilon}, \tag{A23}$$

where $\tilde{c}_{\epsilon,h} = O\left(\dfrac{\epsilon}{h^{d-1} - 1} \right)$.

Remark: Using Lemma A10, we can imply the continuity property, i.e., for all observations $(\mathfrak{X}_m, \mathfrak{Y}_n)$ and $(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'})$, with at least probability $2\,g(\epsilon) - 1$, one obtains

$$\left| \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'}) \right|$$
$$\leq c_{\epsilon,h}^* \left(\#(\mathfrak{X}_m \Delta \, \mathfrak{X}_{m'}) \, \#(\mathfrak{Y}_n \Delta \, \mathfrak{Y}_{n'}) \right)^{1-1/d}, \tag{A24}$$

for given $\epsilon > 0$, $c_{\epsilon,h}^* = O\left(\dfrac{\epsilon}{h^{d-1} - 1} \right)$, $h \geq 2$. Here, $\mathfrak{X}_m \Delta \, \mathfrak{X}_{m'}$ denotes symmetric difference of observations $\mathfrak{X}_m$ and $\mathfrak{X}_{m'}$.

The path to approach the assertions (11) and (12) proceeds via semi-isoperimentic inequality for the $\mathfrak{R}_{m,n}$ involving the Hamming distance.

Lemma A11 (Semi-Isoperimetry). *Let μ be a measure on $[0,1]^d$; μ^n denotes the product measure on space $\left([0,1]^d \right)^n$. In addition, let M_e denotes a median of $\mathfrak{R}_{m,n}$. Set*

$$\mathbb{A} := \left\{ \mathfrak{X}_m \in \left([0,1]^d \right)^m, \mathfrak{Y}_n \in \left([0,1]^d \right)^n; \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq M_e \right\}. \tag{A25}$$

Following the notations in [17], $H(\mathbf{x}, \mathbf{x}') = \#\{i, \mathbf{x}_i \neq \mathbf{x}'_i\}$ and $\phi_{\mathbb{A}}(\mathbf{x}') + \phi_{\mathbb{A}}(\mathbf{y}') = \min\{H(\mathbf{x}, \mathbf{x}') + H(\mathbf{y}, \mathbf{y}') : \mathbf{x}, \mathbf{y} \in \mathbb{A}\}$ and $\phi_{\mathbb{A}}(\mathbf{x}')\,\phi_{\mathbb{A}}(\mathbf{y}') = \min\{H(\mathbf{x}, \mathbf{x}')\,H(\mathbf{y}, \mathbf{y}') : \mathbf{x}, \mathbf{y} \in \mathbb{A}\}$. Then,

$$\mu^{m+n}\left(\left\{\mathbf{x}' \in ([0,1]^d)^m, \mathbf{y}' \in ([0,1]^d)^n : \phi_{\mathbb{A}}(\mathbf{x}')\,\phi_{\mathbb{A}}(\mathbf{y}') \geq t\right\}\right)$$

$$\leq 4\exp\left(\frac{-t}{8(m+n)}\right). \tag{A26}$$

Now, we continue by providing the proof of Theorem 5. Recall (A25) and denote

$$\mathbb{F}_{\mathbf{x}} := \left\{\mathbf{x}_i, i = 1, \ldots, m, \mathbf{x}_i = \mathbf{x}'_i\right\},$$

$$\mathbb{F}_{\mathbf{y}} := \left\{\mathbf{y}_j, j = 1, \ldots, n, \mathbf{y}_j = \mathbf{y}'_j\right\},$$

and

$$\mathbb{G}_{\mathbf{x}} := \left\{\mathbf{x}_i, i = 1, \ldots, m, \mathbf{x}_i \neq \mathbf{x}'_i\right\},$$

$$\mathbb{G}_{\mathbf{y}} := \left\{\mathbf{y}_j, j = 1, \ldots, n, \mathbf{y}_j \neq \mathbf{y}'_j\right\}.$$

In addition, for given integer h, define events $\mathbb{B}, \mathbb{B}'$ by

$$\mathbb{B} := \left\{\left|\mathfrak{R}_{m,n}(\mathcal{X}'_m, \mathcal{Y}'_n) - \mathfrak{R}(\mathbb{F}_{\mathbf{x}}, \mathbb{F}_{\mathbf{y}})\right| \leq c_{\epsilon,h}\left(\#\mathbb{G}_{\mathbf{x}}\,\#\mathbb{G}_{\mathbf{y}}\right)^{1-1/d}\right\},$$

$$\mathbb{B}' := \left\{\left|\mathfrak{R}(\mathbb{F}_{\mathbf{x}}, \mathbb{F}_{\mathbf{y}}) - \mathfrak{R}_{m,n}(\mathcal{X}_m, \mathcal{Y}_n)\right| \leq c_{\epsilon,h}\left(\#\mathbb{G}_{\mathbf{x}}\,\#\mathbb{G}_{\mathbf{y}}\right)^{1-1/d}\right\},$$

where $c_{\epsilon,h}$ is a constant. By virtue of smoothness property, Lemma A10, for $\epsilon \geq h^2 \delta^h_{m,n}$, we know $P(\mathbb{B}) \geq 2g(\epsilon) - 1$ and $P(\mathbb{B}') \geq 2g(\epsilon) - 1$. On the other hand, we have

$$\mathfrak{R}_{m,n}(\mathcal{X}'_m, \mathcal{Y}'_n) \leq \left|\mathfrak{R}_{m,n}(\mathcal{X}'_m, \mathcal{Y}'_n) - \mathfrak{R}(\mathbb{F}_{\mathbf{x}}, \mathbb{F}_{\mathbf{y}})\right|$$

$$+ \left|\mathfrak{R}(\mathbb{F}_{\mathbf{x}}, \mathbb{F}_{\mathbf{y}}) - \mathfrak{R}_{m,n}(\mathcal{X}_m, \mathcal{Y}_n)\right| + \mathfrak{R}_{m,n}(\mathcal{X}_m, \mathcal{Y}_n).$$

$$= |\varpi'| + |\varpi| + \mathfrak{R}_{m,n}(\mathcal{X}_m, \mathcal{Y}_n) \quad \text{(say)}.$$

Moreover, $P(\mathfrak{R}_{m,n}(\mathcal{X}_m, \mathcal{Y}_n) \leq M_{\mathrm{e}}) \geq 1/2$. Therefore, we can write

$$1/2 \leq P\left(\mathfrak{R}_{m,n}(\mathcal{X}'_m, \mathcal{Y}'_n) \leq M_{\mathrm{e}} + |\varpi'| + |\varpi|\right)$$

$$\leq P\left(\mathfrak{R}_{m,n}(\mathcal{X}'_m, \mathcal{Y}'_n) \leq M_{\mathrm{e}} + |\varpi'| + |\varpi|\,\big|\,\mathbb{B} \cap \mathbb{B}'\right) P(\mathbb{B} \cap \mathbb{B}') \tag{A27}$$

$$+ P(\mathbb{B}^{\mathrm{c}} \cup \mathbb{B}'^{\mathrm{c}}).$$

Thus, we obtain

$$P\left(\mathfrak{R}_{m,n}(\mathcal{X}'_m, \mathcal{Y}'_n) \leq M_{\mathrm{e}} + 4\epsilon\,\left(\#\mathbb{G}_{\mathbf{x}}\,\#\mathbb{G}_{\mathbf{y}}\right)^{1-1/d}\right)$$

$$\geq \left(1/2 - 1 + P(\mathbb{B} \cap \mathbb{B}')\right)/P(\mathbb{B} \cap \mathbb{B}')$$

$$= 1 - \left(\left(2\,P(\mathbb{B} \cap \mathbb{B}')\right)^{-1}\right).$$

Note that $P(\mathbb{B} \cap \mathbb{B}') = P(\mathbb{B})\,P(\mathbb{B}') \geq \left(2\,g(\epsilon) - 1\right)^2$. Now, we easily claim that

$$1 - \left(\left(2\,P(\mathbb{B} \cap \mathbb{B}')\right)^{-1}\right) \geq 1 - \left(\left(2\,(2\,g(\epsilon) - 1)^2\right)^{-1}\right). \tag{A28}$$

Thus,

$$P\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \leq M_e + 4\epsilon \left(\#\mathbb{G}_\mathbf{x} \,\#\mathbb{G}_\mathbf{y}\right)^{1-1/d}\right) \geq 1 - \left(\left(2\,(2\,g(\epsilon)-1)^2\right)^{-1}\right).$$

On the other word, calling $\phi_\mathbb{A}(\mathbf{x}')$ and $\phi_\mathbb{A}(\mathbf{y}')$ in Lemma A11, we get

$$P\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \leq M_e + 4\epsilon \left(\phi_\mathbb{A}(\mathbf{x}')\,\phi_\mathbb{A}(\mathbf{y}')\right)^{1-1/d}\right) \geq 1 - \left(\left(2\,(2\,g(\epsilon)-1)^2\right)^{-1}\right). \tag{A29}$$

Furthermore, denote event

$$\mathbb{C} := \left\{\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \leq M_e + 4\epsilon \left(\phi_\mathbb{A}(\mathbf{x}')\,\phi_\mathbb{A}(\mathbf{y}')\right)^{1-1/d}\right\}.$$

Then, we have

$$P\left(\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \geq M_e + t\right) = \mu^{m+n}\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\right)$$

$$= \mu^{m+n}\left(\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\right)\big|\mathbb{C}\right)P(\mathbb{C})$$

$$+ \mu^{m+n}\left(\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\right)\big|\mathbb{C}^\mathsf{c}\right)P(\mathbb{C}^\mathsf{c})$$

$$\leq \mu^{m+n}\left(\left(\phi_\mathbb{A}(\mathbf{x}')\,\phi_\mathbb{A}(\mathbf{y}')\right)^{1-1/d} \geq \frac{t}{4\epsilon}\right)P(\mathbb{C})$$

$$+ \mu^{m+n}\left(\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\right)\big|\mathbb{C}^\mathsf{c}\right)P(\mathbb{C}^\mathsf{c}). \tag{A30}$$

Using $P(\mathbb{C}) = 1 - P(\mathbb{C}^\mathsf{c})$

$$= \mu^{m+n}\left(\left(\phi_\mathbb{A}(\mathbf{x}')\,\phi_\mathbb{A}(\mathbf{y}')\right)^{1-1/d} \geq \frac{t}{4\epsilon}\right)$$

$$+ P(\mathbb{C}^\mathsf{c})\left\{\mu^{m+n}\left(\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\right)\big|\mathbb{C}^\mathsf{c}\right)\right.$$

$$\left. - \mu^{m+n}\left(\left(\phi_\mathbb{A}(\mathbf{x}')\,\phi_\mathbb{A}(\mathbf{y}')\right)^{1-1/d} \geq \frac{t}{4\epsilon}\right)\right\}.$$

Define set $\mathbb{K}_t = \left\{\left(\phi_\mathbb{A}(\mathbf{x}')\,\phi_\mathbb{A}(\mathbf{y}')\right)^{1-1/d} \geq \frac{t}{4\epsilon}\right\}$, so

$$\mu^{m+n}\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\big|\mathbb{C}^\mathsf{c}\right)$$

$$= \mu^{m+n}\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\big|\mathbb{C}^\mathsf{c}, \mathbb{K}_t\right)\mu^{m+n}(\mathbb{K}_t) + \mu^{m+n}\left(\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\right)\big|\mathbb{C}^\mathsf{c}, \mathbb{K}_t^\mathsf{c}\right)\mu^{m+n}(\mathbb{K}_t^\mathsf{c}).$$

Since

$$\mu^{m+n}\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\big|\mathbb{C}^\mathsf{c}, \mathbb{K}_t\right) = 1,$$

and

$$\mu^{m+n}\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\big|\mathbb{C}^\mathsf{c}, \mathbb{K}_t^\mathsf{c}\right) = \mu^{m+n}\left(\mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n) \geq M_e + t\right).$$

Consequently, from (A30), one can write

$$
P\big(\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n) \geq M_e + t\big)
$$

$$
\leq \mu^{m+n}\Big(\big(\phi_{\mathbb{A}}(\mathbf{x}')\,\phi_{\mathbb{A}}(\mathbf{y}')\big)^{1-1/d} \geq \frac{t}{4\epsilon}\Big)
$$

$$
+ P(\mathbb{C}^c)\Big\{\mu^{m+n}\big(\mathfrak{R}_{m,n}(\mathfrak{X}'_m,\mathfrak{Y}'_n) \geq M_e + t\big)\mu^{m+n}(\mathbb{K}^c_t)\Big\} \tag{A31}
$$

$$
\leq \mu^{m+n}\Big(\big(\phi_{\mathbb{A}}(\mathbf{x}')\,\phi_{\mathbb{A}}(\mathbf{y}')\big)^{1-1/d} \geq \frac{t}{4\epsilon}\Big)
$$

$$
+ \Big(\big(2\,(2\,g(\epsilon)-1)^2\big)^{-1}\Big)P\big(\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n) \geq M_e + t\big).
$$

The last inequality implies by owing to (A29) and $\mu^{m+n}(\mathbb{K}^c_t) \leq 1$. For $g(\epsilon) \geq 1/2 + 1/(2\sqrt{2})$, we have

$$
1 - \Big(\big(2\,(2\,g(\epsilon)-1)^2\big)^{-1}\Big) \geq 0,
$$

or equivalently this holds true when $\epsilon \geq (2h\sqrt{2}\,\delta^h_{m,n})/(\sqrt{2}-1)$. Furthermore, for $h \geq 7$, we have

$$
h^2\delta^h_{m,n} \geq (2h\sqrt{2}\,\delta^h_{m,n})/(\sqrt{2}-1), \tag{A32}
$$

therefore $P\big(\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n) \geq M_e + t\big)$ is less than and equal to

$$
\Big(1 - \big((2\,(2\,g(\epsilon)-1)^2)^{-1}\big)\Big)^{-1} \mu^{m+n}\Big(\big(\phi_{\mathbb{A}}(\mathbf{x}')\,\phi_{\mathbb{A}}(\mathbf{y}')\big)^{1-1/d} \geq \frac{t}{4\epsilon}\Big). \tag{A33}
$$

By virtue of Lemma A11, finally we obtain

$$
P\big(\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n) \geq M_e + t\big) \leq 4\left(1 - \big((2\,(2\,g(\epsilon)-1)^2)^{-1}\big)\right)^{-1}\exp\Big(\frac{-t^{d/(d-1)}}{8(4\epsilon)^{d/d-1}(m+n)}\Big). \tag{A34}
$$

Similarly, we can derive the same bound on $P\big(\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n) \leq M_e - t\big)$, so we obtain

$$
P\Big(\big|\mathfrak{R}_{m,n} - M_e\big| \geq t\Big) \leq C'_{m,n}(\epsilon,h)\exp\Big(\frac{-t^{d/(d-1)}}{8(4\epsilon)^{d/(d-1)}(m+n)}\Big), \tag{A35}
$$

where

$$
C'_{m,n}(\epsilon,h) = 8\left(1 - 2^{-1}\Big(1 - \frac{2h\,O\big(h^{d-1}(m+n)^{1/d}\big)}{\epsilon}\Big)^{-2}\right)^{-1}. \tag{A36}
$$

We will analyze (A35) together with Theorem 6. The next lemma will be employed in Theorem 6's proof.

Lemma A12 (Deviation of the Mean and Median). *Consider M_e as a median of $\mathfrak{R}_{m,n}$. Then, for $\epsilon \geq h^2\delta^h_{m,n}$ and given $h \geq 7$, we have*

$$
\Big|\mathbb{E}\big[\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n)\big] - M_e\Big| \leq C_{m,n}(\epsilon,h)\,(m+n)^{(d-1)/d}, \tag{A37}
$$

where $C_{m,n}(\epsilon,h)$ is a constant depending on ϵ, h, m, and n by

$$
C_{m,n}(\epsilon,h) = C\left(1 - \big((2\,(2\,g(\epsilon)-1)^2)^{-1}\big)\right)^{-1}, \tag{A38}
$$

where C is a constant and

$$\delta_{m,n}^h = O\big(h^{d-1}(m+n)^{1/d}\big), \quad \text{and} \quad g(\epsilon) = 1 - \frac{h\,\delta_{m,n}^h}{\epsilon}.$$

We conclude this part by pursuing our primary intension which has been the Theorem 6's proof. Observe from Theorem 5, (11) that

$$P\Big(\big|\Re_{m,n} - \mathbb{E}[\Re_{m,n}]\big| \geq t + C_{m,n}(\epsilon,l)\,(m+n)^{(d-1)/d}\Big)$$

$$\leq P\Big(\big|\Re_{m,n} - M_e\big| + \big|\mathbb{E}[\Re_{m,n}] - M_e\big|$$

$$\geq t + C_{m,n}(\epsilon,l)\,(m+n)^{(d-1)/d}\Big)$$

$$\leq P\Big(\big|\Re_{m,n} - M_e\big| \geq t\Big)$$

$$\leq 8\left(1 - \Big((2\,(2\,g(\epsilon)-1)^2)^{-1}\Big)\right)^{-1} \exp\left(\frac{-t^{d/(d-1)}}{8(4\epsilon)^{d/d-1}(m+n)}\right).$$

Note that the last bound is derived by (11). The rest of the proof is as the following: When $t \geq 2C_{m,n}(\epsilon,h)\,(m+n)^{(d-1)/d}$, we use

$$\Big(t - C_{m,n}(\epsilon,h)\,(m+n)^{(d-1)/d}\Big)^{d/(d-1)} \geq \big(t/2\big)^{d/(d-1)}.$$

Therefore, it turns out that

$$P\Big(\big|\Re_{m,n} - \mathbb{E}[\Re_{m,n}]\big| \geq t\Big)$$

$$\leq 8\left(1 - \Big((2\,(2\,g(\epsilon)-1)^2)^{-1}\Big)\right)^{-1} \exp\left(\frac{-t^{d/(d-1)}}{8(8\epsilon)^{d/(d-1)}(m+n)}\right). \tag{A39}$$

In other words, there exist constants $C_{m,n}'(\epsilon,h)$ depending on m, n, ϵ, and h such that

$$P\Big(\big|\Re_{m,n} - \mathbb{E}[\Re_{m,n}]\big| \geq t\Big) \leq C_{m,n}'(\epsilon,h)\exp\left(\frac{-(t/(2\epsilon))^{d/(d-1)}}{(m+n)\,\tilde{C}}\right), \tag{A40}$$

where $\tilde{C} = 8(4)^{d/(d-1)}$.

To verify the behavior of bound (A40) in terms of ϵ, observe (A35) first; it is not hard to see that this function is decreasing in ϵ. However, the function

$$\exp\left(\frac{-(t/(2\epsilon))^{d/(d-1)}}{(m+n)\tilde{C}}\right)$$

increases in ϵ. Therefore, one can not immediately infer that the bound in (12) is monotonic with respect to ϵ. For fixed $N = n + m$, d, and h, the first and second derivatives of the bound (12) with respect to ϵ are quite complicated functions. Thus, deriving an explicit optimal solution for the minimization problem with the objective function (12) is not feasible. However, in the sequel, we discuss that under conditions when t is not much larger than $N = m + n$, this bound becomes convex with respect to ϵ. Set

$$K(\epsilon) = C_{m,n}'(\epsilon,h)\,\exp\left(\frac{-B(t)}{\epsilon^{d/(d-1)}}\right), \tag{A41}$$

where $C'_{m,n}$ is given in (10) and

$$B(t) = \frac{t^{d/(d-1)}}{8 \, (8)^{d/(d-1)}(N)}.$$

By taking the derivative with respect to ϵ, we have

$$\frac{dK(\epsilon)}{d\epsilon} = K(\epsilon)\left(\frac{d}{d\epsilon}\left(\log C'_{m,n}\right) + \frac{B(t)\, d/(d-1)}{\epsilon^{(2d-1)/(d-1)}}\right), \tag{A42}$$

where

$$\frac{d}{d\epsilon}\left(\log C'_{m,n}\right) = \frac{-4\, a_h\, \epsilon}{(\epsilon - 2a_h)(8a_h^2 - 8\epsilon a_h + \epsilon^2)}, \tag{A43}$$

where $a_h = h\delta_{m,n}^h$. The second derivative $K(\epsilon)$ with respect to ϵ after simplification is given as

$$\frac{d^2}{d\epsilon^2}K(\epsilon) = \left(\frac{-4\, a_h\, \epsilon}{(\epsilon - 2a_h)(8a_h^2 - 8\epsilon a_h + \epsilon^2)} + \frac{B(t)\, \bar{d}}{\epsilon^{\bar{d}+1}}\right)^2$$

$$+ K(\epsilon)\left(\frac{8a_h\,(8a_h^3 + \epsilon^2(\epsilon - 5a_h))}{(8a_h^2 - 8a_h\epsilon + \epsilon^2)^2(\epsilon - 2a_h)^2} - \frac{B(t)\bar{d}(\bar{d}+1)}{\epsilon^{\bar{d}+2}}\right), \tag{A44}$$

where $\bar{d} = d/(d-1)$. The first term in (A44) and $K(\epsilon)$ are non-negative, so $K(\epsilon)$ is convex if the second term in the second line of (A44) is non-negative. We know that $\epsilon \geq h^2\delta_{m,n}^h = h\, a_h$, when $h = 7$, we can parameterize ϵ by setting it equal to γa_h, where $\gamma \geq 7$. After simplification, $K(\epsilon)$ is convex if

$$a_h^{\bar{d}-1}\left(\gamma^{\bar{d}-1} + 3\gamma^{\bar{d}-2}\right) + B(t)\bar{d}(\bar{d}+1)$$

$$\times\left\{a_h^{-1}\left(-32\gamma^{-6} + 64\gamma^{-5} - 48\gamma^{-4} + 8\gamma^{-3} - \frac{7}{2}\gamma^{-2} + 2\gamma^{-1} - \frac{1}{8}\right)\right.$$

$$\left.+ a_h^{-2}\left(32\gamma^{-6} - 64\gamma^{-5} + 40\gamma^{-4} + 8\gamma^{-3} + \frac{1}{2}\gamma^{-2}\right)\right\} \geq 0. \tag{A45}$$

This is implied if

$$0 \leq B(t)\bar{d}(\bar{d}+1)\, a_h^{-1}$$

$$\times\left(-32\gamma^{-6} + 64\gamma^{-5} - 48\gamma^{-4} + 8\gamma^{-3} - \frac{7}{2}\gamma^{-2} + 2\gamma^{-1} - \frac{1}{8}\right), \tag{A46}$$

such that $\gamma \geq 7$. One can easily check that, as $\gamma \to \infty$, then (A46) tends to $-\frac{1}{8}B(t)\bar{d}(\bar{d}+1)\, a_h^{-1}$. This term can be negligible unless we have t that is much larger than $N = m + n$ with the threshold depending on d. Here, by setting $B(t)/a_h = 1$, a rough threshold $t = O\left(7^{d-1}(m+n)^{1-1/d^2}\right)$ depending on d, $m + n$ is proposed. Therefore, minimizing (A35) and (A40) with respect to ϵ when optimal $h = 7$ is a convex optimization problem. Denote ϵ^* the solution of the convex optimization problem (9). By plugging optimal h ($h = 7$) and ϵ ($\epsilon = \epsilon^*$) in (A35) and (A40), we derive (11) and (12), respectively.

In this appendix, we also analyze the bound numerically. By simulation, we observed that lower h i.e., $h = 7$ is the optimal value experimentally. Indeed, this can be verified by Theorem 11's proof. We address the reader to Lemma A8 in Appendix D and Appendix E where, as h increases, the lower bound for the probability increases too. In other words, for fixed $N = m + n$ and d, the lowest h implies the maximum bound in (A92). For this, we set $h = 7$ in our experiments. We vary the dimension d and sample size $N = m + n$ in relatively large and small ranges. In Table A1, we solve (9) for various values of d and $N = m + n$. We also compute the lower bound for ϵ i.e., $7^{d+1}N^{1/d}$ per experiment. In Table A1, we observe that as we have higher dimension the optimal value ϵ^* equals the ϵ lower bound $h^{d+1}N^{1/d}$, but this is not true for smaller dimensions with even relatively large sample size.

Table A1. d, N, ϵ^* are dimension, total sample size $m + n$, and optimal ϵ for the bound in (12). The column $h^{d+1}N^{1/d}$ represents approximately the lower bound for ϵ which is our constraint in the minimization problem and our assumption in Theorems 5 and 6. Here, we set $h = 7$.

			Concentration Bound (11)		
d	$N = m + n$	ϵ^*	t_0	$h^{d+1}N^{1/d}$	Optimal (11)
2	10^3	1.1424×10^4	2×10^7	1.0847×10^4	0.3439
4	10^4	1.7746×10^5	3×10^{10}	168,070	0.0895
5	550	4.7236×10^5	10^{10}	4.1559×10^5	0.9929
6	10^4	3.8727×10^6	2×10^{12}	3.8225×10^6	0.1637
8	1200	9.7899×10^7	12×10^{12}	9.7899×10^7	0.7176
10	3500	4.4718×10^9	2×10^{15}	4.4718×10^9	0.4795
15	10^8	1.1348×10^{14}	10^{24}	1.1348×10^{14}	0.9042

To validate our proposed bound in (12), we again set $h = 7$ and for $d = 4, 5, 7$ we ran experiments with sample sizes $N = m + n = 9000, 1100, 140$, respectively. Then, we solved the minimization problem to derive optimal bound for t in the range $10^{10}[1, 3]$. Note that we chose this range to have a non-trivial bound for all three curves; otherwise, the bounds partly become one. Figure A1 shows that when t increases in the given range, the optimal curves approach zero.
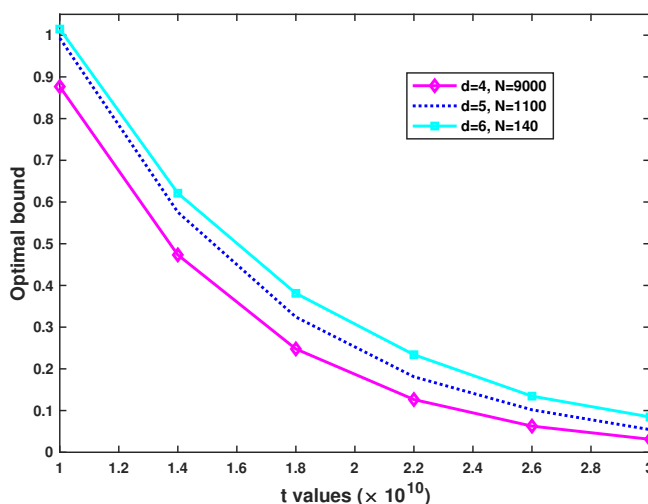

Figure A1. Optimal bound for (12), when $h = 7$ versus $t \in 10^{10}[1, 3]$. The bound decreases as t grows.

To prove the Theorem 7 in the concentration of $\mathfrak{R}_{m,n}$, Theorem 6, let

$$\delta = C'_{m,n}(\epsilon^*) \exp\left(\frac{-(t/(2\epsilon^*))^{d/(d-1)}}{(m + n)\,\tilde{C}}\right),$$

this implies $t = O\left(\epsilon^*\,(m + n)^{(d-1)/d}\left(\log\left(C'_{m,n}(\epsilon^*)/\delta\right)\right)^{(d-1)/d}\right)$ and the proofs are completed.

Appendix E. Additional Proofs

Lemma A3: Let $g(\mathbf{x})$ be a density function with support $[0, 1]^d$ and belong to the Hölder class $\Sigma_d(\eta, L)$, $0 < \eta \leq 1$, expressed in Definition 1. In addition, assume that $P(\mathbf{x})$ is a η-Hölder smooth function, such that its absolute value is bounded from above by some constants c. Define the quantized density function with parameter l and constants ϕ_i as

$$\widehat{g}(\mathbf{x}) = \sum_{i=1}^{M} \phi_i \mathbf{1}\{\mathbf{x} \in Q_i\}, \quad \text{where } \phi_i = l^d \int_{Q_i} g(\mathbf{x}) \, d\mathbf{x}, \tag{A47}$$

and $M = l^d$ and $Q_i = \{\mathbf{x}, \mathbf{x}_i : \|\mathbf{x} - \mathbf{x}_i\| < l^{-d}\}$. Then,

$$\int \left\| (g(\mathbf{x}) - \widehat{g}(\mathbf{x})) P(\mathbf{x}) \right\| \, d\mathbf{x} \le O(l^{-d\eta}). \tag{A48}$$

Proof. By the mean value theorem, there exist points $\epsilon_i \in Q_i$ such that

$$\phi_i = l^d \int_{Q_i} g(\mathbf{x}) \, d\mathbf{x} = g(\epsilon_i).$$

Using the fact that $g \in \Sigma_d(\eta, L)$ and $P(\mathbf{x})$ is a bounded function, we have

$$\begin{aligned}
\int \left\| g(\mathbf{x}) - \widehat{g}(\mathbf{x}) \right) P(\mathbf{x}) \right\| \, d\mathbf{x} &= \sum_{i=1}^{M} \int_{Q_i} \left\| (g(\mathbf{x}) - \Phi_i) P(\mathbf{x}) \right\| \, d\mathbf{x} \\
&= \sum_{i=1}^{M} \int_{Q_i} \left\| (g(\mathbf{x}) - g(\epsilon_i)) P(\mathbf{x}) \right\| \, d\mathbf{x} \\
&\le c L \sum_{i=1}^{M} \int_{Q_i} \left\| \mathbf{x} - \epsilon_i \right\|^{\eta} \, d\mathbf{x}.
\end{aligned}$$

Here, L is the Hölder constant. As $\mathbf{x}, \epsilon_i \in Q_i$, a sub-cube with edge length l^{-1}, then $\left\| \mathbf{x} - \epsilon_i \right\|^{\eta} = O(l^{-d\eta})$ and $\sum_{i=1}^{M} \int_{Q_i} d\mathbf{x} = 1$. This concludes the proof. $\square$

    Lemma A4: Let $\Delta(\mathbf{x}, \mathcal{S})$ denote the degree of vertex $\mathbf{x} \in \mathcal{S}$ in the *MST* over set $\mathcal{S} \subset \mathbb{R}^d$ with the n number of vertices. For given function $P(\mathbf{x}, \mathbf{x})$, one yields

$$\int P(\mathbf{x}, \mathbf{x}) g(\mathbf{x}) \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S})] \, d\mathbf{x} = 2 \int P(\mathbf{x}, \mathbf{x}) g(\mathbf{x}) \, d\mathbf{x} + \varsigma_\eta(l, n), \tag{A49}$$

where for constant $\eta > 0$,

$$\varsigma_\eta(l, n) = \left(O(l/n) - 2 l^d/n \right) \int g(\mathbf{x}) P(\mathbf{x}, \mathbf{x}) \, d\mathbf{x} + O(l^{-d\eta}). \tag{A50}$$

Proof. Recall notations in Lemma A3 and

$$\left| \int g(\mathbf{x}) P(\mathbf{x}) \, d\mathbf{x} - \int \widehat{g}(\mathbf{x}) P(\mathbf{x}) \, d\mathbf{x} \right| \le \int \left| (g(\mathbf{x}) - \widehat{g}(\mathbf{x})) P(\mathbf{x}) \right| \, d\mathbf{x}.$$

Therefore, by substituting $\widehat{g}$, defined in (A47), into g with considering its error, we have

$$\begin{aligned}
\int P&(\mathbf{x}, \mathbf{x}) g(\mathbf{x}) \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S})] \, d\mathbf{x} \\
&= \int P(\mathbf{x}, \mathbf{x}) \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S})] \sum_{i=1}^{M} \phi_i \mathbf{1}\{\mathbf{x} \in Q_i\} \, d\mathbf{x} + O(l^{-d\eta}) \\
&= \sum_{i=1}^{M} \phi_i \int_{Q_i} P(\mathbf{x}, \mathbf{x}) \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S})] \, d\mathbf{x} + O(l^{-d\eta}).
\end{aligned} \tag{A51}$$

Here, Q_i represents as before in Lemma A3, so the RHS of (A51) becomes

$$
\begin{aligned}
&\sum_{i=1}^{M} \phi_i \int_{Q_i} P(\mathbf{x}, \mathbf{x}) \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S} \cap Q_i)] \, d\mathbf{x} + \sum_{i=1}^{M} \phi_i \int_{Q_i} P(\mathbf{x}, \mathbf{x}) O(l^{1-d}/n) + O(l^{-d\eta}) \\
&= \sum_{i=1}^{M} \phi_i P(\mathbf{x}_i, \mathbf{x}_i) \frac{1}{M} \int_{Q_i} M \, \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S} \cap Q_i)] \, d\mathbf{x} + \sum_{i=1}^{M} \phi_i \int_{Q_i} P(\mathbf{x}, \mathbf{x}) O(l^{1-d}/n) + 2 \, O(l^{-d\eta}).
\end{aligned}
\tag{A52}
$$

Now, note that $\int_{Q_i} M \, \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S} \cap Q_i)] \, d\mathbf{x}$ is the expectation of $\mathbb{E}[\Delta(\mathbf{x}, S \cap Q_i)]$ over the nodes in Q_i, which is equal to $2 - \dfrac{2}{k_i}$, where $k_i = \dfrac{n}{M}$. Consequently, we have

$$
\begin{aligned}
\int P(\mathbf{x}, \mathbf{x}) g(\mathbf{x}) \mathbb{E}[\Delta(\mathbf{x}, \mathcal{S})] \, d\mathbf{x} &= \left(2 - \frac{2M}{n}\right) \sum_{i=1}^{M} \phi_i \, P(\mathbf{x}_i, \mathbf{x}_i) \frac{1}{M} + O\left(\frac{l^{1-d}}{n}\right) \sum_{i=1}^{M} \phi_i \, P(\mathbf{x}_i, \mathbf{x}_i) + 3 \, O(l^{-d\eta}) \\
&= 2 \int g(\mathbf{x}) P(\mathbf{x}, \mathbf{x}) \, d\mathbf{x} + 5 \, O(l^{-d\eta}) + M \left(O\left(\frac{l^{1-d}}{n}\right) - \left(\frac{2}{n}\right) \right) \int g(\mathbf{x}) P(\mathbf{x}, \mathbf{x}) \, d\mathbf{x}.
\end{aligned}
\tag{A53}
$$

This gives the assertion (A49). □

Lemma A5: Assume that, for given k, $g_k(\mathbf{x})$ is a bounded function belong to $\Sigma_d(\eta, L)$. Let $P : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0, 1]$ be a symmetric, smooth, jointly measurable function, such that, given k, for almost every $\mathbf{x} \in \mathbb{R}^d$, $P(\mathbf{x}, .)$ is measurable with $\mathbf{x}$ a Lebesgue point of the function $g_k(.)P(\mathbf{x}, .)$. Assume that the first derivative P is bounded. For each k, let $\mathbf{Z}_1^k, \mathbf{Z}_2^k, \ldots, \mathbf{Z}_k^k$ be independent d-dimensional variable with common density function g_k. Set $\mathfrak{Z}_k = \{\mathbf{Z}_1^k, \mathbf{Z}_2^k \ldots, \mathbf{Z}_k^k\}$ and $\mathfrak{Z}_k^{\mathbf{x}} = \{\mathbf{x}, \mathbf{Z}_2^k, \mathbf{Z}_3^k \ldots, \mathbf{Z}_k^k\}$. Then,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{j=2}^{k} P(\mathbf{x}, \mathbf{Z}_j^k) \mathbf{1}\{(\mathbf{x}, \mathbf{Z}_j^k) \in MST(\mathfrak{Z}_k^{\mathbf{x}})\} \right] \\
= P(\mathbf{x}, \mathbf{x}) \, \mathbb{E}\left[\Delta(\mathbf{x}, \mathfrak{Z}_k^{\mathbf{x}})\right] + \left\{ O(k^{-\eta/d}) + O(k^{-1/d}) \right\}.
\end{aligned}
\tag{A54}
$$

Proof. Let $\mathbb{B}(\mathbf{x}, r) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\|_d \leq r\}$. For any positive K, we can obtain:

$$
\begin{aligned}
&\mathbb{E} \sum_{j=2}^{k} \left| P(\mathbf{x}, \mathbf{Z}_j^k) - P(\mathbf{x}, \mathbf{x}) \right| \mathbf{1}\{\mathbf{Z}_j^k \in \mathbb{B}(\mathbf{x}, Kk^{-1/d})\} \\
&= (k-1) \int_{\mathbb{B}(\mathbf{x}; Kk^{-1/d})} \left| (P(\mathbf{x}, \mathbf{y}) g_k(\mathbf{y}) - P(\mathbf{x}, \mathbf{x}) g_k(\mathbf{x})) + P(\mathbf{x}, \mathbf{x})(g_k(\mathbf{x}) - g_k(\mathbf{y})) \right| \, d\mathbf{y} \\
&\leq (k-1) \left[\int_{\mathbb{B}(\mathbf{x}; Kk^{-1/d})} \left| (P(\mathbf{x}, \mathbf{y}) g_k(\mathbf{y}) - P(\mathbf{x}, \mathbf{x}) g_k(\mathbf{x})) \right| d\mathbf{y} + O(k^{-\eta/d}) \mathbf{V}(\mathbb{B}(\mathbf{x}, Kk^{-1/d})) \right],
\end{aligned}
\tag{A55}
$$

where $\mathbf{V}$ is the volume of space $\mathbb{B}$ which equals $O(k^{-1})$. Note that the above inequality appears because $g_k(\mathbf{x}) \in \Sigma_d(\eta, L)$ and $P(\mathbf{x}, \mathbf{x}) \in [0, 1]$. The first order Taylor series expansion of $P(\mathbf{x}, \mathbf{y})$ around $\mathbf{x}$ is

$$
\begin{aligned}
P(\mathbf{x}, \mathbf{y}) &= P(\mathbf{x}, \mathbf{x}) + P^{(1)}(\mathbf{x}, \mathbf{x}) \|\mathbf{y} - \mathbf{x}\| + o(\|\mathbf{y} - \mathbf{x}\|^2) \\
&= P(\mathbf{x}, \mathbf{x}) + O(k^{-1/d}) + o(k^{-2/d}).
\end{aligned}
$$

Then, by recalling the Hölder class, we have

$$
\begin{aligned}
\left| P(\mathbf{x}, \mathbf{y}) g_k(\mathbf{y}) - P(\mathbf{x}, \mathbf{x}) g_k(\mathbf{x}) \right| &= \left| (P(\mathbf{x}, \mathbf{x}) + O(k^{-1/d}))(g_k(\mathbf{x}) + O(k^{-\eta/d})) - P(\mathbf{x}, \mathbf{x}) g_k(\mathbf{x}) \right| \\
&= O(k^{-\eta/d}) + O(k^{-1/d}).
\end{aligned}
$$

Hence, the RHS of (A55) becomes

$$(k-1)\Big[\big(O(k^{-\eta/d})+O(k^{-1/d})\big)\mathbf{V}\big(\mathbb{B}(\mathbf{x},Kk^{-1/d})\big)+O(k^{-\eta/d})\mathbf{V}\big(\mathbb{B}(\mathbf{x},Kk^{-1/d})\big)\Big]$$
$$=(k-1)\Big[O(k^{-1-\eta/d})+O(k^{-1-1/d})\Big].$$

The expression in (A54) can be obtained by choice of K. $\square$

Lemma A6: Consider the notations and assumptions in Lemma A5. Then,

$$\Big|k^{-1}\sum_{1\leq i<j\leq k}\sum P(\mathbf{Z}_i^k,\mathbf{Z}_j^k)\mathbf{1}\{(\mathbf{Z}_i^k,\mathbf{Z}_j^k)\in MST(\mathfrak{Z}_k)-\int_{\mathbb{R}^d}P(\mathbf{x},\mathbf{x})g_k(\mathbf{x})\,\mathrm{d}\mathbf{x}\Big|$$
$$\leq \varsigma_\eta(l,k)+O(k^{-\eta/d})+O(k^{-1/d}). \tag{A56}$$

Here, $MST(\mathcal{S})$ denotes the MST graph over nice and finite set $\mathcal{S}\subset\mathbb{R}^d$ and η is the smoothness Hölder parameter. Note that $\varsigma_\eta(l,k)$ is given as before in (A50).

Proof. Following notations in [49], let $\Delta(\mathbf{x},\mathcal{S})$ denote the degree of vertex $\mathbf{x}$ in the $MST(\mathcal{S})$ graph. Moreover, let $\mathbf{x}$ be a Lebesgue point of g_k with $g_k(\mathbf{x})>0$. In addition, let $\mathfrak{Z}_k^{\mathbf{x}}$ be the point process $\{\mathbf{x},\mathbf{Z}_2^k,\mathbf{Z}_3^k,\dots,\mathbf{Z}_k^k\}$. Now, by virtue of (A55) in Lemma A5, we can write

$$\mathbb{E}\Big[\sum_{j=2}^k P(\mathbf{x},\mathbf{Z}_j^k)\mathbf{1}\{(\mathbf{x},\mathbf{Z}_j^k)\in MST(\mathfrak{Z}_k^{\mathbf{x}})\}\Big]=P(\mathbf{x},\mathbf{x})\,\mathbb{E}\big[\Delta(\mathbf{x},\mathfrak{Z}_k^{\mathbf{x}})\big]+\Big\{O(k^{-\eta/d})+O(k^{-1/d})\Big\}. \tag{A57}$$

On the other hand, it can be seen that

$$k^{-1}\mathbb{E}\Big[\sum_{1\leq i<j\leq k}\sum P(\mathbf{Z}_i^k,\mathbf{Z}_j^k)\mathbf{1}\{(\mathbf{Z}_i^k,\mathbf{Z}_j^k)\in MST(\mathfrak{Z}_k)\}\Big]$$
$$=\frac{1}{2}\mathbb{E}\Big[\sum_{j=2}^k P(\mathbf{Z}_1^k,\mathbf{Z}_j^k)\mathbf{1}\{(\mathbf{Z}_i^k,\mathbf{Z}_j^k)\in MST(\mathfrak{Z}_k)\}\Big] \tag{A58}$$
$$=\frac{1}{2}\int g_k(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathbb{E}\Big[\sum_{j=2}^k P(\mathbf{x},\mathbf{Z}_j^k)\mathbf{1}\{(\mathbf{x},\mathbf{Z}_j^k)\in MST(\mathfrak{Z}_k)\}\Big].$$

Recalling (A57),

$$=\frac{1}{2}\int g_k(\mathbf{x})P(\mathbf{x},\mathbf{x})\mathbb{E}\big[\Delta(\mathbf{x},\mathfrak{Z}_k^{\mathbf{x}})\big]\,\mathrm{d}\mathbf{x}+O(k^{-\eta/d})+O(k^{-1/d}). \tag{A59}$$

By virtue of Lemma A4, (A49) can be substituted into expression (A59) to obtain (A56). $\square$

Theorem A1: Assume $\mathfrak{R}_{m,n}:=\mathfrak{R}(\mathfrak{X}_m,\mathfrak{Y}_n)$ denotes the FR test statistic as before. Then, the rate for the bias of the $\mathfrak{R}_{m,n}$ estimator for $0<\eta\leq 1$, $d\geq 2$ is of the form:

$$\Big|\frac{\mathbb{E}[\mathfrak{R}_{m,n}]}{m+n}-2pq\int\frac{f_0(\mathbf{x})f_1(\mathbf{x})}{pf_0(\mathbf{x})+qf_1(\mathbf{x})}\,\mathrm{d}\mathbf{x}\Big|\leq O\big(l^d(m+n)^{-\eta/d}\big)+O(l^{-d\eta}). \tag{A60}$$

Here, η is the Holder smoothness parameter. A more explicit form for the bound on the RHS is given in (A61) below:

$$\left| \frac{\mathbb{E}\left[\mathfrak{R}'_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\right]}{m+n} - \int \frac{2pq f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})} \, d\mathbf{x} \right| \leq O\left(l^d (m+n)^{-\eta/d}\right)$$

$$+ O\left(l^d (m+n)^{-1/2}\right) + 2\, c_1\, l^{d-1} (m+n)^{(1/d)-1} + c_d\, 2^d\, (m+n)^{-1}$$

$$- 2\, l^d (m+n)^{-1} \int \frac{2pq f_0(\mathbf{x}) f_1(\mathbf{x})}{p\, f_0(\mathbf{x}) + q\, f_1(\mathbf{x})} \, d\mathbf{x} + c_2\, (m+n)^{-1} l^d$$

$$+ O(l)(m+n)^{-1} \sum_{i=1}^{M} l^d (a_i)^{-1} \int \frac{2 f_0(\mathbf{x}) f_1(\mathbf{x})}{p\, f_0(\mathbf{x}) + q\, f_1(\mathbf{x})} \, d\mathbf{x} + O(l^{-d\eta}) \tag{A61}$$

$$+ O(l) \sum_{i=1}^{M} l^{d/2} \frac{\sqrt{b_i}}{a_i^2} \int \frac{2 f_0(\mathbf{x}) f_1(\mathbf{x}) \left(f_0(\mathbf{x}) \sqrt{m} + f_1(\mathbf{x}) \sqrt{n} \right)}{\left(m f_0(\mathbf{x}) + n f_1(\mathbf{x}) \right)^2} \, d\mathbf{x}$$

$$+ \sum_{i=1}^{M} 2\, l^{-d/2} \frac{\sqrt{b_i}}{a_i^2} \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x}) \left(\alpha_i \beta_i \left(m a_i f_0^2(\mathbf{x}) + n b_i f_1^2(\mathbf{x}) \right) \right)^{1/2}}{\left(m f_0(\mathbf{x}) + n f_1(\mathbf{x}) \right)^2 (m+n)} \, d\mathbf{x}.$$

Proof. Assume M_m and N_n be Poisson variables with mean m and n, respectively, one independent of another and of $\{\mathbf{X}_i\}$ and $\{\mathbf{Y}_j\}$. Let also $\mathfrak{X}'_m$ and $\mathfrak{Y}'_n$ be the Poisson processes $\{\mathbf{X}_1, \dots, \mathbf{X}_{M_m}\}$ and $\{\mathbf{Y}_1, \dots, \mathbf{Y}_{N_n}\}$. Set $\mathfrak{R}'_{m,n} := \mathfrak{R}_{m,n}(\mathfrak{X}'_m, \mathfrak{Y}'_n)$. Applying Lemma 1, and (12) cf. [49], we can write

$$\left| \mathfrak{R}'_{m,n} - \mathfrak{R}_{m,n} \right| \leq K_d \left(|M_m - m| + |N_n - n| \right). \tag{A62}$$

Here, K_d denotes the largest possible degree of any vertex of the MST graph in $\mathbb{R}^d$. Moreover, by the matter of Poisson variable fact and using Stirling approximation [51], we have

$$\mathbb{E}\left[|M_m - m|\right] = e^{-m} \frac{m^{m+1}}{m!} \leq e^{-m} \frac{m^{m+1}}{\sqrt{2\pi} m^{m+1/2} e^{-m}} = O\left(m^{1/2}\right). \tag{A63}$$

Similarly, $\mathbb{E}\left[|N_n - n|\right] = O(n^{1/2})$. Therefore, by (A62), one yields

$$\mathbb{E}[\mathfrak{R}_{m,n}] = \mathbb{E}\left[\mathfrak{R}_{m,n} - \mathfrak{R}'_{m,n}\right] + \mathbb{E}\left[\mathfrak{R}'_{m,n}\right] = O\left((m+n)^{1/2}\right) + \mathbb{E}\left[\mathfrak{R}'_{m,n}\right]. \tag{A64}$$

Therefore,

$$\frac{\mathbb{E}[\mathfrak{R}_{m,n}]}{m+n} = \frac{\mathbb{E}\left[\mathfrak{R}'_{m,n}\right]}{m+n} + O\left((m+n)^{-1/2}\right). \tag{A65}$$

Hence, it will suffice to obtain the rate of convergence of $\mathbb{E}\left[\mathfrak{R}'_{m,n}\right]/(m+n)$ in the RHS of (A65). For this, let m_i, n_i denote the number of Poisson process samples $\mathfrak{X}'_m$ and $\mathfrak{Y}'_n$ with the FR statistic $\mathfrak{R}'_{m,n}$, falling into partitions Q'_i with FR statistic $\mathfrak{R}'_{m_i,n_i}$. Then, by virtue of Lemma 4, we can write

$$\mathbb{E}\left[\mathfrak{R}'_{m,n}\right] \leq \sum_{i=1}^{M} \mathbb{E}\left[\mathfrak{R}'_{m_i,n_i}\right] + 2\, c_1\, l^{d-1} (m+n)^{1/d}.$$

Note that the Binomial RVs m_i, n_i are independent with marginal distributions $m_i \sim B(m, a_i l^{-d})$, $n_i \sim B(n, b_i l^{-d})$, where a_i, b_i are non-negative constants satisfying, $\forall i$, $a_i \le b_i$ and $\sum_{i=1}^{l^d} a_i l^{-d} = \sum_{i=1}^{l^d} b_i l^{-d} = 1$. Therefore,

$$\mathbb{E}\left[\mathfrak{R}'_{m,n}\right] \le \sum_{i=1}^{M} \mathbb{E}\left[\mathbb{E}\left[\mathfrak{R}'_{m_i,n_i}|m_i,n_i\right]\right] + 2\, c_1\, l^{d-1}(m+n)^{1/d}. \tag{A66}$$

Let us first compute the internal expectation given m_i, n_i. For this reason, given m_i, n_i, let $Z_1^{m_i,n_i}, Z_2^{m_i,n_i}, \ldots$ be independent variables with common densities $g_{m_i,n_i}(\mathbf{x}) = \left(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})\right)/(m_i + n_i)$, $\mathbf{x} \in \mathbb{R}^d$. Moreover, let L_{m_i,n_i} be an independent Poisson variable with mean $m_i + n_i$. Denote $\mathfrak{F}'_{m_i,n_i} = \{Z_1^{m_i,n_i}, \ldots, Z_{L_{m_i,n_i}}^{m_i,n_i}\}$ a non-homogeneous Poisson of rate $m_i f_0 + n_i f_1$. Let $\mathfrak{F}_{m_i,n_i}$ be the non-Poisson point process $\{Z_1^{m_i,n_i}, \ldots Z_{m_i+n_i}^{m_i,n_i}\}$. Assign a mark from the set $\{1,2\}$ to each points of $\mathfrak{F}'_{m_i,n_i}$. Let $\widetilde{\mathfrak{X}}'_{m_i}$ be the sets of points marked 1 with each probability $m_i f_0(\mathbf{x})/\left(m_i f_0(\mathbf{x}) + n_i f_i(\mathbf{x})\right)$ and let $\widetilde{\mathfrak{Y}}'_{n_i}$ be the set points with mark 2. Note that owing to the marking theorem [52], $\widetilde{\mathfrak{X}}'_{m_i}$ and $\widetilde{\mathfrak{Y}}'_{n_i}$ are independent Poisson processes with the same distribution as $\mathfrak{X}'_{m_i}$ and $\mathfrak{Y}'_{n_i}$, respectively. Considering $\widetilde{R}'_{m_i,n_i}$ as FR statistic over nodes in $\widetilde{\mathfrak{X}}'_{m_i} \cup \widetilde{\mathfrak{Y}}'_{n_i}$ we have

$$\mathbb{E}\left[\mathfrak{R}'_{m_i,n_i}|m_i,n_i\right] = \mathbb{E}\left[\widetilde{\mathfrak{R}}'_{m_i,n_i}|m_i,n_i\right].$$

Again using Lemma 1 and analogous arguments in [49] along with the fact that $\mathbb{E}\left[|M_m + N_n - m - n|\right] = O((m+n)^{1/2})$, we have

$$\mathbb{E}\left[\widetilde{\mathfrak{R}}'_{m_i,n_i}|m_i,n_i\right] = \mathbb{E}\left[\mathbb{E}\left[\widetilde{\mathfrak{R}}'_{m_i,n_i}|\mathfrak{F}'_{m_i,n_i}\right]\right]$$

$$= \mathbb{E}\left[\sum_{s<j<m_i+n_i} P_{m_i,n_i}(Z_s^{m_i,n_i}, Z_j^{m_i,n_i})\mathbf{1}\left\{(Z_s^{m_i,n_i}, Z_j^{m_i,n_i}) \in \mathfrak{F}_{m_i,n_i}\right\}\right] + O((m_i + n_i)^{1/2})).$$

Here,

$$P_{m_i,n_i}(\mathbf{x},\mathbf{y}) := P_r\{\text{mark } x \ne \text{mark } y, (\mathbf{x},\mathbf{y}) \in \mathfrak{F}'_{m_i,n_i}\}$$

$$= \frac{m_i f_0(\mathbf{x}) n_i f_1(\mathbf{y}) + n_i f_1(\mathbf{x}) m_i f_0(\mathbf{y})}{(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))(m_i f_0(\mathbf{y}) + n_i f_1(\mathbf{y}))}.$$

By owing to Lemma A6, we obtain

$$\sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\mathbb{E}\left[\sum_{s<j<m_i+n_i} P_{m_i,n_i}(Z_s^{m_i,n_i}, Z_j^{m_i,n_i})\mathbf{1}\left\{(Z_s^{m_i,n_i}, Z_j^{m_i,n_i}) \in \mathfrak{F}_{m_i,n_i}\right\}\right] + \sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\left[O((m_i+n_i))^{1/2}\right]$$

$$= \sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\left[(m_i+n_i)\int g_{m_i,n_i}(\mathbf{x},\mathbf{x})P_{m_i,n_i}(\mathbf{x},\mathbf{x})\,\mathrm{d}\mathbf{x} + \left(\varsigma_\eta(l,m_i,n_i) + O((m_i+n_i)^{-\eta/d})\right.\right. \tag{A67}$$

$$\left.\left.+O((m_i+n_i)^{-1/d}))(m_i+n_i)\right] + \sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\left[O((m_i+n_i)^{1/2})\right],$$

where

$$\varsigma_\eta(l,m_i,n_i) = \left(O(l/(m_i+n_i)) - 2\, l^d/(m_i+n_i)\right)\int g_{m_i,n_i}(\mathbf{x})P_{m_i,n_i}(\mathbf{x},\mathbf{x})\,\mathrm{d}\mathbf{x} + O(l^{-d\eta}).$$

The expression in (A67) equals

$$\sum_{i=1}^{M} \int \mathbb{E}_{m_i, n_i} \left[\frac{2m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})} \right] \, d\mathbf{x} + \sum_{i=1}^{M} \mathbb{E}_{m_i, n_i} \left[(m_i + n_i) \, \varsigma_\eta(l, m_i, n_i) \right]$$

$$+ O\left(l^d (m+n)^{1-\eta/d} \right) + O\left(l^d (m+n)^{1/2} \right). \tag{A68}$$

Because of Jensen inequality for concave function:

$$\sum_{i=1}^{M} \mathbb{E}_{m_i, n_i} \left[O\left((m_i + n_i)^{1/2} \right) \right] = \sum_{i=1}^{M} O\left(\mathbb{E}[m_i] + \mathbb{E}[n_i] \right)^{1/2}$$

$$= \sum_{i=1}^{M} O\left(m a_i l^{-d} + n b_i l^{-d} \right)^{1/2} = O\left(l^d (m+n)^{1/2} \right).$$

In addition, similarly since $\eta < d$, we have

$$\sum_{i=1}^{M} \mathbb{E}_{m_i, n_i} \left[O\left((m_i + n_i)^{1-\eta/d} \right) \right] = O\left(l^d (m+n)^{1-\eta/d} \right), \tag{A69}$$

and, for $d \geq 2$, one yields

$$\sum_{i=1}^{M} \mathbb{E}_{m_i, n_i} \left[O\left((m_i + n_i)^{1-1/d} \right) \right] = O\left(l^d (m+n)^{1-1/d} \right) = O\left(l^d (m+n)^{1/2} \right). \tag{A70}$$

Next, we state the following lemma (Lemma 1 from [30,31]), which will be used in the sequel:

Lemma A13. *Let $k(x)$ be a continuously differential function of $x \in \mathbb{R}$ which is convex and monotone decreasing over $x \geq 0$. Set $k'(x) = \dfrac{dk(x)}{dx}$. Then, for any $x_0 > 0$, we have*

$$k(x_0) + \frac{k(x_0)}{x_0} |x - x_0| \geq k(x) \geq k(x_0) - k'(x_0) |x - x_0|. \tag{A71}$$

Next, continuing the proof of (A60), we attend to find an upper bound for

$$\mathbb{E}_{m_i, n_i} \left[\frac{m_i n_i}{m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})} \right]. \tag{A72}$$

In order to pursue this aim, in Lemma A13, consider $k(x) = \dfrac{1}{x}$ and $x_0 = \mathbb{E}_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right]$, therefore as the function $k(x)$ is decreasing and convex, one can write

$$\frac{1}{m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})} \leq \frac{1}{\mathbb{E}_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right]} + \frac{\left| m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) - \mathbb{E}_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right] \right|}{\mathbb{E}^2_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right]}. \tag{A73}$$

Using the Hölder inequality implies the following inequality:

$$\mathbb{E}_{m_i, n_i} \left[\frac{m_i n_i}{m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})} \right] \leq \frac{\mathbb{E}_{m_i, n_i} [m_i n_i]}{\mathbb{E}_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right]}$$

$$+ \frac{\left(\mathbb{E}_{m_i, n_i} \left[m_i^2 n_i^2 \right] \right)^{1/2}}{\mathbb{E}^2_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right]} \times \left(\mathbb{E}_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) - \mathbb{E}_{m_i, n_i} \left[m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right] \right]^2 \right)^{1/2}. \tag{A74}$$

As random variables m_i, n_i are independent, and because of $\mathbb{V}[m_i] \leq ma_i l^{-d}$, $\mathbb{V}[n_i] \leq nb_i l^{-d}$, we can claim that the RHS of (A74) becomes less than and equal to

$$\frac{mna_i b_i l^{-2d}}{ma_i l^{-d} f_0(\mathbf{x}) + nb_i l^{-d} f_1(\mathbf{x})} + \frac{\left(\alpha_i \beta_i \left(ma_i l^{-d} f_0^2(\mathbf{x}) + nb_i l^{-d} f_1^2(\mathbf{x})\right)\right)^{1/2}}{\left(ma_i f_0(\mathbf{x}) + nb_i f_1(\mathbf{x})\right)^2}, \tag{A75}$$

where

$$\alpha_i = ma_i l^d \left(1 - a_i l^{-d}\right) + m^2 a_i^2,$$

$$\beta_i = nb_i l^d \left(1 - b_i l^{-d}\right) + n^2 b_i^2.$$

Going back to (A66), we have

$$\mathbb{E}\left[\mathfrak{R}'_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\right] \leq \sum_{i=1}^{M} a_i b_i l^{-d} \int \frac{2\, mn f_0(\mathbf{x}) f_1(\mathbf{x})}{ma_i f_0(\mathbf{x}) + nb_i f_1(\mathbf{x})}\, d\mathbf{x}$$

$$+ \sum_{i=1}^{M} 2 \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x}) \left(\alpha_i \beta_i \left(ma_i l^{-d} f_0^2(\mathbf{x}) + nb_i l^{-d} f_1^2(\mathbf{x})\right)\right)^{1/2}}{\left(ma_i f_0(\mathbf{x}) + nb_i f_1(\mathbf{x})\right)^2}\, d\mathbf{x} \tag{A76}$$

$$+ \sum_{i=1}^{M} \mathbb{E}_{m_i, n_i}\left[(m_i + n_i)\, \varsigma_\eta(l, m_i, n_i)\right] + O\left(l^d (m+n)^{1-\eta/d}\right)$$

$$+ O\left(l^d (m+n)^{1/2}\right) + 2c_1\, l^{d-1} (m+n)^{1/d}.$$

Finally, owing to $a_i \leq b_i$ and $\sum_{i=1}^{M} b_i l^{-d} = 1$, when $\frac{m}{m+n} \to p$, we have

$$\frac{\mathbb{E}\left[\mathfrak{R}'_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\right]}{m+n} \leq \int \frac{2\, pq f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})}\, d\mathbf{x}$$

$$+ \sum_{i=1}^{M} 2 \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x}) \left(\alpha_i \beta_i \left(ma_i l^{-d} f_0^2(\mathbf{x}) + nb_i l^{-d} f_1^2(\mathbf{x})\right)\right)^{1/2}}{\left(ma_i f_0(\mathbf{x}) + nb_i f_1(\mathbf{x})\right)^2 (m+n)}\, d\mathbf{x} \tag{A77}$$

$$+ \frac{1}{m+n} \sum_{i=1}^{M} \mathbb{E}_{m_i, n_i}\left[(m_i + n_i)\, \varsigma_\eta(l, m_i, n_i)\right] + O\left(l^d (m+n)^{-\eta/d}\right)$$

$$+ O\left(l^d (m+n)^{-1/2}\right) + 2c_1\, l^{d-1} (m+n)^{(1/d)-1}.$$

Passing to Definition 2, MST*, and Lemma A2, a similar discussion as above, consider the Poisson processes samples and the FR statistic under the union of samples, denoted by $\mathfrak{R}'^{*}_{m,n}$, and superadditivity of dual $\mathfrak{R}^{*}_{m,n}$, we have

$$\mathbb{E}\left[\mathfrak{R}'^{*}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\right] \geq \sum_{i=1}^{M} \mathbb{E}\left[\mathfrak{R}'^{*}_{m_i, n_i}\left((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i\right)\right] - c_2\, l^d$$

$$= \sum_{i=1}^{M} \mathbb{E}_{m_i, n_i}\left[\mathbb{E}\left[\mathfrak{R}'^{*}_{m_i, n_i}\left((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i\right) \big| m_i, n_i\right]\right] - c_2\, l^d \tag{A78}$$

$$\geq \sum_{i=1}^{M} \mathbb{E}_{m_i, n_i}\left[\mathbb{E}\left[\mathfrak{R}'_{m_i, n_i}\left((\mathfrak{X}_m, \mathfrak{Y}_n) \cap Q_i\right) \big| m_i, n_i\right]\right] - c_2\, l^d,$$

the last line is derived from Lemma A2, (ii), inequality (A8). Owing to the Lemma A6, (A69), and (A70), one obtains

$$
\mathbb{E}\Big[\mathfrak{R}'^{*}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n)\Big] \geq \sum_{i=1}^{M} \int \mathbb{E}_{m_i,n_i}\left[\frac{2m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})}\right] d\mathbf{x}
$$
$$
- \sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\big[(m_i + n_i)\, \varsigma_\eta(l, m_i, n_i)\big] - O\big(l^d (m+n)^{1-\eta/d}\big) - O\big(l^d (m+n)^{1/2}\big) - c_2\, l^d.
\tag{A79}
$$

Furthermore, by using the Jenson's inequality, we get

$$
\mathbb{E}_{m_i,n_i}\left[\frac{m_i n_i}{m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})}\right] \geq \frac{\mathbb{E}[m_i]\mathbb{E}[n_i]}{\mathbb{E}[m_i] f_0(\mathbf{x}) + \mathbb{E}[n_i] f_1(\mathbf{x})} = \frac{l^{-d}\big(m a_i n b_i\big)}{m a_i f_0(\mathbf{x}) + n b_i f_1(\mathbf{x})}.
$$

Therefore, since $a_i \leq b_i$, we can write

$$
\mathbb{E}_{m_i,n_i}\left[\frac{m_i n_i}{m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})}\right] \geq \frac{l^{-d} mn\, a_i b_i}{b_i\big(m f_0(\mathbf{x}) + n f_1(\mathbf{x})\big)} = \frac{l^{-d} mn\, a_i}{\big(m f_0(\mathbf{x}) + n f_1(\mathbf{x})\big)}.
\tag{A80}
$$

Consequently, the RHS of (A79) becomes greater than or equal to

$$
\sum_{i=1}^{M} a_i\, l^{-d} \int \frac{2mn f_0(\mathbf{x}) f_1(\mathbf{x})}{m f_0(\mathbf{x}) + n f_1(\mathbf{x})}\, d\mathbf{x}
$$
$$
- \sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\big[(m_i + n_i)\, \varsigma_\eta(l, m_i, n_i)\big] - O\big(l^d (m+n)^{1-\eta/d}\big) - O\big(l^d (m+n)^{1/2}\big) - c_2\, l^d.
\tag{A81}
$$

Finally, since $\displaystyle\sum_{i=1}^{M} a_i l^{-d} = 1$ and $\dfrac{m}{m+n} \to p$, we have

$$
\frac{\mathbb{E}\Big[\mathfrak{R}'^{*}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n)\Big]}{m+n} \geq \int \frac{2pq f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})}\, d\mathbf{x} - (m+n)^{-1} \sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\big[(m_i + n_i)\, \varsigma(l, m_i, n_i)\big]
$$
$$
- O\big(l^d (m+n)^{-\eta/d}\big) - O\big(l^d (m+n)^{-1/2}\big) - c_2\, l^d (m+n)^{-1}.
\tag{A82}
$$

By definition of the dual $\mathfrak{R}^{*}_{m,n}$ and (i) in Lemma A2,

$$
\frac{\mathbb{E}\Big[\mathfrak{R}'_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n)\Big]}{m+n} + \frac{c_d\, 2^d}{m+n} \geq \frac{\mathbb{E}\Big[\mathfrak{R}'^{*}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n)\Big]}{m+n},
\tag{A83}
$$

we can imply

$$
\frac{\mathbb{E}\Big[\mathfrak{R}'_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n)\Big]}{m+n} \geq \int \frac{2pq f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})}\, d\mathbf{x} - (m+n)^{-1} \sum_{i=1}^{M} \mathbb{E}_{m_i,n_i}\big[(m_i + n_i)\, \varsigma_\eta(l, m_i, n_i)\big]
$$
$$
- O\big(l^d (m+n)^{-\eta/d}\big) - O\big(l^d (m+n)^{-1/2}\big) - c_2\, l^d (m+n)^{-1} - c_d\, 2^d\, (m+n)^{-1}.
\tag{A84}
$$

The combination of two lower and upper bounds (A84) and (A77) yields the following result

$$
\left| \frac{\mathbb{E}\left[\mathfrak{R}'_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \right]}{m+n} - \int \frac{2pq f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})} \, d\mathbf{x} \right|
$$

$$
\leq O\left(l^d (m+n)^{-\eta/d} \right) + O\left(l^d (m+n)^{-1/2} \right) + 2\, c_1\, l^{d-1}\, (m+n)^{(1/d)-1}
$$

$$
+ c_d\, 2^d\, (m+n)^{-1} + c_2\, (m+n)^{-1}\, l^d + \frac{1}{m+n} \sum_{i=1}^{M} \mathbb{E}_{m_i, n_i} \left[(m_i + n_i)\, \varsigma_\eta (l, m_i, n_i) \right]
$$

$$
+ \sum_{i=1}^{M} 2 \int \frac{f_0(\mathbf{x}) f_1(\mathbf{x}) \left(\alpha_i \beta_i \left(m a_i l^{-d} f_0^2(\mathbf{x}) + n b_i l^{-d} f_1^2(\mathbf{x}) \right) \right)^{1/2}}{\left(m a_i f_0(\mathbf{x}) + n b_i f_1(\mathbf{x}) \right)^2 (m+n)} \, d\mathbf{x}. \tag{A85}
$$

Recall $\varsigma_\eta(l, m_i, n_i)$, then we obtain

$$
\sum_{i=1}^{M} \mathbb{E}_{m_i, n_i} \left[(m_i + n_i)\, \varsigma_\eta (l, m_i, n_i) \right] = \sum_{i=1}^{M} O(l) \int \mathbb{E} \left[\frac{2 m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i + n_i)(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right] d\mathbf{x}
$$

$$
- 2\, l^d \sum_{i=1}^{M} \int \mathbb{E} \left[\frac{2 m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i + n_i)(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right] d\mathbf{x} + O(l^{-\eta}) \sum_{i=1}^{M} \mathbb{E}_{m_i, n_i} [m_i + n_i]. \tag{A86}
$$

In addition, we have

$$
\mathbb{E}_{m_i, n_i} \left[\frac{2 m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i + n_i)(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right] \geq \frac{1}{m+n} \mathbb{E}_{m_i, n_i} \left[\frac{2 m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right]. \tag{A87}
$$

This implies

$$
\sum_{i=1}^{M} \int \mathbb{E} \left[\frac{2 m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i + n_i)(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right] d\mathbf{x} \geq \int \frac{2pq f_0(\mathbf{x}) f_1(\mathbf{x})}{p f_0(\mathbf{x}) + q f_1(\mathbf{x})} \, d\mathbf{x}. \tag{A88}
$$

Note that the above inequality is derived from (A80) and $\dfrac{m}{m+n} \to p$. Furthermore,

$$
\frac{1}{m+n} \sum_{i=1}^{M} O(l) \int \mathbb{E}_{m_i, n_i} \left[\frac{2 m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i + n_i)(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right] d\mathbf{x}
$$

$$
\leq \sum_{i=1}^{M} O(l) \int \mathbb{E}_{m_i, n_i} \left[\frac{2 m_i n_i f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i + n_i)^2 (m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right] d\mathbf{x} \tag{A89}
$$

$$
\leq \sum_{i=1}^{M} O(l) \int \mathbb{E}_{m_i, n_i} \left[\frac{2 f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))} \right] d\mathbf{x}.
$$

The last line holds because of $m_i n_i \leq (m_i + n_i)^2$. Going back to (A73), we can give an upper bound for the RHS of above inequality as

$$
\mathbb{E}_{m_i, n_i} \left[\left(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) \right)^{-1} \right] \leq \left(m a_i l^{-d} f_0(\mathbf{x}) + n b_i l^{-d} f_1(\mathbf{x}) \right)^{-1}
$$

$$
+ \left(\mathbb{E}_{m_i, n_i} \left| m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}) - \left(\mathbb{E}[m_i] f_0(\mathbf{x}) + \mathbb{E}[n_i] f_1(\mathbf{x}) \right| \right) \right/ \left(m a_i l^{-d} f_0(\mathbf{x}) + n b_i l^{-d} f_1(\mathbf{x}) \right)^2.
$$

Note that we have assumed $a_i \leq b_i$ and by using Hölder inequality we write

$$
\mathbb{E}_{m_i,n_i}\left[\left(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x})\right)^{-1}\right] \leq l^d(a_i)^{-1}\left(m f_0(\mathbf{x}) + n f_1(\mathbf{x})\right)^{-1}
$$

$$
+\left(f_0(\mathbf{x})\sqrt{\mathbb{V}(m_i)} + f_1(\mathbf{x})\sqrt{\mathbb{V}(n_i)}\right)\bigg/\left(a_i^2 l^{-d}(m f_0(\mathbf{x}) + n f_1(\mathbf{x}))^2\right) \leq l^d(a_i)^{-1}\left(m f_0(\mathbf{x}) + n f_1(\mathbf{x})\right)^{-1} \quad \text{(A90)}
$$

$$
+ l^{-d/2}\sqrt{b_i}\left(f_0(\mathbf{x})\sqrt{m} + f_1(\mathbf{x})\sqrt{n}\right)\bigg/\left(a_i^2 l^{-d}(m f_0(\mathbf{x}) + n f_1(\mathbf{x}))^2\right).
$$

As result, we have

$$
\sum_{i=1}^{M} O(l) \int \mathbb{E}_{m_i,n_i}\left[\frac{2 f_0(\mathbf{x}) f_1(\mathbf{x})}{(m_i f_0(\mathbf{x}) + n_i f_1(\mathbf{x}))}\right] d\mathbf{x}
$$

$$
\leq \sum_{i=1}^{M} O(l) \int l^d(a_i)^{-1} \frac{2 f_0(\mathbf{x}) f_1(\mathbf{x})}{m f_0(\mathbf{x}) + n f_1(\mathbf{x})} d\mathbf{x} \quad \text{(A91)}
$$

$$
+ \sum_{i=1}^{M} O(l) \int l^{-d/2}\sqrt{b_i}\, \frac{2 f_0(\mathbf{x}) f_1(\mathbf{x})\left(f_0(\mathbf{x})\sqrt{m} + f_1(\mathbf{x})\sqrt{n}\right)}{a_i^2 l^{-d}\left(m f_0(\mathbf{x}) + n f_1(\mathbf{x})\right)^2} d\mathbf{x}.
$$

As a consequence, owing to (A85), for $0 < \eta \leq 1$, $d \geq 2$, which implies $\eta \leq d - 1$, we can derive (A61). Thus, the proof can be concluded by giving the summarized bound in (A60). □

Lemma A8: For $h = 1, 2, \ldots$, let $\delta_{m,n}^h$ be the function $c\, h^{d-1}(m+n)^{1/d}$. Then, for $\epsilon > 0$, we have

$$
P\left(\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_{m_i}, \mathfrak{Y}_{n_i}) + 2\epsilon\right) \geq \frac{\epsilon - \delta_{m,n}^h}{\epsilon}. \quad \text{(A92)}
$$

Note that in case $\epsilon \leq \delta_{m,n}^h$ the above claimed inequality is trivial.

Proof. Consider the cardinality of the set of all edges of $\mathrm{MST}\left(\bigcup_{i=1}^{h^d} Q_i\right)$ which intersect two different subcubes Q_i and Q_j, $|D|$. Using the Markov inequality, we can write

$$
P\left(|D| \geq \epsilon\right) \leq \frac{\mathbb{E}(|D|)}{\epsilon},
$$

where $\epsilon > 0$. Since $\mathbb{E}|D| \leq c\, h^{d-1}(m+n)^{1/d} := \delta_{m,n}^h$, therefore for $\epsilon > \delta_{m,n}^h$ and $h = 1, 2, \ldots$:

$$
P\left(|D| \geq \epsilon\right) \leq \frac{\delta_{m,n}^h}{\epsilon}.
$$

In addition, if Q_i, $i = 1, \ldots h^d$ is a partition of $[0,1]^d$ into congruent subcubes of edge length $1/h$, then

$$
P\left(\sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i) + 2|D| \geq \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i) + 2\epsilon\right) \leq \frac{\delta_{m,n}^h}{\epsilon}. \quad \text{(A93)}
$$

This implies

$$
P\left(\sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i) + 2|D| \leq \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i) + 2\epsilon\right) \geq 1 - \frac{\delta_{m,n}^h}{\epsilon}. \quad \text{(A94)}
$$

By subadditivity (A6), we can write

$$
\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i) + 2|D|,
$$

and this along with (A94) establishes (A92).　□

Lemma A9: (Growth bounds for $\mathfrak{R}_{m,n}$) Let $\mathfrak{R}_{m,n}$ be the FR statistic. Then, for given non-negative ϵ, such that $\epsilon \geq h^2 \, \delta^h_{m,n}$, with at least probability $g(\epsilon) := 1 - \dfrac{h \, \delta^h_{m,n}}{\epsilon}$, $h = 2, 3, \ldots$, we have

$$
\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq c''_{\epsilon,h} \left(\#\mathfrak{X}_m \, \#\mathfrak{Y}_n \right)^{1-1/d}. \tag{A95}
$$

Here, $c''_{\epsilon,h} = O\left(\dfrac{\epsilon}{h^{d-1} - 1} \right)$ depending only on ϵ, h. Note that, for $\epsilon < h^2 \, \delta^h_{m,n}$, the claim is trivial.

Proof. Without loss of generality, consider the unit cube $[0,1]^d$. For given h, if Q_i, $i = 1, \ldots h^d$ is a partition of $[0,1]^d$ into congruent subcubes of edge length $1/h$, then, by Lemma A8, we have

$$
P\left(\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_{m_i}, \mathfrak{Y}_{n_i}) + 2\epsilon \right) \geq \frac{\epsilon - \delta^h_{m,n}}{\epsilon}. \tag{A96}
$$

We apply the induction methodology on $\#\mathfrak{X}_m$ and $\#\mathfrak{Y}_n$. Set $c := \sup\limits_{\mathbf{x}, \mathbf{y} \in [0,1]^d} \mathfrak{R}_{m,n}(\{\mathbf{x}, \mathbf{y}\})$ which is finite according to assumption. Moreover, set $c_2 := \dfrac{2\epsilon}{h^{d-1} - 1}$ and $c_1 := c + d \, h^{d-1} c_2$. Therefore, it is sufficient to show that for all $(\mathfrak{X}_m, \mathfrak{Y}_n) \in [0,1]^d$ with at least probability $g(\epsilon)$

$$
\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq c_1 \left(\#\mathfrak{X}_m \, \#\mathfrak{Y}_n \right)^{(d-1)/d}. \tag{A97}
$$

Alternatively, as for the induction hypothesis, we assume the stronger bound

$$
\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq c_1 \left(\#\mathfrak{X}_m \, \#\mathfrak{Y}_n \right)^{(d-1)/d} - c_2 \tag{A98}
$$

holds whenever $\#\mathfrak{X}_m < m$ and $\#\mathfrak{Y}_n < n$ with at least probability $g(\epsilon)$. Note that $d \geq 2$, $\epsilon > 0$ and c_1, c_2 both depend on ϵ, h. Hence,

$$
c_1 - c_2 = c + c_2 \left(d \, h^{d-1} - 1 \right) \geq c + c_2 \left(h^{d-1} - 1 \right) = c + 2\epsilon \geq c,
$$

which implies $P(\mathfrak{R}_{m,n} \leq c_1 - c_2) \geq P(\mathfrak{R}_{m,n} \leq c)$. In addition, we know that $P(\mathfrak{R}_{m,n} \leq c) = 1 \geq g(\epsilon)$; therefore, the induction hypothesis holds particularly $\#\mathfrak{X}_m = 1$ and $\#\mathfrak{Y}_n = 1$. Now, consider the partition Q_i of $[0,1]^d$; therefore, for all $1 \leq i \leq h^d$, we have $m_i := \#(\mathfrak{X}_m \cap Q_i) < m$ and $n_i := \#(\mathfrak{Y}_n \cap Q_i) < n$ and thus, by induction hypothesis, one yields with at least probability $g(\epsilon)$

$$
\mathfrak{R}_{m_i,n_i}(\mathfrak{X}_m, \mathfrak{Y}_n \cap Q_i) \leq c_1 \, (m_i \, n_i)^{1-1/d} - c_2. \tag{A99}
$$

Set $\mathbb{B}$ the event $\{ \text{all } i \ : \ \mathfrak{R}_{m_i,n_i} \leq c_1 \, (m_i \, n_i)^{1-1/d} - c_2 \}$ and $\mathbb{B}_i$ stands with the event $\{ \mathfrak{R}_{m_i,n_i} \leq c_1 \, (m_i \, n_i)^{1-1/d} - c_2 \}$. From (A96) and since Q_i's are partitions, which implies

$$
P(\mathbb{B}) = \left(P(\mathbb{B}_i) \right)^{h^d} \leq P(\mathbb{B}_i), \quad P(\mathbb{B}^c) = P\left(\bigcup_{i=1}^{l^d} \mathbb{B}_i^c \right) \leq \sum_{i=1}^{h^d} P(\mathbb{B}_i^c) \leq h^d \left(1 - g(\epsilon) \right),
$$

$$
\text{and} \quad P(\mathbb{B}) = \prod_{i=1}^{h^d} P(\mathbb{B}_i) \geq \left(g(\epsilon) \right)^{h^d},
$$

we thus obtain

$$
\frac{\epsilon - \delta_{m,n}^h}{\epsilon} \le P\Big(\mathfrak{R}_{m,n} \le \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_{m_i}, \mathfrak{Y}_{n_i}) + 2\epsilon\,\big|\mathbb{B}\Big)P(\mathbb{B}) + P\Big(\mathfrak{R}_{m,n} \le \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_{m_i}, \mathfrak{Y}_{n_i}) + 2\epsilon\,\big|\mathbb{B}^c\Big)P(\mathbb{B}^c)
$$

$$
\le P\Big(\mathfrak{R}_{m,n} \le \sum_{i=1}^{l^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_{m_i}, \mathfrak{Y}_{n_i}) + 2\epsilon\,\big|\mathbb{B}\Big)P(\mathbb{B}) + P(\mathbb{B}^c).
$$

Equivalently,

$$
P\Big(\mathfrak{R}_{m,n} \le \sum_{i=1}^{h^d} \mathfrak{R}_{m_i,n_i}(\mathfrak{X}_{m_i}, \mathfrak{Y}_{n_i}) + 2\epsilon\,\big|\mathbb{B}\Big) \ge \Big(1 - \frac{\delta_{m,n}^h}{\epsilon} - 1 + P(\mathbb{B})\Big)\big/P(\mathbb{B}) = 1 - \frac{\delta_{m,n}^h}{\epsilon\,P(\mathbb{B})}.
$$

In fact, in this stage, we want to show that

$$
1 - \frac{\delta_{m,n}^h}{\epsilon\,P(\mathbb{B})} \ge g(\epsilon) \quad \text{or} \quad P(\mathbb{B}) \ge \frac{\delta_{m,n}^h}{\epsilon\,(1 - g(\epsilon))}.
$$

Since $P(\mathbb{B}) \ge \big(g(\epsilon)\big)^{h^d}$, therefore it is sufficient to derive that $\big(g(\epsilon)\big)^{h^d} \ge \dfrac{\delta_{m,n}^h}{\epsilon\,(1 - g(\epsilon))}$. Indeed, for given $g(\epsilon) = \Big(\dfrac{\epsilon - h\,\delta_{m,n}^h}{\epsilon}\Big)$, we have $g(\epsilon) \le \dfrac{\epsilon - \delta_{m,n}^h}{\epsilon}$ hence $\dfrac{\delta_{m,n}^h}{\epsilon\,(1 - g(\epsilon))} = \dfrac{1}{h} \le 1$. Furthermore, we know $\dfrac{1}{h} \le 1 - \dfrac{1}{h^{(1/h^d)}}$ and since $\epsilon \ge h^2\,\delta_{m,n}^h$ this implies $\dfrac{h\,\delta_{m,n}^h}{\epsilon} \le \dfrac{1}{h}$ and consequently

$$
\frac{h\,\delta_{m,n}^h}{\epsilon} \le 1 - \frac{1}{h^{h^{-d}}}
$$

or

$$
g(\epsilon)^{h^d} = \Big(\frac{\epsilon - h\,\delta_{m,n}^h}{\epsilon}\Big)^{h^d} \ge \frac{1}{h} = \frac{\delta_{m,n}^h}{\epsilon\,(1 - g(\epsilon))}.
$$

This implies the fact that for $\epsilon \ge h^2\delta_{m,n}^h$

$$
P\Big(\mathfrak{R}_{m,n} \le \sum_{i=1}^{h^d} \big(c_1(m_i n_i)^{1-1/d} - c_2\big) + 2\epsilon\Big) \ge g(\epsilon), \quad \text{where} \quad g(\epsilon) = \frac{\epsilon - h\,\delta_{m,n}^h}{\epsilon}.
$$

Now, let $\gamma := \#\{i : m_i, n_i > 0\}$ and using Hölder inequality gives

$$
P\Big(\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \le c_1\gamma^{1/d}(m\,n)^{1-1/d} - \gamma c_2 + c_2\,(h^{d-1} - 1)\Big) \ge g(\epsilon). \tag{A100}
$$

Next, we just need to show that $c_1\gamma^{1/d}(m\,n)^{1-1/d} - \gamma c_2 + c_2\,(h^{d-1} - 1)$ in (A100) is less than or equal to $c_1(m\,n)^{1-1/d} - c_2$, which is equivalent to show

$$
c_2\big(h^{d-1} - \gamma\big) \le c_1(m\,n)^{1-1/d}(1 - \gamma^{1/d}).
$$

We know that $m, n \ge 1$ and $c_1 \ge d\,h^{d-1}c_2$, so it is sufficient to get

$$
c_2\big(h^{d-1} - \gamma\big) \le d\,h^{d-1}c_2(1 - \gamma^{1/d}), \tag{A101}
$$

choose t as $\gamma = t\,h^d$, then $0 < t \le 1$, so (A101) becomes

$$
(h^{-1} - t) \ge d\,h^{-1}\big(1 - h\,t^{1/d}\big). \tag{A102}
$$

Note that the function $d\, h^{-1}\left(1 - h\, t^{1/d}\right) + t - h^{-1}$ has a minimum at $t = 1$ which implies (A101) and subsequently (A95). Hence, the proof is completed. $\square$

Lemma A10: (Smoothness for $\mathfrak{R}_{m,n}$) Given observations of

$$\mathfrak{X}_m := (\mathfrak{X}_{m'}, \mathfrak{X}_{m''}) = \{\mathbf{X}_1, \ldots, \mathbf{X}_{m'}, \mathbf{X}_{m'+1}, \ldots, \mathbf{X}_m\},$$

such that $m' + m'' = m$ and $\mathfrak{Y}_n := (\mathfrak{Y}_{n'}, \mathfrak{Y}_{n''}) = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n'}, \mathbf{Y}_{n'+1}, \ldots, \mathbf{Y}_n\}$, where $n' + n'' = n$, denote $\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)$ as before, the number of edges of $\mathrm{MST}(\mathfrak{X}_m, \mathfrak{Y}_n)$ which connect a point of $\mathfrak{X}_m$ to a point of $\mathfrak{Y}_n$. Then, for integer $h \geq 2$, for all $(\mathfrak{X}_n, \mathfrak{Y}_m) \in [0,1]^d$, $\epsilon \geq h^2\, \delta_{m,n}^h$, where $\delta_{m,n}^h = O\left(h^{d-1}(m+n)^{1/d}\right)$, we have

$$P\left(\left|\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'})\right| \leq \tilde{c}_{\epsilon,h}\left(\#\mathfrak{X}_{m''}\ \#\mathfrak{Y}_{n''}\right)^{1-1/d}\right) \geq 1 - \frac{2h\, \delta_{m,n}^h}{\epsilon}, \qquad (\mathrm{A}103)$$

where $\tilde{c}_{\epsilon,h} = O\left(\dfrac{\epsilon}{h^{d-1}-1}\right)$. For the case $\epsilon < h^2\, \delta_{m,n}^h$, this holds trivially.

Proof. We begin with removing the edges which contain a vertex in $\mathfrak{X}_{m''}$ and $\mathfrak{Y}_{n''}$ in minimal spanning tree on $(\mathfrak{X}_m, \mathfrak{Y}_n)$. Now, since each vertex has bounded degree, say c_d, we can generate a subgraph in which has at most $c_d(\#\mathfrak{X}_{m''} + \#\mathfrak{Y}_{n''})$ components. Next, choose one vertex from each component and form the minimal spanning tree on these vertices, assuming all of them can be considered in FR test statistic, we can write

$$\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq \mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'}) + c''_{\epsilon,h}\left(c_d^2\, \#\mathfrak{X}_{m''}\ \#\mathfrak{Y}_{n''}\right)^{1-1/d},$$

or equivalently

$$\leq \mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'}) + c_{\epsilon 1}^h\left(\#\mathfrak{X}_{m''}\ \#\mathfrak{Y}_{n''}\right)^{1-1/d},$$

$$(\mathrm{A}104)$$

with probability at least $g(\epsilon)$, where $g(\epsilon)$ is as in Lemma A9. Note that this expression is obtained from Lemma A9. In this stage, it remains to show that with at least probability $g(\epsilon)$

$$\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \geq \mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'}) - \tilde{c}_{\epsilon,h}\left(\#\mathfrak{X}_{m''}\ \#\mathfrak{Y}_{n''}\right)^{1-1/d}, \qquad (\mathrm{A}105)$$

which, again by using the method before, with at least probability $g(\epsilon)$, one derives

$$\mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'}) \leq \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) + \hat{c}_{\epsilon,h}\left(c_d^2\left(\#\mathfrak{X}_{m''}\ \#\mathfrak{Y}_{n''}\right)\right)^{1-1/d},$$

orequivalently

$$\leq \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) + c_{\epsilon 2}^h\left(\#\mathfrak{X}_{m''}\ \#\mathfrak{Y}_{n''}\right)^{1-1/d}.$$

Letting $\tilde{c}_{\epsilon,h} = \max\{c_{\epsilon 1}^h, c_{\epsilon 2}^h\}$ implies (A105). Thus,

$$P\left(\left|\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) - \mathfrak{R}_{m',n'}(\mathfrak{X}_{m'}, \mathfrak{Y}_{n'})\right| \geq \tilde{c}_{\epsilon,h}\left(\#\mathfrak{X}_{m''}\ \#\mathfrak{Y}_{n''}\right)^{1-1/d}\right) \leq 2 - 2\, g(\epsilon), \qquad (\mathrm{A}106)$$

Hence, the smoothness is given with at least probability $2\ g(\epsilon) - 1$ as in the statement of Lemma A10. $\square$

Lemma A11: (Semi-Isoperimetry) Let μ be a measure on $[0,1]^d$; μ^n denotes the product measure on space $([0,1]^d)^n$. In addition, let M_e denotes a median of $\mathfrak{R}_{m,n}$. Set

$$\mathbb{A} := \left\{\mathfrak{X}_m \in \left([0,1]^d\right)^m, \mathfrak{Y}_n \in \left([0,1]^d\right)^n; \mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n) \leq M_e\right\}. \qquad (\mathrm{A}107)$$

Then,

$$\mu^{m+n}\left(\left\{\mathbf{x}' \in ([0,1]^d)^m, \mathbf{y}' \in ([0,1]^n) : \phi_{\mathbb{A}}(\mathbf{x}')\,\phi_{\mathbb{A}}(\mathbf{y}') \geq t\right\}\right) \leq 4\exp\left(\frac{-t}{8(m+n)}\right). \tag{A108}$$

Proof. Let $\phi_{\mathbb{A}}(\mathbf{z}') = \min\{H(\mathbf{z}, \mathbf{z}'), \mathbf{z} \in \mathbb{A}\}$. Using Proposition 6.5 in [17], isoperimetric inequality, we have

$$\mu^{m+n}\left(\left\{\mathbf{z}' \in ([0,1]^d)^{m+n} : \phi_{\mathbb{A}}(\mathbf{z}') \geq t\right\}\right) \leq 4\exp\left(\frac{-t^2}{8(m+n)}\right). \tag{A109}$$

Furthermore, we know that

$$\left(\phi_{\mathbb{A}}(\mathbf{x}') + \phi_{\mathbb{A}}(\mathbf{y}')\right)^2 \geq \phi_{\mathbb{A}}(\mathbf{x}')\,\phi_{\mathbb{A}}(\mathbf{y}'),$$

hence

$$\mu^{m+n}\left(\left\{(\mathbf{x}' \in ([0,1]^d)^m, \mathbf{y}' \in ([0,1]^n) : \phi_{\mathbb{A}}(\mathbf{x}')\phi_{\mathbb{A}}(\mathbf{y}') \geq t\right\}\right)$$

$$\leq \mu^{m+n}\left(\left\{(\mathbf{x}' \in ([0,1]^d)^m, \mathbf{y}' \in ([0,1]^n) : \left(\phi_{\mathbb{A}}(\mathbf{x}') + \phi_{\mathbb{A}}(\mathbf{y}')\right)^2 \geq t\right\}\right) \tag{A110}$$

$$= \mu^{m+n}\left(\left\{(\mathbf{x}' \in ([0,1]^d)^m, \mathbf{y}' \in ([0,1]^n) : \phi_{\mathbb{A}}(\mathbf{x}') + \phi_{\mathbb{A}}(\mathbf{y}') \geq \sqrt{t}\right\}\right).$$

The last equality in (A110) achieves because of $\phi_{\mathbb{A}}(\mathbf{x}'), \phi_{\mathbb{A}}(\mathbf{y}') \geq 0$ and note that $\phi_{\mathbb{A}}(\mathbf{z}') \geq \phi_{\mathbb{A}}(\mathbf{x}') + \phi_{\mathbb{A}}(\mathbf{y}')$. Therefore,

$$\mu^{m+n}\left(\left\{(\mathbf{x}' \in ([0,1]^d)^m, \mathbf{y}' \in ([0,1]^n) : \phi_{\mathbb{A}}(\mathbf{x}') + \phi_{\mathbb{A}}(\mathbf{y}') \geq \sqrt{t}\right\}\right)$$

$$\leq \mu^{m+n}\left(\left\{(\mathbf{z}' \in ([0,1]^d)^{m+n} : \phi_{\mathbb{A}}(\mathbf{z}') \geq \sqrt{t}\right\}\right).$$

By recalling (A109), we derive the bound (A108). $\square$

Lemma A12: (Deviation of the Mean and Median) Consider M_e as a median of $\mathfrak{R}_{m,n}$. Then, for given $g(\epsilon) = 1 - \dfrac{h\,\delta_{m,n}^h}{\epsilon}$, and $\delta_{m,n}^h = O\left(h^{d-1}(m+n)^{1/d}\right)$ such that for $h \geq 7$, $\epsilon \geq h^2\delta_{m,n}^h$, we have

$$\left|\mathbb{E}\left[\mathfrak{R}_{m,n}(\mathfrak{X}_m, \mathfrak{Y}_n)\right] - M_e\right| \leq C_{m,n}(\epsilon, h)\,(m+n)^{(d-1)/d}, \tag{A111}$$

where $C_{m,n}(\epsilon, h)$ stands with a form depends on ϵ, h, m, n as

$$C_{m,n}(\epsilon, h) = C\left(1 - \left(\left(2\,(2\,g(\epsilon) - 1)^2\right)^{-1}\right)\right)^{-1}, \tag{A112}$$

where C is a constant.

Proof. Following the analogous arguments in [17,53], we have

$$\left|\mathbb{E}\left[\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n)\right] - M_e\right| \leq \mathbb{E}\left|\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n) - M_e\right| = \int_0^\infty P\left(\left|\mathfrak{R}_{m,n}(\mathfrak{X}_m,\mathfrak{Y}_n) - M_e\right| \geq t\right)\,\mathrm{d}t$$

$$\leq 8\left(1 - \left(1\big/\left(2\,(2\,g(\epsilon)-1)^2\right)\right)\right)^{-1}\int_0^\infty \exp\left(\frac{-t^{d/(d-1)}}{8(4\epsilon)^{d/d-1}(m+n)}\right)\,\mathrm{d}t$$

$$= C\left(1 - \left(\left(2\,(2\,g(\epsilon)-1)^2\right)^{-1}\right)\right)^{-1}(m+n)^{(d-1)/d},$$

(A113)

where $g(\epsilon) = 1 - \left(h\,O\!\left(h^{d-1}(m+n)^{1/d}\right)\right)\big/\epsilon$. The inequality in (A113) is implied from Theorem 5. Hence, the proof is completed. $\square$

References

1. Xuan, G.; Chia, P.; Wu, M. Bhattacharyya distance feature selection. In Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 25–29 August 1996; Volume 2, pp. 195–199.
2. Hamza, A.; Krim, H. Image registration and segmentation by maximizing the Jensen-Renyi divergence. In *Energy Minimization Methods in Computer Vision and Pattern Recognition. EMMCVPR 2003*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 147–163.
3. Hild, K.E.; Erdogmus, D.; Principe, J. Blind source separation using Renyi's mutual information. *IEEE Signal Process. Lett.* **2001**, *8*, 174–176. [CrossRef]
4. Basseville, M. Divergence measures for statistical data processing–An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633. [CrossRef]
5. Battacharyya, A. On a measure of divergence between two multinomial populations. *Sankhy ā Indian J. Stat.* **1946**, *7*, 401–406.
6. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
7. Berisha, V.; Hero, A. Empirical non-parametric estimation of the Fisher information. *IEEE Signal Process. Lett.* **2015**, *22*, 988–992. [CrossRef]
8. Berisha, V.; Wisler, A.; Hero, A.; Spanias, A. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Trans. Signal Process.* **2016**, *64*, 580–591. [CrossRef]
9. Moon, K.; Hero, A. Multivariate f-divergence estimation with confidence. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2420–2428.
10. Moon, K.; Hero, A. Ensemble estimation of multivariate f-divergence. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Honolulu, HI, USA, 29 June–4 July 2014; pp. 356–360.
11. Moon, K.; Sricharan, K.; Greenewald, K.; Hero, A. Improving convergence of divergence functional ensemble estimators. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1133–1137.
12. Moon, K.; Sricharan, K.; Greenewald, K.; Hero, A. Nonparametric ensemble estimation of distributional functionals. *arXiv* **2016**, arXiv:1601.06884v2.
13. Noshad, M.; Moon, K.; Yasaei Sekeh, S.; Hero, A. Direct Estimation of Information Divergence Using Nearest Neighbor Ratios. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017.
14. Yasaei Sekeh, S.; Oselio, B.; Hero, A. A Dimension-Independent discriminant between distributions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
15. Noshad, M.; Hero, A. Rate-optimal Meta Learning of Classification Error. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.

16. Wisler, A.; Berisha, V.; Wei, D.; Ramamurthy, K.; Spanias, A. Empirically-estimable multi-class classification bounds. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.

17. Yukich, J. *Probability Theory of Classical Euclidean Optimization*; Lecture Notes in Mathematics; Springer: Berlin, Germany, 1998; Volume 1675.

18. Steele, J. An Efron–Stein inequality for nonsymmetric statistics. *Ann. Stat.* **1986**, *14*, 753–758. [CrossRef]

19. Aldous, D.; Steele, J.M. Asymptotic for Euclidean minimal spanning trees on random points. *Probab. Theory Relat. Fields* **1992**, *92*, 247–258. [CrossRef]

20. Ma, B.; Hero, A.; Gorman, J.; Michel, O. Image registration with minimal spanning tree algorithm. In Proceedings of the IEEE International Conference on Image Processing, Vancouver, BC, Canada, 10–13 September 2000; pp. 481–484.

21. Neemuchwala, H.; Hero, A.; Carson, P. Image registration using entropy measures and entropic graphs. *Eur. J. Signal Process.* **2005**, *85*, 277–296. [CrossRef]

22. Hero, A.; Ma, B., M.O.; Gorman, J. Applications of entropic spanning graphs. *IEEE Signal Process. Mag.* **2002**, *19*, 85–95. [CrossRef]

23. Hero, A.; Michel, O. Estimation of Rényi information divergence via pruned minimal spanning trees. In Proceedings of the IEEE Workshop on Higher Order Statistics, Caesarea, Isreal, 16 June 1999.

24. Smirnov, N. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Mosc. Univ.* **1939**, *2*, 3–6.

25. Wald, A.; Wolfowitz, J. On a test whether two samples are from the same population. *Ann. Math. Stat.* **1940**, *11*, 147–162. [CrossRef]

26. Gibbons, J. *Nonparametric Statistical Inference*; McGraw-Hill: New York, NY, USA, 1971.

27. Singh, S.; Póczos, B. *Probability Theory and Combinatorial Optimization*; CBMF-NSF Regional Conference in Applied Mathematics; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 1997; Volume 69.

28. Redmond, C.; Yukich, J. Limit theorems and rates of convergence for Euclidean functionals. *Ann. Appl. Probab.* **1994**, *4*, 1057–1073. [CrossRef]

29. Redmond, C.; Yukich, J. Asymptotics for Euclidean functionals with power weighted edges. *Stoch. Process. Their Appl.* **1996**, *6*, 289–304. [CrossRef]

30. Hero, A.; Costa, J.; Ma, B. Convergence Rates of Minimal Graphs with Random Vertices. Available online: https://pdfs.semanticscholar.org/7817/308a5065aa0dd44098319eb66f81d4fa7a14.pdf (accessed on 18 November 2019).

31. Hero, A.; Costa, J.; Ma, B. *Asymptotic Relations between Minimal Graphs and Alpha-Entropy*; Tech. Rep.; Communication and Signal Processing Laboratory (CSPL), Department EECS, University of Michigan: Ann Arbor, MI, USA, 2003.

32. Lorentz, G. *Approximation of Functions*; Holt, Rinehart and Winston: New York, NY, USA, 1996.

33. Talagrand, M. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de i'I. H. E. S.* **1995**, *81*, 73–205. [CrossRef]

34. Kullback, S.; Leibler, R. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]

35. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, USA, 20 June–30 July 1961; pp. 547–561.

36. Ali, S.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *28*, 131–142. [CrossRef]

37. Cha, S. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math. Models Methods Appl. Sci.* **2007**, *1*, 300–307.

38. Rukhin, A. Optimal estimator for the mixture parameter by the method of moments and information affinity. In Proceedings of the 12th Prague Conference on Information Theory, Prague, Czech Republic, 29 August–2 September 1994; pp. 214–219.

39. Toussaint, G. The relative neighborhood graph of a finite planar set. *Pattern Recognit.* **1980**, *12*, 261–268. [CrossRef]

40. Zahn, C. Graph-theoretical methods for detecting and describing Gestalt clusters. *IEEE Trans. Comput.* **1971**, *100*, 68–86. [CrossRef]

41. Banks, D.; Lavine, M.; Newton, H. The minimal spanning tree for nonparametric regression and structure discovery. In *Computing Science and Statistics, Proceedings of the 24th Symposium on the Interface*; Joseph Newton, H., Ed.; Interface Foundation of North America: Fairfax Station, FA, USA, 1992; pp. 370–374.
42. Hoffman, R.; Jain, A. A test of randomness based on the minimal spanning tree. *Pattern Recognit. Lett.* **1983**, *1*, 175–180. [CrossRef]
43. Efron, B.; Stein, C. The jackknife estimate of variance. *Ann. Stat.* **1981**, *9*, 586–596. [CrossRef]
44. Singh, S.; Póczos, B. Generalized exponential concentration inequality for Rényi divergence estimation. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), Bejing, China, 22–24 June 2014; pp. 333–341.
45. Singh, S.; Póczos, B. Exponential concentration of a density functional estimator. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 3032–3040.
46. Lichman, M. UCI Machine Learning Repository. 2013. Available online: https://www.re3data.org/repository/r3d100010960 (accessed on 18 November 2019).
47. Bhatt, R.B.; Sharma, G.; Dhall, A.; Chaudhury, S. Efficient skin region segmentation using low complexity fuzzy decision tree model. In Proceedings of the IEEE-INDICON, Ahmedabad, India, 16–18 December 2009; pp. 1–4.
48. Steele, J.; Shepp, L.; Eddy, W. On the number of leaves of a euclidean minimal spanning tree. *J. Appl. Prob.* **1987**, *24*, 809–826. [CrossRef]
49. Henze, N.; Penrose, M. On the multivarite runs test. *Ann. Stat.* **1999**, *27*, 290–298.
50. Rhee, W. A matching problem and subadditive Euclidean funetionals. *Ann. Appl. Prob.* **1993**, *3*, 794–801. [CrossRef]
51. Whittaker, E.;Watson, G. *A Course in Modern Analysis*, 4th ed.; Cambridge University Press: New York, NY, USA, 1996.
52. Kingman, J. *Poisson Processes*; Oxford Univ. Press: Oxford, UK, 1993.
53. Pál, D.; Póczos, B.; Szapesvári, C. Estimation of Renyi entropy andmutual information based on generalized nearest-neighbor graphs. In Proceedings of the 23th International Conference on Neural Information Processing Systems (NIPS 2010), Vancouver, BC, Canada, 6–9 December 2010.