
Estimation of Rényi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs

Dávid Pál

Department of Computing Science
University of Alberta
Edmonton, AB, Canada
dpal@cs.ualberta.ca

Barnabás Póczos

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
poczos@ualberta.ca

Csaba Szepesvári

Department of Computing Science
University of Alberta
Edmonton, AB, Canada
szepesva@ualberta.ca

Abstract

We present simple and computationally efficient nonparametric estimators of Rényi entropy and mutual information based on an i.i.d. sample drawn from an unknown, absolutely continuous distribution over \mathbb{R}^d . The estimators are calculated as the sum of p -th powers of the Euclidean lengths of the edges of the ‘generalized nearest-neighbor’ graph of the sample and the empirical copula of the sample respectively. For the first time, we prove the almost sure consistency of these estimators and upper bounds on their rates of convergence, the latter of which under the assumption that the density underlying the sample is Lipschitz continuous. Experiments demonstrate their usefulness in independent subspace analysis.

1 Introduction

We consider the nonparametric problem of estimating Rényi α -entropy and mutual information (MI) based on a finite sample drawn from an unknown, absolutely continuous distribution over \mathbb{R}^d . There are many applications that make use of such estimators, of which we list a few to give the reader a taste: Entropy estimators can be used for goodness-of-fit testing (Vasicek, 1976; Goria et al., 2005), parameter estimation in semi-parametric models (Wolsztynski et al., 2005), studying fractal random walks (Alemany and Zanette, 1994), and texture classification (Hero et al., 2002b,a). Mutual information estimators have been used in feature selection (Peng and Ding, 2005), clustering (Aghagolzadeh et al., 2007), causality detection (Hlaváčková-Schindler et al., 2007), optimal experimental design (Lewi et al., 2007; Póczos and Lőrincz, 2009), fMRI data processing (Chai et al., 2009), prediction of protein structures (Adami, 2004), or boosting and facial expression recognition (Shan et al., 2005). Both entropy estimators and mutual information estimators have been used for independent component and subspace analysis (Learned-Miller and Fisher, 2003; Póczos and Lőrincz, 2005; Hulle, 2008; Szabó et al., 2007), and image registration (Kybic, 2006; Hero et al., 2002b,a). For further applications, see Leonenko et al. (2008); Wang et al. (2009a).

In a naïve approach to Rényi entropy and mutual information estimation, one could use the so called “plug-in” estimates. These are based on the obvious idea that since entropy and mutual information are determined solely by the density f (and its marginals), it suffices to first estimate the density using one’s favorite density estimate which is then “plugged-in” into the formulas defining entropy

and mutual information. The density is, however, a nuisance parameter which we do *not* want to estimate. Density estimators have tunable parameters and we may need cross validation to achieve good performance.

The entropy estimation algorithm considered here is *direct*—it does not build on density estimators. It is based on k -nearest-neighbor (NN) graphs with a fixed k . A variant of these estimators, where each sample point is connected to its k -th nearest neighbor only, were recently studied by Goria et al. (2005) for Shannon entropy estimation (*i.e.* the special case $\alpha = 1$) and Leonenko et al. (2008) for Rényi α -entropy estimation. They proved the *weak* consistency of their estimators under certain conditions. However, their proofs contain some errors, and it is not obvious how to fix them. Namely, Leonenko et al. (2008) apply the generalized Helly-Bray theorem, while Goria et al. (2005) apply the inverse Fatou lemma under conditions when these theorems do not hold. This latter error originates from the article of Kozachenko and Leonenko (1987), and this mistake can also be found in Wang et al. (2009b).

The first main contribution of this paper is to give a correct proof of consistency of these estimators. Employing a very different proof techniques than the papers mentioned above, we show that these estimators are, in fact, *strongly* consistent provided that the unknown density f has bounded support and $\alpha \in (0, 1)$. At the same time, we allow for more general nearest-neighbor graphs, wherein as opposed to connecting each point only to its k -th nearest neighbor, we allow each point to be connected to an arbitrary subset of its k nearest neighbors. Besides adding generality, our numerical experiments seem to suggest that connecting each sample point to all its k nearest neighbors improves the rate of convergence of the estimator.

The second major contribution of our paper is that we prove a finite-sample high-probability bound on the error (*i.e.* the rate of convergence) of our estimator provided that f is Lipschitz. According to the best of our knowledge, this is the very first result that gives a rate for the estimation of Rényi entropy. The closest to our result in this respect is the work by Tsybakov and van der Meulen (1996) who proved the root- n consistency of an estimator of the *Shannon* entropy and only in *one* dimension.

The third contribution is a *strongly* consistent estimator of Rényi mutual information that is based on NN graphs and the empirical copula transformation (Dedecker et al., 2007). This result is proved for $d \geq 3$ ¹ and $\alpha \in (1/2, 1)$. This builds upon and extends the previous work of Póczos et al. (2010) where instead of NN graphs, the minimum spanning tree (MST) and the shortest tour through the sample (*i.e.* the traveling salesman problem, TSP) were used, but it was only conjectured that NN graphs can be applied as well.

There are several advantages of using k -NN graph over MST and TSP (besides the obvious conceptual simplicity of k -NN): On a serial computer the k -NN graph can be computed somewhat faster than MST and much faster than the TSP tour. Furthermore, in contrast to MST and TSP, computation of k -NN can be easily parallelized. Secondly, for different values of α , MST and TSP need to be recomputed since the distance between two points is the p -th power of their Euclidean distance where $p = d(1 - \alpha)$. However, the k -NN graph does not change for different values of p , since p -th power is a monotone transformation, and hence the estimates for multiple values of α can be calculated without the extra penalty incurred by the recomputation of the graph. This can be advantageous *e.g.* in intrinsic dimension estimators of manifolds (Costa and Hero, 2003), where p is a free parameter, and thus one can calculate the estimates efficiently for a few different parameter values.

The fourth major contribution is a proof of a finite-sample high-probability error bound (*i.e.* the rate of convergence) for our mutual information estimator which holds under the assumption that the copula of f is Lipschitz. According to the best of our knowledge, this is the first result that gives a rate for the estimation of Rényi mutual information.

The toolkit for proving our results derives from the deep literature of Euclidean functionals, see, (Steele, 1997; Yukich, 1998). In particular, our strong consistency result uses a theorem due to Redmond and Yukich (1996) that essentially states that any quasi-additive power-weighted Euclidean functional can be used as a strongly consistent estimator of Rényi entropy (see also Hero and Michel 1999). We also make use of a result due to Koo and Lee (2007), who proved a rate of convergence result that holds under more stringent conditions. Thus, the main thrust of the present work is show-

¹Our result for Rényi entropy estimation holds for $d = 1$ and $d = 2$, too.

ing that these conditions hold for p -power weighted nearest-neighbor graphs. Curiously enough, up to now, no one has shown this, except for the case when $p = 1$, which is studied in Section 8.3 of (Yukich, 1998). However, the condition $p = 1$ gives results only for $\alpha = 1 - 1/d$.

All proofs and supporting lemmas can be found in the appendix. In the main body of the paper, we focus on clear explanation of Rényi entropy and mutual information estimation problems, the estimation algorithms and the statements of our converge results.

Additionally, we report on two numerical experiments. In the first experiment, we compare the empirical rates of convergence of our estimators with our theoretical results and plug-in estimates. Empirically, the NN methods are the clear winner. The second experiment is an illustrative application of mutual information estimation to an Independent Subspace Analysis (ISA) task.

The paper is organized as follows: In the next section, we formally define Rényi entropy and Rényi mutual information and the problem of their estimation. Section 3 explains the ‘generalized nearest neighbor’ graphs. This graph is then used in Section 4 to define our Rényi entropy estimator. In the same section, we state a theorem containing our convergence results for this estimator (strong consistency and rates). In Section 5, we explain the copula transformation, which connects Rényi entropy with Rényi mutual information. The copula transformation together with the Rényi entropy estimator from Section 4 is used to build an estimator of Rényi mutual information. We conclude this section with a theorem stating the convergence properties of the estimator (strong consistency and rates). Section 6 contains the numerical experiments. We conclude the paper by a detailed discussion of further related work in Section 7, and a list of open problems and directions for future research in Section 8.

2 The Formal Definition of the Problem

Rényi entropy and Rényi mutual information of d real-valued random variables² $\mathbf{X} = (X^1, X^2, \dots, X^d)$ with joint density $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and marginal densities $f_i : \mathbb{R} \rightarrow \mathbb{R}$, $1 \leq i \leq d$, are defined for any real parameter α assuming the underlying integrals exist. For $\alpha \neq 1$, Rényi entropy and Rényi mutual information are defined respectively as³

$$H_\alpha(\mathbf{X}) = H_\alpha(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} f^\alpha(x^1, x^2, \dots, x^d) d(x^1, x^2, \dots, x^d), \quad (1)$$

$$I_\alpha(\mathbf{X}) = I_\alpha(f) = \frac{1}{\alpha-1} \log \int_{\mathbb{R}^d} f^\alpha(x^1, x^2, \dots, x^d) \left(\prod_{i=1}^d f_i(x^i) \right)^{1-\alpha} d(x^1, x^2, \dots, x^d). \quad (2)$$

For $\alpha = 1$ they are defined by the limits $H_1 = \lim_{\alpha \rightarrow 1} H_\alpha$ and $I_1 = \lim_{\alpha \rightarrow 1} I_\alpha$. In fact, Shannon (differential) entropy and the Shannon mutual information are just special cases of Rényi entropy and Rényi mutual information with $\alpha = 1$.

The goal of this paper is to present estimators of Rényi entropy (1) and Rényi information (2) and study their convergence properties. To be more explicit, we consider the problem where we are given i.i.d. random variables $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ where each $\mathbf{X}_j = (X_j^1, X_j^2, \dots, X_j^d)$ has density $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and marginal densities $f_i : \mathbb{R} \rightarrow \mathbb{R}$ and our task is to construct an estimate $\hat{H}_\alpha(\mathbf{X}_{1:n})$ of $H_\alpha(f)$ and an estimate $\hat{I}_\alpha(\mathbf{X}_{1:n})$ of $I_\alpha(f)$ using the sample $\mathbf{X}_{1:n}$.

3 Generalized Nearest-Neighbor Graphs

The basic tool to define our estimators is the generalized nearest-neighbor graph and more specifically the sum of the p -th powers of Euclidean lengths of its edges.

Formally, let V be a finite set of points in an Euclidean space \mathbb{R}^d and let S be a finite non-empty set of positive integers; we denote by k the maximum element of S . We define the *generalized*

²We use superscript for indexing dimension coordinates.

³The base of the logarithms in the definition is not important; any base strictly bigger than 1 is allowed. Similarly as with Shannon entropy and mutual information, one traditionally uses either base 2 or e . In this paper, for definitiveness, we stick to base e .

nearest-neighbor graph $NN_S(V)$ as a directed graph on V . The edge set of $NN_S(V)$ contains for each $i \in S$ an edge from each vertex $\mathbf{x} \in V$ to its i -th nearest neighbor. That is, if we sort $V \setminus \{\mathbf{x}\} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{|V|-1}\}$ according to the Euclidean distance to \mathbf{x} (breaking ties arbitrarily): $\|\mathbf{x} - \mathbf{y}_1\| \leq \|\mathbf{x} - \mathbf{y}_2\| \leq \dots \leq \|\mathbf{x} - \mathbf{y}_{|V|-1}\|$ then \mathbf{y}_i is the i -th nearest-neighbor of \mathbf{x} and for each $i \in S$ there is an edge from \mathbf{x} to \mathbf{y}_i in the graph.

For $p \geq 0$ let us denote by $L_p(V)$ the sum of the p -th powers of Euclidean lengths of its edges. Formally,

$$L_p(V) = \sum_{(\mathbf{x}, \mathbf{y}) \in E(NN_S(V))} \|\mathbf{x} - \mathbf{y}\|^p, \quad (3)$$

where $E(NN_S(V))$ denotes the edge set of $NN_S(V)$. We intentionally hide the dependence on S in the notation $L_p(V)$. For the rest of the paper, the reader should think of S as a fixed but otherwise arbitrary finite non-empty set of integers, say, $S = \{1, 3, 4\}$.

The following is a basic result about L_p . The proof can be found in the appendix.

Theorem 1 (Constant γ). *Let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ be an i.i.d. sample from the uniform distribution over the d -dimensional unit cube $[0, 1]^d$. For any $p \geq 0$ and any finite non-empty set S of positive integers there exists a constant $\gamma > 0$ such that*

$$\lim_{n \rightarrow \infty} \frac{L_p(\mathbf{X}_{1:n})}{n^{1-p/d}} = \gamma \quad a.s. \quad (4)$$

The value of γ depends on d, p, S and, except for special cases, an analytical formula for its value is not known. This causes a minor problem since the constant γ appears in our estimators. A simple and effective way to deal with this problem is to generate a large i.i.d. sample $\mathbf{X}_{1:n}$ from the uniform distribution over $[0, 1]^d$ and estimate γ by the empirical value of $L_p(\mathbf{X}_{1:n})/n^{1-p/d}$.

4 An Estimator of Rényi Entropy

We are now ready to present an estimator of Rényi entropy based on the generalized nearest-neighbor graph. Suppose we are given an i.i.d. sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ from a distribution μ over \mathbb{R}^d with density f . We estimate entropy $H_\alpha(f)$ for $\alpha \in (0, 1)$ by

$$\widehat{H}_\alpha(\mathbf{X}_{1:n}) = \frac{1}{1-\alpha} \log \frac{L_p(\mathbf{X}_{1:n})}{\gamma n^{1-p/d}} \quad \text{where } p = d(1-\alpha), \quad (5)$$

and $L_p(\cdot)$ is the sum of p -th powers of Euclidean lengths of edges of the nearest-neighbor graph $NN_S(\cdot)$ for some finite non-empty $S \subset \mathbb{N}^+$ as defined by equation (3). The constant γ is the same as in Theorem 1.

The following theorem is our main result about the estimator \widehat{H}_α . It states that \widehat{H}_α is strongly consistent and gives upper bounds on the rate of convergence. The proof of theorem is in the appendix.

Theorem 2 (Consistency and Rate for \widehat{H}_α). *Let $\alpha \in (0, 1)$. Let μ be an absolutely continuous distribution over \mathbb{R}^d with bounded support and let f be its density. If $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ is an i.i.d. sample from μ then*

$$\lim_{n \rightarrow \infty} \widehat{H}_\alpha(\mathbf{X}_{1:n}) = H_\alpha(f) \quad a.s. \quad (6)$$

Moreover, if f is Lipschitz then for any $\delta > 0$ with probability at least $1 - \delta$,

$$\left| \widehat{H}_\alpha(\mathbf{X}_{1:n}) - H_\alpha(f) \right| \leq \begin{cases} O \left(n^{-\frac{d-p}{d(2d-p)}} (\log(1/\delta))^{1/2-p/(2d)} \right), & \text{if } 0 < p < d-1; \\ O \left(n^{-\frac{d-p}{d(d+1)}} (\log(1/\delta))^{1/2-p/(2d)} \right), & \text{if } d-1 \leq p < d. \end{cases} \quad (7)$$

5 Copulas and Estimator of Mutual Information

Estimating mutual information is slightly more complicated than estimating entropy. We start with a basic property of mutual information which we call *rescaling*. It states that if $h_1, h_2, \dots, h_d : \mathbb{R} \rightarrow \mathbb{R}$ are arbitrary strictly increasing functions, then

$$I_\alpha(h_1(X^1), h_2(X^2), \dots, h_d(X^d)) = I_\alpha(X^1, X^2, \dots, X^d). \quad (8)$$

A particularly clever choice is $h_j = F_j$ for all $1 \leq j \leq d$, where F_j is the cumulative distribution function (c.d.f.) of X^j . With this choice, the marginal distribution of $h_j(X^j)$ is the uniform distribution over $[0, 1]$ assuming that F_j , the c.d.f. of X^j , is continuous. Looking at the definition of H_α and I_α we see that

$$I_\alpha(X^1, X^2, \dots, X^d) = I_\alpha(F_1(X^1), F_2(X^2), \dots, F_d(X^d)) = -H_\alpha(F_1(X^1), F_2(X^2), \dots, F_d(X^d)).$$

In other words, calculation of mutual information can be reduced to the calculation of entropy provided that marginal c.d.f.'s F_1, F_2, \dots, F_d are known. The problem is, of course, that these are not known and need to be estimated from the sample. We will use empirical c.d.f.'s $(\widehat{F}_1, \widehat{F}_2, \dots, \widehat{F}_d)$ as their estimates. Given an i.i.d. sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ from distribution μ and with density f , the empirical c.d.f.'s are defined as

$$\widehat{F}_j(x) = \frac{1}{n} |\{i : 1 \leq i \leq n, x \leq X_i^j\}| \quad \text{for } x \in \mathbb{R}, 1 \leq j \leq d.$$

Introduce the compact notation $\mathbf{F} : \mathbb{R}^d \rightarrow [0, 1]^d$, $\widehat{\mathbf{F}} : \mathbb{R}^d \rightarrow [0, 1]^d$,

$$\mathbf{F}(x^1, x^2, \dots, x^d) = (F_1(x^1), F_2(x^2), \dots, F_d(x^d)) \quad \text{for } (x^1, x^2, \dots, x^d) \in \mathbb{R}^d; \quad (9)$$

$$\widehat{\mathbf{F}}(x^1, x^2, \dots, x^d) = (\widehat{F}_1(x^1), \widehat{F}_2(x^2), \dots, \widehat{F}_d(x^d)) \quad \text{for } (x^1, x^2, \dots, x^d) \in \mathbb{R}^d. \quad (10)$$

Let us call the maps \mathbf{F} , $\widehat{\mathbf{F}}$ the *copula transformation*, and the *empirical copula transformation*, respectively. The joint distribution of $\mathbf{F}(\mathbf{X}) = (F_1(X^1), F_2(X^2), \dots, F_d(X^d))$ is called the copula of μ , and the sample $(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \dots, \widehat{\mathbf{Z}}_n) = (\widehat{\mathbf{F}}(\mathbf{X}_1), \widehat{\mathbf{F}}(\mathbf{X}_2), \dots, \widehat{\mathbf{F}}(\mathbf{X}_n))$ is called the empirical copula (Dedecker et al., 2007). Note that j -th coordinate of $\widehat{\mathbf{Z}}_i$ equals

$$\widehat{Z}_i^j = \frac{1}{n} \text{rank}(X_i^j, \{X_1^j, X_2^j, \dots, X_n^j\}),$$

where $\text{rank}(x, A)$ is the number of element of A less than or equal to x . Also, observe that the random variables $\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \dots, \widehat{\mathbf{Z}}_n$ are not even independent! Nonetheless, the empirical copula $(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \dots, \widehat{\mathbf{Z}}_n)$ is a good approximation of an i.i.d. sample $(\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n) = (\mathbf{F}(\mathbf{X}_1), \mathbf{F}(\mathbf{X}_2), \dots, \mathbf{F}(\mathbf{X}_n))$ from the copula of μ . Hence, we estimate the Rényi mutual information I_α by

$$\widehat{I}_\alpha(\mathbf{X}_{1:n}) = -\widehat{H}_\alpha(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \dots, \widehat{\mathbf{Z}}_n), \quad (11)$$

where \widehat{H}_α is defined by (5). The following theorem is our main result about the estimator \widehat{I}_α . It states that \widehat{I}_α is strongly consistent and gives upper bounds on the rate of convergence. The proof of this theorem can be found in the appendix.

Theorem 3 (Consistency and Rate for \widehat{I}_α). *Let $d \geq 3$ and $\alpha = 1 - p/d \in (1/2, 1)$. Let μ be an absolutely continuous distribution over \mathbb{R}^d with density f . If $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ is an i.i.d. sample from μ then*

$$\lim_{n \rightarrow \infty} \widehat{I}_\alpha(\mathbf{X}_{1:n}) = I_\alpha(f) \quad \text{a.s.}$$

Moreover, if the density of the copula of μ is Lipschitz, then for any $\delta > 0$ with probability at least $1 - \delta$,

$$\left| \widehat{I}_\alpha(\mathbf{X}_{1:n}) - I_\alpha(f) \right| \leq \begin{cases} O \left(\max \left\{ n^{-\frac{d-p}{d(2d-p)}}, n^{-p/2+p/d} \right\} (\log(1/\delta))^{1/2} \right), & \text{if } 0 < p \leq 1; \\ O \left(\max \left\{ n^{-\frac{d-p}{d(2d-p)}}, n^{-1/2+p/d} \right\} (\log(1/\delta))^{1/2} \right), & \text{if } 1 \leq p \leq d-1; \\ O \left(\max \left\{ n^{-\frac{d-p}{d(d+1)}}, n^{-1/2+p/d} \right\} (\log(1/\delta))^{1/2} \right), & \text{if } d-1 \leq p < d. \end{cases}$$

6 Experiments

In this section we show two numerical experiments to support our theoretical results about the convergence rates, and to demonstrate the applicability of the proposed Rényi mutual information estimator, \widehat{I}_α .

6.1 The Rate of Convergence

In our first experiment (Fig. 1), we demonstrate that the derived rate is indeed an upper bound on the convergence rate. Figure 1a-1c show the estimation error of \widehat{I}_α as a function of the sample size. Here, the underlying distribution was a 3D uniform, a 3D Gaussian, and a 20D Gaussian with randomly chosen nontrivial covariance matrices, respectively. In these experiments α was set to 0.7. For the estimation we used $S = \{3\}$ (kth) and $S = \{1, 2, 3\}$ (knn) sets. Our results also indicate that these estimators achieve better performances than the histogram based plug-in estimators (hist). The number and the sizes of the bins were determined with the rule of Scott (1979). The histogram based estimator is not shown in the 20D case, as in this large dimension it is not applicable in practice. The figures are based on averaging 25 independent runs, and they also show the theoretical upper bound (Theoretical) on the rate derived in Theorem 3. It can be seen that the theoretical rates are rather conservative. We think that this is because the theory allows for quite irregular densities, while the densities considered in this experiment are very nice.

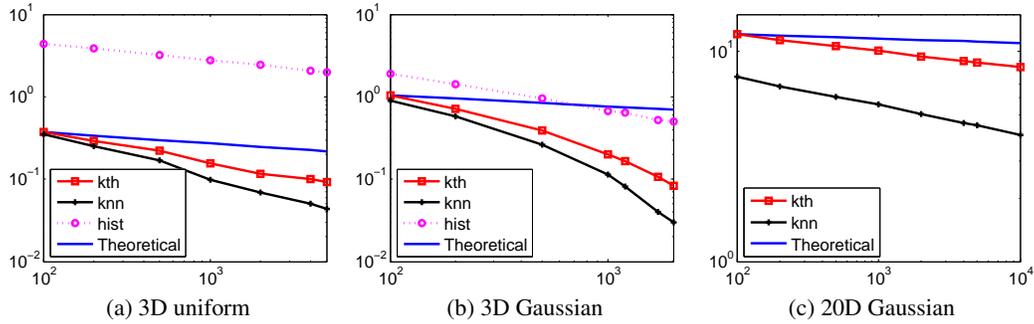


Figure 1: Error of the estimated Rényi informations in the number of samples.

6.2 Application to Independent Subspace Analysis

An important application of dependence estimators is the Independent Subspace Analysis problem (Cardoso, 1998). This problem is a generalization of the Independent Component Analysis (ICA), where we assume the independent sources are multidimensional vector valued random variables. The formal description of the problem is as follows. We have $\mathbf{S} = (\mathbf{S}^1; \dots; \mathbf{S}^m) \in \mathbb{R}^{dm}$, m independent d -dimensional sources, *i.e.* $\mathbf{S}^i \in \mathbb{R}^d$, and $I(\mathbf{S}^1, \dots, \mathbf{S}^m) = 0$.⁴ In the ISA statistical model we assume that \mathbf{S} is hidden, and only n i.i.d. samples from $\mathbf{X} = \mathbf{A}\mathbf{S}$ are available for observation, where $\mathbf{A} \in \mathbb{R}^{q \times dm}$ is an unknown invertible matrix with full rank and $q \geq dm$. Based on n i.i.d. observation of \mathbf{X} , our task is to estimate the hidden sources \mathbf{S}^i and the mixing matrix \mathbf{A} . Let the estimation of \mathbf{S} be denoted by $\mathbf{Y} = (\mathbf{Y}^1; \dots; \mathbf{Y}^m) \in \mathbb{R}^{dm}$, where $\mathbf{Y} = \mathbf{W}\mathbf{X}$. The goal of ISA is to calculate $\text{argmin}_{\mathbf{W}} I(\mathbf{Y}^1, \dots, \mathbf{Y}^m)$, where $\mathbf{W} \in \mathbb{R}^{dm \times q}$ is a matrix with full rank. Following the ideas of Cardoso (1998), this ISA problem can be solved by first preprocessing the observed quantities \mathbf{X} by a traditional ICA algorithm which provides us \mathbf{W}_{ICA} estimated separation matrix⁵, and then simply grouping the estimated ICA components into ISA subspaces by maximizing the sum of the MI in the estimated subspaces, that is we have to find a permutation matrix $\mathbf{P} \in \{0, 1\}^{dm \times dm}$ which solves

$$\max_{\mathbf{P}} \sum_{j=1}^m I(Y_1^j, Y_2^j, \dots, Y_d^j). \quad (12)$$

where $\mathbf{Y} = \mathbf{P}\mathbf{W}_{ICA}\mathbf{X}$. We used the proposed copula based information estimation, \widehat{I}_α with $\alpha = 0.99$ to approximate the Shannon mutual information, and we chose $S = \{1, 2, 3\}$. Our experiment shows that this ISA algorithm using the proposed MI estimator can indeed provide good

⁴Here we need the generalization of MI to multidimensional quantities, but that is obvious by simply replacing the 1D marginals by d -dimensional ones.

⁵for simplicity we used the FastICA algorithm in our experiments (Hyvärinen et al., 2001)

estimation of the ISA subspaces. We used a standard ISA benchmark dataset from Szabó et al. (2007); we generated 2,000 i.i.d. sample points on 3D geometric wireframe distributions from 6 different sources independently from each other. These sampled points can be seen in Fig. 2a, and they represent the sources, \mathbf{S} . Then we mixed these sources by a randomly chosen invertible matrix $\mathbf{A} \in \mathbb{R}^{18 \times 18}$. The six 3-dimensional projections of $\mathbf{X} = \mathbf{A}\mathbf{S}$ observed quantities are shown in Fig. 2b. Our task was to estimate the original sources \mathbf{S} using the sample of the observed quantity \mathbf{X} only. By estimating the MI in (12), we could recover the original subspaces as it can be seen in Fig. 2c. The successful subspace separation is shown in the form of Hinton diagrams as well, which is the product of the estimated ISA separation matrix $\mathbf{W} = \mathbf{P}\mathbf{W}_{ICA}$ and \mathbf{A} . It is a block permutation matrix if and only if the subspace separation is perfect (Fig. 2d).

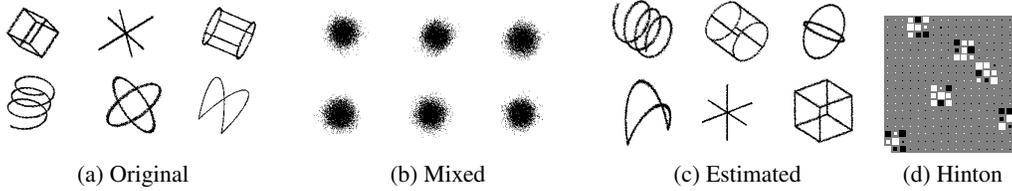


Figure 2: ISA experiment for six 3-dimensional sources.

7 Further Related Works

As it was pointed out earlier, in this paper we heavily built on the results known from the theory of Euclidean functionals (Steele, 1997; Redmond and Yukich, 1996; Koo and Lee, 2007). However, now we can be more precise about earlier work concerning nearest-neighbor based Euclidean functionals: The closest to our work is Section 8.3 of Yukich (1998), where the case of NN_S graph based p -power weighted Euclidean functionals with $S = \{1, 2, \dots, k\}$ and $p = 1$ was investigated.

Nearest-neighbor graphs have first been proposed for Shannon entropy estimation by Kozachenko and Leonenko (1987). In particular, in the mentioned work only the case of NN_S graphs with $S = \{1\}$ was considered. More recently, Gorja et al. (2005) generalized this approach to $S = \{k\}$ and proved the resulting estimator's weak consistency under some conditions on the density. The estimator in this paper has a form quite similar to that of ours:

$$\tilde{H}_1 = \log(n-1) - \psi(k) + \log\left(\frac{2\pi^{d/2}}{d\Gamma(d/2)}\right) + \frac{d}{n} \sum_{i=1}^n \log \|\mathbf{e}_i\|.$$

Here ψ stands for the digamma function, and \mathbf{e}_i is the directed edge pointing from \mathbf{X}_i to its k^{th} nearest-neighbor. Comparing this with (5), unsurprisingly, we find that the main difference is the use of the logarithm function instead of $|\cdot|^p$ and the different normalization. As mentioned before, Leonenko et al. (2008) proposed an estimator that uses the NN_S graph with $S = \{k\}$ for the purpose of estimating the Rényi entropy. Their estimator takes the form

$$\tilde{H}_\alpha = \frac{1}{1-\alpha} \log\left(\frac{n-1}{n} V_d^{1-\alpha} C_k^{1-\alpha} \sum_{i=1}^n \frac{\|\mathbf{e}_i\|^{d(1-\alpha)}}{(n-1)^\alpha}\right),$$

where Γ stands for the Gamma function, $C_k = \left[\frac{\Gamma(k)}{\Gamma(k+1-\alpha)}\right]^{1/(1-\alpha)}$ and $V_d = \pi^{d/2}\Gamma(d/2+1)$ is the volume of the d -dimensional unit ball, and again \mathbf{e}_i is the directed edge in the NN_S graph starting from node \mathbf{X}_i and pointing to the k -th nearest node. Comparing this estimator with (5), it is apparent that it is (essentially) a special case of our NN_S based estimator. From the results of Leonenko et al. (2008) it is obvious that the constant γ in (5) can be found in analytical form when $S = \{k\}$. However, we kindly warn the reader again that the proofs of these last three cited articles (Kozachenko and Leonenko, 1987; Gorja et al., 2005; Leonenko et al., 2008) contain a few errors, just like the Wang et al. (2009b) paper for KL divergence estimation from two samples. Kraskov et al. (2004) also proposed a k -nearest-neighbors based estimator for the Shannon mutual information estimation, but the theoretical properties of their estimator are unknown.

8 Conclusions and Open Problems

We have studied Rényi entropy and mutual information estimators based on NN_S graphs. The estimators were shown to be strongly consistent. In addition, we derived upper bounds on their convergence rate under some technical conditions. Several open problems remain unanswered:

An important open problem is to understand how the choice of the set $S \subset \mathbb{N}^+$ affects our estimators. Perhaps, there exists a way to choose S as a function of the sample size n (and d, p) which strikes the optimal balance between the bias and the variance of our estimators.

Our method can be used for estimation of Shannon entropy and mutual information by simply using α close to 1. The open problem is to come up with a way of choosing α , approaching 1, as a function of the sample size n (and d, p) such that the resulting estimator is consistent and converges as rapidly as possible. An alternative is to use the logarithm function in place of the power function. However, the theory would need to be changed significantly to show that the resulting estimator remains strongly consistent.

In the proof of consistency of our mutual information estimator \hat{I}_α we used Kiefer-Dvoretzky-Wolfowitz theorem to handle the effect of the inaccuracy of the empirical copula transformation. Our particular use of the theorem seems to restrict α to the interval $(1/2, 1)$ and the dimension to values larger than 2. Is there a better way to estimate the error caused by the empirical copula transformation and prove consistency of the estimator for a larger range of α 's and $d = 1, 2$?

Finally, it is an important open problem to prove bounds on converge rates for densities that have higher order smoothness (*i.e.* β -Hölder smooth densities). A related open problem, in the context of theory of Euclidean functionals, is stated in Koo and Lee (2007).

Acknowledgements

This work was supported in part by AICML, AITF (formerly iCore and AIF), NSERC, the PASCAL2 Network of Excellence under EC grant no. 216886 and by the Department of Energy under grant number DESC0002607. Cs. Szepesvári is on leave from SZTAKI, Hungary.

References

- C. Adami. Information theory in molecular biology. *Physics of Life Reviews*, 1:3–22, 2004.
- M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh. A hierarchical clustering based on mutual information maximization. In *IEEE ICIP*, pages 277–280, 2007.
- P. A. Alemany and D. H. Zanette. Fractal random walks from a variational formalism for Tsallis entropies. *Phys. Rev. E*, 49(2):R956–R958, Feb 1994.
- N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, 2nd edition, 2000.
- J. Cardoso. Multidimensional independent component analysis. *Proc. ICASSP'98, Seattle, WA.*, 1998.
- B. Chai, D. B. Walther, D. M. Beck, and L. Fei-Fei. Exploring functional connectivity of the human brain using multivariate information analysis. In *NIPS*, 2009.
- J. A. Costa and A. O. Hero. Entropic graphs for manifold learning. In *IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2003.
- J. Dedecker, P. Doukhan, G. Lang, J.R. Leon, S. Louhichi, and C Prieur. *Weak Dependence: With Examples and Applications*, volume 190 of *Lecture notes in Statistics*. Springer, 2007.
- L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.
- D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.
- M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17: 277–297, 2005.
- A. O. Hero and O. J. Michel. Asymptotic theory of greedy approximations to minimal k -point random graphs. *IEEE Trans. on Information Theory*, 45(6):1921–1938, 1999.
- A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval, 2002a. Communications and Signal Processing Laboratory Technical Report CSPL-328.

- A. O. Hero, B. Ma, O. Michel, and J. Gorman. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002b.
- D. Hilbert. Über die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen*, 38:459–460, 1891.
- K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.
- M. M. Van Hulle. Constrained subspace ICA based on mutual information optimization directly. *Neural Computation*, 20:964–973, 2008.
- A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley, New York, 2001.
- Y. Koo and S. Lee. Rates of convergence of means of Euclidean functionals. *Journal of Theoretical Probability*, 20(4):821–841, 2007.
- L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 23:9–16, 1987.
- A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.
- J. Kybic. Incremental updating of nearest neighbor-based high-dimensional entropy estimation. In *Proc. Acoustics, Speech and Signal Processing*, 2006.
- E. Learned-Miller and J. W. Fisher. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.
- N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 2008.
- J. Lewi, R. Butera, and L. Paninski. Real-time adaptive information-theoretic optimization of neurophysiology experiments. In *Advances in Neural Information Processing Systems*, volume 19, 2007.
- S. C. Milne. Peano curves and smoothness of functions. *Advances in Mathematics*, 35:129–157, 1980.
- G. Peano. Sur une qui remplit toute une aire plane. *Mathematische Annalen*, 36(1):157–160, 1890.
- H. Peng and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans On Pattern Analysis and Machine Intelligence*, 27, 2005.
- B. Póczos and A. Lőrincz. Independent subspace analysis using geodesic spanning trees. In *ICML*, pages 673–680, 2005.
- B. Póczos and A. Lőrincz. Identification of recurrent neural networks by Bayesian interrogation techniques. *Journal of Machine Learning Research*, 10:515–554, 2009.
- B. Póczos, S. Kirshner, and Cs. Szepesvári. REGO: Rank-based estimation of Rényi information using Euclidean graph optimization. In *AISTATS 2010*, 2010.
- C. Redmond and J. E. Yukich. Asymptotics for Euclidean functionals with power-weighted edges. *Stochastic processes and their applications*, 61(2):289–304, 1996.
- D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.
- C. Shan, S. Gong, and P. W. Mcowan. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2005.
- J. M. Steele. *Probability Theory and Combinatorial Optimization*. Society for Industrial and Applied Mathematics, 1997.
- Z. Szabó, B. Póczos, and A. Lőrincz. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.
- M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l’IHES*, 81(1):73–205, 1995.
- A. B. Tsybakov and E. C. van der Meulen. Root- n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*, 23:75–83, 1996.
- O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, 38:54–59, 1976.
- Q. Wang, S. R. Kulkarni, and S. Verdú. Universal estimation of information measures for analog sources. *Foundations and Trends in Communications and Information Theory*, 5(3):265–352, 2009a.
- Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k -nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009b.
- E. Wolsztynski, E. Thierry, and L. Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Process.*, 85(5):937–949, 2005.
- J. E. Yukich. *Probability Theory of Classical Euclidean Optimization Problems*. Springer, 1998.

A Quasi-Additive and Very Strong Euclidean Functionals

The basic tool to prove convergence properties of our estimators is the theory of quasi-additive Euclidean functionals developed by Yukich (1998); Steele (1997); Redmond and Yukich (1996); Koo and Lee (2007) and others. We apply this machinery to the nearest neighbor functional L_p defined in equation (3).

In particular, we use the axiomatic definition of a quasi-additive Euclidean functional from Yukich (1998) and the definition of a very strong Euclidean functional from Koo and Lee (2007) who add two extra axioms. We then use the results of Redmond and Yukich (1996) and Koo and Lee (2007) which hold for these kinds of functionals. These results determine the limit behavior of the functionals on a set of points chosen i.i.d. from an absolutely continuous distribution over \mathbb{R}^d . As we show in the following sections, the nearest neighbor functional L_p defined by equation (3) is a very strong Euclidean functional and thus both of these results apply to it.

Technically, a quasi-additive Euclidean functional is a pair of real non-negative functionals $(L_p(V), L_p^*(V, B))$ where $B \subset \mathbb{R}^d$ is a d -dimensional cube and $V \subset B$ is a finite set of points. Here, a d -dimensional *cube* is a set of the form $\prod_{i=1}^d [a^i, a^i + s]$ where $(a^1, a^2, \dots, a^d) \in \mathbb{R}^d$ is its ‘‘lower-left’’ corner and $s > 0$ is its side. The functional L_p^* is called the *boundary functional*. The common practice is to neglect L_p^* and refer to the pair $(L_p(V), L_p^*(V, B))$ simply as L_p . We provide a boundary functional L_p^* for the nearest neighbor functional L_p in the next section.

Definition 4 (Quasi-additive Euclidean functional). L_p is a quasi-additive Euclidean functional of power p if it satisfies axioms (A1)–(A7) below.

Definition 5 (Very strong Euclidean functional). L_p is a very strong Euclidean functional of power p if it satisfies axioms (A1)–(A9) below.

Axioms. For all cubes $B \subseteq \mathbb{R}^d$, any finite $V \subseteq B$, all $\mathbf{y} \in \mathbb{R}^d$, all $t > 0$,

$$L_p(\emptyset) = 0 ; \quad L_p^*(\emptyset, B) = 0 ; \quad (\text{A1})$$

$$L_p(\mathbf{y} + V) = L_p(V) ; \quad L_p^*(\mathbf{y} + V, \mathbf{y} + B) = L_p^*(V, B) ; \quad (\text{A2})$$

$$L_p(tV) = t^p L_p(V) ; \quad L_p^*(tV, tB) = t^p L_p^*(V, B) ; \quad (\text{A3})$$

$$L_p(V) \geq L_p^*(V, B) . \quad (\text{A4})$$

For all $V \subseteq [0, 1]^d$ and a partition $\{Q_i : 1 \leq i \leq m^d\}$ of $[0, 1]^d$ into m^d subcubes of side $1/m$

$$L_p(V) \leq \sum_{i=1}^{m^d} L_p(V \cap Q_i) + O(m^{d-p}) , \quad L_p^*(V, [0, 1]^d) \geq \sum_{i=1}^{m^d} L_p^*(V \cap Q_i, [0, 1]^d) - O(m^{d-p}) . \quad (\text{A5})$$

For all finite $V, V' \subseteq [0, 1]^d$,

$$|L_p(V') - L_p(V)| \leq O(|V' \Delta V|^{1-p/d}) ; \quad |L_p^*(V', [0, 1]^d) - L_p^*(V, [0, 1]^d)| \leq O(|V' \Delta V|^{1-p/d}) \quad (\text{A6})$$

For a set \mathcal{U}_n of n points drawn i.i.d. from the uniform distribution over $[0, 1]^d$,

$$|\mathbb{E} L_p(\mathcal{U}_n) - \mathbb{E} L_p^*(\mathcal{U}_n, [0, 1]^d)| \leq o(n^{1-p/d}) ; \quad (\text{A7})$$

$$|\mathbb{E} L_p(\mathcal{U}_n) - \mathbb{E} L_p^*(\mathcal{U}_n, [0, 1]^d)| \leq O(\max(n^{1-p/d-1/d}, 1)) ; \quad (\text{A8})$$

$$|\mathbb{E} L_p(\mathcal{U}_n) - \mathbb{E} L_p(\mathcal{U}_{n+1})| \leq O(n^{-p/d}) . \quad (\text{A9})$$

Axiom (A2) is translation invariance, axiom (A3) is scaling. First part of (A5) is subadditivity of L_p and second part is super-additivity of L_p^* . Axiom (A6) is smoothness and we call (A7) quasi-additivity. Axiom (A8) is a strengthening of (A7) with an explicit rate. Axiom (A9) is the add-one bound. The axioms in Koo and Lee (2007) are slightly different, however it is a routine to check that they are implied by our set of axioms.

We will use two fundamental results about Euclidean functionals. The first is (Redmond and Yukich, 1996, Theorem 2.2) and the second is essentially (Koo and Lee, 2007, Theorem 4).

Theorem 6 (Redmond-Yukich). *Let L_p be quasi-additive Euclidean functional of power $0 < p < d$. Let \mathcal{V}_n consist of n points drawn i.i.d. from an absolutely continuous distribution over $[0, 1]^d$ with common probability density function $f : [0, 1]^d \rightarrow \mathbb{R}$. Then,*

$$\lim_{n \rightarrow \infty} \frac{L_p(\mathcal{V}_n)}{n^{1-p/d}} = \gamma \int_{[0,1]^d} f^{1-p/d}(\mathbf{x}) \, d\mathbf{x} \quad \text{a.s.},$$

where $\gamma := \gamma(L_p, d)$ is a constant depending only on the functional L_p and d .

Theorem 7 (Koo-Lee). *Let L_p be a very strong Euclidean functional of power $0 < p < d$. Let \mathcal{V}_n consist of n points drawn i.i.d. from an absolutely distribution over $[0, 1]^d$ with common probability density function $f : [0, 1]^d \rightarrow \mathbb{R}$. If f is Lipschitz⁶, then*

$$\left| \frac{\mathbb{E} L_p(\mathcal{V}_n)}{n^{1-p/d}} - \gamma \int_{[0,1]^d} f^{1-p/d}(\mathbf{x}) \, d\mathbf{x} \right| \leq \begin{cases} O\left(n^{-\frac{d-p}{d(2d-p)}}\right), & \text{if } 0 < p < d-1; \\ O\left(n^{-\frac{d-p}{d(d+1)}}\right), & \text{if } d-1 \leq p < d, \end{cases}$$

where γ is the constant from Theorem 6.

Theorem 7 differs from its original statement (Koo and Lee, 2007, Theorem 4) in two ways. First, our version is restricted to Lipschitz densities. Koo and Lee prove a generalization of Theorem 7 for β -Hölder smooth density functions. The coefficient β then appears in the exponent of n in the rate. However, their result holds only for β in the interval $(0, 1]$ which does not make it very interesting. The case $\beta = 1$ corresponds to Lipschitz densities and is perhaps the most important in this range. Second, Theorem 7 has slight improvement in the rate. Koo and Lee have an extraneous $\log(n)$ factor which we remove by “correcting” their axiom (A8).

In the next section, we prove that the nearest neighbor functional L_p defined by (3) is a very strong Euclidean functional. First, in section B, we provide a boundary functional L_p^* for L_p . Then, in section C, we verify that (L_p, L_p^*) satisfy axioms (A1)–(A9). Once the verification is done, Theorem 1 follows from Theorem 6.

Theorem 2 will follow from Theorem 7 and a concentration result. We prove the concentration result in Section D and finish that section with the proof of Theorem 2. Proof of Theorem 3 requires more work—we need to deal with the effect of empirical copula transformation. We handle this in Section E by employing the classical Kiefer-Dvoretzky-Wolfowitz theorem.

B The Boundary Functional L_p^*

We start by constructing the nearest neighbor boundary functional L_p^* . For that we will need to introduce an auxiliary graph, which we call the nearest-neighbor graph *with boundary*. This graph is related to NN_S and will be useful later.

Let B be a d -dimensional cube, $V \subset B$ be finite, and $S \subset \mathbb{N}^+$ be non-empty and finite. We define nearest-neighbor graph with boundary $NN_S^*(V, B)$ to be a directed graph, with possibly parallel edges, on vertex set $V \cup \partial B$, where ∂B denotes the boundary of B . Roughly speaking, for every vertex $\mathbf{x} \in V$ and every $i \in S$ there is an edge to its “ i -th nearest-neighbor” in $V \cup \partial B$.

More precisely, we define the edges from $\mathbf{x} \in V$ as follows: Let $\mathbf{b} \in \partial B$ be the boundary point closest to \mathbf{x} . (If there are multiple boundary points that are the closest to \mathbf{x} we choose one arbitrarily.) If $(\mathbf{x}, \mathbf{y}) \in E(NN_S(V))$ and $\|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{b}\|$ then (\mathbf{x}, \mathbf{y}) also belongs to $E(NN_S^*(V, B))$. For each $(\mathbf{x}, \mathbf{y}) \in E(NN_S(V))$ such that $\|\mathbf{x} - \mathbf{y}\| > \|\mathbf{x} - \mathbf{b}\|$ we create in $NN_S^*(V, B)$ one copy of the edge (\mathbf{x}, \mathbf{b}) . In other words, there is a bijection between edge sets $E(NN_S(V))$ and $E(NN_S^*(V, B))$. An example of a graph $NN_S(V)$ and a corresponding graph $NN_S^*(V, B)$ are shown in Figure 3.

Analogously, we define $L_p^*(V, B)$ as the sum of p -powered edges of $NN_S^*(V, B)$. Formally,

$$L_p^*(V, B) = \sum_{(\mathbf{x}, \mathbf{y}) \in E(NN_S^*(V, B))} \|\mathbf{x} - \mathbf{y}\|^p. \quad (13)$$

⁶Recall that a function f is Lipschitz if there exists a constant $C > 0$ such that $|f(x) - f(y)| \leq C\|x - y\|$ for all x, y in the domain of f .

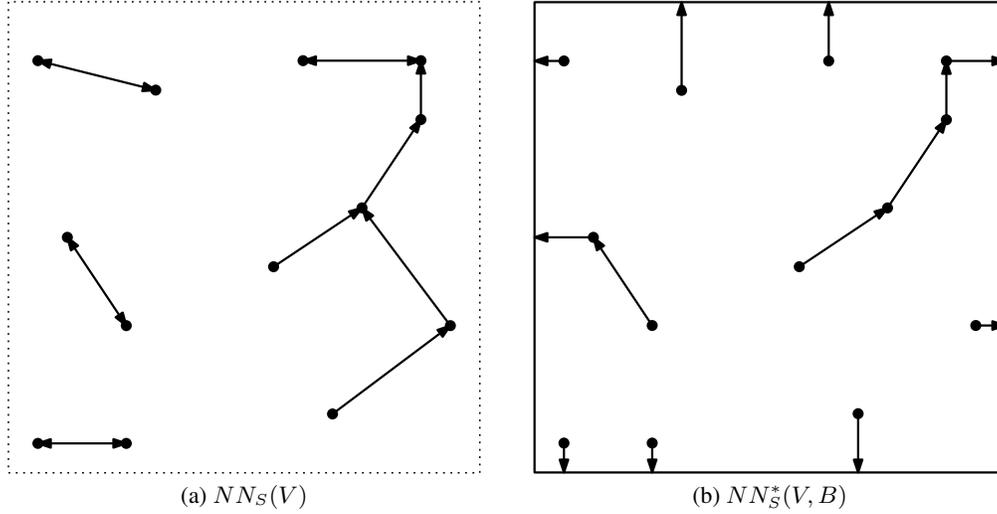


Figure 3: Figure (a) shows an example of a nearest neighbor graph $NN_S(V)$ in two dimensions and a corresponding boundary nearest neighbor graph $NN_S^*(V, B)^*$ is shown in Figure (b). We have used $S = \{1\}$, $B = [0, 1]^2$ and a set V consisting of 13 points in B .

We will need some basic geometric properties of $NN_S^*(V, B)$ and $L_p^*(V, B)$. By construction, the edges of $NN_S^*(V, B)$ are shorter than the corresponding edges of $NN_S(V)$. As an immediate consequence we get the following proposition.

Proposition 8 (Upper Bound). *For any cube B , any $p \geq 0$ and any finite set $V \subset B$, $L_p^*(V, B) \leq L_p(V)$.*

C Verification of Axioms (A1)–(A9) for (L_p, L_p^*)

It is easy to see that the nearest neighbor functional L_p and its boundary functional L_p^* satisfy axioms (A1)–(A3). Axiom (A4) is verified by Proposition 8. It thus remains to verify axioms (A5)–(A9) which we do in subsections C.1, C.2 and C.3. We start with two simple lemmas.

Lemma 9 (In-Degree). *For any finite $V \subseteq \mathbb{R}^d$ the in-degree of any vertex in $NN_S(V)$ is $O(1)$.*

Proof. Fix a vertex $\mathbf{x} \in V$. We show that the in-degree of \mathbf{x} is bounded by some constant that depends only on d and $k = \max S$. For any unit vector $\mathbf{u} \in \mathbb{R}^d$ we consider the convex open cone $C(\mathbf{x}, \mathbf{u})$ with apex at \mathbf{x} , rotationally symmetric about its axis \mathbf{u} and angle 30° :

$$Q(\mathbf{x}, \mathbf{u}) = \left\{ \mathbf{y} \in \mathbb{R}^d : \mathbf{u} \cdot (\mathbf{y} - \mathbf{x}) < \frac{\sqrt{3}}{2} \|\mathbf{u} - \mathbf{x}\| \right\}.$$

As it is well known, $\mathbb{R}^d \setminus \{\mathbf{x}\}$ can be written as a union of finitely many, possibly overlapping, cones $Q(\mathbf{x}, \mathbf{u}_1), Q(\mathbf{x}, \mathbf{u}_2), \dots, Q(\mathbf{x}, \mathbf{u}_B)$, where B depends only on the dimension d . We show that the in-degree of \mathbf{x} is at most kB .

Suppose, by contradiction, that the in-degree of \mathbf{x} is larger than kB . Then, by pigeonhole principle, there is a cone $Q(\mathbf{x}, \mathbf{u})$ containing $k + 1$ vertices of the graph with an incoming edge to \mathbf{x} . Denote these vertices $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{k+1}$ and assume that they are indexed so that $\|\mathbf{x} - \mathbf{y}_1\| \leq \|\mathbf{x} - \mathbf{y}_2\| \leq \dots \leq \|\mathbf{x} - \mathbf{y}_{k+1}\|$.

By a simple calculation, we can verify that $\|\mathbf{x} - \mathbf{y}_{k+1}\| > \|\mathbf{y}_i - \mathbf{y}_{k+1}\|$ for all $1 \leq i \leq k$. Indeed, by the law of cosines

$$\begin{aligned} \|\mathbf{y}_i - \mathbf{y}_{k+1}\|^2 &= \|\mathbf{x} - \mathbf{y}_i\|^2 + \|\mathbf{x} - \mathbf{y}_{k+1}\|^2 - 2(\mathbf{x} - \mathbf{y}_i) \cdot (\mathbf{x} - \mathbf{y}_{k+1}) \\ &< \|\mathbf{x} - \mathbf{y}_i\|^2 + \|\mathbf{x} - \mathbf{y}_{k+1}\|^2 - \|\mathbf{x} - \mathbf{y}_i\| \|\mathbf{x} - \mathbf{y}_{k+1}\| \leq \|\mathbf{x} - \mathbf{y}_{k+1}\|^2, \end{aligned}$$

where the sharp inequality follows from that $\mathbf{y}_{k+1}, \mathbf{y}_i \in Q(\mathbf{x}, \mathbf{u})$ and so the angle between vectors $(\mathbf{x} - \mathbf{y}_i)$ and $(\mathbf{x} - \mathbf{y}_{k+1})$ is strictly less than 60° , and the second inequality follows from $\|\mathbf{x} - \mathbf{y}_i\| \leq \|\mathbf{x} - \mathbf{y}_{k+1}\|$. Thus, \mathbf{x} cannot be among the k nearest-neighbors of \mathbf{y}_{k+1} which contradicts the existence of the edge $(\mathbf{y}_{k+1}, \mathbf{x})$. \square

Lemma 10 (Growth Bound). *For any $p \geq 0$ and finite $V \subset [0, 1]^d$, $L_p(V) \leq O(\max(|V|^{1-p/d}, 1))$.*

Proof. An elegant way to prove the lemma is with the use of space-filling curves.⁷ Since Peano (1890) and Hilbert (1891), it is known that there exists a continuous function ψ from the unit interval $[0, 1]$ onto the cube $[0, 1]^d$ (i.e. a surjection). For obvious reason ψ is called a space-filling curve. Moreover, there are space-filling curves which are $(1/d)$ -Hölder; see Milne (1980). In other words, we can assume that there exists a constant $C > 0$ such that

$$\|\psi(x) - \psi(y)\| \leq C|x - y|^{1/d} \quad \forall x, y \in [0, 1]. \quad (14)$$

Since ψ is a surjective function we can consider a right inverse $\psi^{-1} : [0, 1]^d \rightarrow [0, 1]$ i.e. a function such that $\psi(\psi^{-1}(x)) = x$ and we let $W = \psi^{-1}(V)$. Let $0 \leq w_1 < w_2 < \dots < w_{|V|} \leq 1$ be the points of W sorted in the increasing order. We construct a “nearest neighbor” graph G on W . For every $1 \leq j \leq |V|$ and every $i \in S$ we create a directed edge (w_j, w_{j+i}) , where the addition $i + j$ is taken modulo $|V|$. It is not hard to see that the total length of the edges of G is

$$\sum_{(x,y) \in E(G)} |x - y| \leq O(k^2) = O(1) \quad (15)$$

To see more clearly why (15) holds, note that every line segment $[w_i, w_{i+1}]$, $1 \leq i < |V|$ belongs to at most $O(k^2)$ edges and the total length of the line segments is $\sum_{i=1}^{|V|-1} (w_{i+1} - w_i) \leq 1$.

Let H be a graph on $V \subset [0, 1]^d$ isomorphic to G , where for each edge $(w_i, w_j) \in E(G)$ there is a corresponding edge $(\psi(w_i), \psi(w_j)) \in E(H)$. By the construction of H

$$L_p(V) \leq \sum_{(\mathbf{x}, \mathbf{y}) \in E(H)} \|\mathbf{x} - \mathbf{y}\|^p = \sum_{(x,y) \in E(G)} \|\psi(x) - \psi(y)\|^p. \quad (16)$$

Hölder property of ψ implies that

$$\sum_{(x,y) \in E(G)} \|\psi(x) - \psi(y)\|^p \leq C \sum_{(x,y) \in E(G)} |x - y|^{p/d}. \quad (17)$$

If $p \geq d$ then $|x - y|^{p/d} \leq |x - y|$ since $|x - y| \in [0, 1]$ and thus

$$\sum_{(x,y) \in E(G)} |x - y|^{p/d} \leq \sum_{(x,y) \in E(G)} |x - y|.$$

Chaining the last inequality with (16), (17) and (15) we obtain that $L_p(V) \leq O(1)$ for $p \geq d$.

If $0 < p < d$ we use the inequality between arithmetic and (p/d) -mean. It states that for positive numbers a_1, a_2, \dots, a_n

$$\left(\frac{\sum_{i=1}^n a_i^{p/d}}{n} \right)^{d/p} \leq \frac{\sum_{i=1}^n a_i}{n} \quad \text{or equivalently} \quad \sum_{i=1}^n a_i^{p/d} \leq n^{1-p/d} \left(\sum_{i=1}^n a_i \right)^{p/d}.$$

In our case a_i 's are the edge length of G and $n \leq k|V|$, and we have

$$\sum_{(x,y) \in E(G)} |x - y|^{p/d} \leq (k|V|)^{1-p/d} \left(\sum_{(x,y) \in E(G)} |x - y| \right)^{p/d}.$$

Combining the last inequality with (16), (17) and (15) we get that $L_p(V) \leq O(|V|^{1-p/d})$ for $0 < p < d$.

Finally, for $p = 0$, $L_p(V) \leq k|V| = O(|V|)$. \square

⁷There is an elementary proof, too, based on a discretization argument. However, this proof introduces an extraneous logarithmic factor when $p = d$.

C.1 Smoothness

In this section, we verify axiom (A6).

Lemma 11 (Smoothness of L_p). *For $p \geq 0$ and finite disjoint $V, V' \subset [0, 1]^d$, $|L_p(V' \cup V) - L_p(V')| \leq O(\max(|V|^{1-p/d}, 1))$.*

Proof. For $p \geq d$ the lemma trivially follows from the growth bound $L_p(V') = O(1)$, $L_p(V' \cup V) = O(1)$. For $0 \leq p < d$, we need to prove two inequalities:

$$L_p(V' \cup V) \leq L_p(V') + O(|V|^{1-p/d}) \quad \text{and} \quad L_p(V') \leq L_p(V' \cup V) + O(|V|^{1-p/d}).$$

We start with the first inequality. We use the obvious property of L_p that $L_p(V' \cup V) \leq L_p(V') + L_p(V) + O(1)$. Combined with the growth bound (Lemma 10) for V we get

$$L_p(V' \cup V) \leq L_p(V') + L_p(V) + O(1) \leq L_p(V') + O(|V|^{1-p/d}) + O(1) \leq L_p(V') + O(|V|^{1-p/d}).$$

The second inequality is a bit more tricky to prove. We introduce a generalized nearest-neighbor graph $NN_S(W, W')$ for any pair of finite sets W, W' such that $W \subseteq W' \subset \mathbb{R}^d$. We define $NN_S(W, W')$ as the subgraph of $NN_S(W')$ where all edges from $W' \setminus W$ are deleted. Similarly, we define $L_p(W, W')$ as the sum p -powered lengths of edges of $NN_S(W, W')$:

$$L_p(W, W') = \sum_{(x,y) \in E(NN_S(W, W'))} \|x - y\|^p.$$

We will use two obvious properties of $L_p(W, W')$ valid for any finite $W \subseteq W' \subset \mathbb{R}^d$:

$$L_p(W, W) = L_p(W) \quad \text{and} \quad L_p(W, W') \leq L_p(W) + O(1). \quad (18)$$

Let $U \subseteq V'$ be the set of vertices x such that in $NN_S(V' \cup V)$ there exists an edge from x to a vertex V . Using the two observations and the growth bound we have

$$\begin{aligned} L_p(V') &= L_p(V', V') = L_p(U, V') + L_p(V' \setminus U, V') \leq L_p(U) + O(1) + L_p(V' \setminus U, V') \\ &\leq O(|U|^{1-p/d}) + L_p(V' \setminus U, V'). \end{aligned}$$

The term $L_p(V' \setminus U, V')$ can be upper bounded by $L_p(V' \cup V)$ since by the choice of U the graph $NN_S(V' \setminus U, V')$ is a subgraph of $NN_S(V' \cup V)$. The term $O(|U|^{1-p/d})$ is at most $O(|V|^{1-p/d})$ since $|U|$ is upper bounded by the number of edges of $NN_S(V' \cup V)$ ending in V and, in turn, the number of these edges is by the in-degree lemma at most $O(|V|)$. \square

Corollary 12 (Smoothness of L_p). *For $p \geq 0$ and finite $V, V' \subset [0, 1]^d$,*

$$|L_p(V') - L_p(V)| \leq O(\max(|V' \Delta V|^{1-p/d}, 1)),$$

where $V' \Delta V$ denotes the symmetric difference.

Proof. Applying the previous lemma twice

$$\begin{aligned} |L_p(V') - L_p(V)| &\leq |L_p(V') - L_p(V' \cup V)| + |L_p(V' \cup V) - L_p(V)| \\ &= |L_p(V') - L_p(V' \cup (V \setminus V'))| + |L_p(V \cup (V' \setminus V)) - L_p(V)| \\ &\leq O(\max(|V \setminus V'|^{1-p/d}, 1)) + O(\max(|V' \setminus V|^{1-p/d}, 1)) \\ &= O(\max(|V' \Delta V|^{1-p/d}, 1)). \end{aligned}$$

\square

Lemma 13 (Smoothness of L_p^*). *For $p \geq 0$ and finite disjoint $V, V' \subset [0, 1]^d$,*

$$|L_p^*(V' \cup V, [0, 1]^d) - L_p^*(V', [0, 1]^d)| \leq O(\max(|V|^{1-p/d}, 1)).$$

Proof. The proof of the lemma is identical to the proof of Lemma 11 if we replace $L_p(\cdot)$ by $L_p^*(\cdot, [0, 1]^d)$, $NN_S^*(\cdot)$ by $NN_S(\cdot, [0, 1]^d)$, $L_p(\cdot, \cdot)$ by $L_p^*(\cdot, \cdot, [0, 1]^d)$ and $NN_S(\cdot, \cdot)$ by $NN_S^*(\cdot, \cdot, [0, 1]^d)$. We, of course, need to explain what $NN_S^*(V, W, [0, 1]^d)$ and $L_p^*(V, W, [0, 1]^d)$ mean. For $V \subseteq W$, we define $NN_S^*(V, W, [0, 1]^d)$ as the subgraph of $NN_S^*(W, [0, 1]^d)$, where the edges starting in $W \setminus V$ are removed, and $L_p^*(V, W, [0, 1]^d)$ is the sum the p -th powers of Euclidean lengths of edges of $NN_S^*(V, W, [0, 1]^d)$. \square

Corollary 14 (Smoothness of L_p^*). *For $p \geq 0$ and finite $V, V' \subset [0, 1]^d$,*

$$|L_p^*(V', [0, 1]^d) - L_p^*(V, [0, 1]^d)| \leq O(\max(|V' \Delta V|^{1-p/d}, 1)),$$

where $V' \Delta V$ denotes the symmetric difference.

Proof. The corollary is proved in exactly the same way as Corollary 12, where $L_p(\cdot)$ is replaced by $L_p^*(\cdot, [0, 1]^d)$. \square

C.2 Subadditivity and Superadditivity

In this section, we verify axiom (A5).

Lemma 15 (Subadditivity). *Let $p \geq 0$. For $m \in \mathbb{N}^+$ consider the partition $\{Q_i : 1 \leq i \leq m^d\}$ of the cube $[0, 1]^d$ into m^d disjoint subcubes⁸ of side $1/m$. For any finite $V \subset [0, 1]^d$,*

$$L_p(V) \leq \sum_{i=1}^{m^d} L_p(V \cap Q_i) + O(\max(m^{d-p}, 1)). \quad (19)$$

Proof. Consider a subcube Q_i which contains at least $k+1$ points. Using the “ $L_p(W, W')$ notation” from the proof of Lemma 11

$$L_p(V \cap Q_i, V) \leq L_p(V \cap Q_i, V \cap Q_i) = L_p(V \cap Q_i).$$

Let R be the union subcubes that contain at most k points. Clearly $|V \cap R| \leq km^d$. Then

$$\begin{aligned} L_p(V) &= L_p(V, V) \\ &= L_p(V \cap R, V) + \sum_{\substack{1 \leq i \leq m^d \\ |V \cap Q_i| \geq k+1}} L_p(V \cap Q_i, V) \\ &\leq L_p(V \cap R) + O(1) + \sum_{i=1}^{m^d} L_p(V \cap Q_i), \end{aligned}$$

where we have used the second part of (18). The proof is finished by applying the growth bound $L_p(V \cap R) \leq O(\max(|V \cap R|^{1-p/d}, 1)) \leq O(\max(m^{d-p}, 1))$. \square

Lemma 16 (Superadditivity of L_p^*). *Let $p \geq 0$. For $m \in \mathbb{N}^+$ consider a partition $\{Q_i : 1 \leq i \leq m\}$ of $[0, 1]^d$ into m^d disjoint subcubes of side $1/m$. For any finite $V \subset [0, 1]^d$,*

$$\sum_{i=1}^{m^d} L_p^*(V \cap Q_i, Q_i) \leq L_p^*(V, [0, 1]^d).$$

Proof. We construct a new graph \hat{G} by modifying the graph $NN_S^*(V, [0, 1]^d)$. Consider any edge (\mathbf{x}, \mathbf{y}) such that $\mathbf{x} \in Q_i$ and $\mathbf{y} \notin Q_i$ for some $1 \leq i \leq m^d$. Let \mathbf{z} be the point where ∂Q_i and the line segment from \mathbf{x} to \mathbf{y} intersect. In \hat{G} , we replace (\mathbf{x}, \mathbf{y}) by (\mathbf{x}, \mathbf{z}) . Note that the all edges of \hat{G} lie completely in one of the subcubes Q_i and they are shorter or equal to the corresponding edges in $NN_S^*(V, [0, 1]^d)$.

⁸In order the subcubes to be pairwise disjoint, most of them need to be semi-open and some of them closed.

Let $\hat{L}_{i,p}$ be the sum of p -th powers of the Euclidean length of the edges of \hat{G} lying in Q_i . Since edges in \hat{G} are shorter than in $NN_S^*(V, [0, 1]^d)$, $\sum_{i=1}^{m^d} \hat{L}_{p,i} \leq L_p^*(V, [0, 1]^d)$. To finish the proof it remains to show that $L_p^*(V \cap Q_i, Q_i) \leq \hat{L}_{i,p}$ for all $1 \leq i \leq m^d$.

For any edge (\mathbf{x}, \mathbf{z}) in \hat{G} from $\mathbf{x} \in V \cap Q_i$ to $\mathbf{z} \in \partial Q_i$, the point $\mathbf{z} \in \partial Q_i$ is not necessarily the closest to \mathbf{x} . Therefore, any edge in $NN_S^*(V \cap Q_i, Q_i)$ is shorter than the corresponding edge in \hat{G} . \square

C.3 Uniformly Distributed Points

Axiom (A7) is a direct consequence of axiom (A8). Hence, we are left with verifying axioms (A8) and (A9). In this section, \mathcal{U}_n denotes a set of n points chosen independently uniformly at random from $[0, 1]^d$.

Lemma 17 (Average Edge Length). *Assume $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are chosen i.i.d. uniformly at random from $[0, 1]^d$. Let k be a fixed positive integer. Let Z be the distance from \mathbf{X}_1 to k -th nearest-neighbor in $\{\mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_n\}$. For any $p \geq 0$,*

$$\mathbb{E}[Z^p \mid \mathbf{X}_1] \leq O(n^{-p/d}).$$

Proof. We denote by $B(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{y}\| \leq r\}$ the ball of radius of $r \geq 0$ centered at a point $\mathbf{x} \in \mathbb{R}^d$. Since Z lies in the interval $[0, \sqrt{d}]$ is non-negative,

$$\begin{aligned} \mathbb{E}[Z^p \mid \mathbf{X}_1] &= \int_0^\infty \Pr[Z^p > t \mid \mathbf{X}_1] dt \\ &= p \int_0^{\sqrt{d}} u^{p-1} \Pr[Z > u \mid \mathbf{X}_1] du \\ &= p \int_0^{\sqrt{d}} u^{p-1} \Pr[|\{\mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_n\} \cap B(\mathbf{X}_1, u)| < k \mid \mathbf{X}_1] du \\ &= p \int_0^{\sqrt{d}} \sum_{j=0}^{k-1} \binom{n-1}{j} u^{p-1} [\mathbf{Vol}(B(\mathbf{X}_1, u) \cap [0, 1]^d)]^j \\ &\quad \cdot [1 - \mathbf{Vol}(B(\mathbf{X}_1, u) \cap [0, 1]^d)]^{n-1-j} du \\ &\leq p \int_0^{2\sqrt{d}} \sum_{j=0}^{k-1} \binom{n-1}{j} u^{p-1} [\mathbf{Vol}(\mathbf{X}_1, u)]^j \left[1 - \left(\frac{u}{2\sqrt{d}}\right)^d\right]^{n-1-j} du. \end{aligned}$$

The last inequality follows from the obvious bound $\mathbf{Vol}(B(\mathbf{X}_1, u) \cap [0, 1]^d) \leq \mathbf{Vol}(B(\mathbf{X}_1, u))$ and that for $u \in [0, \sqrt{d}]$ the intersection $B(\mathbf{X}_1, u) \cap [0, 1]^d$ contains a cube of side at least $\frac{u}{2\sqrt{d}}$. To simplify this complicated integral, we note that $\mathbf{Vol}(B(\mathbf{X}_1, u)) = \mathbf{Vol}(B(\mathbf{X}_1, 1))u^d$ and make substitution $s = (\frac{u}{2\sqrt{d}})^d$. The last integral can be bounded by a constant multiple of

$$\sum_{j=0}^{k-1} \binom{n-1}{j} \int_0^1 s^{p/d+j-1} (1-s)^{n-1-j} ds.$$

Since $\binom{n-1}{j} = O(n^j)$ and the sum consists of only constant number of terms, it remains to show that the inner integral is $O(n^{-p/d-j})$. We can express the inner integral using the gamma function. Then, we use the asymptotic relation $\binom{n}{\epsilon} = \Theta(n^\epsilon)$ for generalized binomial coefficients $\binom{a}{b} =$

$\frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)}$ to upper-bound the result:

$$\begin{aligned} \int_0^1 s^{p/d+j-1}(1-s)^{n-1-j} ds &= \frac{\Gamma(p/d+j)\Gamma(n-j)}{\Gamma(n+p/d)} \\ &= \frac{1}{(p/d+j)\binom{n+p/d-1}{p/d+j}} \\ &= O(n^{-p/d-j}). \end{aligned} \quad \square$$

Lemma 18 (Add-One Bound). *For any $p \geq 0$, $|\mathbb{E}[L_p(\mathcal{U}_n)] - \mathbb{E}[L_p(\mathcal{U}_{n+1})]| \leq O(n^{-p/d})$.*

Proof. Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \mathbf{X}_{n+1}$ be i.i.d. points from the uniform distribution over $[0, 1]^d$. We couple \mathcal{U}_n and \mathcal{U}_{n+1} in the obvious way $\mathcal{U}_n = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$ and $\mathcal{U}_{n+1} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{n+1}\}$. Let Z be the distance from \mathbf{X}_{n+1} to k -th closest neighbor in \mathcal{U}_n . The inequality

$$L_p(\mathcal{U}_{n+1}) \leq L_p(\mathcal{U}_n) + |S|Z^p$$

holds since $|S|Z^p$ accounts for the edges from \mathbf{X}_{n+1} and since the edges from \mathcal{U}_n are shorter (or equal) in $NN_S(\mathcal{U}_{n+1})$ than the corresponding edges in \mathcal{U}_{n+1} . Taking expectations and using Lemma 17 we get

$$\mathbb{E}[L_p(\mathcal{U}_{n+1})] \leq \mathbb{E}[L_p(\mathcal{U}_n)] + O(n^{-p/d}).$$

To show the other direction of the inequality, let Z_i be the distance from \mathbf{X}_i its $(k+1)$ -th nearest point in \mathcal{U}_{n+1} . (Recall that $k = \max S$.) Let $N(j) = \{\mathbf{X}_i : (\mathbf{X}_i, \mathbf{X}_j) \in E(NN_S(\mathcal{U}_{n+1}))\}$ be the incoming neighborhood of \mathbf{X}_j . Now if we remove \mathbf{X}_j from $NN_S(V)$, the vertices in $N(j)$ lose \mathbf{X}_j as their neighbor and they need to be connected to a new neighbor in $\mathcal{U}_{n+1} \setminus \{\mathbf{X}_j\}$. This neighbor is not farther than their $(k+1)$ -th nearest-neighbor in \mathcal{U}_{n+1} . Therefore,

$$L_p(\mathcal{U}_{n+1} \setminus \{\mathbf{X}_j\}) \leq L_p(\mathcal{U}_{n+1}) + \sum_{\mathbf{X}_i \in N(j)} Z_i^p.$$

Summing over all $j = 1, 2, \dots, n+1$ we have

$$\sum_{j=1}^{n+1} L_p(\mathcal{U}_{n+1} \setminus \{\mathbf{X}_j\}) \leq (n+1)L_p(\mathcal{U}_{n+1}) + \sum_{j=1}^{n+1} \sum_{\mathbf{X}_i \in N(j)} Z_i^p.$$

The double sum on the right hand side is simply the sum over all edges of $NN_S(\mathcal{U}_{n+1})$ and so we can write

$$\sum_{j=1}^{n+1} L_p(\mathcal{U}_{n+1} \setminus \{\mathbf{X}_j\}) \leq (n+1)L_p(\mathcal{U}_{n+1}) + |S| \sum_{i=1}^{n+1} Z_i^p.$$

Taking expectations and using Lemma 17 to bound $\mathbb{E}[Z_i^p]$ we arrive at

$$(n+1)\mathbb{E}[L_p(\mathcal{U}_n)] \leq (n+1)\mathbb{E}[L_p(\mathcal{U}_{n+1})] + (n+1)O(n^{-p/d}).$$

The proof is finished by dividing through by $(n+1)$. □

Lemma 19 (Quasi-additivity). $0 \leq \mathbb{E}[L_p(\mathcal{U}_n)] - \mathbb{E}[L_p^*(\mathcal{U}_n, [0, 1]^d)] \leq O(\max(n^{1-p/d-1/d}, 1))$ for any $p \geq 0$.

Proof. The first inequality follows from Proposition 8 by taking expectation. The proof of the second inequality is much more involved. Consider the (random) subset of points $\hat{\mathcal{U}}_n \subseteq \mathcal{U}_n$ which are connected to the boundary in $NN_S^*(\mathcal{U}_n, [0, 1]^d)$ by at least one edge. We use the notation $L_p(W, W')$ for any $W \subseteq W'$ and its two properties expressed by Eq. (18) and a third obvious property $L_p(W, W') \leq L_p(W')$. We have

$$L_p(\mathcal{U}_n) = L_p(\mathcal{U}_n, \mathcal{U}_n) = L_p(\hat{\mathcal{U}}_n, \mathcal{U}_n) + L_p(\mathcal{U}_n \setminus \hat{\mathcal{U}}_n, \mathcal{U}_n) \leq L_p(\hat{\mathcal{U}}_n) + O(1) + L_p^*(\mathcal{U}_n, [0, 1]^d),$$

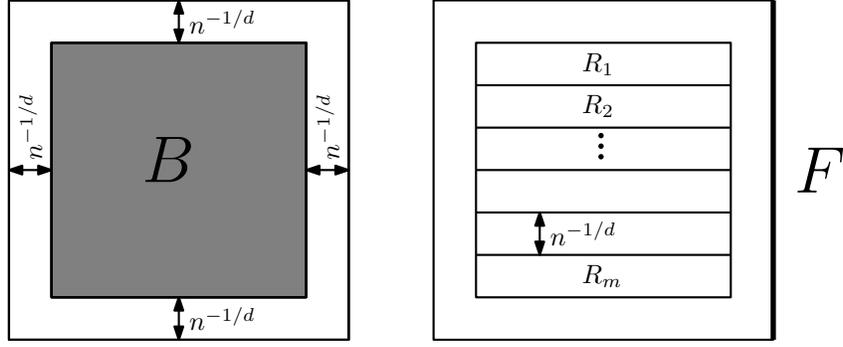


Figure 4: The left drawing shows the box $B = [n^{-1/d}, 1 - n^{-1/d}]^d \subset [0, 1]^d$ shown in gray. The right drawing shows partition of B into rectangles R_1, R_2, \dots, R_m . The diameter of the projection of each rectangle R_i on the right side F has diameter (at most) $n^{-1/d}$. In each rectangle R_i at most k points are connected to F by an edge.

where in the last step we have used that $L_p(\mathcal{U}_n \setminus \widehat{\mathcal{U}}_n, \mathcal{U}_n) \leq L_p^*(\mathcal{U}_n, [0, 1]^d)$ which holds since the edges from vertices $\mathcal{U}_n \setminus \widehat{\mathcal{U}}_n$ are the same in both graphs $NN_S(\mathcal{U}_n)$ and $NN_S^*(\mathcal{U}_n, [0, 1]^d)$. If we take expectation, we get

$$\mathbb{E}[L_p(\mathcal{U}_n)] - \mathbb{E}[L_p^*(\mathcal{U}_n, [0, 1]^d)] \leq \mathbb{E}[L_p(\widehat{\mathcal{U}}_n)] + O(1)$$

and we see that we are left to show that $\mathbb{E}[L_p(\widehat{\mathcal{U}}_n)] \leq O(\max(n^{1-p/d-1/d}, 1))$. In order to do that, we start by showing that

$$\mathbb{E}[|\widehat{\mathcal{U}}_n|] \leq O(n^{1-1/d}). \quad (20)$$

Consider the cube $B = [n^{-1/d}, 1 - n^{-1/d}]^d$. We bound $\mathbb{E}[|\widehat{\mathcal{U}}_n \cap B|]$ and $\mathbb{E}[|\widehat{\mathcal{U}}_n \cap ([0, 1]^d \setminus B)|]$ separately. The latter is easily bounded by $O(n^{1-1/d})$ since there are n points and the probability that a point lies in $[0, 1]^d \setminus B$ is $\text{Vol}([0, 1]^d \setminus B) \leq O(n^{-1/d})$. We now bound $|\widehat{\mathcal{U}}_n \cap B|$. Consider a face of F . Partition B into $m = \Theta(n^{1-1/d})$ rectangles R_1, R_2, \dots, R_m such that the perpendicular projection of any rectangle R_i , $1 \leq i \leq m$, on F has diameter at most $n^{-1/d}$ and its $(d-1)$ -dimensional volume is $\Theta(n^{1/d-1})$; see Figure 4. It is not hard to see that, in $\mathcal{U}_n \cap R_i$, only the k closest points to F can be connected to F by an edge in $NN_S^*(\mathcal{U}_n, [0, 1]^d)$. There are $2d$ faces and m rectangles and hence $|\widehat{\mathcal{U}}_n \cap B| \leq 2dkm = O(n^{1-1/d})$. We have thus proved (20).

The second key component that we need is that the expected sum of p -th powers of lengths of edges of $NN_S^*(\mathcal{U}_n, [0, 1]^d)$ that connect points in \mathcal{U}_n to $\partial[0, 1]^d$ is “small”. More precisely, for any point $\mathbf{x} \in [0, 1]^d$ let $\mathbf{b}_\mathbf{x} \in \partial[0, 1]^d$ be the boundary point closest to \mathbf{x} . We show that

$$\mathbb{E} \left[\sum_{\mathbf{X} \in \widehat{\mathcal{U}}_n} \|\mathbf{X} - \mathbf{b}_\mathbf{X}\|^p \right] \leq O(n^{1-p/d-1/d}). \quad (21)$$

We decompose the task as

$$\mathbb{E} \left[\sum_{\mathbf{X} \in \widehat{\mathcal{U}}_n} \|\mathbf{X} - \mathbf{b}_\mathbf{X}\|^p \right] = \mathbb{E} \left[\sum_{\mathbf{X} \in \widehat{\mathcal{U}}_n \cap B} \|\mathbf{X} - \mathbf{b}_\mathbf{X}\|^p \right] + \mathbb{E} \left[\sum_{\mathbf{X} \in \widehat{\mathcal{U}}_n \cap ([0, 1]^d \setminus B)} \|\mathbf{X} - \mathbf{b}_\mathbf{X}\|^p \right].$$

Clearly, the second term is bounded by $n^{-p/d} \mathbb{E}[|\widehat{\mathcal{U}}_n \cap ([0, 1]^d \setminus B)|] = O(n^{1-1/d-1/p})$. To bound the first term, consider a face F of the cube $[0, 1]^d$ and a rectangle R_i in the decomposition of B into R_1, R_2, \dots, R_m mentioned above. Let Z be the distance of the k -th closest point in $\widehat{\mathcal{U}}_n \cap R_i$ to F . (If $\widehat{\mathcal{U}}_n \cap R_i$ contains less than k points, we define Z to be $1 - n^{-1/d}$.) Recall that only the k closest points of $\widehat{\mathcal{U}}_n \cap R_i$ can be connected to F and this distance is bounded by Z . There are $2d$ faces, $m = O(n^{1-1/d})$ rectangles and at most k points in each rectangle connected to a face. If we

can show that $\mathbb{E}[Z^p] = O(n^{-p/d})$, we can upper bound the second term by $2dkm \cdot O(n^{-1/p}) = O(n^{1-p/d-1/d})$ from which (21) follows.

We now prove that $\mathbb{E}[Z^p] = O(n^{-p/d})$. Let $Y = Z - n^{-1/d}$. Since $\mathbb{E}[Z^p] \leq 2^p \mathbb{E}[Y^p] + 2^p n^{-p/d}$ it suffices to show that $\mathbb{E}[Y^p] = O(n^{-p/d})$. Let q be the $(d-1)$ -dimensional volume of the projection of R_i to F . Recall that $q = \Theta(n^{1/d-1})$. Since $Y \in [0, 1 - 2n^{-1/d}]$ we have

$$\begin{aligned}
\mathbb{E}[Y^p] &= p \int_0^{1-2n^{-1/d}} t^{p-1} \Pr[Y > t] dt \\
&= p \int_0^{1-2n^{-1/d}} t^{p-1} \sum_{j=0}^{k-1} \binom{n}{j} (qt)^j (1-qt)^{n-j} dt \\
&\leq pq^{-p} \int_0^1 x^{p-1} \sum_{j=0}^{k-1} \binom{n}{j} x^j (1-x)^{n-j} dx \\
&= pq^{-p} \sum_{j=0}^{k-1} \binom{n}{j} \frac{\Gamma(p+j)\Gamma(n-j+1)}{\Gamma(n+p+1)} \\
&= pq^{-p} \sum_{j=0}^{k-1} \frac{1}{(p+j)} \binom{n}{j} / \binom{n+p}{p+j} \\
&= \Theta(q^{-p} n^{-p}) = \Theta(n^{-p/d}).
\end{aligned}$$

We now use (20) and (21) to show that $\mathbb{E}[L_p(\widehat{\mathcal{U}}_n)] \leq O(\max(n^{1-p/d-1/d}, 1))$ which will finish the proof. For any point $\mathbf{X} \in \widehat{\mathcal{U}}_n$ consider the point $\mathbf{b}_\mathbf{X}$ lying on the boundary. Let $\widehat{\mathcal{V}}_n = \{\mathbf{b}_\mathbf{X} : \mathbf{X} \in \widehat{\mathcal{U}}_n\}$ and let $NN_S(\widehat{\mathcal{V}}_n)$ be its nearest-neighbor graph. Since $\widehat{\mathcal{V}}_n$ lies in a union of $(d-1)$ -dimensional faces, by the growth bound $L_p(\widehat{\mathcal{V}}_n) \leq O(\max(|\widehat{\mathcal{V}}_n|^{1-p/(d-1)}, 1))$. Thus, if $0 \leq p < d-1$ we use that $x \mapsto x^{1-p/(d-1)}$ is concave and (20), and we have

$$\begin{aligned}
\mathbb{E}[L_p(\widehat{\mathcal{V}}_n)] &\leq O\left(\mathbb{E}\left[|\widehat{\mathcal{V}}_n|^{1-p/(d-1)}\right]\right) = O\left(\mathbb{E}\left[|\widehat{\mathcal{U}}_n|^{1-p/(d-1)}\right]\right) \\
&\leq O\left(\mathbb{E}\left[|\widehat{\mathcal{U}}_n|^{1-p/(d-1)}\right]\right) \leq O(n^{1-1/d})^{1-p/(d-1)} \leq O(n^{1-p/d-1/d}).
\end{aligned}$$

If $p \geq d-1$ then $L_p(\widehat{\mathcal{V}}_n) = O(1)$. Therefore, for any $p \geq 0$

$$\mathbb{E}[L_p(\widehat{\mathcal{V}}_n)] \leq O(\max(n^{1-p/d-1/d}, 1)) \quad (22)$$

We construct a nearest-neighbor graph \widehat{G} on $\widehat{\mathcal{U}}_n$ by lifting $NN_S(\widehat{\mathcal{V}}_n)$. For every edge, $(\mathbf{b}_\mathbf{X}, \mathbf{b}_\mathbf{Y})$ in $NN_S(\widehat{\mathcal{V}}_n)$ we create an edge (\mathbf{X}, \mathbf{Y}) . Clearly, $L_p(\widehat{\mathcal{U}}_n)$ is at most the sum of p -the powers of the edges lengths of \widehat{G} . By triangle inequality, for any $p > 0$

$$\begin{aligned}
\|\mathbf{X} - \mathbf{Y}\|^p &\leq (\|\mathbf{X} - \mathbf{b}_\mathbf{X}\| + \|\mathbf{b}_\mathbf{X} - \mathbf{b}_\mathbf{Y}\| + \|\mathbf{b}_\mathbf{Y} - \mathbf{Y}\|)^p \\
&\leq 3^p (\|\mathbf{X} - \mathbf{b}_\mathbf{X}\|^p + \|\mathbf{b}_\mathbf{X} - \mathbf{b}_\mathbf{Y}\|^p + \|\mathbf{b}_\mathbf{Y} - \mathbf{Y}\|^p).
\end{aligned}$$

In-degrees and out-degrees of \widehat{G} are $O(1)$ and so if we sum over all edges of (\mathbf{X}, \mathbf{Y}) of \widehat{G} and take expectation, we get

$$\mathbb{E}[L_p(\widehat{\mathcal{U}}_n)] \leq \mathbb{E}[L_p(\widehat{\mathcal{V}}_n)] + O\left(\mathbb{E}\left[\sum_{\mathbf{X} \in \widehat{\mathcal{U}}_n} \|\mathbf{X} - \mathbf{b}_\mathbf{X}\|^p\right]\right).$$

To upper the right hand side we use (21) and (22), which proves that $\mathbb{E}[L_p(\widehat{\mathcal{U}}_n)] \leq O(\max(n^{1-p/d-1/d}, 1))$ and finishes the proof. \square

D Concentration and Estimator of Entropy

In this section, we show that if \mathcal{V}_n is a set of n points drawn i.i.d. from any distribution over $[0, 1]^d$ then $L_p(\mathcal{V}_n)$ is tightly concentrated. That is, we show that with high probability $L_p(\mathcal{V}_n)$ is within $O(n^{1/2-p/(2d)})$ its expected value. We use this result at the end of this section to give a proof of Theorem 2.

It turns out that in order to derive the concentration result, the properties of the distribution generating the points are irrelevant (even the existence of density is not necessary). The only property that we exploit is smoothness of L_p . As a technical tool, we use the isoperimetric inequality for Hamming distance and product measures. This inequality is, in turn, a simple consequence of Talagrand's isoperimetric inequality, see e.g. Dubhashi and Panconesi (2009); Alon and Spencer (2000); Talagrand (1995). To phrase the isoperimetric inequality, we use Hamming distance $H(\mathbf{x}_{1:n}, \mathbf{y}_{1:n})$ between two tuples $\mathbf{x}_{1:n} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, $\mathbf{y}_{1:n} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$ which is defined as the number of elements in which $\mathbf{x}_{1:n}$ and $\mathbf{y}_{1:n}$ disagree.

Theorem 20 (Isoperimetric Inequality). *Let $A \subset \Omega^n$ be a subset of an n -fold product of a probability space equipped with a product measure. For any $t \geq 0$ let $A_t = \{\mathbf{x}_{1:n} \in \Omega^n : \exists \mathbf{y}_{1:n} \in \Omega^n \text{ s.t. } H(\mathbf{x}_{1:n}, \mathbf{y}_{1:n}) \leq t\}$ be an expansion of A . Then, for any $t \geq 0$,*

$$\Pr[A] \Pr[\overline{A}_t] \leq \exp\left(-\frac{t^2}{4n}\right),$$

where \overline{A}_t denotes the complement of A_t with respect to Ω^n .

Theorem 21 (Concentration Around the Median). *Let \mathcal{V}_n consists of n points drawn i.i.d. from an absolutely continuous probability distribution over $[0, 1]^d$, let $0 \leq p \leq d$. For any $t > 0$,*

$$\Pr[|L_p(\mathcal{V}_n) - M(L_p(\mathcal{V}_n))| > t] \leq e^{-\Theta(t^{2d/(d-p)}/n)},$$

where $M(\cdot)$ denotes the median of a random variable.

Proof. Let $\Omega = [0, 1]^d$ and $\mathcal{V}_n = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$, where $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent. To emphasize that we are working in a product space, we use the notations $L_p(\mathbf{x}) := L_p(\{x_1, x_2, \dots, x_n\})$, $L_p(\mathbf{X}_{1:n}) := L_p(\mathcal{V}_n) = L_p(\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\})$ and $M := M(L_p(\mathbf{X}_{1:n}))$. Let $A = \{\mathbf{x} \in \Omega^n : L_p(\mathbf{x}) \leq M\}$. By smoothness of L_p there exists a constant $C > 0$ such that

$$L_p(\mathbf{x}) \leq L_p(\mathbf{y}) + C \cdot H(\mathbf{x}, \mathbf{y})^{1-p/d}.$$

Therefore, $L_p(\mathbf{x}) > M + t$ implies that $\mathbf{x} \in \overline{A_{(t/C)^{d/(d-p)}}$. Hence for a random $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$

$$\Pr[L_p(\mathbf{X}_{1:n}) > M + t] \leq \Pr[\mathbf{X}_{1:n} \in \overline{A_{(t/C)^{d/(d-p)}}}] \leq \frac{1}{\Pr[A]} e^{-\Theta(t^{2d/(d-p)}/n)}$$

by the isoperimetric inequality. Similarly, we set $B = \overline{A}$ and note that by smoothness we have also the reversed inequality

$$L_p(\mathbf{y}) \leq L_p(\mathbf{x}) + C \cdot H(\mathbf{x}, \mathbf{y})^{1-p/d}.$$

Therefore, $L_p(\mathbf{x}) < M + t$ implies that $\mathbf{x} \in \overline{B_{(t/C)^{d/(d-p)}}$. By the same argument as before

$$\Pr[L_p(\mathbf{X}_{1:n}) < M + t] \leq \Pr[\mathbf{X}_{1:n} \in \overline{B_{(t/C)^{d/(d-p)}}}] \leq \frac{1}{\Pr[B]} e^{-\Theta(t^{2d/(d-p)}/n)}.$$

The theorem follows by the union bound and the fact that $\Pr[A] = \Pr[B] = 1/2$. \square

Corollary 22 (Deviation of the Mean and the Median). *Let \mathcal{V}_n consists of n points drawn i.i.d. from an absolutely continuous probability distribution over $[0, 1]^d$, let $0 \leq p \leq d$ and $S \subset \mathbb{N}^+$ a finite set. Then*

$$|\mathbb{E}[L_p(\mathcal{V}_n)] - M(L_p(\mathcal{V}_n))| \leq O(n^{1/2-p/(2d)}).$$

Proof. For conciseness let $L_p = L_p(\mathcal{V}_n)$ and $M = M(L_p(\mathcal{V}_n))$. We have

$$\begin{aligned} |\mathbb{E}[L_p] - M| &\leq \mathbb{E}|L_p - M| \\ &= \int_0^\infty \Pr[|L_p - M| > t] dt \\ &\leq \int_0^\infty e^{-\Theta(t^{2d/(d-p)}/n)} dt \\ &= \Theta(n^{1/2-p/(2d)}) . \end{aligned} \quad \square$$

Putting these pieces together we arrive at what we wanted to prove:

Corollary 23 (Concentration). *Let \mathcal{V}_n consists of n points drawn i.i.d. from an absolutely continuous probability distribution over $[0, 1]^d$, let $0 \leq p \leq d$ and $S \subset \mathbb{N}^+$ and finite. For any $\delta > 0$ with probability at least $1 - \delta$,*

$$|\mathbb{E}[L_p(\mathcal{V}_n)] - L_p(\mathcal{V}_n)| \leq O(n \log(1/\delta))^{1/2-p/(2d)} . \quad (23)$$

Proof of Theorem 2. By scaling and translation, we can assume that the support of μ is contained in the unit cube $[0, 1]^d$. The first part of the theorem follows immediately from Theorem 6. To prove the second part observe from (23) that for any $\delta > 0$ with probability at least $1 - \delta$,

$$\left| \frac{\mathbb{E}[L_p(\mathcal{V}_n)]}{\gamma n^{1-p/d}} - \frac{L_p(\mathcal{V}_n)}{\gamma n^{1-p/d}} \right| \leq O\left(n^{-1/2+p/(2d)} (\log(1/\delta))^{1/2-p/(2d)}\right) . \quad (24)$$

It is easy to see that if $0 < p \leq d - 1$ then $-1/2 + p/(2d) < -\frac{d-p}{d(2d-p)} < 0$, and if $d - 1 \leq p < d$ then $-1/2 + p/(2d) < -\frac{d-p}{d(d+1)} < 0$. Now using (24), Theorem 7 and the triangle inequality, we have that for any $\delta > 0$ with probability at least $1 - \delta$,

$$\begin{aligned} \left| \frac{L_p(\mathcal{V}_n)}{\gamma n^{1-p/d}} - \int_{[0,1]^d} f^{1-p/d}(\mathbf{x}) d\mathbf{x} \right| &\leq \left| \frac{\mathbb{E}[L_p(\mathcal{V}_n)]}{\gamma n^{1-p/d}} - \frac{L_p(\mathcal{V}_n)}{\gamma n^{1-p/d}} \right| \\ &\quad + \left| \frac{\mathbb{E}[L_p(\mathcal{V}_n)]}{\gamma n^{1-p/d}} - \int_{[0,1]^d} f^{1-p/d}(\mathbf{x}) d\mathbf{x} \right| \\ &\leq \begin{cases} O\left(n^{-\frac{d-p}{d(2d-p)}} (\log(1/\delta))^{1/2-p/(2d)}\right), & \text{if } 0 < p < d - 1 ; \\ O\left(n^{-\frac{d-p}{d(d+1)}} (\log(1/\delta))^{1/2-p/(2d)}\right), & \text{if } d - 1 \leq p < d . \end{cases} \end{aligned}$$

To finish the proof of (7) exploit the fact that $\log(1 \pm x) = \pm O(x)$ for $x \rightarrow 0$. □

E Copulas and Estimator of Mutual Information

The goal of this section is to prove Theorem 3 on convergence of the estimator \widehat{I}_α . The main additional problem that we need to deal with in the proof is the effect of the empirical copula transformation. A version of the classical Kiefer-Dvoretzky-Wolfowitz theorem due to Massart gives a convenient way to do it; see e.g. Devroye and Lugosi (2001).

Theorem 24 (Kiefer-Dvoretzky-Wolfowitz). *Let X_1, X_2, \dots, X_n be an i.i.d. sample from a probability distribution over \mathbb{R} with c.d.f. $F : \mathbb{R} \rightarrow [0, 1]$. Define the empirical c.d.f.*

$$\widehat{F}(x) = \frac{1}{n} |\{i : 1 \leq i \leq n, X_i \leq x\}| \quad \text{for } x \in \mathbb{R} .$$

Then, for any $t \geq 0$,

$$\Pr \left[\sup_{x \in \mathbb{R}} |F(x) - \widehat{F}(x)| > t \right] \leq 2e^{-2nt^2} .$$

As a simple consequence of the Kiefer-Dvoretzky-Wolfowitz theorem, we can derive that $\widehat{\mathbf{F}}$ is a good approximation of \mathbf{F} .

Lemma 25 (Convergence of Empirical Copula). *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be an i.i.d. sample from a probability distribution over \mathbb{R}^d with marginal c.d.f.'s F_1, F_2, \dots, F_d . Let \mathbf{F} be the copula defined by (9) and let $\widehat{\mathbf{F}}$ be the empirical copula transformation defined by (10). Then, for any $t \geq 0$,*

$$\Pr \left[\sup_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x})\|_2 > t \right] \leq 2de^{-2ndt^2}.$$

Proof. Using $\|\cdot\|_2 \leq \sqrt{d}\|\cdot\|_\infty$ in \mathbb{R}^d and union-bound we have

$$\begin{aligned} \Pr \left[\sup_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x})\|_2 > t \right] &\leq \Pr \left[\sup_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x})\|_\infty > t\sqrt{d} \right] \\ &= \Pr \left[\sup_{x \in \mathbb{R}} \max_{1 \leq j \leq d} |F_j(x) - \widehat{F}_j(x)| > t\sqrt{d} \right] \\ &\leq \sum_{i=1}^d \Pr \left[\sup_{x \in \mathbb{R}} |F_i(x) - \widehat{F}_i(x)| > t\sqrt{d} \right] \\ &\leq 2de^{-2ndt^2}. \quad \square \end{aligned}$$

The following corollary is an obvious consequence of this lemma:

Corollary 26 (Convergence of Empirical Copula). *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be an i.i.d. sample from a probability distribution over \mathbb{R}^d with marginal c.d.f.'s F_1, F_2, \dots, F_d . Let \mathbf{F} be the copula defined by (9), and let $\widehat{\mathbf{F}}$ be the empirical copula transformation defined by (10). Then, for any $\delta > 0$,*

$$\Pr \left[\max_{1 \leq i \leq n} \|\mathbf{F}(\mathbf{X}_i) - \widehat{\mathbf{F}}(\mathbf{X}_i)\| < \sqrt{\frac{\log(2d/\delta)}{2nd}} \right] \geq 1 - \delta. \quad (25)$$

Proposition 27 (Order statistics). *Let a_1, a_2, \dots, a_m and b_1, b_2, \dots, b_m be real numbers. Let $a_{(1)} \leq a_{(2)} \leq \dots \leq a_{(m)}$ and $b_{(1)} \leq b_{(2)} \leq \dots \leq b_{(m)}$ be the same numbers sorted in ascending order. Then, $|a_{(i)} - b_{(i)}| \leq \max_j |a_j - b_j|$, for all $1 \leq i \leq m$.*

Proof. The proof is left as an exercise for the reader. □

Lemma 28 (Perturbation). *Consider points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n \in \mathbb{R}^d$ such that $\|\mathbf{x}_i - \mathbf{y}_i\| < \epsilon$ for all $1 \leq i \leq n$. Then,*

$$|L_p(\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) - L_p(\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\})| \leq \begin{cases} O(n\epsilon^p), & \text{if } 0 < p < 1; \\ O(n\epsilon), & \text{if } 1 \leq p. \end{cases}$$

Proof. Let $k = \max S$, $A = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ and $B = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$. Let $w_A(i, j) = \|\mathbf{x}_i - \mathbf{x}_j\|^p$ and $w_B(i, j) = \|\mathbf{y}_i - \mathbf{y}_j\|^p$ be the edge weights defined by A and B respectively. Let $a_{(j)}^i$ be the p -th power of the distance from \mathbf{x}_i to its j -th nearest-neighbor in A , for $1 \leq i \leq n, 1 \leq j \leq n-1$. Similarly, let $b_{(j)}^i$ be the p -th power of the distance from \mathbf{y}_i to its j -th nearest-neighbor in B . Note that for any i , if we sort the real numbers $w_A(i, 1), \dots, w_A(i, i-1), w_A(i, i+1), \dots, w_A(i, n)$, then we get $a_{(1)}^i \leq a_{(2)}^i \leq \dots \leq a_{(n-1)}^i$. Similarly for w_B 's and $b_{(j)}^i$'s. Using these notations we

can write

$$\begin{aligned}
|L_p(A) - L_p(B)| &= \left| \sum_{i=1}^n \sum_{j \in S} a_{(j)}^i - b_{(j)}^i \right| \\
&\leq \sum_{i=1}^n \sum_{j \in S} |a_{(j)}^i - b_{(j)}^i| \\
&\leq \sum_{i=1}^n \sum_{j \in S} \max_{1 \leq i, j \leq n} |a_{(j)}^i - b_{(j)}^i| \\
&\leq \sum_{i=1}^n \sum_{j \in S} \max_{i, j} |w_A(i, j) - w_B(i, j)| \\
&\leq kn \max_{1 \leq i, j \leq n} |w_A(i, j) - w_B(i, j)|.
\end{aligned}$$

The third inequality follows from Proposition 27. It remains to bound $|w_A(i, j) - w_B(i, j)|$. We consider two cases:

Case $0 < p < 1$. Using $|u^p - v^p| \leq |u - v|^p$ valid for any $u, v \geq 0$ and the triangle inequality

$$\left| \|\mathbf{a} - \mathbf{b}\| - \|\mathbf{c} - \mathbf{d}\| \right| \leq \|\mathbf{a} - \mathbf{c}\| + \|\mathbf{b} - \mathbf{d}\| \quad (26)$$

valid for any $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} \in \mathbb{R}^d$ we have

$$\begin{aligned}
|w_A(i, j) - w_B(i, j)| &= \left| \|\mathbf{x}_i - \mathbf{x}_j\|^p - \|\mathbf{y}_i - \mathbf{y}_j\|^p \right| \\
&\leq \left| \|\mathbf{x}_i - \mathbf{x}_j\| - \|\mathbf{y}_i - \mathbf{y}_j\| \right|^p \\
&\leq (\|\mathbf{x}_i - \mathbf{y}_i\| + \|\mathbf{x}_j - \mathbf{y}_j\|)^p \\
&\leq 2^p \epsilon^p.
\end{aligned}$$

Case $p \geq 1$. Consider the function $f(u) = u^p$ on interval $[0, \sqrt{d}]$. On this interval $|f'(u)| \leq pd^{(p-1)/2}$ and so f is Lipschitz with constant $pd^{(p-1)/2}$. In other words, for any $u, v \in [0, \sqrt{d}]$, $|u^p - v^p| \leq pd^{(p-1)/2}|u - v|$. Thus

$$\begin{aligned}
|w_A(i, j) - w_B(i, j)| &= \left| \|\mathbf{x}_i - \mathbf{x}_j\|^p - \|\mathbf{y}_i - \mathbf{y}_j\|^p \right| \\
&\leq pd^{(p-1)/2} \left| \|\mathbf{x}_i - \mathbf{x}_j\| - \|\mathbf{y}_i - \mathbf{y}_j\| \right| \\
&\leq pd^{(p-1)/2} (\|\mathbf{x}_i - \mathbf{y}_i\| + \|\mathbf{x}_j - \mathbf{y}_j\|) \\
&\leq 2\epsilon pd^{(p-1)/2},
\end{aligned}$$

where the second inequality follows from (26). \square

Corollary 29 (Copula Perturbation). *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be an i.i.d. sample from a probability distribution over \mathbb{R}^d with marginal c.d.f.'s F_1, F_2, \dots, F_d . Let \mathbf{F} be the copula defined by (9) and let $\widehat{\mathbf{F}}$ be the empirical copula transformation defined by (10). Let $\mathbf{Z}_i = \mathbf{F}(\mathbf{X}_i)$ and $\widehat{\mathbf{Z}}_i = \widehat{\mathbf{F}}(\mathbf{X}_i)$. Then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\left| \frac{L_p(\mathbf{Z}_{1:n})}{\gamma n^{1-p/d}} - \frac{L_p(\widehat{\mathbf{Z}}_{1:n})}{\gamma n^{1-p/d}} \right| \leq \begin{cases} O(n^{p/d-p/2}(\log(1/\delta))^{p/2}), & \text{if } 0 < p < 1; \\ O(n^{p/d-1/2}(\log(1/\delta))^{1/2}), & \text{if } 1 \leq p. \end{cases}$$

Proof. It follows immediately from Corollary 26 and Lemma 28 that with probability at least $1 - \delta$,

$$|L_p(\{\mathbf{Z}_{1:n}\}) - L_p(\widehat{\mathbf{Z}}_{1:n})| \leq \begin{cases} O(n^{1-p/2}(\log(1/\delta))^{p/2}), & \text{if } 0 < p < 1; \\ O(n^{1/2}(\log(1/\delta))^{1/2}), & \text{if } 1 \leq p. \end{cases}$$

\square

We are now ready to give the proof of Theorem 3.

Proof of Theorem 3. Let g denote the density of the copula of μ . The first part follows from (6), Corollary 29 and a standard Borel-Cantelli argument with $\delta = 1/n^2$. Corollary 29 puts the restrictions $d \geq 3$ and $1/2 < \alpha < 1$.

The second part can be proved along the same lines. From (7) we have that for any $\delta > 0$ with probability at least $1 - \delta$,

$$\left| \frac{L_p(\mathbf{Z}_{1:n})}{\gamma n^{1-p/d}} - \int_{[0,1]^d} g^{1-p/d}(\mathbf{x}) \, d\mathbf{x} \right| \leq \begin{cases} O\left(n^{-\frac{d-p}{d(2d-p)}} (\log(1/\delta))^{1/2-p/(2d)}\right), & \text{if } 0 < p < d-1; \\ O\left(n^{-\frac{d-p}{d(d+1)}} (\log(1/\delta))^{1/2-p/(2d)}\right), & \text{if } d-1 \leq p < d. \end{cases}$$

Hence using the triangle inequality again, and exploiting that $(\log(1/\delta))^{1/2-p/(2d)} < (\log(1/\delta))^{1/2}$ if $0 < p, \delta < 1$, we have that with probability at least $1 - \delta$,

$$\left| \frac{L_p(\widehat{\mathbf{Z}}_{1:n})}{\gamma n^{1-p/d}} - \int_{[0,1]^d} g^{1-p/d}(\mathbf{x}) \, d\mathbf{x} \right| \leq \begin{cases} O\left(\max\{n^{-\frac{d-p}{d(2d-p)}}, n^{-p/2+p/d}\} \sqrt{\log(1/\delta)}\right), & \text{if } 0 < p \leq 1; \\ O\left(\max\{n^{-\frac{d-p}{d(2d-p)}}, n^{-1/2+p/d}\} \sqrt{\log(1/\delta)}\right), & \text{if } 1 \leq p \leq d-1; \\ O\left(\max\{n^{-\frac{d-p}{d(d+1)}}, n^{-1/2+p/d}\} \sqrt{\log(1/\delta)}\right), & \text{if } d-1 \leq p < d. \end{cases}$$

To finish the proof exploit that when $x \rightarrow 0$ then $\log(1 \pm x) = \pm O(x)$. □