Journal of Machine Learning Research ? (2011) ? Submitted ?; Published ?

Nonparametric Divergence Estimation and its Applications to Machine Learning

Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Liang Xiong LXIONG@CS.CMU.COM
Jeff Schneider SCHNEIDE@CS.CMU.EDU
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Editor: ?

Abstract

Low-dimensional embedding, manifold learning, clustering, classification, and anomaly detection are among the most important problems in machine learning. Here we consider the setting where each instance of the inputs corresponds to a continuous probability distribution. These distributions are unknown to us, but we are given some i.i.d. samples from each of them. While most of the existing machine learning methods operate on *points*, i.e. finite-dimensional feature vectors, in our setting we study algorithms that operate on *groups*, i.e. sets of feature vectors. For this purpose, we propose new nonparametric, consistent estimators for a large family of divergences and describe how to apply them for machine learning problems. As important special cases, the estimators can be used to estimate Rényi, Tsallis, Kullback-Leibler, Hellinger, Bhattacharyya distance, L_2 divergences, and mutual information. We present empirical results on synthetic data, real word images, and astronomical data sets.

1. Introduction

Consider the following problem where we have several independent groups of people, and the groups might have different size. In each group we make some measurements of the people, for example we measure their blood pressure. Suppose that in each group there is a well-defined distribution of blood pressure, and each measurement is an i.i.d. sample from this distribution. The question we want to study is how different these groups are from each other. In particular, is it possible to arrange the groups into some natural clusters using the measurements? Can we embed the distributions (i.e. the groups) into a small-dimensional space preserving proximity where they would reveal some structure? Can we detect interesting, unusual groups? It can happen that each measurement in a group looks normal, that is the blood pressure values are in the same normal range, but the distributions of these values might be different from those of other groups. Can we detect these anomalous groups? The standard anomaly/novelty detection only focuses on finding individual points (Chandola et al., 2009). Our group anomaly detection task, however, is different; we want to find anomalous groups of points in which each individual point can be normal.

Similar questions arise in many other scientific research areas. Contemporary observatories, such as the Sloan Digital Sky Survey, produce a vast amount of data about galaxies and other celestial objects. It is an important question how to find anomalous clusters of galaxies, where each galaxy in the cluster is normal, but the cluster members together exhibit unusual behavior, i.e., the distribution of the feature vectors in the cluster is different than the feature vectors in other clusters, although each feature vector is normal.

Low-dimensional embedding and manifold learning are well-studied problems; several different algorithms have been proposed for this problem (Roweis and Saul, 2000; Borg and Groenen, 2005; Tenenbaum et al., 2000; Sun et al., 2010; Zhang and Zha, 2004; Belkin and Niyogi, 2003; Donoho and Grimes, 2003). These methods usually consider a fixed-dimensional feature representation and try to embed these feature vectors into a lower-dimensional space. In this paper we generalize this problem and propose a method that is able to embed distributions into lower-dimensional space. In this case the original large-dimensional space is the space of distributions. In contrast to standard manifold learning problems, here the original instances (i.e. distributions) are not known either, only a few i.i.d. samples are given from them. Our goal is to embed them into a lower-dimensional space.

Clustering and classification are also among the most frequent machine learning problems. The most well-known algorithms can only deal with fixed, finite-dimensional representations, and they are not developed to work on sets and distributions. We will show how these problems can be solved using our methods.

To study these kind of questions we need to estimate the distance between distributions. We will show a method that can estimate these "distances" for a large family of divergences including the Rényi-α (Rényi, 1961, 1970), Tsallis-α (Villmann and Haase, 2010), Kullback-Leibler, Hellinger, Bhattacharyya, and L_2 divergences. While the question of how far distributions are from each other is an important and very basic statistical problem, interestingly, we know very little about how to estimate it efficiently. If the distributions are Gaussian mixtures, then there is a closed form expression for the L_2 divergence between them. Nonetheless, we do not have closed form expressions for Rényi, Kullback-Leibler, or many other divergences.

An indirect way to obtain the desired estimates would be to use a "plug-in" estimation scheme—first, apply a consistent density estimator for the underlying densities, and then plug them into the desired formula. The unknown densities, however, are nuisance parameters in the case of divergence estimation, and we would prefer to avoid estimating them. Furthermore, density estimators usually have tunable parameters, and we may need expensive cross validation to achieve good performance. Our proposed estimators, in contrast, avoid density estimation completely, estimating the divergences directly using only simple k-nearest-neighbor statistics. We are nonetheless able to prove that the estimators are consistent under certain conditions. We also describe how to apply these estimators to mutual information.

The main contribution of our work is to introduce new, consistent, nonparametric divergence and mutual information estimators. We also propose new algorithms for clustering, classification, low-dimensional embedding, and anomaly detection when the input consists of distributions or i.i.d. sample sets (groups of points), rather than finite dimensional vectors (points) as it is the case most machine learning algorithms.

This work is an extended version of the Póczos et al. (2011) and Póczos and Schneider (2011) conference papers, where some of these results have been presented without proofs. Here we study the estimation of slightly more general density functionals, and also provide the detailed proofs of the consistency theorems.

Notation: We use the $X_n \to_p X$ and $X_n \to_d X$ notations for the convergence of random variables in probability and in distribution, respectively. $F_n \to_w F$ will denote the weak convergence of distribution functions. $\mathcal{V}(\mathcal{M})$ and $cl(\mathcal{M})$ stand for the volume and the closure of set $\mathcal{M}$, respectively. $L_1(\mathcal{M})$ denotes the set of Lebesgue measurable functions defined on the domain $\mathcal{M}$ and have finite integral over $\mathcal{M}$. We will use the $X_{1:n\setminus i}$ notation to denote the set $\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$.

Organization: In the next section we review some related work. The estimation problem is defined formally in Section 3. There we introduce several different divergence measures, and explain their most important properties. Section 4 proposes estimators for the divergences, and we also present our most important theoretical results about the asymptotic unbiasedness and consistency of the estimators. These theorems are proven in Section 5. The consistency of the estimators are illustrated in Section 6. Section 7 demonstrates the effectiveness of our proposed algorithm on several machine learning problems including clustering, classification, and low-dimensional embedding problems. For the demonstration we use syntectic data, real word images, and astronomical data sets. Finally, we conclude with a discussion of our work. Many details of the proofs can be found in the Appendix. To help understand the proofs, in the Appendix we also provide dependence charts of the main theorems and lemmas.

2. Related Work

Under certain conditions, divergence estimators can also be used to estimate entropy and mutual information. Entropy estimators are important in 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 Dind, 2005), clustering (Aghagolzadeh et al., 2007), causality detection (Hlaváckova-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), and 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), as well as for image registration (Kybic, 2006; Hero et al., 2002b,a). For further applications, see Leonenko et al. (2008a); Wang et al. (2009a). A nice introduction to the Rényi divergence and its applications can be found in van Erven and Harremoës (2010).

The closest existing work most relevant to the topic of this paper is the work of Wang et al. (2009b), who provided an estimator for the KL-divergence.[1] Hero et al. (2002a,b)

1. We note that there is an apparent error in their work; they applied the reverse Fatou lemma under conditions when it does not hold. It is not obvious how this portion of the proof can be remedied. This error originates in the work of Kozachenko and Leonenko (1987) and can also be found in other

also investigated the Rényi divergence estimation problem but assumed that one of the two density functions is known. Gupta and Srivastava (2010) developed algorithms for estimating the Shannon entropy and the KL divergence for certain parametric families.

Recently, Nguyen et al. (2009, 2010) developed methods for estimating f-divergences using their variational characterization properties. They estimate the likelihood ratio of the two underlying densities and plug that into the divergence formulas. This approach involves solving a convex minimization problem over an infinite-dimensional function space. For certain function classes defined by reproducing kernel Hilbert spaces (RKHS), however, they were able to reduce the computational load from solving infinite-dimensional problems to solving n-dimensional problems, where n denotes the sample size. When n is large, solving these convex problems can still be very demanding. Furthermore, choosing an appropriate RKHS also introduces questions regarding model selection. An appealing property of our estimator is that we do not need to solve minimization problems over function classes; we only need to calculate certain k-nearest-neighbor (k-NN) based statistics. Recently, Sricharan et al. (2010) proposed k-nearest-neighbor based methods for estimating non-linear functionals of density, but in contrast to our approach, they were interested in the case where k increases with the sample size.

Our work borrows ideas from Leonenko et al. (2008a) and Goria et al. (2005), who considered Shannon and Rényi-α entropy estimation from a single sample.[2] In contrast, we propose divergence estimators using two independent samples. Recently, Póczos et al. (2010); Pál et al. (2010) proposed a method for consistent Rényi information estimation, but this estimator also uses one sample only and cannot be used for estimating divergences. Further information and useful reviews of several different divergences can be found, e.g., in Villmann and Haase (2010), Cichocki et al. (2009), and Wang et al. (2009a).

Machine learning on distributions have been studied for example in Jebara et al. (2004), who introduced the probability product kernels to define kernels on distributions. Here a parametric family (e.g.exponential family) is fitted to the densities, and these parameters are used to estimate inner products between distributions. The Fisher kernel (Jaakkola and Haussler, 1998) also works on parametric families only. In contrast, the estimator we are going to study is completely nonparametric. Kernels on finite sets have been studied in computer vision problems as well. For example, Lyu (2005) constructed composite set kernels from simpler kernels defined on local features. It has not been studied, however, if these methods can be related to divergences between distributions.

3. Divergences

For the remainder of this work we will assume that $\mathcal{M} \subset \mathbb{R}^d$ is a measurable set with respect to the d-dimensional Lebesgue measure and that p and q are densities on $\mathcal{M}$. The set where they are strictly positive will be denoted by $supp(p)$ and $supp(q)$, respectively. We will need the definition of Csiszár's f divergence (Csiszár, 1967; Liese and Vajda, 2005).

works. Recently, Pérez-Cruz (2008) has proposed an other consistency proof for this estimator, but it also contains some errors: he applies the strong law of large numbers under conditions when it does not hold, and also assumes that convergence in probability implies almost sure convergence.

2. The original presentations of these works contained some errors; Leonenko and Pronzato (2010) provide corrections for some of these theorems.

Definition 1 (Csiszár's f-Divergence) *Let p, q be $\mathbb{R}^d \supseteq \mathcal{M} \to \mathbb{R}$ density functions, and let $f : [0, \infty) \to \mathbb{R}$ be a convex function such that $f(1) = 0$. The Csiszár's f-divergence is defined as*

$$D_f(p\|q) \doteq \int_{\mathcal{M}} f\left(\frac{q(x)}{p(x)}\right) p(x)\mathrm{d}x, \tag{1}$$

with definitions $0 \cdot f(\frac{0}{0}) = 0$, $0 \cdot f(\frac{a}{0}) = \lim_{x\to 0} x f(\frac{a}{x})$ and assuming that the integral in (1) exists.

Lemma 2 (Csiszár's f-Divergence) *For any convex function f such that $f(1) \geq 0$ and the integral in (1) exists, we have that $D_f(p\|q) \geq 0$.*

Proof

$$0 \leq f(1) = f\left(\mathbb{E}_{X\sim p}\left[\frac{q(X)}{p(X)}\right]\right) \leq \mathbb{E}_{X\sim p}\left[f\left(\frac{q(X)}{p(X)}\right)\right] = D_f(p\|q),$$

where in the last inequality we used the Jensen inequality for the convex function f. ■

An important subset of these f-divergences are the so-called α-divergences (Cichocki et al., 2008, 2009):

Definition 3 *Let p, q be $\mathbb{R}^d \supseteq \mathcal{M} \to \mathbb{R}$ density functions, and $\alpha \in \mathbb{R} \setminus \{0, 1\}$. The α-divergence $\tilde{D}_\alpha(p\|q)$ is defined as*

$$\tilde{D}_\alpha(p\|q) \doteq \frac{1}{\alpha(1-\alpha)}\left[1 - \int_{\mathcal{M}} p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x\right] \tag{2}$$

assuming the integral exists.

One can see that this is indeed a special case of the Csiszár divergence by choosing the

$$f(z) \doteq \frac{1}{\alpha(1-\alpha)}(\alpha + (1-\alpha)z - z^{(1-\alpha)}), \quad z \geq 0$$

convex function, and hence $\tilde{D}_\alpha(p\|q)$ is always nonnegative. We have to be careful with the $\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x$ term in (2), since this quantity can easily be infinity for certain αs with large absolute values, even if p and q are strictly positive on the whole domain $\mathcal{M}$. To see this consider e.g., two zero mean Gaussian distributions (Section 6.1). Closely related divergences (but not special cases) to (2) are the Rényi-α (Rényi, 1961), and the Tsallis-α (Villmann and Haase, 2010) divergences.

Definition 4 *Let p, q be $\mathbb{R}^d \supseteq \mathcal{M}\colon \to \mathbb{R}$ density functions and let $\alpha \in \mathbb{R} \setminus \{1\}$. The* Rényi-$\alpha$ divergence *is defined as*

$$R_\alpha(p\|q) \doteq \frac{1}{\alpha - 1} \log \int_{\mathcal{M}} p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x. \tag{3}$$

The Tsallis-α divergence *is defined as*

$$T_\alpha(p\|q) \doteq \frac{1}{\alpha - 1}\left(\int_{\mathcal{M}} p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x - 1\right). \tag{4}$$

Both definitions assume that the corresponding integral exists.

One can see that as $\alpha \to 1$ these divergences converge to the KL-divergence. The following lemma summarizes the behavior of these divergences.

Lemma 5

$$\begin{aligned}
&\alpha < 0 \Rightarrow R_\alpha(p\|q) \leq 0,\, T_\alpha(p\|q) \leq 0 \\
&\alpha = 0 \Rightarrow R_\alpha(p\|q) = T_\alpha(p\|q) = 0 \\
&0 < \alpha \Rightarrow R_\alpha(p\|q) \geq 0,\, T_\alpha(p\|q) \geq 0 \\
&\alpha = 1 \Rightarrow R_\alpha(p\|q) = T_\alpha(p\|q) = KL(p\|q) \geq 0
\end{aligned}$$

We mention two other important divergences that are distances too, i.e. they are symmetric and satisfy the triangle inequality as well.

Definition 6 *L_2 distance*

$$L_2(p\|q) \doteq \left(\int_{\mathcal{M}} p^2(x) - 2p(x)q(x) + q^2(x) \mathrm{d}x \right)^{1/2}.$$

Definition 7 *Hellinger distance*

$$H(p\|q) \doteq 1 - \int_{\mathcal{M}} \sqrt{p(x)q(x)} \mathrm{d}x.$$

A closely related divergence to the Hellinger distance is the Bhattacharyya divergence. This, however, does not obey the triangle inequality.

Definition 8 *Bhattacharyya divergence*

$$B(p\|q) \doteq -\log \int_{\mathcal{M}} \sqrt{p(x)q(x)} \mathrm{d}x.$$

In all of these divergences and distances the most difficult problem is the estimation of the quantity $D_{\alpha,\beta}(p\|q) \doteq \int_{\mathcal{M}} p^\alpha(x) q^\beta(x) p(x) \mathrm{d}x$ for some $\alpha, \beta \in \mathbb{R}$. Given two independent i.i.d. samples from distributions with densities p and q, respectively, we will provide an L_2-consistent estimator for $D_{\alpha,\beta}(p\|q)$.

4. Divergence Estimation

In the remainder of this paper we will heavily exploit some properties of k-NN based density estimators. In the following section we define these estimators and briefly summarize their most important properties.

4.1 k-NN Based Density Estimators

k-NN density estimators operate using only distances between the observations in a given sample and their kth nearest neighbors (breaking ties arbitrarily). Let $X_{1:n} \doteq (X_1, \ldots, X_n)$ be an i.i.d. sample from a distribution with density p, and similarly let $Y_{1:m} \doteq (Y_1, \ldots, Y_m)$ be an i.i.d. sample from a distribution having density q. Let $\rho_k(i)$ denote the Euclidean

distance of the kth nearest neighbor of X_i in the sample $X_{1:n\setminus i}$, and similarly let $\nu_k(i)$ denote the distance of the kth nearest neighbor of X_i in the sample $Y_{1:m}$. Figure 4.1 illustrate these random variables. Let $\mathcal{B}(x, R)$ denote a closed ball around $x \in \mathbb{R}^d$ with radius R, and let $\mathcal{V}\big(\mathcal{B}(x, R)\big) = \bar{c}R^d$ be its volume, where $\bar{c}$ stands for the volume of a d-dimensional unit ball. Loftsgaarden and Quesenberry (1965) define the k-NN based density estimators of p and q at X_i as follows.

Definition 9 (k-NN based density estimators)

$$\hat{p}_k(X_i) \doteq \frac{k/(n-1)}{\mathcal{V}\big(\mathcal{B}(x, \rho_k)\big)} = \frac{k}{(n-1)\bar{c}\rho_k^d(i)}, \tag{5}$$

$$\hat{q}_k(X_i) \doteq \frac{k/m}{\mathcal{V}\big(\mathcal{B}(x, \nu_k)\big)} = \frac{k}{m\bar{c}\nu_k^d(i)}. \tag{6}$$

The following theorems show the consistency of these density estimators.

Theorem 10 (k-NN estimators, convergence in probability) *If $k(n)$ denotes the number of neighbors applied at sample size n, $\lim_{n\to\infty} k(n) = \infty$, and $\lim_{n\to\infty} n/k(n) = \infty$, then $\hat{p}_{k(n)}(x) \to_p p(x)$ for almost all x.*

Theorem 11 (k-NN estimators, convergence in sup norm) *If $\lim_{n\to\infty} k(n)/\log(n) = \infty$ and $\lim_{n\to\infty} n/k(n) = \infty$, then $\lim_{n\to\infty} \sup_x \big|\hat{p}_{k(n)}(x) - p(x)\big| = 0$ almost surely.*

Note that these estimators are consistent only when $k(n) \to \infty$. We will use these density estimators in our proposed divergence estimators. However, we will keep k fixed, which implies that the *density estimators are not consistent*, but we will still be able to prove the *consistency of the divergence estimators*.

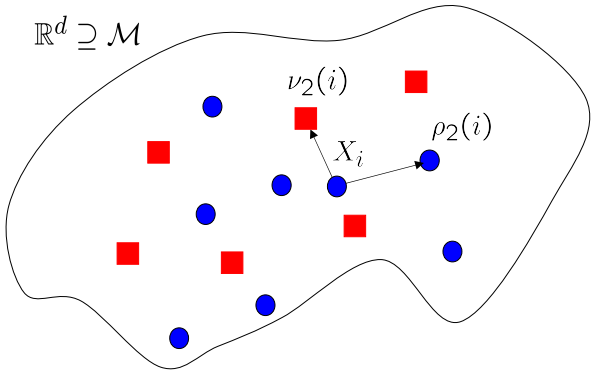


Figure 1: Calculating $\rho_k(i)$ and $\nu_k(i)$. Blue dots and red squares are i.i.d. samples from p and q respectively. We set $k = 2$ in this example.

4.2 Estimator for $D_{\alpha,\beta}(p\|q)$

In this section we introduce our estimator for $D_{\alpha,\beta}(p\|q)$ and claim its L_2 consistency.

$$D_{\alpha,\beta}(p\|q) \doteq \int_{\mathcal{M}} p^{\alpha}(x)q^{\beta}(x)p(x)\mathrm{d}x, \tag{7}$$

where $\mathcal{M} = supp(p)$. If we simply plugged (5) and (6) into (7), then we could estimate $D_{\alpha,\beta}(p\|q)$ with

$$\frac{1}{n}\sum_{i=1}^{n}\frac{k^{\alpha+\beta}}{\bar{c}^{\alpha+\beta}}(n-1)^{-\alpha}m^{-\beta}\rho_k^{-d\alpha}(i)\nu_k^{-d\beta}(i);$$

however, this estimator is asymptotically biased. We will prove that by introducing a multiplicative term the following estimator is asymptotically unbiased under certain conditions:

$$\frac{1}{n}\sum_{i=1}^{n}(n-1)^{-\alpha}m^{-\beta}\rho_k^{-d\alpha}(i)\nu_k^{-d\beta}(i)B_{k,\alpha,\beta}; \tag{8}$$

where $B_{k,\alpha,\beta} \doteq \bar{c}^{-\alpha-\beta}\frac{\Gamma(k)^2}{\Gamma(k-\alpha)\Gamma(k-\beta)}$. Notably, this multiplicative bias does not depend on p or q. The following theorems of this section contain our main results: $\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m})$ is an L_2-consistent estimator for $D_{\alpha,\beta}(p\|q)$, i.e., it is asymptotically unbiased, and the variance of the estimator is asymptotically zero.

In our theorems we will assume that almost all points of $\mathcal{M}$ are in its interior and that $\mathcal{M}$ has the following additional property:

$$\inf_{0<\delta<1}\inf_{x\in\mathcal{M}}\frac{\mathcal{V}\big(\mathcal{B}(x,\delta)\cap\mathcal{M}\big)}{\mathcal{V}\big(\mathcal{B}(x,\delta)\big)} \doteq r_{\mathcal{M}} > 0;$$

we will explain why this condition is needed later. If $\mathcal{M}$ is a finite union of bounded convex sets, then this condition holds.

Theorem 12 (Asymptotic unbiasedness) *Let $-k < \alpha, \beta < k$. If $0 < \alpha < k$, then let p be bounded away from zero and uniformly continuous. If $-k < \alpha < 0$, then let p be bounded. Similarly, If $0 < \beta < k$, then let q be bounded away from zero and uniformly continuous. If $-k < \beta < 0$, then let q be bounded. Under these conditions we have that*

$$\lim_{n,m\to\infty}\mathbb{E}\left[\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m})\right] = D_{\alpha,\beta}(p\|q),$$

i.e., the estimator is asymptotically unbiased.

The following theorems provide conditions under which $\widehat{D}_{\alpha,\beta}$ is L_2 consistent. In the previous theorems we have stated conditions that lead to asymptotically unbiased divergence estimation. In the following theorem we will assume that the estimator is asymptotically unbiased for (α,β) as well as for $(2\alpha,2\beta)$, and also assume that $D_{\alpha,\beta}(p\|q) < \infty, D_{2\alpha,2\beta}(p\|q) < \infty$.

Theorem 13 (L_2 consistency) *Let $k \geq 2$ and $-(k-1)/2 < \alpha, \beta < (k-1)/2$. If $0 < \alpha < (k-1)/2$, then let p be bounded away from zero and uniformly continuous. If $-(k-1)/2 < \alpha < 0$, then let p be bounded. Similarly, If $0 < \beta < (k-1)/2$, then let q be bounded away from zero and uniformly continuous. If $-(k-1)/2 < \beta < 0$, then let q be bounded. Under these conditions we have that*

$$\lim_{n,m\to\infty} \mathbb{E}\left[\left(\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m}) - D_{\alpha,\beta}(p\|q)\right)^2\right] = 0;$$

that is, the estimator is L_2 consistent.

5. Analysis of the Estimators

The proofs of these main theorems will require a couple of lemmas. The next section collects these tools.

5.1 Limit of Moments and Lebesgue Approximation

By the Portmanteau lemma (van der Wart, 2007), we know that the weak convergence of $X_n \to_d X$ implies that $\mathbb{E}\big[g(X_n)\big] \to \mathbb{E}\big[g(X)\big]$ for every continuous bounded function g. However, it is in general not true that if $X_n \to_d X$, then $\mathbb{E}[X_n^\gamma] \to \mathbb{E}[X^\gamma]$. The following lemma provides a sufficient condition under which this does hold.

Lemma 14 (Limit of moments, (van der Wart, 2007)) *Let $X_n \to_d X$, $0 \leq X_n$, $0 \leq X$, and $\gamma \in \mathbb{R}$. If there exists an $\varepsilon > 0$ with $\limsup_{n\to\infty} \mathbb{E}\left[X_n^{\gamma(1+\varepsilon)}\right] < \infty$, then $\lim_{n\to\infty} \mathbb{E}\left[X_n^\gamma\right] = \mathbb{E}\left[X^\gamma\right]$.*

The following lemma of Lebesgue states that any function in $L_1(\mathbb{R}^d)$ restricted to a very small ball approximately looks like a constant function.

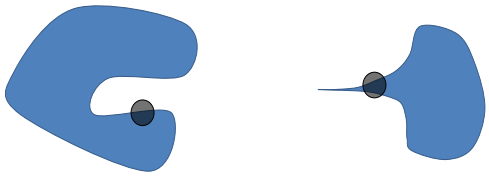
Lemma 15 (Lebesgue (1910)) *If $g \in L_1(\mathbb{R}^d)$, then for any sequence of open balls $\mathcal{B}(x, R_n)$ with radius $R_n \to 0$, and for almost all $x \in \mathbb{R}^d$,*

$$\lim_{n\to\infty} \frac{\int_{\mathcal{B}(x,R_n)} g(t)\mathrm{d}t}{\mathcal{V}\big(\mathcal{B}(x,R_n)\big)} = g(x). \tag{9}$$

This implies that if $\mathcal{M} \subset \mathbb{R}^d$ is a Lebesgue-measurable set, and $g \in L_1(\mathcal{M})$, then for any sequence of $R_n \to 0$, for any $\delta > 0$ and for almost all $x \in \mathcal{M}$, there exists an $n_0(x,\delta) \in \mathbb{Z}^+$ such that if $n > n_0(x,\delta)$, then

$$g(x) - \delta < \frac{\int_{\mathcal{B}(x,R_n)} g(t)\mathrm{d}t}{\mathcal{V}(\mathcal{B}(x,R_n))} < g(x) + \delta. \tag{10}$$

We will later require a stronger property; namely, we will need this to hold uniformly over $x \in \mathcal{M}$. However, for this requirement to hold we must put slight restrictions on the domain $\mathcal{M}$ to avoid effects around its boundary. We will consider only those domains $\mathcal{M}$ that possess the property that the intersection of $\mathcal{M}$ with an arbitrary small ball having



(a) feasible (b) not allowed

Figure 2: A possible allowed and a not-allowed domain $\mathcal{M}$ under the property in (11).

center in $\mathcal{M}$ has volume that cannot be arbitrary small relative to the volume of the ball. To be more formal, we want the following inequality to be satisfied:

$$\inf_{0<\delta<1} \inf_{x\in\mathcal{M}} \frac{\mathcal{V}\big(\mathcal{B}(x,\delta)\cap\mathcal{M}\big)}{\mathcal{V}\big(\mathcal{B}(x,\delta)\big)} \doteq r_{\mathcal{M}} > 0. \tag{11}$$

Figure 2 illustrates this notion by showing example domains that satisfy and violate this constraint. When the following property holds uniformly over $x \in \mathcal{M}$, we say that the function g is uniformly Lebesgue approximable.

Definition 16 (Uniformly Lebesgue-approximable function) *Let $g \in L_1(\mathcal{M})$. g is* uniformly Lebesgue approximable on $\mathcal{M}$ *if for any series $R_n \to 0$ and any $\delta > 0$, there exists an $n = n_0(\delta) \in \mathbb{Z}^+$ (independent of x) such that if $n > n_0$, then for almost all $x \in \mathcal{M}$,*

$$g(x) - \delta < \frac{\int_{\mathcal{B}(x,R_n)\cap\mathcal{M}} g(t)\mathrm{d}t}{\mathcal{V}\big(\mathcal{B}(x,R_n)\cap\mathcal{M}\big)} < g(x) + \delta. \tag{12}$$

This property is a uniform variant of (10). The following lemma provides examples of uniformly Lebesgue-approximable functions.

Lemma 17 *If g is uniformly continuous on $\mathcal{M}$, then it is uniformly Lebesgue approximable on $\mathcal{M}$.*

Finally, as we proceed we will frequently use the following lemma:

Lemma 18 (Moments of the Erlang distribution) *Let*

$$f_{x,k}(u) \doteq \frac{1}{\Gamma(k)}\lambda^k(x)u^{k-1}\exp(-\lambda(x)u)$$

be the density of the Erlang distribution with parameters $\lambda(x) > 0$ and $k \in \mathbb{Z}^+$. Let $\gamma \in \mathbb{R}$ such that $\gamma + k > 0$. The γth moments of this Erlang distribution can be calculated as $\int_0^\infty u^\gamma f_{x,k}(u)\ \mathrm{d}u = \lambda(x)^{-\gamma}\frac{\Gamma(k+\gamma)}{\Gamma(k)}$.

5.2 Proof Outline for Theorems 12-13

We can see from (5) that the k-NN estimation of $1/p(x)$ is simply $(n-1)\bar{c}\rho_k^d(x)/k$. Using Lemma 15, we can prove that the distribution of $(n-1)\bar{c}\rho_k^d(x)$ converges weakly to an Erlang distribution with mean $k/p(x)$, and variance $k/p^2(x)$ (Leonenko et al., 2008a). In turn, if we divide $(n-1)\bar{c}\rho_k^d(x)$ by k, then asymptotically it has mean $1/p(x)$ and variance $1/(kp^2(x))$. It implies that indeed (in accordance with Theorems 10–11) k should converge to infinity in order to get a consistent estimator, otherwise the variance will not disappear. On the other hand, k cannot grow too fast: if say $k = n-1$, then the estimator would be simply $\bar{c}\rho_k^d(x)$, which is a useless estimator since it is asymptotically zero whenever $x \in supp(p)$.

Luckily, in our case we do not need to apply consistent density estimators. The trick is that (7) has a special form; $\int p(x)p^\alpha(x)q^\beta(x)\mathrm{d}x$ form. In (8) this is estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_k(X_i)\right)^\alpha\left(\hat{q}_k(X_i)\right)^\beta B_{k,\alpha,\beta}, \tag{13}$$

where $B_{k,\alpha,\beta}$ is a correction factor that ensures asymptotic unbiasedness. Using Lemma 15, we can prove that the distributions of $\hat{p}_k(X_i)$ and $\hat{q}_k(X_i)$ converge weakly to the Erlang distribution with means $k/p(X_i)$, $k/q(X_i)$ and variances $k/p^2(X_i)$, $k/q^2(X_i)$, respectively (Leonenko et al., 2008a). Furthermore, they are conditionally independent for a given X_i. Therefore, "in the limit" (13) is simply the empirical average of the products of the αth (and βth) powers of independent Erlang distributed variables. These moments can be calculated by Lemma 18. For a fixed k, the k-NN density estimator is not consistent since its variance does not vanish. In our case, however, this variance will disappear thanks to the empirical average in (13) and the law of large numbers.

While the underlying ideas of this proof are simple, there are a couple of serious gaps in it. Most importantly, from the Lebesgue lemma (Lemma 15) we can guarantee only the weak convergence of $\hat{p}_k(X_i)$, $\hat{q}_k(X_i)$ to the Erlang distribution. From this weak convergence we cannot imply that the moments of the random variables converge too. To handle this issue, we will need stronger tools such as the concept of asymptotically uniformly integrable random variables (van der Wart, 2007), and we also need the uniform generalization of the Lebesgue lemma (Definition 16). As a result, we need to put some extra conditions on the densities p and q in Theorems 12–13. We provide the details in the subsequent sections.

5.3 Proving Asymptotic Unbiasedness

The following section contains several specific lemmas and theorems that we will use for proving the consistency of the proposed estimator (8).

5.3.1 Preliminaries

Remember that $\rho_k(j)$ is a random variable which measures the distance between X_j and its kth nearest neighbor in $X_{1:n\setminus j}$.

Lemma 19 *Let $\zeta_{n,k,1} \doteq (n-1)\rho_k^d(1)$ be a random variable, $x \in \mathbb{R}^d$, and let $F_{n,k,x}(u) \doteq \Pr(\zeta_{n,k,1} < u | X_1 = x)$ denote its conditional distribution function. Then*

$$F_{n,k,x}(u) = 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j}, \tag{14}$$

where $P_{n,u,x} \doteq \int_{\mathcal{M} \cap \mathcal{B}(x, R_n(u))} p(t)\, \mathrm{d}t$, and $R_n(u) \doteq (u/(n-1))^{1/d}$.

Here $x \in \mathbb{R}^d$, i.e. the lemma is valid in the $x \notin \mathcal{M} = supp(p)$ case too. The limit distribution of $F_{n,k,x}$ is described by the following lemma, (Leonenko et al., 2008a).

Lemma 20 *For almost all $x \in \mathcal{M}$, $F_{n,k,x} \to_w F_{k,x}$, where $F_{k,x}(u) \doteq 1 - \exp(-\lambda u) \sum_{j=0}^{k-1} \frac{(\lambda u)^j}{j!}$ is the cdf of the Erlang distribution with $\lambda = \bar{c}p(x)$.*

We use the Lebesgue approximation in the proof of this lemma, that is the reason why we can only talk about almost all $x \in \mathcal{M}$ (and not all $x \in \mathcal{M}$).

Lemma 21 *The Lemma 20 holds for almost all $x \notin \mathcal{M}$ too, i.e. when $p(x) = 0$. In this degenerate case, $\lim_{n\to\infty} F_{n,k,x}(u) = F_{k,x}(u) = 0$ for all u.*

Lemma 22 *Let $\xi_{n,k,x}$ and $\xi_{k,x}$ be random variables with $F_{n,k,x}$ and $F_{k,x}$ distribution functions, and let $\gamma \in \mathbb{R}$ be arbitrary. Then for almost all $x \in \mathcal{M}$ $\xi_{n,k,x}^\gamma \to_d \xi_{k,x}^\gamma$. If $x \notin cl(\mathcal{M})$, then $\xi_{n,k,x} \to \infty$ almost surely.*

From Lemma 22 we have that $\xi_{n,k,x} \to \infty$ almost surely when $x \notin cl(\mathcal{M})$. This immediately proves the following theorem.

Theorem 23 *If $x \notin cl(\mathcal{M})$, then*

$$\lim_{n\to\infty} \mathbb{E}\left[(n-1)^\gamma \rho_k^{d\gamma}(1) | X_1 = x\right] = \begin{cases} 0 & \text{if } \gamma < 0 \\ 1 & \text{if } \gamma = 0 \\ \infty & \text{if } 0 < \gamma \end{cases}$$

Similarly, if $x \notin cl(supp(q))$, then

$$\lim_{m\to\infty} \mathbb{E}\left[m^\gamma \nu_k^{d\gamma}(1) | X_1 = x\right] = \begin{cases} 0 & \text{if } \gamma < 0 \\ 1 & \text{if } \gamma = 0 \\ \infty & \text{if } 0 < \gamma \end{cases}$$

Theorem 24 *For almost all $x \in \mathcal{M}$ the following statements hold. If (i) $-k < \gamma < 0$, or (ii) $0 \le \gamma$, and $\int_{\mathcal{M}} \|x-y\|^\gamma p(y) \mathrm{d}y < \infty$, then*

$$\lim_{n\to\infty} \mathbb{E}\left[(n-1)^\gamma \rho_k^{d\gamma}(1) | X_1 = x\right] = (\bar{c}p(x))^{-\gamma} \frac{\Gamma(k+\gamma)}{\Gamma(k)}.$$

Similarly, if (i) $-k < \gamma < 0$, or (ii) $0 \le \gamma$, and $\int_{\mathcal{M}} \|x-y\|^\gamma q(y) \mathrm{d}y < \infty$, then for almost all $x \in supp(q)$ we have that

$$\lim_{m\to\infty} \mathbb{E}\left[m^\gamma \nu_k^{d\gamma}(1) | X_1 = x\right] = (\bar{c}q(x))^{-\gamma} \frac{\Gamma(k+\gamma)}{\Gamma(k)}.$$

Note that the conditions here are different from those given in Leonenko et al. (2008a), Leonenko et al. (2008b), Goria et al. (2005) and Wang et al. (2009b).

Proof We already know from Lemma 22 that $\xi^{\gamma}_{n,k,x} \to_d \xi^{\gamma}_{k,x}$, for almost all $x \in \mathcal{M}$. If from this it follows that $\mathbb{E}[\xi^{\gamma}_{n,k,x}] \to \mathbb{E}[\xi^{\gamma}_{k,x}]$, then

$$\lim_{n\to\infty} \mathbb{E}\left[(n-1)^{\gamma}\rho_k^{d\gamma}(1)|X_1 = x\right] = \lim_{n\to\infty} \mathbb{E}\left[\xi^{\gamma}_{n,k,x}\right] = \mathbb{E}\left[\lim_{n\to\infty}{}_{\mathrm{d}}\, \xi^{\gamma}_{n,k,x}\right] = \mathbb{E}\left[\xi^{\gamma}_{k,x}\right]$$
$$= \int_0^{\infty} u^{\gamma} f_{x,k}(u)\, \mathrm{d}u = (\bar{c}p(x))^{-\gamma}\frac{\Gamma(k+\gamma)}{\Gamma(k)},$$

assuming $k + \gamma > 0$ and using Lemma 18.

For brevity introduce the following notation:

$$f_n(x) \doteq \mathbb{E}\left[(n-1)^{-\alpha}\rho_k^{-d\alpha}(1)\Big|X_1 = x\right] \tag{15}$$
$$g_m(x) \doteq \mathbb{E}\left[m^{-\beta}\nu_k^{-d\beta}(1)\Big|X_1 = x\right]. \tag{16}$$

Theorem 23 and Theorem 24 together describe the limit properties of $f_n(x)$ and $g_m(x)$ for almost all $x \in \mathbb{R}^d$. All that remained is to prove that if $(\xi^{\gamma}_{n,k,x} \to_d \xi^{\gamma}_{k,x})$, then $(\mathbb{E}[\xi^{\gamma}_{n,k,x}] \to \mathbb{E}[\xi^{\gamma}_{k,x}])$. We are going to prove this in Theorem 26. To see this, it is enough to show (according to Theorem 14) that for some $\varepsilon > 0$, and $c(x) < \infty$ it holds that $\limsup_n \mathbb{E}[\xi^{\gamma(1+\varepsilon)}_{n,k,x}] < c(x)$. We do not need to calculate explicitly $\mathbb{E}[\xi^{\gamma(1+\varepsilon)}_{n,k,x}]$; we simply need to find a finite upper bound. ■

5.3.2 Properties of $F_{n,k,x}$

In what follows we will need a couple of properties of the $F_{n,k,x}$ distribution functions. In this section we summarize these properties. Let $\gamma > 0$, $x \in \mathcal{M} = supp(p)$, and define the following functions:

$$H(x,p,\delta,\omega) \doteq \sum_{j=0}^{k-1}\left(\frac{1}{j!}\right)^{\omega}\Gamma(\gamma+j\omega)\left(1+\frac{2\delta}{p(x)-\delta}\right)^{j\omega}\big(p(x)-\delta\big)^{-\gamma}\big((1-\delta)\omega\big)^{-\gamma-j\omega}. \tag{17}$$

$$L(x,\omega,k,\gamma,p,\delta,\delta_1) \doteq \delta_1 + \delta_1\int_{\mathcal{M}}\|x-y\|^{\gamma}p(y)\mathrm{d}y + (\bar{c}r_{\mathcal{M}})^{-\gamma}H(x,p,\delta,\omega). \tag{18}$$

Let p be bounded away from zero, and let $\underline{p} > 0$ and $\delta > 0$ so small such that $\inf_x p(x) > \underline{p} > \delta > 0$. Since p is bounded away from zero, therefore $\mathcal{M} = supp(p)$ is bounded and thus $\int_{\mathcal{M}}\|x-y\|^{\gamma}p(y)\mathrm{d}y$ and L are bounded functions of x. It implies that there exists $\bar{L} = \bar{L}(\omega,k,\gamma,p,\delta,\delta_1)$ such that $\sup_{x\in\mathcal{M}} L(x,\omega,k,\gamma,p,\delta,\delta_1) < \bar{L} < \infty$.

Theorem 25 (Properties of $F_{n,k,x}$)

1. *Let $\gamma > 0$, and let p be uniformly continuous on $\mathcal{M}$ and bounded away from zero. Let $\delta > 0$ so small that $p(x) - \delta > 0$ for all $x \in \mathcal{M}$. If $\delta_1 > 0$, and $\omega \in (0, 1]$, then there exists $n_0 = n_0(\omega, k, \gamma, p, \delta, \delta_1) \in \mathbb{Z}^+$, which is independent of x such that when $n > n_0$, then for almost all $x \in \mathcal{M}$*

$$\int_0^\infty (1 - F_{n,k,x}(u))^\omega u^{\gamma-1} \mathrm{d}u \leq \bar{L}(\omega, k, \gamma, p, \delta, \delta_1) < \infty. \tag{19}$$

For the $\omega = 1$ special case choosing $\gamma = -\alpha > 0$, we have that

$$f_n(x) \doteq \int_0^\infty u^\gamma F_{n,k,x}(\mathrm{d}u) = \gamma \int_0^\infty u^{\gamma-1}(1 - F_{n,k,x}(u))\mathrm{d}u \leq \gamma \bar{L}(1, k, \gamma, p, \delta, \delta_1) < \infty. \tag{20}$$

For brevity, we will simply write $f_n(x) \leq K(p, \gamma) < \infty$.

2. *[Non-unifrom version in x] Let $\gamma > 0$, and let $\delta(x) > 0$ so small that $p(x) - \delta(x) > 0$ for all $x \in \mathcal{M}$. [We do not require p to be bounded away from zero this time!] Let furthermore $\delta_1 > 0$, and $\omega \in (0, 1]$. Then there exists $n_0(x) \in \mathbb{Z}^+$ such that for almost all $x \in \mathcal{M}$, when $n > n_0(x)$ then*

$$\int_0^\infty u^{\gamma-1}(1 - F_{n,k,x}(u))^\omega \mathrm{d}u < L(x, \omega, k, \gamma, p, \delta(x), \delta_1) < \infty, \tag{21}$$

3. *If p is bounded above by $\bar{p}$ on $\mathbb{R}^d$, $0 \leq u \leq b$, then for all $n \in \mathbb{Z}_+$, and for all $x \in \mathbb{R}^d$*

$$F_{n,x,k}(u) \leq u^k \hat{L}(\bar{p}, b), \tag{22}$$

where $\hat{L}(\bar{p}, b) \doteq \bar{p}^k \bar{c}^k \exp(\bar{p}\bar{c}b)$. This implies that if $0 < \gamma < k$, then for all $x \in \mathbb{R}^d$

$$\lim_{u \to 0} \frac{1}{u^\gamma} F_{n,k,x}(u) = 0. \tag{23}$$

4. *If $\gamma < 0 < b$, and $\omega \in (0, 1]$, then for all $x \in \mathbb{R}^d$*

$$\int_b^\infty u^{\gamma-1}(F_{n,k,x}(u))^\omega \mathrm{d}u \leq \frac{-b^\gamma}{\gamma}. \tag{24}$$

5. *Let $-k < \gamma < 0$. If p is bounded above by $\bar{p}$ on $\mathbb{R}^d$, $0 < b$, $\omega \in (0, 1]$, and $0 < k\omega + \gamma$, then for all $n \in \mathbb{Z}_+$, and for all $x \in \mathbb{R}^d$*

$$\int_0^b u^{\gamma-1}(F_{n,k,x}(u))^\omega \mathrm{d}u \leq \hat{L}^\omega(\bar{p}, b) \frac{b^{k\omega+\gamma}}{k\omega + \gamma}. \tag{25}$$

Using this and (24) with $b = 1$ and $\omega = 1$, we have that for all $x \in \mathbb{R}^d$

$$f_n(x) \doteq (-\gamma) \int_0^\infty u^{\gamma-1} F_{n,k,x}(u)\mathrm{d}u \leq (-\gamma) \left[\frac{\hat{L}(\bar{p}, 1)}{k + \gamma} - \frac{1}{\gamma} \right] < \infty. \tag{26}$$

For brevity, we will simply write $f_n(x) \leq K(p, \gamma) < \infty$.

Theorem 26 *For almost all $x \in \mathcal{M}$ we have that (i) if $0 \leq \gamma$, $\int_{\mathcal{M}} \|x-y\|^{\gamma} p(y)\mathrm{d}y < \infty$ and $\xi^{\gamma}_{n,k,x} \to_d \xi^{\gamma}_{k,x}$, or (ii) if $-k < \gamma < 0$, and $\xi^{\gamma}_{n,k,x} \to_d \xi^{\gamma}_{k,x}$, then $\lim_{n\to\infty} \mathbb{E}[\xi^{\gamma}_{n,k,x}] = \mathbb{E}[\xi^{\gamma}_{k,x}]$.*

Proof

Case (i) of Theorem 26. Let $\gamma > 0$. We need to prove that there exists $\varepsilon > 0$ such that $\limsup_n \mathbb{E}[\xi^{\gamma(1+\varepsilon)}_{n,k,x}] < \infty$. Thanks to Lemma 41, this can be rewritten as

$$\limsup_n \int_0^{\infty} \gamma u^{\gamma(1+\varepsilon)-1}(1-F_{n,k,x}(u))\mathrm{d}u \quad < \quad \infty.$$

This, however, follows directly from (21) using $\omega = 1$. (21) holds for any $\gamma > 0$, and thus for $\gamma(1+\varepsilon)$, too, with any $\varepsilon > 0$.

Case (ii) of Theorem 26. Let $-k < \gamma < 0$. We need to prove that there exists $\varepsilon > 0$ such that $\limsup_n \mathbb{E}[\xi^{\gamma(1+\varepsilon)}_{n,k,x}] < \infty$. Thanks to Lemma 42, this can be rewritten as

$$\limsup_n \int_0^{\infty} (-\gamma) u^{\gamma(1+\varepsilon)-1} F_{n,k,x}(u)\mathrm{d}u \quad < \quad \infty.$$

This follows from (26) by choosing an appropriate $\varepsilon > 0$ such that $\gamma(1+\varepsilon) < 0$, and $k + \gamma(1+\varepsilon) > 0$.

■

Now, we are ready to put the pieces together and prove our main theorems on the asymptotic unbiasedness of the estimator (8).

5.3.3 The Proof of Theorem 12

Proof We want to prove that

$$\begin{aligned}\frac{D_{\alpha,\beta}(p\|q)}{B_{k,\alpha,\beta}} &= \lim_{n,m\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(n-1)^{-\alpha}m^{-\beta}\rho_k^{-d\alpha}(i)\nu_k^{-d\beta}(i)\right] \\ &= \lim_{n,m\to\infty} \mathbb{E}\left[(n-1)^{-\alpha}m^{-\beta}\rho_k^{-d\alpha}(1)\nu_k^{-d\beta}(1)\right].\end{aligned}$$

By exploiting the fact that $\rho_k(1)$ and $\nu_k(1)$ are independent if X_1 is given, we can see that the r.h.s. can be rewritten as

$$\lim_{n,m\to\infty} \mathbb{E}_{X_1\sim p}\left[\mathbb{E}\left[(n-1)^{-\alpha}\rho_k^{-d\alpha}(1)\middle|X_1\right]\mathbb{E}\left[m^{-\beta}\nu_k^{-d\beta}(1)\middle|X_1\right]\right] = \lim_{n,m\to\infty} \mathbb{E}_{X_1\sim p}\left[f_n(X_1)g_m(X_1)\right]$$

If we could move the limit inside the expectation, then we could apply Theorem 24 to continue the derivation as follows.

$$\begin{aligned}&\mathbb{E}_{X_1\sim p}\left[\lim_{n\to\infty}\mathbb{E}\left[(n-1)^{-\alpha}\rho_k^{-d\alpha}(1)\middle|X_1\right]\lim_{m\to\infty}\mathbb{E}\left[m^{-\beta}\nu_k^{-d\beta}(1)\middle|X_1\right]\right] \\ &\qquad = \mathbb{E}_{X_1\sim p}\left[(\bar{c}p(X_1))^{\alpha}(\bar{c}q(X_1))^{\beta}\right]\frac{\Gamma(k-\alpha)}{\Gamma(k)}\frac{\Gamma(k-\beta)}{\Gamma(k)}.\end{aligned}$$

This would complete the proof of Theorem 12. In the next section we will discuss conditions under which the outer limit can be moved inside the expectation above. ■

5.3.4 Switching Limit and Expectation

Our goal is to prove that

$$\lim_{n,m} \int_{\mathcal{M}} f_n(x) g_m(x) p(x) \mathrm{d}x = \int_{\mathcal{M}} \lim_{n,m} f_n(x) g_m(x) p(x) \mathrm{d}x, \tag{27}$$

$$\begin{aligned}
&\lim_{n,m} \int_{\mathcal{M}} f_n(x) g_m(x) p(x) \mathrm{d}x \\
&\quad = \lim_{n,m} \int_{supp(p) \setminus supp(q)} f_n(x) g_m(x) p(x) \mathrm{d}x + \lim_{n,m} \int_{supp(p) \cap supp(q)} f_n(x) g_m(x) p(x) \mathrm{d}x
\end{aligned}$$

If $\beta = 0$, then $g_m(x) = 1$ for all x. If $\alpha = 0$, then $f_n(x) = 1$ for all x. These degenerate cases are easy to analyze, thus in what follows we will assume that $\alpha \neq 0$ and $\beta \neq 0$.
Let $\mathcal{M}^* = supp(p) \setminus supp(q)$. First we prove that under the conditions of Theorem 12 the first term, $\lim_{n,m} \int_{\mathcal{M}^*} f_n(x) g_m(x) p(x) \mathrm{d}x = \int_{\mathcal{M}^*} \lim_{n,m} f_n(x) g_m(x) p(x) \mathrm{d}x$, is either zero or infinity. If it is infinity, then $D_{\alpha,\beta} = \infty$. If it is zero then this term has no contribution to $D_{\alpha,\beta}$ and it is enough to calculate the integrals on $supp(p) \cap supp(q)$.

Under the conditions of Theorem 12 we have that $0 \leq f_n(x) \leq K(p, -\alpha) < \infty$ for almost all $x \in \mathcal{M}$. When $0 < \alpha = -\gamma < k$, then this is true because of (20). When $-k < \alpha = -\gamma < 0$ then this follows from (26). When $\alpha = 0$ then $f_n(x) = 1$ for all $x \in \mathbb{R}^d$.

Let $0 < \beta < k$. From (26), we have that $g_m(x) \leq K(q, -\beta) < \infty$ for all $x \in \mathbb{R}^d$ Therefore, $f_n(x) g_m(x) \leq K(p, -\alpha) K(q, -\beta) < \infty$ for all n, m and for almost all $x \in \mathcal{M}^*$. We can use the dominated convergence theorem to switch the limit and the integral:

$$\begin{aligned}
\lim_{n,m} \int_{\mathcal{M}^*} f_n(x) g_m(x) p(x) \mathrm{d}x &= \int_{\mathcal{M}^*} \lim_{n,m} f_n(x) g_m(x) p(x) \mathrm{d}x \\
&\leq K(p, -\alpha) \int_{\mathcal{M}^*} \lim_{m} g_m(x) p(x) \mathrm{d}x = 0,
\end{aligned}$$

where in the last inequality we used Theorem 23.

If $\beta < 0$, then by the Fatou lemma and Theorem 24 we have that

$$\begin{aligned}
\liminf_{n,m} \int_{\mathcal{M}^*} f_n(x) g_m(x) p(x) \mathrm{d}x &\geq \int_{\mathcal{M}^*} \liminf_{n,m} f_n(x) g_m(x) p(x) \mathrm{d}x \\
&= \int_{\mathcal{M}^*} (\bar{c} p(x))^{\alpha} \frac{\Gamma(k - \alpha)}{\Gamma(k)} \liminf_{m} g_m(x) p(x) \mathrm{d}x = \infty,
\end{aligned}$$

assuming $\int_{\mathcal{M}^*} p^{\alpha+1}(x) \mathrm{d}x > 0$ and using the fact that $\liminf_m g_m(x) = \infty$ for almost all $x \in \mathcal{M}^*$ thanks to Theorem 23.

Now, using (20) and (26) again, we have that $f_n(x) g_m(x) \leq K(p, -\alpha) K(q, -\beta) < \infty$, for almost all $x \in supp(p) \cap supp(q)$. Applying Lebesgue's dominated convergence theorem shows that

$$\lim_{n,m} \int_{supp(p) \cap supp(q)} f_n(x) g_m(x) p(x) \mathrm{d}x = \int_{supp(p) \cap supp(q)} \lim_{n,m} f_n(x) g_m(x) p(x) \mathrm{d}x,$$

which finishes the proof of (27), and thus the proofs of Theorem 12 as well.

5.4 The Asymptotic Variance

In this section we prove Theorem 13 by analyzing the asymptotic squared deviation between the estimator $\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m})$ and the true quantity $D_{\alpha,\beta}$.

5.4.1 Preliminaries

We want to show that $\lim_{n,m\to\infty}\mathbb{E}[(\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m})-D_{\alpha,\beta})^2]=0$. For the sake of brevity, let $\tau(i)\doteq(n-1)^{-\alpha}\rho_k^{-d\alpha}(i)m^{-\beta}\nu_k^{-d\beta}(i)$.

$$\begin{aligned}\mathbb{E}[(\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m})-D_{\alpha,\beta})^2] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\tau(i)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right)^2\right]\\ &= \frac{1}{n^2}\mathbb{E}\left[\sum_{i\neq j}^{n}\left(\tau(i)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right)\left(\tau(j)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right)\right]+\frac{1}{n}\mathbb{E}\left[\left(\tau(1)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right)^2\right].\end{aligned}$$

Thus, it is enough to prove that

$$\limsup_{n,m\to\infty}\mathbb{E}\left[\left(\tau(1)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right)^2\right]<\infty, \tag{28}$$

and

$$\lim_{n,m\to\infty}\frac{1}{n^2}\mathbb{E}\left[\sum_{i\neq j}^{n}\left(\tau(i)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right)\left(\tau(j)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right)\right]=0. \tag{29}$$

Investigating (28) we have that

$$\begin{aligned}&\mathbb{E}\left[\left\{\tau(1)B_{k,\alpha,\beta}-D_{\alpha,\beta}\right\}^2\right]\\ &\quad= D_{\alpha,\beta}^2+\mathbb{E}\left[(n-1)^{-2\alpha}\rho_k^{-2d\alpha}(1)m^{-2\beta}\nu_k^{-2d\beta}(1)B_{k,\alpha,\beta}^2\right]-2D_{\alpha,\beta}\mathbb{E}\left[\tau(1)\right]B_{k,\alpha,\beta}.\end{aligned}$$

We already know that the estimator is asymptotically unbiased (i.e., $\lim_{n,m\to\infty}\mathbb{E}\left[\tau(1)\right]B_{k,\alpha,\beta}=D_{\alpha,\beta}$). To see that (28) holds, it is enough to show that the following lemma is true.

Lemma 27

$$\limsup_{n,m\to\infty}\mathbb{E}_{X_1\sim p}\left\{\mathbb{E}\left[(n-1)^{-2\alpha}\rho_k^{-2d\alpha}(1)m^{-2\beta}\nu_k^{-2d\beta}(1)\middle|X_1\right]\right\}<\infty.$$

Proof We can use the same techniques that we used for proving Theorem 12. We just have to replace α and β with 2α and 2β and assume that $D_{2\alpha,2\beta}(p\|q)<\infty$. ■

To see that $\lim_{n,m\to\infty}\mathbb{E}[(\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m})-D_{\alpha,\beta})^2]=0$, all that remained is to prove (29). To be able to prove it, we need to make a couple of observations. We discuss them in the next section.

5.4.2 Properties of F_{n,k,x_1,x_2} Joint Distribution Function

Introduce the $\zeta_{n,k,1} \doteq (n-1)\rho_k^d(1)$, $\zeta_{n,k,2} \doteq (n-1)\rho_k^d(2)$ random variables, and let

$$F_{n,k,x_1,x_2}(u,v) \doteq \Pr(\zeta_{n,k,1} < u \wedge \zeta_{n,k,2} < v | X_1 = x_1, X_2 = x_2) \tag{30}$$

denote their joint distribution for $x_1, x_2 \in \mathbb{R}^d$. In order to see (29), we will need a couple of more lemmas that we list below.

Lemma 28 (Generalization of Lemma 19) *Let $x_1, x_2 \in \mathbb{R}^d$. For brevity let*

$$S_{n,j,l} \doteq (P_{n,u,x_1})^j (P_{n,v,x_2})^l (1 - P_{n,u,x_1} - P_{n,v,x_2})^{n-2-j-l}.$$

If $\max(R_n(u), R_n(v)) \leq \|x_1 - x_2\|$, then

$$F_{n,k,x_1,x_2}(u,v) = \sum_{j=k}^{n-2} \sum_{l=k}^{n-2-j} \binom{n-2}{j} \binom{n-2-j}{l} S_{n,j,l}. \tag{31}$$

It is easy to see that there exists $n_0(x_1, x_2, u_1, u_2) \in \mathbb{Z}^+$ such that for all $n > n_0$ we have that $\mathcal{B}(x_1, R_n(u)) \cap \mathcal{B}(x_2, R_n(v)) = \emptyset$, and thus (31) holds. The following lemma claims that in this case F_{n,k,x_1,x_2} can be rewritten.

Lemma 29 *If (31) holds, then F_{n,k,x_1,x_2} can be rewritten as*

$$F_{n,k,x_1,x_2}(u,v) = \tilde{F}_{n-1,k,x_1}(u) + \tilde{F}_{n-1,k,x_2}(v) - 1 + \sum_{j=0}^{k-1} \sum_{l=0}^{k-1} \binom{n-2}{j} \binom{n-2-j}{l} S_{n,j,l},$$

where

$$\tilde{F}_{n-1,k,x_1}(u) \doteq 1 - \sum_{j=0}^{k-1} \binom{n-2}{j} (P_{n,u,x_1})^j (1 - P_{n,u,x_1})^{n-2-j}.$$

Note that $\lim_{n\to\infty} \tilde{F}_{n-1,k,x_1}(u) = \lim_{n\to\infty} F_{n,k,x_1}(u)$.

Lemma 30 (Generalization of Lemma 20) *For almost all $x_1, x_2 \in \mathcal{M}$ when $x_1 \neq x_2$, then*

$$\lim_{n\to\infty} F_{n,k,x_1,x_2}(u,v) = F_{k,x_1}(u) F_{k,x_2}(v).$$

Lemma 31 (Generalization of Lemma 21) *The Lemma 30 holds for almost all $x_1, x_2 \in \mathbb{R}^d$ (even when they are not in $\mathcal{M}$) In this degenerate case, $\lim_{n\to\infty} F_{n,k,x_1,x_2}(u,v) = F_{k,x_1}(u) F_{k,x_2}(v) = 0$ for all u, v.*

Proof If $x_1 \notin cl(\mathcal{M})$, then there exist $\epsilon > 0$ such that $\rho_k^d(1) > \epsilon > 0$. In turn, $F_{n,k,x_1,x_2}(u,v)$ becomes zero for all fixed u, v if n is large enough thanks to the definition (30). ■

Lemma 32 (Generalization of Lemma 22) *Let $\gamma \in \mathbb{R}$ be arbitrary, $(\xi_{n,k,x_1}, \xi_{n,k,x_2}) \sim F_{n,k,x_1,x_2}(u,v)$, and $(\xi_{k,x_1}, \xi_{k,x_2}) \sim F_{k,x_1}(u)F_{k,x_2}(v)$. Then for almost all $x_1, x_2 \in \mathcal{M}$ $\xi^{\gamma}_{n,k,x_1}\xi^{\gamma}_{n,k,x_2} \to_d \xi^{\gamma}_{k,x_1}\xi^{\gamma}_{k,x_2}$. If $x_1 \notin cl(\mathcal{M})$ or $x_2 \notin cl(\mathcal{M})$, then $\xi_{n,k,x_1}\xi_{n,k,x_2} \to \infty$ almost surely.*

Proof Thanks to Lemma 30, we have that $F_{n,k,x_1,x_2}(u,v) \to F_{k,x_1}(u)F_{k,x_2}(v)$ for almost all $x_1, x_2 \in \mathcal{M}$, and thus $(\xi_{n,k,x_1}, \xi_{n,k,x_2}) \to_d (\xi_{k,x_1}, \xi_{k,x_2})$, and $\xi^{\gamma}_{n,k,x_1}\xi^{\gamma}_{n,k,x_2} \to_d \xi^{\gamma}_{k,x_1}\xi^{\gamma}_{k,x_2}$ by the continuous mapping theorem van der Wart (2007).

$$(\xi_{n,k,x_1}, \xi_{n,k,x_2}) \sim \Pr((n-1)\rho^d_k(1) < u \wedge (n-1)\rho^d_k(2) < v | X_1 = x_1, X_2 = x_2).$$

If $x_1 \notin cl(\mathcal{M})$, then there exists $\epsilon > 0$ such that $\rho^d_k(1) > \epsilon > 0$ almost surely, and thus $\xi_{n,k,x_1} \to \infty$ (a.s.). ξ_{n,k,x_2} converges to ξ_{k,x_2} when $x_2 \in \mathcal{M}$, which has Erlang distribution, or it diverges to infinity when $x_2 \notin cl(\mathcal{M})$. Now, we have that $\Pr(\xi_{n,k,x_1}\xi_{n,k,x_2} > u) \leq \Pr(\epsilon(n-1)\xi_{n,k,x_2} > u) = F_{n,k,x_2}(u/((n-1)\epsilon)) \to 1$ for all $u \geq 0$. ■

Introduce the following shorthands: $F(u,\infty) \doteq \lim_{a\to\infty} F(u,a)$, and $F(\infty,v) \doteq \lim_{a\to\infty} F(a,v)$. The following lemma describes the the marginal distributions of F_{n,k,x_1,x_2}.

Lemma 33 (The marginal distributions of F_{n,k,x_1,x_2})

$$F_{n,k,x_1,x_2}(u,\infty) = \begin{cases} F_{n-1,k,x_1}(u); & \text{if } \|x_1 - x_2\| > R_n(u) \\ F_{n-1,k-1,x_1}(u); & \text{if } \|x_1 - x_2\| \leq R_n(u), \text{ and } k \geq 2 \\ 1; & \text{if } \|x_1 - x_2\| \leq R_n(u), \text{ and } k = 1. \end{cases}$$

Similarly,

$$F_{n,k,x_1,x_2}(\infty,v) = \begin{cases} F_{n-1,k,x_2}(v); & \text{if } \|x_1 - x_2\| > R_n(v) \\ F_{n-1,k-1,x_2}(v); & \text{if } \|x_1 - x_2\| \leq R_n(v), \text{ and } k \geq 2 \\ 1; & \text{if } \|x_1 - x_2\| \leq R_n(v), \text{ and } k = 1. \end{cases}$$

Proof By definition,

$$F_{n,k,x_1,x_2}(u,v) \doteq \Pr(\zeta_{n,k,1} < u \wedge \zeta_{n,k,2} < v | X_1 = x_1, X_2 = x_2).$$

■

Introduce the following notation:

$$f_n(x_1,x_2) \doteq \mathbb{E}\left[(n-1)^{-2\alpha}\rho_k^{-d\alpha}(1)\rho_k^{-d\alpha}(2)\Big| X_1 = x_1, X_2 = x_2\right],$$
$$g_m(x_1,x_2) \doteq \mathbb{E}\left[m^{-2\beta}\nu_k^{-d\beta}(1)\nu_k^{-d\beta}(2)\Big| X_1 = x_1, X_2 = x_2\right].$$

The next theorem generalizes Theorem 25.

Theorem 34 (Properties of $f_n(x_1,x_2)$) *Let $\alpha = -\gamma$ in the definition of $f_n(x_1,x_2)$.*

1. *Let $\gamma > 0$, $\delta_1 > 0$, and let p be uniformly continuous on $\mathcal{M}$, and bounded away from zero. Let $\delta > 0$ so small that $p(x) - \delta > 0$ for all $x \in \mathcal{M}$. Then there exists $n_0 = n_0(k, \gamma, p, \delta, \delta_1) \in \mathbb{Z}^+$, which is independent of x_1, x_2 such that when $n > n_0$, then for almost all $x_1, x_2 \in \mathcal{M}$ we have that*

$$\begin{aligned} f_n(x_1, x_2) &\leq 2\gamma^2 L(x_1, 1/2, k, \gamma, p, \delta, \delta_1) L(x_2, 1/2, k, \gamma, p, \delta, \delta_1) \\ &\leq 2\gamma^2 \bar{L}(1/2, k, \gamma, p, \delta, \delta_1)^2 < \infty. \end{aligned} \tag{32}$$

2. *[Non-unifrom version in x] Let $\gamma > 0$, $\delta_1 > 0$, and let $\delta(x) > 0$ so small that $p(x) - \delta(x) > 0$ for all $x \in \mathcal{M}$. [Here we do not require p to be bounded away from zero.] Then for almost all $x_1, x_2 \in \mathcal{M}$, there exists $n_0 = n_0(x_1, x_2, k, \gamma, p, \delta(x_1), \delta(x_2), \delta_1) \in \mathbb{Z}^+$ such that when $n > n_0$, then*

$$f_n(x_1, x_2) \leq 2\gamma^2 L(x_1, 1/2, k, \gamma, p, \delta(x_1), \delta_1) L(x_2, 1/2, k, \gamma, p, \delta(x_2), \delta_1). \tag{33}$$

3. *If $k \geq 2$, $-(k-1)/2 < \gamma < 0$, and p is bounded above by $\bar{p}$, then for all $x_1, x_2 \in \mathbb{R}^d$*

$$f_n(x_1, x_2) \quad \leq \quad K_1(\bar{p}, \gamma) < \infty. \tag{34}$$

where

$$\begin{aligned} K_1(\bar{p}, \gamma) = &\left[\frac{\hat{L}^{1/2}(\bar{p}, 1)}{(k-1)/2 + \gamma} - \frac{1}{\gamma} \right]^2 + 2 \left[\frac{\hat{L}^{1/2}(\bar{p}, 1)}{(k-1)/2 + \gamma} - \frac{1}{\gamma} \right] \left[\frac{\hat{L}^{1/2}(\bar{p}, 1)}{k/2 + \gamma} - \frac{1}{\gamma} \right] \\ &+ \left[\frac{\hat{L}^{1/2}(\bar{p}, 1)}{k/2 + \gamma} - \frac{1}{\gamma} \right]^2. \end{aligned}$$

4. *If $-1/2 < \gamma < 0$, $k = 1$, p is bounded above by $\bar{p}$, then for all $x_1, x_2 \in \mathbb{R}^d$*

$$f_n(x_1, x_2) \leq K_2(x_1, x_2, \bar{p}, \gamma), \tag{35}$$

where

$$\begin{aligned} &K_2(x_1, x_2, \bar{p}, \gamma) \\ &\quad = \left[\frac{\hat{L}^{1/2}(\bar{p}, 1)}{k/2 + \gamma} - \frac{1}{\gamma} \right]^2 + \frac{1}{\gamma^2} \|x_1 - x_2\|^{2d\gamma} - \frac{2}{\gamma} \|x_1 - x_2\|^{d\gamma} \left[\frac{\hat{L}^{1/2}(\bar{p}, 1)}{1 + \gamma} - \frac{1}{\gamma} \right]. \end{aligned}$$

Theorem 35 (Generalization of Theorem 23) *If $x_1 \notin cl(\mathcal{M})$ or $x_2 \notin cl(\mathcal{M})$, then*

$$\lim_{n \to \infty} f_n(x_1, x_2) = \begin{cases} 0 & \text{if } 0 < \alpha \\ 1 & \text{if } \alpha = 0 \\ \infty & \text{if } \alpha < 0 \end{cases}$$

If $x_1 \notin cl(supp(q))$ or $x_2 \notin cl(supp(q))$, then

$$\lim_{m \to \infty} g_m(x_1, x_2) = \begin{cases} 0 & \text{if } 0 < \beta \\ 1 & \text{if } \beta = 0 \\ \infty & \text{if } 0 < \beta \end{cases}$$

Proof From Lemma 32 we have that $\xi_{n,k,x_1}\xi_{n,k,x_2} \to_d \infty$. ■

We will need the lemma below for the generalization of Theorem 24.

Lemma 36

(i) If $0 < \gamma$, *and* $\int_{\mathcal{M}} \|x-y\|p(y)\mathrm{d}y < \infty$,

(ii) or $-(k-1)/2 < \gamma < 0$, $k \geq 2$, *and* p *is bounded above by* $\bar{p}$,

then there exists $\varepsilon > 0$ *such that for almost all* $x_1, x_2 \in \mathcal{M}$ *we have*

$$\limsup_n \int_0^\infty \int_0^\infty u^{\gamma(1+\varepsilon)} v^{\gamma(1+\varepsilon)} F_{n,k,x_1,x_2}(\mathrm{d}u, \mathrm{d}v) < \infty.$$

Proof The case (i) follows from (33) and using the definition $f_n(x_1,x_2) = \int_0^\infty \int_0^\infty u^\gamma v^\gamma F_{n,k,x_1,x_2}(\mathrm{d}u,\mathrm{d}v)$. We can see that under these conditions we can always increase γ to $\gamma(1+\varepsilon)$ with an appropriate ε such that $\int_0^\infty \int_0^\infty u^{\gamma(1+\varepsilon)} v^{\gamma(1+\varepsilon)} F_{n,k,x_1,x_2}(\mathrm{d}u,\mathrm{d}v)$ will be bounded above by a number independent of n. Case (ii) follows from (34) using the same argument. ■

Theorem 37 (Generalization of Theorem 24) *If there exists* $\varepsilon > 0$ *such that*

$$\limsup_n \int_0^\infty \int_0^\infty u^{\gamma(1+\varepsilon)} v^{\gamma(1+\varepsilon)} F_{n,k,x_1,x_2}(\mathrm{d}u,\mathrm{d}v) = \limsup_n \mathbb{E}[(\xi_{n,k,x_1}^\gamma \xi_{n,k,x_2}^\gamma)^{1+\varepsilon}] < \infty,$$

then

$$\lim_{n\to\infty} \mathbb{E}\left[(n-1)^{2\gamma}\rho_k^{d\gamma}(1)\rho_k^{d\gamma}(2)\middle| X_1 = x_1, X_2 = x_2\right] = (\bar{c}p(x_1))^{-\gamma}(\bar{c}p(x_2))^{-\gamma}\frac{\Gamma^2(k+\gamma)}{\Gamma^2(k)}. \quad (36)$$

Proof According to Lemma 14 we have that $\mathbb{E}[\xi_{n,k,x_1}^\gamma \xi_{n,k,x_2}^\gamma] \to \mathbb{E}[\xi_{k,x_1}^\gamma \xi_{k,x_2}^\gamma]$, and thus

$$\begin{aligned}
\lim_{n\to\infty} f_n(x_1,x_2) &= \lim_{n\to\infty} \mathbb{E}\left[(n-1)^{2\gamma}(\rho_k(1))^{d\gamma}(\rho_k(2))^{d\gamma}|X_1 = x_1, X_2 = x_2\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\xi_{n,k,x_1}^\gamma \xi_{n,k,x_2}^\gamma\right] = \mathbb{E}\left[\lim_{n\to\infty} \xi_{n,k,x_1}^\gamma \xi_{n,k,x_2}^\gamma\right] = \mathbb{E}\left[\xi_{k,x_1}^\gamma \xi_{k,x_2}^\gamma\right] \\
&= \left[\int_0^\infty u^\gamma f_{k,x_1}(u)\mathrm{d}u\right]\left[\int_0^\infty v^\gamma f_{k,x_2}(v)\mathrm{d}v\right] \\
&= (\bar{c}p(x_1))^{-\gamma}(\bar{c}p(x_2))^{-\gamma}\frac{\Gamma^2(k+\gamma)}{\Gamma^2(k)}.
\end{aligned}$$

Here we also used Lemma 18 and Lemma 30. ■

Now we are ready to put the pieces together and prove the L_2 consistency of the estimator $\widehat{D}_{\alpha,\beta}$. To see that $\lim_{n,m\to\infty} \mathbb{E}[(\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m}) - D_{\alpha,\beta})^2] = 0$,

all that remained is to prove (29). We assumed that $D_{\alpha,\beta}(p\|q) < \infty$, and $\lim_{n,m\to\infty} \mathbb{E}[\widehat{D}_{\alpha,\beta}(X_{1:n}\|Y_{1:m})] = D_{\alpha,\beta}(p|q)$, that is the estimator is asymptotically unbiased. To see (29) holds it is enough to prove that

$$\lim_{n,m\to\infty} \mathbb{E}\left[\tau(1)\tau(2)\right] B_{k,\alpha,\beta}^2 + D_{\alpha,\beta}^2 - \lim_{n,m\to\infty} 2\mathbb{E}\left[\tau(1)\right] B_{k,\alpha,\beta} D_{\alpha,\beta} = 0. \tag{37}$$

In turn, it is enough to prove that $\lim_{n,m\to\infty} \mathbb{E}[\tau(1)\tau(2)] B_{k,\alpha,\beta}^2 = D_{\alpha,\beta}^2$, that is,

$$\lim_{n,m\to\infty} \mathbb{E}\left[(n-1)^{-\alpha}\rho_k^{-d\alpha}(1)m^{-\beta}\nu_k^{-d\beta}(1)(n-1)^{-\alpha}\rho_k^{-d\beta}(2)m^{-\beta}\nu_k^{-d\beta}(2)B_{k,\alpha,\beta}^2\right] = D_{\alpha,\beta}^2.$$

In other words, we want to prove that

$$\frac{D_{\alpha,\beta}^2}{B_{k,\alpha,\beta}^2} = \lim_{n,m\to\infty} \mathbb{E}_{\substack{X_1\sim p \\ X_2\sim p}} \left[f_n(X_1,X_2)g_m(X_1,X_2)\right].$$

In the same way as we did for proving asymptotic unbiasedness, using Theorem 34 we can uniformly upper bound $f_n(x_1,x_2)g_m(x_1,x_2)$ by a finite quantity. The application of the Lebesgue's dominated convergence theorem enables us to move the limit inside the expectation, which completes the proof of Theorem 13.

6. Illustration of Consistency

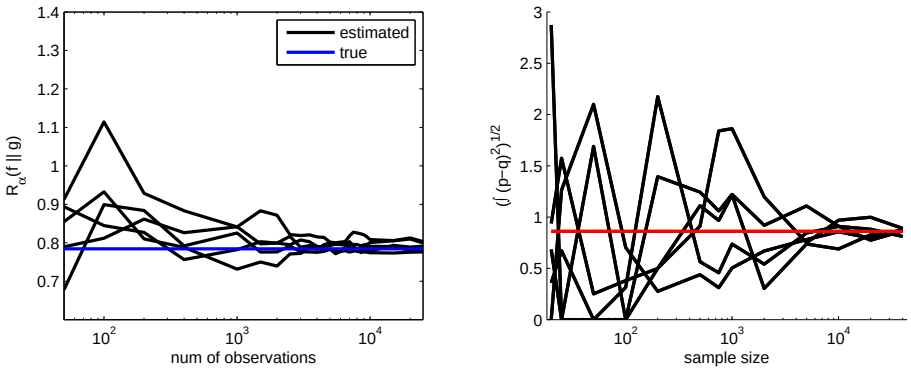
In this section we present a few numerical experiments to demonstrate the consistency of the proposed divergence estimators. We run experiments on beta distributions, where the domains are bounded, and we also study normal distributions, which have unbounded domains. We chose these distributions because in these cases the studied divergences have known closed-form expressions, and thus it is easy to evaluate our methods. We will also demonstrate that the proposed divergence estimators can be applied to estimate mutual information.

6.1 Normal Distributions

We begin our discussion by investigating the performance of our divergence estimators on normal distributions. Note that when $\alpha \notin [0,1]$, the divergences can easily become unbounded; see the Appendix for the details.

In Figure 3(a) we display the performances of the proposed $\widehat{R}_\alpha$ Rényi-α divergence estimator when the underlying densities were zero-mean Gaussians with randomly chosen 5-dimensional covariance matrices. Our results demonstrate that when we increase the sample sizes n and m, then the $\widehat{R}_\alpha$ value converges to the true value. For simplicity, in our experiments we always set $n = m$. The figure shows five independent experiments; the number of instances were varied between 50 and 25 000. The number of nearest neighbors k was set to 8, and α to 0.8.

In Figure 3(b) we show results for the L_2 divergence estimators when the underlying densities were zero-mean 1-dimensional Gaussians with randomly chosen covariance matrices. As we can see from the figure, the estimator converges to true value when we increase the sample sizes n and m.



(a) $R_\alpha(p\|q)$ (b) $L_2(p\|q)$

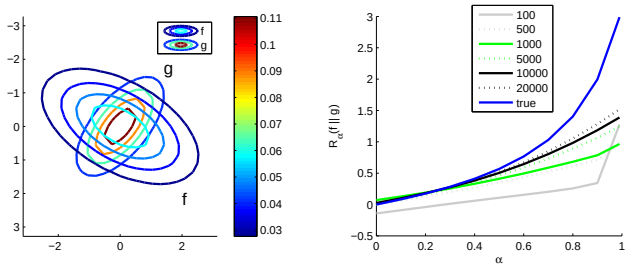
Figure 3: Estimated vs. true divergence as a function of the sample size. The plots show the results of five independent experiments converging to $R_\alpha(p\|q)$ and $L_2(p\|q)$, respectively.

The next experiment (Figure 4) demonstrates that even when f and g are normal distributions with zero means, the difficulty of the estimations of $R_\alpha(f\|g)$ can be significantly different for the different values of α, depending on the structure of the covariance matrices of f and g. In this experiment we set f and g to 2-dimensional Gaussians with zero means and different covariance matrices. Their contour plots are shown in the left column of Figure 4. The right column shows the estimation of R_α. We varied both α and the number of instances $n = m$. The first row show a case with randomly chosen covariance matrices, while in the second row we simply switched f and g from the first row. Our results indicate that depending on the structure of the covariance matrices of f and g, the parameter α which gives the fastest rate for the estimation of R_α can be anywhere in the interval [0,1].

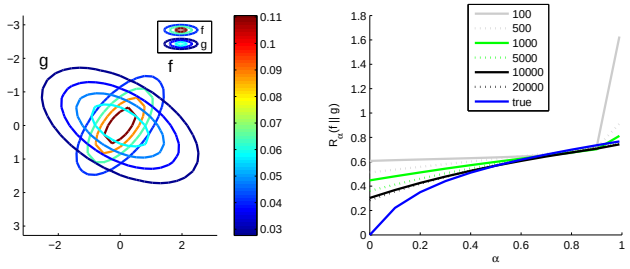
6.2 Beta Distributions

We were also interested in examining the performance of our estimators on beta distributions. To be able to study multidimensional cases, we construct d-dimensional distributions with independent 1-dimensional beta distributions as marginals. For a closed-form expression of the true divergence in this case, see the Appendix.

Our first experiment (Figure 5(a)) demonstrates that the estimators are consistent when $d = 2$. As we increase the number of instances, the estimators converge to the true $R_\alpha(f\|g)$ value. The figures show five independent experiments. We varied the sample sizes between 100 and 10 000. α was set to 0.4, and we used $k = 4$ nearest neighbors in the estimators. The parameters of the beta distributions were chosen independently and uniformly random from $[1, 2]$. We repeated this experiment in $5d$ as well. The $5d$ results, shown in Figure 5(b), show that the estimators were consistent in this case as well.



(a) Contour plot of f and g (b) $R_\alpha(f\|g)$



(c) Contour plot of f and g (d) $R_\alpha(f\|g)$

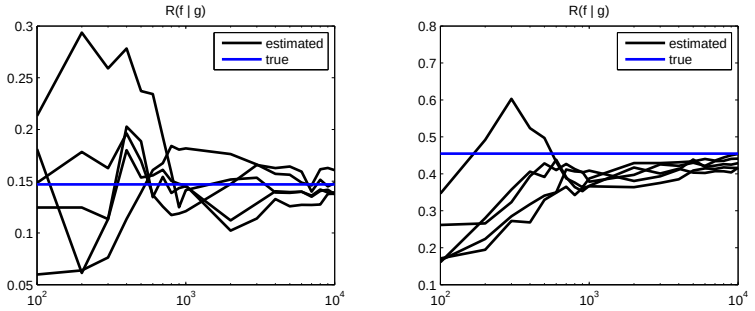
Figure 4: Estimated vs. true divergence in the number of observations. The left column shows the contour plots of two 2-dimensional normal distributions with different covariance matrices and zero means. The right column displays results of estimating $R_\alpha(f\|g)$. We varied α in $[0, 1]$ and also used different sample sizes between (100, 20 000). k was always set to 8. Each curve shows one single experiment.

6.3 Mutual Information Estimation

In this section we demonstrate that the proposed divergence estimators can also be used to estimate mutual information. Let $f = (f_1, \ldots, f_d) \in \mathbb{R}^d$ be the density of a d-dimensional distribution. The mutual information $I_\alpha(f)$ is the divergence between f and the product of the marginal variables. For the Rényi divergence we have $I_\alpha(f) = R_\alpha(f\|\prod_{i=1}^d f_i)$. Therefore, if we are given a sample $X_1, \ldots, X_{2n}$ from f, we may estimate the mutual information as follows. We form one set of size n by setting aside the first n samples. We build another sample set by randomly permuting the coordinates of the remaining n observations independently for each coordinate. They can be considered as n independent instances sampled from $\prod_{i=1}^d f_i$. Using these two sets, we can estimate $I_\alpha(f)$. Figure 6(a) shows the results of applying this procedure for a $2d$ Gaussian distribution with a randomly chosen covariance matrix.[3] The subfigure shows the true R_α values, as well as their estimations using different sample sizes. k was set to 8, and α was 0.8.

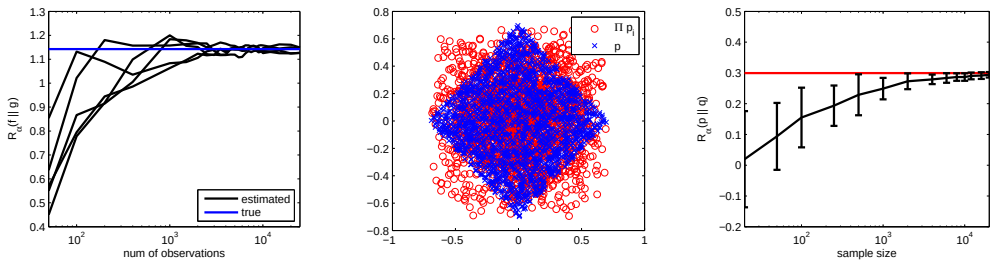
Figures 6(b)–6(c) show the results of repeating the previous experiment with two alterations. In this case we estimated the Shannon (rather than Rényi) information, and for this purpose we selected a $2d$ uniform distribution on $[-1/2, 1/2]^2$ rotated by $\pi/4$. Due to this rotation, the marginal distributions are no longer independent. Because our goal was

3. $\Sigma = CC^T$, where $C_{i,j} \sim U[0, 1]$.



(a) $R_\alpha(f\|g)$ (b) $R_\alpha(f\|g)$

Figure 5: Estimated vs. true divergence for the beta distribution experiments as a function of the number of observations. The figures show the results of five independent experiments for estimating the $R_\alpha(f\|g)$ divergence. (a): f and g were the densities of two $2d$ beta distributions—the marginal distributions were independent $1d$ betas with randomly chosen parameters. (b): The same as (a), but here f and g were the densities of two $5d$ beta distributions with independent marginals.



(a) $R_\alpha(f\|g)$ (b) samples from p and $\prod_{i=1}^d p_i$ (c) $R_\alpha(p\|\prod_{i=1}^d p_i)$

Figure 6: Estimated vs. true Rényi information as a function of sample size. (a) MI estimation for Gaussian distribution (b-c) MI estimation for rotated uniform distribution. (c) The error bars are calculated from 50 independent runs.

to estimate the Shannon information, we set α to 0.9999. The number of nearest neighbors used was $k = 8$, and the sample size was varied between 500 and 40 000. The estimators gave satisfactory results for the Shannon mutual information. Figure 6(b) shows the original samples as well as the independent samples from the product of the marginal distributions. Figure 6(c) demonstrates the consistency of the algorithm; as we increase the sample size, the estimator approaches the Shannon information.

7. Machine Learning on Distributions

In this section, we test the performance of the proposed divergence estimators on several machine learning problems. In these experiments, we first estimate the divergences and then use them for machine learning tasks including embedding, clustering, classification, and group anomaly detection.

We mainly compare our nonparametric (*NP*) methods to the alternative in which we first fit a parametric Gaussian mixture model (GMM) to each group, and then calculate the divergences between these GMMs. This alternative method is called *GMM estimation*. An even simpler method is to fit a single Gaussian to each group and then calculate the divergences between these Gaussians. We call this method *Gaussian estimation*. For GMMs, we have the analytical result for their L_2 divergences (Jian and Vemuri, 2005). However, there is no closed formula for the Rényi divergence between two GMMs, hence we resort to MCMC methods to approximate this quantity. In all of our experiments we use $\alpha = 0.5$ for the Rényi divergence. We also symmetrize both the Rényi and L_2 divergences by taking the average of the two way estimations: $D_{sym}(p|q) = (D(p\|q) + D(q\|p))/2$.

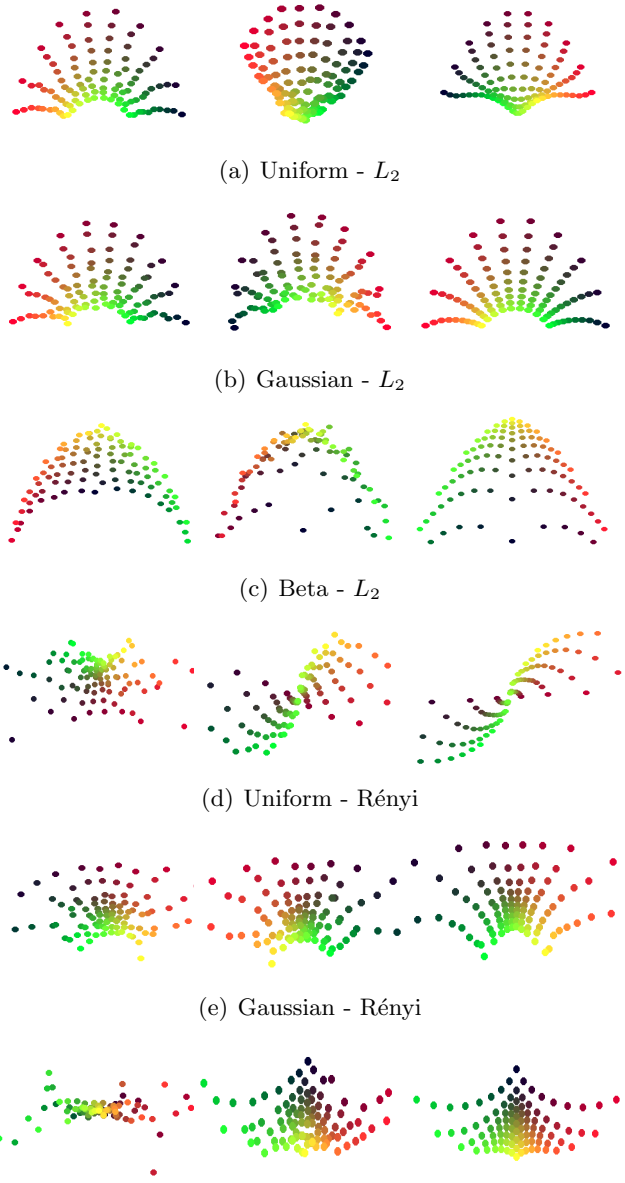
7.1 Embedding of Distributions

In this experiment we use synthetic data to demonstrate how the proposed estimators can be used to embed simple distributions including uniform, beta, and 1-dimensional Gaussian. For each type of distribution, we realize many distributions using different parameters, and each realization generates a group of data. Then we estimate divergences between these groups and embed them into a low-dimensional space using multidimensional scaling (Borg and Groenen, 2005). Finally, we visualize the embedded groups. As we will see, the underlying structure of the parameters are captured by the embedding.

For uniform and Gaussian distributions, the parameters are the mean and standard deviation. We selected the parameters from a uniform 10×10 grid, where the mean and standard deviation vary within $[0, 1]$ and $[0.3, 0.7]$ respectively. For beta distributions, we use the canonical parametrization with parameters α, β, and select their values from a uniform 10×10 grid on $[0.7, 3] \times [0.7, 3]$. To visualize the results, we color the embedded groups according to the above parameters. For each group we generate 2 000 samples. For the nonparametric estimators we use $k = 20$ nearest neighbors.

We compare our NP estimators to the Gaussian estimations. As the ground truth, we also calculate the embedding using the true divergences between the underlying distributions. Results using both the L_2 divergence (Figure 7 (a)-(c)) and Rényi divergence (Figure 7 (d)-(f)) are shown. We can see that the NP estimator can reveal the structure of the underlying parameters, and always produces embeddings that are similar to the ground truth. On the other hand, the embeddings by simple Gaussian estimation can be quite poor when the distribution is very different from Gaussian.

Next we show how embedding can reveal the structure of more complex distributions. To generate the data for groups, we first uniformly sample 3 000 points from sine curves $y = sin(\theta x)$, where $x \in [0, 2\pi]$, and θ is selected uniformly over $[2, 4]$. Then we added



(a) Uniform - L_2

(b) Gaussian - L_2

(c) Beta - L_2

(d) Uniform - Rényi

(e) Gaussian - Rényi

(f) Beta - Rényi

Figure 7: (a)-(c): Embeddings of Uniform, Gaussian and Beta distributions using L_2 divergence. (d)-(f): Results using Rényi divergence. Gaussian and Uniform distributions are colored by their means and variances (the red and green color component respectively). Beta distributions are colored by the two parameters. From left to right, the embeddings are produced by the Gaussian estimation, NP estimation, and the true divergence.

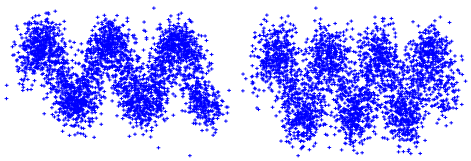


Figure 8: Two groups of the simulated noisy sine data.

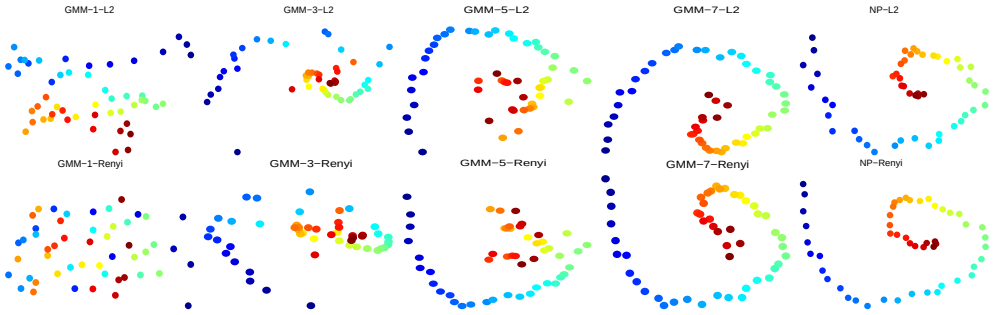


Figure 9: Embeddings of the 2-dimensional noisy sine data using GMMs and our nonparametric estimators. Points are colored by their underlying frequency. Column 1 to 4 show the embeddings by GMMs with increasing number of Gaussian components. The last column shows the embedding by our nonparametric estimators. The first row uses the L_2 divergence and the second row uses the Rényi divergence.

Gaussian noise from $\mathcal{N}(0, 0.3^2)$ to each (x, y) pair we sampled. Two groups of data are shown in Figure 8.

We embedded the groups into a 2D space using the proposed estimators with $k = 20$ nearest neighbors. We also performed the embedding using GMM estimations. Results using both methods are shown in Figure 9. The nonparametric approach correctly reveals the 1-dimensional nature of these distributions and orders the groups by their frequency. On the other hand, GMM estimation fails when the number of components is small. Although with enough components GMM can eventually work, it involves excessive computation and parameter tuning that are not needed in NP estimators.

7.2 Image Clustering and Classification

We can also use the proposed estimators to facilitate clustering and classification tasks by feeding the estimated divergences to algorithms that only need the dissimilarities between instances. In our experiments we use k-nearest-neighbors (k-NN) based clustering algorithms and classifiers.

We test the performance on the image data from Fei-Fei and Perona (2005). We adopt the "bag-of-words" representation for the images. Each image is a group of local patches, and each patch has a feature vector. We assume that each image has an inherent distribution to generate its patches, and these patches are i.i.d. In other words, each image is a distribution, and its patches are samples from this distribution. Then, we can measure

the dissimilarity between images by estimating the divergences between the corresponding distributions. Note that we do *not* quantize the patches as in Fei-Fei and Perona (2005), but rather use the original real-valued features and deal with their distributions directly.

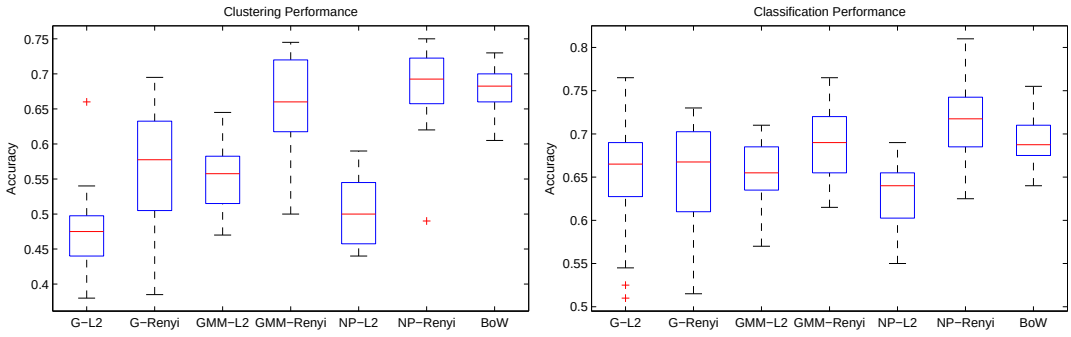
In this experiment, we use the categories "MITmountain", "MITcoast", "MIThighway", and "MITinsidecity" from the data set. From each category we randomly select 50 images. Features are extracted as in Fei-Fei and Perona (2005): Points are sampled on a uniform grid with interval 5, and then at each point we extract the 128-dimensional SIFT features (Lowe, 2004) and then reduce their dimension to 2 using PCA. In the end, we have 200 groups (images), each of which contains about 1 600 2-dimensional points (patches).

We compare the NP estimators to the Gaussian estimator and the 5-component-GMM estimator. We also compare them to the algorithm described by Bosch et al. (2006), which reflects the performance of a conventional "bag-of-word" (BoW) approach. In this BoW method, patches are quantized to 100 "visual words", and each image is represented as a 100-dimensional histogram of these words. Then *probabilistic latent semantic analysis* (pLSA) by Hofmann (1999) is applied to embed the images into a latent semantic space to get low-dimensional representations called topic distributions (here 20 topics are used, and thus each image is converted to a 20-dimensional probability vector). Finally, Euclidean distances between these topic distributions are used to measure the dissimilarities between images.

To cluster these images, we feed the divergences to the *spectral clustering* algorithm by Zelnik-Manor and Perona (2004). To evaluate the clustering results, we first form a confusion matrix from the category labels and the cluster labels, then permute the columns to maximize the trace of this matrix, which is equal to the number of correctly identified groups. We repeat 20 random runs and report the results in Figure 10(a).

We can observe that the Rényi divergence performs better than the L_2 divergence for this data set. We can also see that the Gaussian estimator is clearly inadequate. The GMM estimator improves over the single Gaussian one but is still slightly worse than the NP. The standard BoW approach also produces a slightly worse results than NP. Paired t-tests show that the difference between GMM and NP is significant, but the difference between NP and BoW is not significant (p-value is 6×10^{-3} for GMM-Rényi vs. NP-Rényi, and 0.94 for NP-Rényi vs. BoW).

We can also use the divergences for classification of distributions. Here we adopt a simple k-NN strategy: a group's label is predicted based on votes from the labeled groups that are closest, i.e., have the smallest divergence. We use $k = 11$ nearest neighbors for this classifier. In each run, we conduct 10-fold cross-validation on the randomly selected images and report the classification accuracy. The results from 20 random runs are reported in Figure 10(b). Similar results can be observed as in the clustering task; the nonparametric Rényi divergence estimator achieves the best performance among the competitors. Paired t-test gives p-value 5.33×10^{-4} for the difference between GMM-Rényi and NP-Rényi, and 0.15 for NP-Rényi vs. BoW. We also note that the nonparametric L_2 estimators produced poor results in this experiment. As Figure 3(b) demonstrates, for small sample size the L_2 estimator might have larger bias and variance than the that of the Rényi divergence estimator. This can result in poor performance.



(a) Spectral clustering performance.

(b) k-NN classification performance.

Figure 10: Clustering and classification performance using different divergence estimators. The columns correspond to the Gaussian estimation (G), GMM estimation, and nonparametric divergence estimation (NP) using Rényi and L_2 divergences, respectively. Finally, the column on the right displays the performance of the BoW method.

7.3 Group Anomaly Detection

One novel application of our divergence estimators is the detection of anomalous groups of data points. Note that unlike traditional anomaly detection methods that focus on unusual points, a group may have an anomalous distribution of points even if none of the individual points are unusual.

We use a simple detection algorithm based on nearest neighbors (Zhao and Saligrama, 2009). In this case the anomaly score of a group (distribution) is just the divergence between this group and its kth nearest neighbors. We apply this detector and compare the performance of different divergence estimators. The performance is measured by the *area under the ROC curve* (AUC).

We experimented with the different divergence estimators on the images we used before. The normal data was defined to be images from the categories in the previous experiment. In addition, we used images from two other categories "MITforest" and "livingroom" as anomalies. We used a random 75% of the normal images as training data and the rest 25% for testing. We also add some anomalous images to the test set to make it half normal and half abnormal. Then we asked the anomaly detector to find the anomalies, i.e., "livingroom" and "forest" images from this mixture. Those test groups that were the furthest away from their nearest neighbors in the training set were selected as anomalous groups.

We again compare the NP, Gaussian, and GMM estimators in this task. We use 5 Gaussian components in the GMM. The anomaly score of a test group is the divergence between the group and its 5th nearest neighbor in the training set. The results from 100 random runs are shown in Figure 11. Our NP estimator for Rényi divergence produces the best results, and the L_2 divergence again performs poorly. It is also interesting to see that the GMM estimator failed as well. Various reasons can cause this result, e.g., the inherent

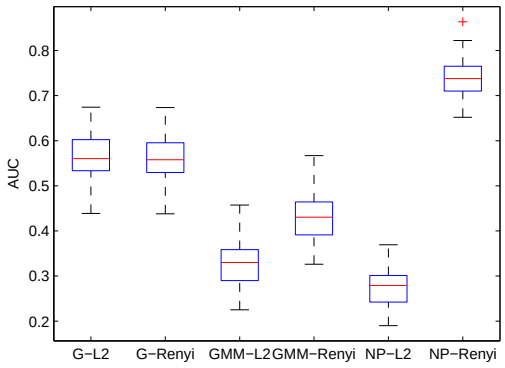


Figure 11: Anomalous image detection performance.

difference between normal and abnormal images can influence the estimators, and the GMM may overfit the data.

In the next experiment, we detect anomalous galaxy clusters in the astronomical data set from Sloan Digital Sky Survey[4] (SDSS). SDSS contains about 7×10^5 galaxies, each of which has a 4 000-dimensional continuum of the spectrum. We downsampled the continuum to get a 500-dimensional feature vector for each galaxy.

The "friends-of-friends" method (Garcia, 1993) was used to find spatial clusters (groups of nearby galaxies). 505 groups (7 530 galaxies) were found, each of which contains about 10–50 galaxies. In each group we used PCA to reduce the 500-dimensional continuum to 2-dimensional features preserving 95% of the variance. Note that this data set could be difficult for the NP estimators since the group sizes are small.

Due to the lack of labels, we use artificially injected anomalies to get statistically meaningful results. These injected groups are synthesized in the way such that each group consists of normal galaxies, but the distribution of the galaxies' features are rare in real galaxy clusters. In each run we injected 10 such random anomalies, and the whole data set contained 515 groups.

The AUC results from 20 random runs are shown in Figure 12. In this problem, the NP L_2 estimator achieves the best performance, and the NP estimators clearly outperform the parametric alternatives.

8. Discussion and Conclusion

We developed a new framework and proposed algorithms for several machine learning problems defined on the space of distributions. These problems include low-dimensional embedding, clustering, classification and outlier/anomaly detection. Most of the machine learning algorithms operate on fixed finite-dimensional feature representation. Kernel methods might transform the instances temporarily to an infinite-dimensional space, but the ultimate goal is still the same: to solve the classification, clustering, outlier detection, low-dimensional embedding problems in the original finite-dimensional feature space. In our setting, the

4. `http://www.sdss.org`

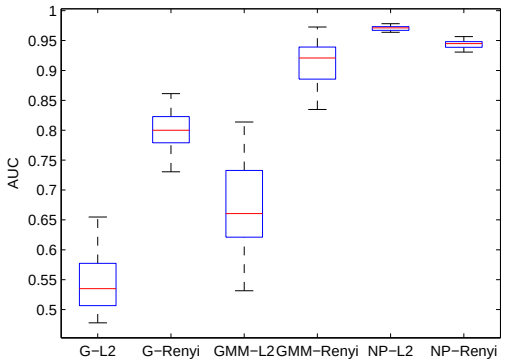


Figure 12: Anomalous galaxy cluster detection performance.

space of our features (continuous distributions) is infinite-dimensional. Furthermore, in contrast to standard machine learning problems we cannot observe them directly, only a few i.i.d. samples are available for us to represent these distributions.

In this paper we used nonparametric Rényi, Tsallis, Hellinger, Bhattacharyya, and L_2 divergence estimators to estimate the deviation between distributions. Under certain conditions we showed the consistency of these estimators and how they can be applied to estimate mutual information.

This new framework has many potential applications from bioinformatics to astronomy. It is useful anywhere where we take measurements of objects and our goal is to differentiate the divergences between the distributions of these measurements. We demonstrated the applicability of our framework both on synthetic toy problems and on real world problems including computer vision and anomaly detection in astronomical data.

We also compared our nonparametric estimators with a few competitors including a parametric estimator that assumes the distributions to be Gaussians, and a more complex estimator that first fits a mixture of Gaussian to the data and then estimates the divergences between these mixtures. We found that our nonparametric estimators outperform the competitors under various conditions. If the data does not match the parametric assumptions, then parametric approaches can lead to poor divergence estimators. Even though many distributions can be well-approximated by mixture of Gaussians, this approach might be too slow and sensitive to the number of Gaussian components in the model. The L_2 divergence can be easily calculated between two mixtures of Gaussians; however, it is challenging to calculate the Rényi divergence between them. Empirically we observed that the L_2 and Rényi estimators exhibit different behaviors and their performances depend on the actual distributions. We also found that the Rényi divergence is usually easier to estimate than the L_2 divergence, in which case the estimator seems to have higher variance.

There are several open questions left waiting for answers. Currently, the convergence rates of our divergence estimators are unknown, and we also do not know if the estimators are asymptotically normal.

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 *in Proc. of IEEE International Conference on Image Processing*, 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. doi: 10.1103/PhysRevE.49.R956.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag, New York, 2005.

A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *ECCV*, 2006.

B. Chai, D. Walther, D. Beck, and L. Fei-Fei. Exploring functional connectivity of the human brain using multivariate information analysis. In *NIPS*, 2009.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41-3, 2009.

A. Cichocki, H. Lee, Y. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 2008.

A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations*. John Wiley and Sons, 2009.

I. Csiszár. Information-type measures of differences of probability distributions and indirect observations. *Studia Sci. Math. Hungarica*, 2:299–318, 1967.

D. Donoho and C. Grimes. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *PNAS*, 100:5591–5596, 2003.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *IEEE Conf. CVPR*, pages 524–531, 2005.

W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 2*. Wiley series in probability and mathematical statistics, 1965.

A. Garcia. General study of group membership. ii. determination of nearby groups. *Astronomy & Astrophysics Suppliment Series*, 100:47–90, 1993.

M. Goria, N. Leonenko, V. Mergel, and N. Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17:277–297, 2005.

M. Gupta and S. Srivastava. Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, 12:818–843, 2010.

A. 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. Hero, B. Ma, O. Michel, and J. Gorman. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002b.

K. Hlaváckova-Schindler, M. Palušb, M. Vejmelkab, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, 1999.

M. Van Hulle. Constrained subspace ICA based on mutual information optimization directly. *Neural Computation*, 20:964–973, 2008.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493. MIT Press, 1998.

T. Jebara, R. Kondor, A. Howard, K. Bennett, and N. Cesa-bianchi. Probability product kernels. *JMLR*, 5:819–844, 2004.

B. Jian and B. Vemuri. A robust algorithm for point set registration using mixture of gaussians. In *ICCV*, 2005.

L. Kozachenko and N. Leonenko. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 23:9–16, 1987.

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. Fisher. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

N. Leonenko and L. Pronzato. Correction of 'a class of Rényi information estimators for mulitidimensional densities' Ann. Statist., 36(2008) 2153-2182, 2010.

N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 2008a.

N. Leonenko, L. Pronzato, and V. Savani. Estimation of entropies and divergences via nearest neighbours. *Tatra Mt. Mathematical Publications*, 39, 2008b.

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.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transaction on Information Theory*, 52:4394–4412, 2005.

D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics*, 36:1049–1051, 1965.

D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91 – 110, 2004.

Siwei Lyu. Mercer kernels for object recognition with local features. In *CVPR*, pages 223–229, 2005.

X. Nguyen, M. Wainwright, and M. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37:876–904, 2009.

X. Nguyen, M. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory, To appear.*, 2010.

D. Pál, B. Póczos, and Cs. Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *NIPS 2010*, 2010.

H. Peng and C. Dind. 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.

F. Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *NIPS-2008*, volume 21, 2008.

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 and J. Schneider. On the estimation of α-divergences. In *14th International Conference on AI and Statistics, Ft. Lauderdale, FL, USA*, 2011.

B. Póczos, S. Kirshner, and Cs. Szepesvári. REGO: Rank-based estimation of Rényi information using Euclidean graph optimization. In *AISTATS 2010*, 2010.

B. Póczos, L. Xiong, and J. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *27th Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain*, 2011.

A. Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, 1961.

A. Rényi. *Probability Theory*. North-Holland Publishing Company, Amsterdam, 1970.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

C. Shan, S. Gong, and P. Mcowan. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2005.

K. Sricharan, R. Raich, and A. Hero. Empirical estimation of entropy functionals with confidence. Technical Report, `http://arxiv.org/abs/1012.4188`, 2010.

J. Sun, C. Fyfe, and M. Crowe. Curvilinear component analysis and bregman divergences. In *ESANN 2010*, 2010.

Z. Szabó, B. Póczos, and A. Lőrincz. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.

J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

A. van der Wart. *Asymptotic Statistics*. Cambridge University Press, 2007.

T. van Erven and P. Harremoës. Rényi divergence and its properties. *CoRR*, abs/1001.4448, 2010.

O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, 38:54–59, 1976.

T. Villmann and S. Haase. Mathematical aspects of divergence based vector quantization using Frechet-derivatives, 2010. University of Applied Sciences Mittweida.

Q. Wang, S. 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. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5), 2009b.

E. Wolsztynski, E. Thierry, and L. Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Process.*, 85(5):937–949, 2005. ISSN 0165-1684. doi: http://dx.doi.org/10.1016/j.sigpro.2004.11.028.

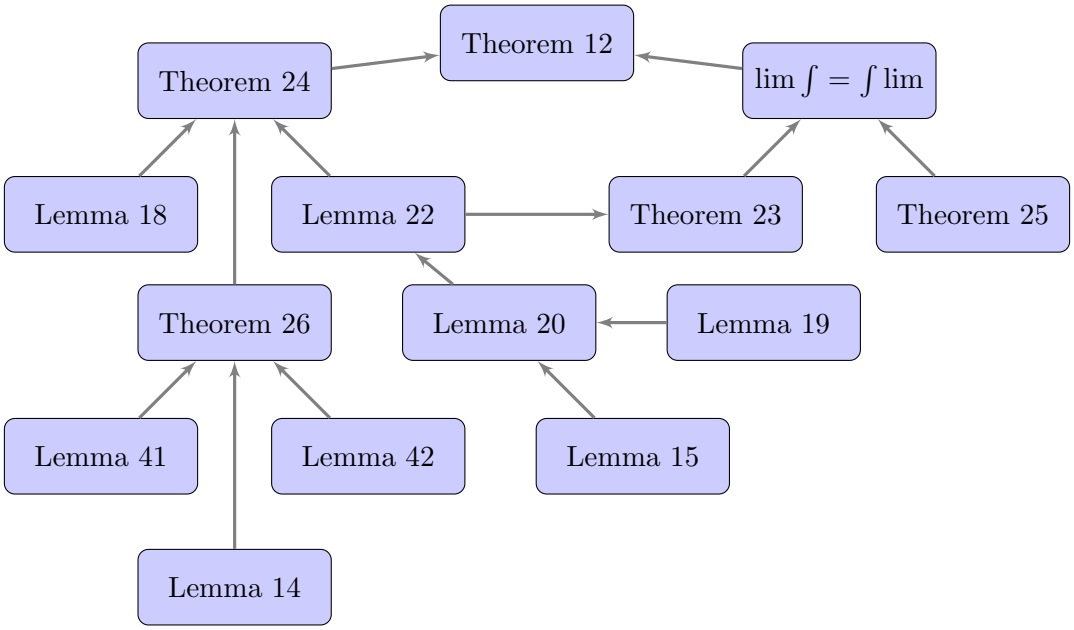
L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS 17*, 2004.

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Scientific Computing*, 26:313–338, 2004.

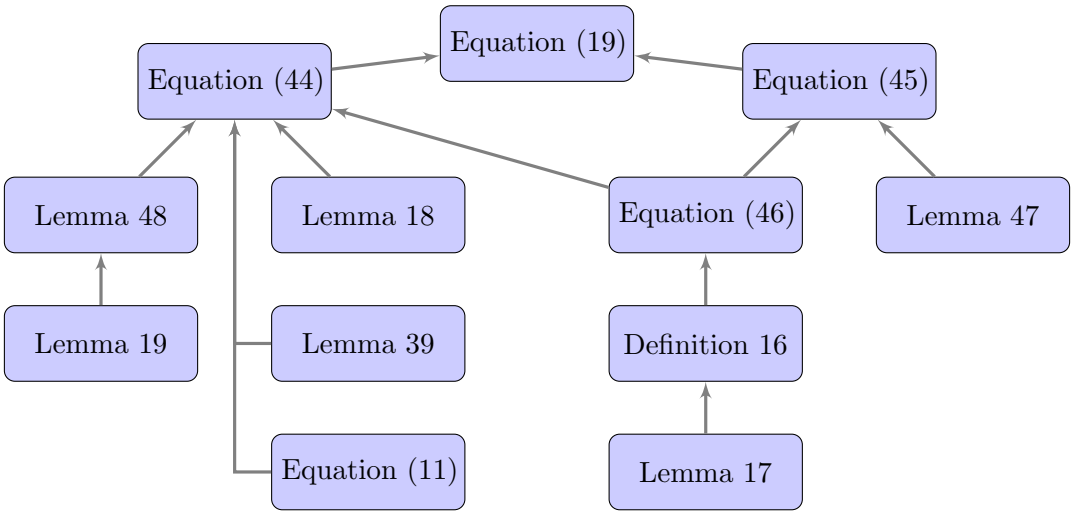
M. Zhao and V. Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In *NIPS*, 2009.

Appendix A. Dependency Charts of the Main Theorems and Lemmas

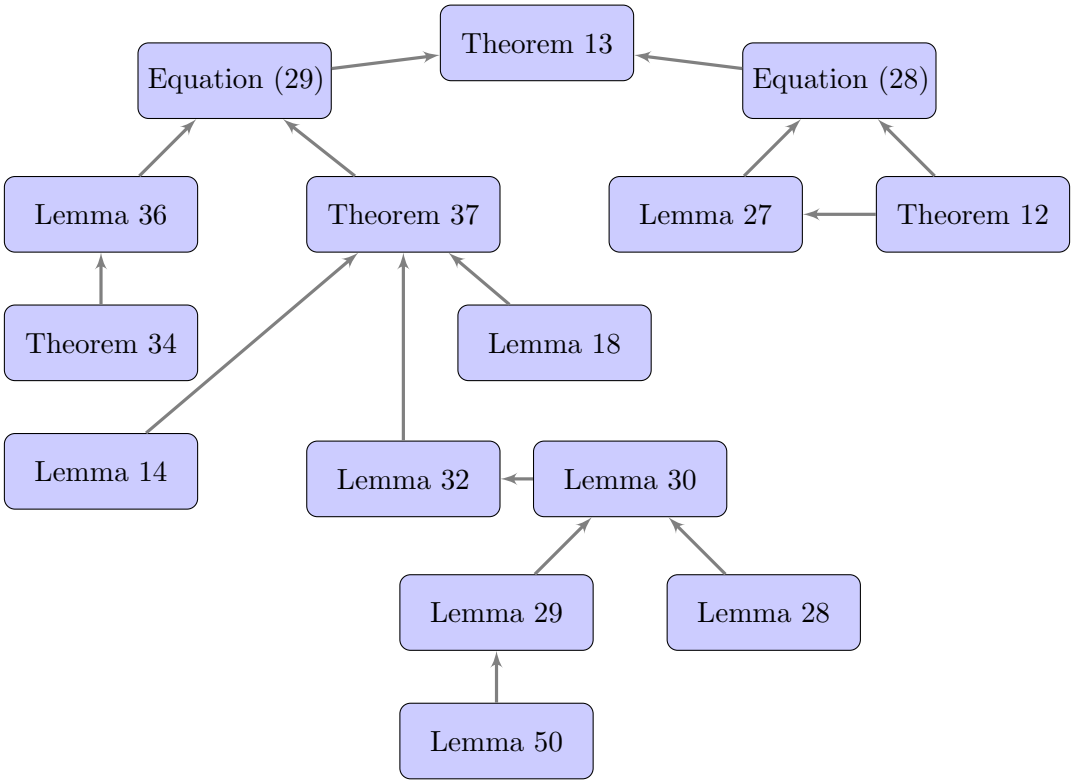
Proving Theorem 12



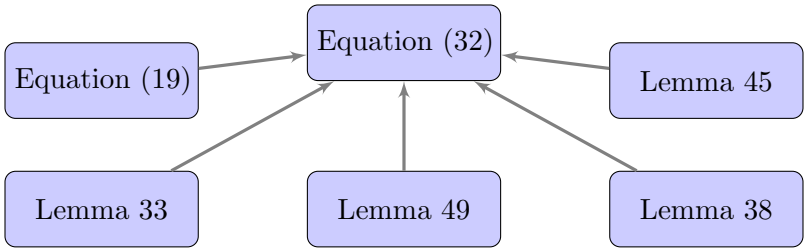
Proving Equation (19) of Theorem 25



Proving Theorem 13



Proving Equation (32) in Theorem 34



Appendix B. General Tools

Lemma 38 *If* $a, b \in [0, 1]$, *then* $\min(a, b) - ab \leq \sqrt{1-a}\sqrt{1-b}$.

Proof Without restricting the general case let $a \leq b$. We need to prove that

$$\begin{aligned} a - ab &\leq \sqrt{1-a}\sqrt{1-b} \\ a(1-b) &\leq \sqrt{1-a}\sqrt{1-b} \\ \sqrt{1-b} &\leq \frac{\sqrt{1-a}}{a} \\ 1-b &\leq \frac{1-a}{a^2} \end{aligned}$$

$$1 - \frac{1-a}{a^2} \leq b.$$

Since $a \leq b$, thus it enough to prove that

$$\begin{aligned} 1 - \frac{1-a}{a^2} &\leq a \\ a^2 - 1 + a &\leq a^3 \\ (1-a^2)(a-1) &\leq 0, \end{aligned}$$

which is true since $0 \leq a \leq 1$ by assumption. ■

Lemma 39 (Minkowski inequality) *If $0 \leq a, b$ and $0 \leq \omega \leq 1$, then $(a+b)^\omega \leq a^\omega + b^\omega$.*

Proof It is enough to prove that

$$\begin{aligned} a + b &\leq (a^\omega + b^\omega)^{1/\omega} \\ \left(\frac{a^\omega}{a^\omega + b^\omega}\right)^{1/\omega} + \left(\frac{b^\omega}{a^\omega + b^\omega}\right)^{1/\omega} &\leq 1. \end{aligned}$$

When $0 \leq c, d$, and $c + d = 1$, then $c^{1/\omega} \leq c \leq 1$, $d^{1/\omega} \leq d \leq 1$, thus $c^{1/\omega} + d^{1/\omega} \leq 1$. ■

Lemma 40 (Reverse triangle inequality) *If $0 \leq a, b$, and $0 \leq \alpha \leq 1$, then $|a^\alpha - b^\alpha| \leq |a-b|^\alpha$*

Proof It is enough to prove the $a \geq b$ case, that is

$$\begin{aligned} a^\alpha &\leq (a-b)^\alpha + b^\alpha \\ (c+b)^\alpha &\leq c^\alpha + b^\alpha, \end{aligned}$$

using the $c = a - b > 0$ notation. This immediately follows from Lemma 39. ■

Lemma 41 *Let $\gamma > 0$, $F : \mathbb{R} \to [0,1]$ distribution function. Then*

$$\int_a^\infty u^\gamma F(\mathrm{d}u) = a^\gamma(1 - F(a)) + \int_a^\infty \gamma u^{\gamma-1}(1 - F(u))\mathrm{d}u, \tag{38}$$

In the $a = 0$ case:

$$\int_0^\infty u^\gamma F(\mathrm{d}u) = \gamma \int_0^\infty u^{\gamma-1}(1 - F(u))\mathrm{d}u, \tag{39}$$

in the sense that the integral of either side exists and finite iff the integral on the other side exists and finite, too.

Proof The proof can be found in (Feller, 1965), here we just give a sketch. Integrate by parts. We have that for all $\lambda \in \mathbb{R}$ the equations below hold:

$$\begin{aligned}\int_a^b u^\gamma \underbrace{F(\mathrm{d}u)}_{f(u)\mathrm{d}u} &= [u^\gamma F(u)]_a^b - [u^\gamma \lambda]_a^b + \underbrace{[u^\gamma \lambda]_a^b}_{\int_a^b \lambda\gamma u^{\gamma-1}\mathrm{d}u} - \int_a^b \gamma u^{\gamma-1} F(u)\mathrm{d}u \\ &= b^\gamma(F(b)-\lambda) - a^\gamma(F(a)-\lambda) + \int_a^b \gamma u^{\gamma-1}(\lambda - F(u))\mathrm{d}u.\end{aligned}$$

Now, we can choose $\lambda = 1$ and analyze the behavior of the integrals when $a \to 0$ and $b \to \infty$. ■

Lemma 42 *Let $\gamma > 0$, $F : \mathbb{R} \to [0,1]$ distribution function. Then*

$$\int_a^\infty u^{-\gamma} F(\mathrm{d}u) = -a^{-\gamma}F(a) + \gamma \int_a^\infty u^{-\gamma-1} F(u)\mathrm{d}u.$$

In the $a = 0$ case:

$$\int_0^\infty u^{-\gamma} F(\mathrm{d}u) = \gamma \int_0^\infty u^{-\gamma-1} F(u)\mathrm{d}u. \tag{40}$$

Proof The proof can be found in Leonenko et al. (2008a). ■

We will also need the 2-dimensional generalizations of Lemmas 41–42. They can be proved in the same way.

Lemma 43 *If $\gamma > 0$, then*

$$\int_0^\infty \int_0^\infty u^{-\gamma} v^{-\gamma} F(\mathrm{d}u, \mathrm{d}v) = \gamma^2 \int_0^\infty \int_0^\infty u^{-\gamma-1} v^{-\gamma-1} F(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

Lemma 44 *Let $F(u,v) \doteq F_{U,V}(u,v)$ be a 2-dimensional distribution function with $f(u,v) \doteq f_{U,V}(u,v)$ density and $F_U(u) = F_{U,V}(u,\infty)$, $F_V(v) = F_{U,V}(\infty,v)$ marginal distribution functions. For $\gamma > 0$, $\lambda \in \mathbb{R}$ arbitrary, the following equation holds:*

$$\int_a^b u^\gamma f(u,v)\mathrm{d}u = b^\gamma\left(\frac{\partial F(b,v)}{\partial v} - \lambda\right) + a^\gamma\left(\lambda - \frac{\partial F(a,v)}{\partial v}\right) + \int_a^b \gamma u^{\gamma-1}\left(\lambda - \frac{\partial F(a,v)}{\partial v}\right)\mathrm{d}u.$$

Hence, when we study the $b \to \infty$ limit case, then $\lambda \doteq \lambda(v) = \frac{\partial F(\infty,v)}{\partial v} = f_V(v)$ will be a useful choice:

$$\int_0^\infty u^\gamma f(u,v)\mathrm{d}u = \int_0^\infty \gamma u^{\gamma-1}\left(f_V(v) - \frac{\partial F(a,v)}{\partial v}\right)\mathrm{d}u. \tag{41}$$

Lemma 45 *Let $\gamma > 0$, and let $F(u,v) \doteq F_{U,V}(u,v)$ be a 2-dimensional distribution function with $f(u,v) \doteq f_{U,V}(u,v)$ density and $F_U(u), F_V(v)$ marginal distributions. The following equation holds:*

$$\int_0^\infty \int_0^\infty u^\gamma v^\gamma F(\mathrm{d}u, \mathrm{d}v) = \int_0^\infty \int_0^\infty \gamma^2 u^{\gamma-1} v^{\gamma-1} [1 - F_U(u) - F_V(v) + F(u,v)] \mathrm{d}u \mathrm{d}v.$$

Proof

$$\begin{aligned}
&\int_0^\infty \int_0^\infty u^\gamma v^\gamma f(u,v) \mathrm{d}u \mathrm{d}v \\
&\quad = \lim_{\substack{d\to\infty \\ c\to 0}} \int_c^d v^\gamma \left[\int_0^\infty u^\gamma f(u,v) \mathrm{d}u \right] \mathrm{d}v \\
&\quad = \lim_{\substack{d\to\infty \\ c\to 0}} \int_c^d v^\gamma \left[\int_0^\infty \gamma u^{\gamma-1} \left(f_V(v) - \frac{\partial F(a,v)}{\partial v} \right) \mathrm{d}u \right] \mathrm{d}v \\
&\quad = \int_0^\infty \gamma u^{\gamma-1} \left[\lim_{\substack{d\to\infty \\ c\to 0}} \int_c^d v^\gamma \left(f_V(v) - \frac{\partial F(a,v)}{\partial v} \right) \mathrm{d}v \right] \mathrm{d}u \\
&\quad = \int_0^\infty \gamma u^{\gamma-1} \left[\underbrace{\int_0^\infty v^\gamma f_V(v) \mathrm{d}v}_{\int_0^\infty \gamma v^{\gamma-1}(1-F_V(v))\mathrm{d}v} - \underbrace{\int_0^\infty v^\gamma \frac{\partial F(a,v)}{\partial v} \mathrm{d}v}_{\int_0^\infty \gamma v^{\gamma-1}(F_U(u)-F(u,v))\mathrm{d}v} \right] \mathrm{d}u.
\end{aligned}$$

Here we chose $\lambda = F_U(u) = F(u,\infty)$ and applied the lemmas above. ■

Appendix C. Proofs of Section 5.1

Proof of Lemma 14

Proof We want to prove that X_n^γ is asymptotically uniformly integrable, that is

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}[X_n^\gamma 1_{\{X_n^\gamma > M\}}] = 0.$$

This follows from the observation that for all $M > 0$

$$\mathbb{E}[X_n^\gamma 1_{\{X_n^\gamma > M\}}] \leq \mathbb{E}\left[(X_n^\gamma) \frac{(X_n^\gamma)^\varepsilon}{M^\varepsilon} 1_{\{X_n^\gamma > M\}} \right] \leq \frac{1}{M^\varepsilon} \mathbb{E}[(X_n^\gamma)^{1+\varepsilon}].$$

For all $\varepsilon > 0$ we have that $\lim_{M\to\infty} \frac{1}{M^\varepsilon} = 0$, and by assumption $\limsup_{n\to\infty} \mathbb{E}\left[X_n^{\gamma(1+\varepsilon)} \right] = K < \infty$, thus

$$\begin{aligned}
0 &\leq \lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}[X_n^\gamma 1_{\{X_n^\gamma > M\}}] \leq \lim_{M\to\infty} \limsup_{n\to\infty} \frac{1}{M^\varepsilon} \mathbb{E}[(X_n^\gamma)^{1+\varepsilon}] \\
&\leq \lim_{M\to\infty} \frac{1}{M^\varepsilon} K = 0.
\end{aligned}$$

■

Proof of Lemma 17

Proof If g is uniformly continuous on $\mathcal{M}$ then for all $\delta > 0$ there exists $R_\delta > 0$ such that if $x, y \in \mathcal{M}$, $\|x-y\| < R_\delta$, then $|g(x)-g(y)| < \delta$. Thus, $g(x)-\delta < g(\mathcal{B}(x, R_\delta)\cap\mathcal{M}) < g(x)+\delta$, and furthermore if $R_n < R_\delta$, then

$$(g(x) - \delta)\mathcal{V}(\mathcal{B}(x, R_n) \cap \mathcal{M}) < \int_{\mathcal{B}(x,R_n)\cap\mathcal{M}} g(t)\ \mathrm{d}t < (g(x) + \delta)\mathcal{V}(\mathcal{B}(x, R_n) \cap \mathcal{M}).$$

■

Proof of Lemma 18

Proof By the definition of the Gamma function we have that

$$\int_0^\infty u^{\beta-1} \exp(-\lambda u)\mathrm{d}u = \lambda^{-\beta}\Gamma(\beta). \tag{42}$$

Thus,

$$\begin{aligned}
\int_0^\infty u^\gamma f_{x,k}(u)\ \mathrm{d}u &= \int_0^\infty u^\gamma \frac{\lambda^k u^{k-1}\exp(-\lambda u)}{\Gamma(k)}\ \mathrm{d}u \\
&= \frac{\lambda^k}{\Gamma(k)}\int_0^\infty u^{\gamma+k-1}\exp(-\lambda u)\ \mathrm{d}u \\
&= \frac{\lambda^k}{\Gamma(k)}\frac{1}{\lambda}\int_0^\infty \left(\frac{y}{\lambda}\right)^{k+\gamma-1}\exp(-y)\ \mathrm{d}y \\
&= \frac{\lambda^{-\gamma}}{\Gamma(k)}\int_0^\infty y^{k+\gamma-1}\exp(-y)\ \mathrm{d}y = \lambda^{-\gamma}\frac{\Gamma(k+\gamma)}{\Gamma(k)}.
\end{aligned}$$

Here we used the $y = \lambda u$, $\mathrm{d}y = \lambda \mathrm{d}u$ integral transform. ■

Appendix D. Proving Asymptotic Unbiasedness

Proof of Lemma 19

Proof Calculate this distribution in a "closed" form.

$$\begin{aligned}
F_{n,k,x}(u) &\doteq \Pr(\zeta_{n,k,1} < u | X_1 = x) \\
&= \Pr((n-1)\rho_k^d(1) < u | X_1 = x) \\
&= \Pr\left(\rho_k(1) < (u/(n-1))^{1/d}\,|X_1 = x\right) \\
&= \Pr\left(\rho_k(1) < R_n(u) | X_1 = x\right) \\
&= \Pr(k \text{ elements or more in} \{X_2, \ldots, X_n\} \in \mathcal{B}(x, R_n(u)) \cap \mathcal{M} | X_1 = x) \\
&= \Pr(k \text{ elements or more in } \{X_2, \ldots, X_n\} \in \mathcal{B}(x, R_n(u)) \cap \mathcal{M})
\end{aligned} \tag{43}$$

$$= \sum_{j=k}^{n-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j}$$
$$= 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j}.$$

■

Corollary 46 $F_{n,1,x}(u) = 1 - (1 - P_{n,u,x})^{n-1}$.

Proof of Lemma 20

Proof

Let $u > 0$ be fixed, and $R_n(u) \doteq (u/(n-1))^{1/d}$. Let $x \in \mathcal{M} = supp(p)$, and let $\delta > 0$ be so small such that $p(x) > \delta > 0$ holds. From (10), we know that for all $\delta > 0$ and almost all $x \in \mathcal{M}$, there exists $n_0(x, \delta, u) \in \mathbb{Z}^+$ such that if $n > n_0(x, \delta, u)$, then $\mathcal{B}(x, R_n(u)) = \mathcal{M} \cap \mathcal{B}(x, R_n(u))$ (since almost all points are inner points in $\mathcal{M}$ and $\lim_{n\to\infty} R_n(u) = 0$), and

$$p(x) - \delta < \frac{\int_{\mathcal{B}(x,R_n(u))} p(t)\ \mathrm{d}t}{\mathcal{V}(\mathcal{B}(x, R_n(u)))} < p(x) + \delta,$$
$$p(x) - \delta < \frac{P_{n,u,x}}{\frac{u\bar{c}}{n-1}} < p(x) + \delta.$$

Introduce the following shorthands: $\bar{s} \doteq (p(x) + \delta)\bar{c}$, $\underline{s} \doteq (p(x) - \delta)\bar{c} > 0$. Now, if $n > n_0(x, \delta, u)$, then

$$\begin{aligned}
F_{n,k,x}(u) &\doteq 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j} \\
&\geq 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} \left(\frac{u\bar{s}}{n-1}\right)^j \left(1 - \frac{u\underline{s}}{n-1}\right)^{n-1-j} \\
&= 1 - \sum_{j=0}^{k-1} \frac{(n-1)!}{j!(n-1-j)!} \left(\frac{u\bar{s}}{n-1}\right)^j \left(1 - \frac{u\underline{s}}{n-1}\right)^{n-1-j} \\
&= 1 - \sum_{j=0}^{k-1} \frac{1}{j!} \underbrace{\frac{(n-1)!}{(n-1-j)!(n-1)^j}}_{\to 1} (u\bar{s})^j \underbrace{\left(1 - \frac{u\underline{s}}{n-1}\right)^{n-1-j}}_{\to \exp(-u\underline{s})}.
\end{aligned}$$

Thus for all $p(x) > \delta > 0$ and for almost all $x \in \mathcal{M}$, we have that

$$\liminf_{n\to\infty} F_{n,k,x}(u) \geq 1 - \sum_{j=0}^{k-1} \frac{1}{j!} (u\bar{c}[p(x) + \delta])^j \exp(-u\bar{c}[p(x) - \delta]),$$

and hence by choosing $\delta \to 0$, we can see that for almost all $x \in \mathcal{M}$

$$\liminf_{n\to\infty} F_{n,k,x}(u) \geq 1 - \sum_{j=0}^{k-1} \frac{1}{j!} (u\lambda)^j \exp(-u\lambda) \doteq F_{k,x}(u),$$

where $\lambda \doteq \bar{c}p(x)$. Using similar arguments we can also prove that for almost all $x \in \mathcal{M}$,

$$\limsup_{n\to\infty} F_{n,k,x}(u) \leq 1 - \sum_{j=0}^{k-1} \frac{1}{j!} (u\lambda)^j \exp(-u\lambda) \doteq F_{k,x}(u).$$

This completes the proof of the lemma. ■

Proof of Lemma 21

Proof If $x \notin cl(\mathcal{M})$, then $P_{n,u,x} \doteq \int_{\mathcal{M}\cap\mathcal{B}(x,R_n(u))} p(t)\, \mathrm{d}t$ becomes zero for all fixed u if n is large enough. For these n values $F_{n,k,x}(u) = 0$. ■

Proof of Lemma 22

Proof We already know by Lemma 20 that for almost all $x \in \mathcal{M}$ $F_{n,k,x}(\cdot) \to_w F_{k,x}(\cdot)$ as $n \to \infty$. From this, it follows by definition that $\xi_{n,k,x} \to_d \xi_{k,x}$ $(a.a.\, x \in \mathcal{M})$. Now, since the $(\cdot)^\gamma$ function is continuous on $(0,\infty)$, and $X_i \in (0,\infty)$ almost surely, thus by the continuous mapping theorem (van der Wart, 2007) the lemma follows.

$\xi_{n,k,x} \sim \Pr((n-1)\rho_k^d(1)|X_1 = x)$. Therefore, if $x \notin cl(\mathcal{M})$, then there exists $\epsilon > 0$ such that $\rho_k^d(1) > \epsilon > 0$ almost surely, and thus $\xi_{n,k,x} \to \infty$. ■

Proof of Theorem 25

Proof [Proof of (19) in Theorem 25] First we will prove that there exists n_0 independent of x such that for all $n > n_0$ it holds that

$$\begin{aligned}
&\int_0^{\sqrt{n-1}} (1 - F_{n,k,x}(u))^\omega u^{\gamma-1} \mathrm{d}u \qquad (44)\\
&\quad \leq \sum_{j=0}^{k-1} \left(\frac{1}{j!}\right)^\omega (\bar{c}r_{\mathcal{M}})^{-\gamma} \Gamma(\gamma + j\omega) \left(\frac{p(x)+\delta}{p(x)-\delta}\right)^{j\omega} (p(x)-\delta)^{-\gamma} ((1-\delta)\omega)^{-\gamma-j\omega}\\
&\quad = (\bar{c}r_{\mathcal{M}})^{-\gamma} H(x,p,\delta,\omega).
\end{aligned}$$

Then we will see that for $n > n_0$, it also holds that

$$\int_{\sqrt{n-1}}^{\infty} (1 - F_{n,k,x}(u))^\omega u^{\gamma-1} \mathrm{d}u$$

$$\leq (kn^k)^\omega \left[\frac{(n-1)^\gamma}{\gamma} + \int \|x-y\|^\gamma p(y)\mathrm{d}y\right] \exp\left[-((n-k)\omega - 1)(p(x)-\delta)\frac{\bar{c}r(x)}{\sqrt{n-1}}\right]$$
$$\leq \delta_1 + \delta_1 \int \|x-y\|^\gamma p(y)\mathrm{d}y. \tag{45}$$

Assume that p is bounded away from zero, that is, there exists a $\delta > 0$ such that $0 < p(x) - \delta$, for almost all $x \in \mathcal{M}$. Furthermore, let p be uniformly Lebesgue approximable (see definition 16), i.e. for all $\delta > 0$ there exists $n_0(\delta)$ such that if $\tilde{n} - 1 > n_0(\delta)$ then for almost all $x \in \mathcal{M}$ we have that

$$p(x) - \delta < \frac{\int_{\mathcal{M}\cap\mathcal{B}\left(x,\left(\frac{1}{\sqrt{\tilde{n}-1}}\right)^{1/d}\right)} p(t)\ \mathrm{d}t}{\mathcal{V}\left(\mathcal{M}\cap\mathcal{B}\left(x,\left(\frac{1}{\sqrt{\tilde{n}-1}}\right)^{1/d}\right)\right)} < p(x) + \delta.$$
$$p(x) - \delta < \frac{\int_{\mathcal{M}\cap\mathcal{B}\left(x,\left(\frac{1}{\sqrt{\tilde{n}-1}}\right)^{1/d}\right)} p(t)\ \mathrm{d}t}{\frac{\bar{c}}{\sqrt{\tilde{n}-1}}r(x)} < p(x) + \delta, \tag{46}$$

where

$$r(x) \doteq \frac{\mathcal{V}\left(\mathcal{M}\cap\mathcal{B}\left(x,\left(\frac{1}{\sqrt{\tilde{n}-1}}\right)^{1/d}\right)\right)}{\mathcal{V}\left(\mathcal{B}\left(x,\left(\frac{1}{\sqrt{\tilde{n}-1}}\right)^{1/d}\right)\right)} \in [r_{\mathcal{M}}, 1].$$

Usually $r(x) = 1$ in $\mathcal{M}$, however, close to the boundary of $\mathcal{M}$ its value can be less. Nonetheless, according to our conditions it is always at least as large as $r_{\mathcal{M}} > 0$. By definition $P_{n,u,x} \doteq \int_{M\cap\mathcal{B}\left(x,\left(\frac{u}{n-1}\right)^{1/d}\right)} p(t)\mathrm{d}t$. Let $n - 1 > n_0(\delta)$, and $\beta < u < \sqrt{n-1}$. If we define $\tilde{n} \doteq 1 + \left(\frac{n-1}{u}\right)^2$, then $\tilde{n} - 1 = \left(\frac{n-1}{u}\right)^2 > n - 1 > n_0(\delta)$, and thus

$$\begin{aligned} 0 < p(x) - \delta &< \frac{\int_{M\cap\mathcal{B}\left(x,\left(\frac{1}{\sqrt{\tilde{n}-1}}\right)^{1/d}\right)} p(t)\ \mathrm{d}t}{\frac{\bar{c}r(x)}{\sqrt{\tilde{n}-1}}} \\ &= \frac{\int_{\mathcal{M}\cap\mathcal{B}\left(x,\left(\frac{u}{n-1}\right)^{1/d}\right)} p(t)\ \mathrm{d}t}{\frac{\bar{c}r(x)u}{n-1}} = \frac{P_{n,u,x}}{\frac{\bar{c}r(x)u}{n-1}} < p(x) + \delta. \end{aligned} \tag{47}$$

Assume also that $n_0(\delta)$ is so large that $(n-k)/n > 1 - \delta$ when $n > n_0$. Hence, when $n - 1 > n_0(\delta)$, where n_0 is independent from x, then

$$\int_0^{\sqrt{n-1}} (1 - F_{n,k,x}(u))^\omega u^{\gamma-1}\mathrm{d}u = \int_0^{\sqrt{n-1}} u^{\gamma-1} \left(\sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1-P_{n,u,x})^{n-1-j}\right)^\omega \mathrm{d}u$$

by applying Lemma 19. For brevity, introduce the following notations: $\bar{s} \doteq (p(x)+\delta)\bar{c}r(x)$, $\underline{s} \doteq (p(x)-\delta)\bar{c}r(x)$. Now using (47) we can continue the inequality as follows

$$\int_0^{\sqrt{n-1}} (1 - F_{n,k,x}(u))^\omega u^{\gamma-1}\mathrm{d}u \leq$$

$$
\begin{aligned}
&\leq \int_0^{\sqrt{n-1}} u^{\gamma-1} \left(\sum_{j=0}^{k-1} \binom{n-1}{j} \frac{\bar{s}^j u^j}{(n-1)^j} \left[1 - \frac{\underline{s}u}{n-1}\right]^{n-1-j} \right)^{\omega} \mathrm{d}u; \\
&\leq \int_0^{\sqrt{n-1}} u^{\gamma-1} \left(\sum_{j=0}^{k-1} \binom{n-1}{j} \frac{\bar{s}^j u^j}{(n-1)^j} \exp\left[-(n-1-j)\frac{\underline{s}u}{n-1}\right] \right)^{\omega} \mathrm{d}u \\
&\leq \sum_{j=0}^{k-1} \left(\frac{1}{j!}\right)^{\omega} \bar{s}^{j\omega} \int_0^{\sqrt{n-1}} u^{\gamma+j\omega-1} \underbrace{\exp\left[-\frac{n-1-j}{n-1}\omega\underline{s}u\right]}_{\leq \exp[-\omega(1-\delta)\underline{s}u] \text{ if } n > n_1(\delta)} \mathrm{d}u; \quad [\omega s(1-\delta) \geq 0, \gamma + j\omega \geq 0] \\
&\leq \sum_{j=0}^{k-1} \left(\frac{1}{j!}\right)^{\omega} \bar{s}^{j\omega} \left[(1-\delta)\omega\underline{s}\right]^{-(\gamma+j\omega)} \Gamma(\gamma + j\omega); \quad [\text{using (42)}] \\
&= (\bar{c}r(x))^{-\gamma} (p(x)-\delta)^{-\gamma} \sum_{j=0}^{k-1} \left(\frac{1}{j!}\right)^{\omega} \left(\frac{p(x)+\delta}{p(x)-\delta}\right)^{j\omega} [(1-\delta)\omega]^{-(\gamma+j\omega)} \Gamma(\gamma+j\omega) \\
&= (\bar{c}r(x))^{-\gamma} H(x,p,\delta,\omega) \leq (\bar{c}r_{\mathcal{M}})^{-\gamma} H(x,p,\delta,\omega).
\end{aligned}
$$

In the proof we also used Lemma 39. The proof of (44) is finished.

One might wonder if instead of using this approach, we could try the following cruder way, too:

$$
\begin{aligned}
\int_0^{\sqrt{n-1}} (1 - F_{n,k,x}(u))^{\omega} u^{\gamma-1} \mathrm{d}u &\leq \int_0^{\sqrt{n-1}} u^{\gamma-1} k n^k (1 - P_{n,u,x})^{n-k} \mathrm{d}u \\
&\leq k n^k \int_0^{\sqrt{n-1}} u^{\gamma-1} \exp[-\frac{n-k}{n-1}(p(x)-\delta)u\bar{c}r(x)] \mathrm{d}u \to \infty.
\end{aligned}
$$

But this approximation is too crude for us since the right hand side diverges to infinity, and hence it does not lead to useful upper bound. That is why in this case we also need to upper bound $P_{n,u,x}$. Nonetheless, this approach will work for bounding (45), where lower bounding of $P_{n,u,x}$ will be enough. Let us see now the proof of (45). In this case we have to upper bound

$$
\int_{\sqrt{n-1}}^{\infty} (1 - F_{n,k,x}(u))^{\omega} u^{\gamma-1} \mathrm{d}u = \int_{\sqrt{n-1}}^{\infty} u^{\gamma-1} \left[\sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j} \right]^{\omega} \mathrm{d}u. \tag{48}
$$

Let $\mathcal{A} \subseteq \mathbb{R}^d$ be an arbitrary measurable set, and introduce the $I(\mathcal{A}) = \int_{\mathcal{A}} p(t)\mathrm{d}t$ notations. We start with an easy observation.

Lemma 47 *If $u \geq \sqrt{n-1}$, $\omega \in (0,1]$, and n is at least as large that $\omega(n-k)-1 > 0$ holds, then*

$$
(1 - P_{n,u,x})^{\omega(n-k)-1} \leq \left[1 - I\left(\mathcal{M} \cap \mathcal{B}\left(x, \left\{\frac{1}{\sqrt{n-1}}\right\}^{1/d}\right)\right)\right]^{\omega(n-k)-1}. \tag{49}
$$

Proof If $u \geq \sqrt{n-1}$, then

$$I\left(\mathcal{M} \cap \mathcal{B}\left(x, \left(\frac{1}{\sqrt{n-1}}\right)^{1/d}\right)\right) \leq I\left(\mathcal{M} \cap \mathcal{B}\left(x, \left(\frac{u}{n-1}\right)^{1/d}\right)\right) \doteq P_{n,u,x}.$$

■

Assume again that p is uniformly Lebesgue approximable (Eq. (46)). We cannot use this time the previous approach, since u can diverge to ∞. However, we can at least lower bound $P_{n,u,x}$. Let $\delta > 0$, $\sqrt{n-1} \leq u$, and p be uniformly Lebesgue approximable with the corresponding $n_0(\delta)$ threshold number, which is independent of x, and let $n > n_0$. Then for almost all $x \in \mathcal{M}$ we have that

$$\begin{aligned}(p(x)-\delta)\frac{\bar{c}r(x)}{\sqrt{n-1}} &\leq I\left(\mathcal{M} \cap \mathcal{B}\left(x, \left(\frac{1}{\sqrt{n-1}}\right)^{1/d}\right)\right) \\ &\leq I\left(\mathcal{M} \cap \mathcal{B}\left(x, \left(\frac{u}{n-1}\right)^{1/d}\right)\right) \doteq P_{n,u,x}. \qquad (50)\end{aligned}$$

In turn, for almost all $x \in \mathcal{M}$, when $0 < \delta < p(x)$, $n > n_0$, and $\omega(n-k)-1 > 0$, then (thanks to $P_{n,u,x} \leq 1$, and $\binom{n-1}{j}$) it holds that

$$\begin{aligned}&\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}\left[\sum_{j=0}^{k-1}\binom{n-1}{j}(P_{n,u,x})^j(1-P_{n,u,x})^{n-1-j}\right]^{\omega}\mathrm{d}u \\ &\leq \int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}\left[kn^k(1-P_{n,u,x})^{n-k}\right]^{\omega}\mathrm{d}u \\ &= \left(kn^k\right)^{\omega}\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})^{(n-k)\omega-1}(1-P_{n,u,x})\mathrm{d}u \\ &\leq \left(kn^k\right)^{\omega}\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}\left(1-I(\widetilde{M})\right)^{(n-k)\omega-1}(1-P_{n,u,x})\mathrm{d}u \quad \text{[by (49)]} \\ &\leq \left(kn^k\right)^{\omega}\left(1-I(\widetilde{M})\right)^{(n-k)\omega-1}\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u \\ &\leq \left(kn^k\right)^{\omega}\exp\left[\frac{-((n-k)\,\omega-1)\underline{s}}{\sqrt{n-1}}\right]\times\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u \qquad (51)\end{aligned}$$

[using $n > n_0$ in (50)].

Here we introduced the $\widetilde{\mathcal{M}} \doteq \mathcal{M} \cap \mathcal{B}\left(x, \left(1/\sqrt{n-1}\right)^{1/d}\right)$ shorthands. We have to upper bound its last term, $\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u$, as well. We want to prove that asymptotically $\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u < v(n)$, where $v(n)$ is an appropriate polynomial. The previous bounding "trick" of $P_{n,u,x}$ would not be good enough now, since all what we would get is

as follows:

$$\begin{aligned}\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u &\leq \int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}\left(1-(p(x)-\delta)\frac{\bar{c}}{\sqrt{n-1}}\right)\mathrm{d}u \\ &= \left(1-(p(x)-\delta)\frac{\bar{c}}{\sqrt{n-1}}\right)\underbrace{\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}\mathrm{d}u}_{\infty}.\end{aligned}$$

Hence, we have to follow another route, and make stronger assumptions. The following lemma will show that

$$\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u < \frac{(n-1)^{\gamma}}{\gamma} + \int \|x-y\|^{\gamma}p(y)\ \mathrm{d}y$$

for almost all $x \in \mathcal{M}$. To finish the proof, let n_0 be so large that when $n > n_0$, then

$$\left(kn^k\right)^{\omega}\exp\left[-\big(\,(n-k)\,\omega-1\big)(p(x)-\delta)\frac{\bar{c}r(x))}{\sqrt{n-1}}\right] < \min\left(\delta_1, \frac{\delta_1}{(n-1)^{\gamma}/\gamma}\right).$$

Since $0 < r_{\mathcal{M}} < r(x)$ and p is bounded away from zero, there exists this n_0 threshold number and this is independent from x. ■

Lemma 48 *If $\gamma > 0$ and $\int \|x-y\|^{\gamma}p(y)\ \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$, then*

$$\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u \leq \frac{(n-1)^{\gamma}}{\gamma} + \int \|x-y\|^{\gamma}p(y)\ \mathrm{d}y.$$

Proof By using the $\mathrm{d}u = \mathrm{d}t(n-1)$, $u = t(n-1)$ integral transformation, it is easy to see that

$$\begin{aligned}&\int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}(1-P_{n,u,x})\mathrm{d}u \qquad (52)\\ &= \int_{\sqrt{n-1}}^{\infty} u^{\gamma-1}\left(1-I\left(\mathcal{M}\cap\mathcal{B}(x,\left(\frac{u}{n-1}\right)^{1/d})\right)\right)\mathrm{d}u \\ &= (n-1)^{\gamma-1}\int_{\frac{1}{\sqrt{n-1}}}^{\infty} t^{\gamma-1}\left(1-I\left(\mathcal{M}\cap\mathcal{B}(x,t^{1/d})\right)\right)(n-1)\mathrm{d}t \\ &= (n-1)^{\gamma}\int_{\frac{1}{\sqrt{n-1}}}^{1} t^{\gamma-1}\left(1-I\left(\mathcal{M}\cap\mathcal{B}(x,t^{1/d})\right)\right)\mathrm{d}t \qquad (53)\\ &\qquad +(n-1)^{\gamma}\int_{1}^{\infty} t^{\gamma-1}\left(1-I\left(\mathcal{M}\cap\mathcal{B}(x,t^{1/d})\right)\right)\mathrm{d}t. \qquad (54)\end{aligned}$$

We upper bound (53) first:

$$(n-1)^{\gamma}\int_{\frac{1}{\sqrt{n-1}}}^{1} t^{\gamma-1}\underbrace{\left(1-\int_{\mathcal{M}\cap\mathcal{B}(x,t^{1/d})} p(y)\ \mathrm{d}y\right)}_{\leq 1}\mathrm{d}t$$

$$
\begin{aligned}
&\leq (n-1)^{\gamma} \int_{\frac{1}{\sqrt{n-1}}}^{1} t^{\gamma-1} \mathrm{d}t \\
&\leq (n-1)^{\gamma} \left[\frac{t^{\gamma}}{\gamma}\right]_{\frac{1}{\sqrt{n-1}}}^{1} = (n-1)^{\gamma} \left[\frac{1}{\gamma} - \frac{(n-1)^{-\gamma/2}}{\gamma}\right] \\
&= \frac{(n-1)^{\gamma}}{\gamma} - \frac{(n-1)^{\gamma/2}}{\gamma} \leq \frac{(n-1)^{\gamma}}{\gamma}.
\end{aligned}
$$

Now we upper bound (54). Thanks to (14), we have that

$$
F_{2,1,x}(t) = 1 - (1 - P_{2,t,x})^1 = P_{2,t,x} = \int_{\mathcal{M} \cap \mathcal{B}(x, t^{1/d})} p(y)\ \mathrm{d}y, \tag{55}
$$

and thus

$$
\begin{aligned}
&\gamma \int_1^{\infty} t^{\gamma-1} \left(1 - \int_{\mathcal{M} \cap \mathcal{B}(x, t^{1/d})} p(y)\ \mathrm{d}y\right) \mathrm{d}t \\
&\quad = \gamma \int_1^{\infty} t^{\gamma-1} \left(1 - F_{2,1,x}(t)\right) \mathrm{d}t; \quad \text{[using (55)]} \\
&\quad = \int_1^{\infty} t^{\gamma}\ \mathrm{d}F_{2,1,x}(t) - (1 - F_{2,1,x}(1)); \quad \text{[using (38)]} \\
&\quad \leq \int_0^{\infty} t^{\gamma}\ \mathrm{d}F_{2,1,x}(t) - (1 - F_{2,1,x}(1)) \\
&\quad = \mathbb{E}\left[\|X_1 - X_2\|^{\gamma} | X_1 = x\right] - (1 - F_{2,1,x}(1)); \quad \text{[by the def. of } F_{n,k,x}\text{]} \\
&\quad = -(1 - F_{2,1,x}(1)) + \int \|x - y\|^{\gamma} p(y)\ \mathrm{d}y \\
&\quad \leq \int \|x - y\|^{\gamma} p(y)\ \mathrm{d}y < \infty; \quad \text{[by assumption]}.
\end{aligned}
$$

$F_{2,1,x} \in [0,1]$, $\forall x$, since this is a cdf. The $\int \|x-y\|^{\gamma} p(y) \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$ was our assumption here, and is similar to that of in Wang et al. (2009b). This finishes the proof of Lemma 48. ■

Proof [Proof of (21) in Theorem 25] The proof is almost the same as in proving (19). However, in this case instead of assuming the uniformly Lebesgue approximability and "bounded away from zero" properties of p, we use the Lebesgue lemma locally only, that is, we use a different $\delta(x)$ in each $x \in \mathcal{M}$. As a result, there will not exist a global threshold number n_0; these $n_0(\delta(x))$ threshold numbers will be different in each point of $x \in \mathcal{M}$. ■

Proof [Proof of (22) in Theorem 25]

Note that if p is bounded, $p(x) \leq \bar{p}$, then

$$
\int_{\mathcal{M} \cap \mathcal{B}\left(x, \left(\frac{u}{n-1}\right)^{1/d}\right)} p(t)\ \mathrm{d}t \doteq P_{n,u,x} \leq \frac{\bar{p} u}{n-1} \bar{c}, \quad \forall x \in \mathcal{M}, \forall n, \forall u > 0. \tag{56}
$$

Therefore,

$$\begin{aligned}
\frac{F_{n,k,x}(u)}{u^k} &= \frac{1}{u^k}\sum_{j=k}^{n-1}\binom{n-1}{j}(P_{n,u,x})^j(1-P_{n,u,x})^{n-1-j}; \quad \text{[thanks to (14)]} \\
&\leq \frac{1}{u^k}\sum_{j=k}^{n-1}\binom{n-1}{j}\left(\bar{p}\bar{c}\frac{u}{n-1}\right)^j \\
&\leq \sum_{j=k}^{n-1}\frac{1}{j!}\left(\bar{p}\bar{c}\right)^j u^{j-k}; \quad \left[\text{since } \binom{n-1}{j}\frac{1}{(n-1)^j} \leq \frac{1}{j!}\right] \\
&\leq \left(\bar{p}\bar{c}\right)^k \sum_{j=k}^{n-1}\frac{1}{(j-k)!}\left(\bar{p}\bar{c}\right)^{j-k}\beta^{j-k}; \quad \text{by assumption } u<\beta \\
&\leq \left(\bar{p}\bar{c}\right)^k \exp(\bar{p}\bar{c}\beta) = \hat{L}(\bar{p},\beta).
\end{aligned}$$

■

Proof [Proof of (24) in Theorem 25] Use that $F_{n,k,x} \leq 1$ and $\gamma < 0$. Thus,

$$\int_\beta^\infty u^{\gamma-1}\left(F_{n,k,x}(u)\right)^\omega \mathrm{d}u \leq \int_\beta^\infty u^{\gamma-1} = \left[\frac{u^\gamma}{\gamma}\right]_\beta^\infty = \frac{-\beta^\gamma}{\gamma}.$$

■

Proof [Proof of (25) in Theorem 25] According to (22), if $u \leq \beta$, then $F^\omega_{n,k,x}(u) \leq u^{k\omega}\hat{L}^\omega(\bar{p},\beta)$. In turn,

$$\begin{aligned}
\int_0^\beta u^{\gamma-1}\left(F_{n,k,x}(u)\right)^\omega \mathrm{d}u &\leq \int_0^\beta u^{\gamma-1}\left(u^k\hat{L}(\bar{p},\beta)\right)^\omega \mathrm{d}u \\
&= \hat{L}^\omega(\bar{p},\beta)\int_0^\beta u^{\gamma+k\omega-1}\mathrm{d}u = \hat{L}^\omega(\bar{p},\beta)\frac{\beta^{\gamma+k\omega}}{\gamma+k\omega},
\end{aligned}$$

assuming $\gamma + k\omega > 0$. ■

Appendix E. Proofs for the Asymptotic Variance

We will need the following observation:

Lemma 49

$$F_{n,k,x}(u) \leq F_{n,k-1,x}(u) \leq \ldots \leq F_{n,0,x}(u) = 1.$$

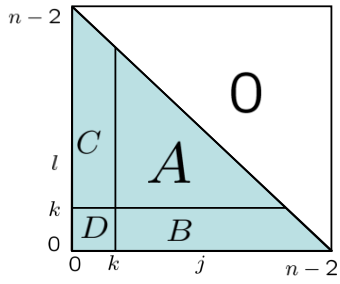


Figure 13: Explanation of the proof of Lemma 29. $0 \leq j, l$, and $j + l \leq n - 2$.

Proof of Lemma 28

Proof Follow the proof of Lemma 19 and calculate this distribution in a "closed" form.

$$\begin{aligned}
F_{n,k,x_1,x_2}(u,v) &= \Pr(\zeta_{n,k,1} < u \wedge \zeta_{n,k,2} < v | X_1 = x_1, X_2 = x_2) \\
&= \Pr((n-1)\rho_k^d(1) < u \wedge (n-1)\rho_k^d(2) < v | X_1 = x_1, X_2 = x_2) \\
&= \Pr\left(\rho_k(1) < R_n(u) \wedge \rho_k(2) < R_n(v) | X_1 = x_1, X_2 = x_2\right) \\
&= \Pr\left(k \text{ elements or more} \in \mathcal{B}(x_1, R_n(u))\right. \\
&\qquad \left. \wedge k \text{ elements or more} \in \mathcal{B}(x_2, R_n(v)) | X_1 = x_1, X_2 = x_2\right) \\
&= \sum_{j=k}^{n-2} \sum_{l=k}^{n-2-j} \binom{n-2}{j} \binom{n-2-j}{l} S_{n,j,l},
\end{aligned}$$

where

$$S_{n,j,l} \doteq (P_{n,u,x_1})^j (P_{n,v,x_2})^l (1 - P_{n,u,x_1} - P_{n,v,x_2})^{n-2-j-l},$$

and $P_{n,u,x_i} \doteq \int_{\mathcal{M} \cap \mathcal{B}(x_i, R_n(u))} p(t) \mathrm{d}t$. In the last equation we used that $\max(R_n(u), R_n(v)) \leq \|x_1 - x_2\|$, and thus $x_2 \notin \mathcal{B}(x_1, R_n(u))$, and $x_1 \notin \mathcal{B}(x_2, R_n(v))$.

■

Proof of Lemma 29

Proof The lemma below can be easily proven by induction.

Lemma 50 *Let $0 \leq p_1, p_2$, and $p_1 + p_2 \leq 1$. Then*

$$\sum_{l=0}^{n} \binom{n}{l} p_2^l (1 - p_1 - p_2)^{n-2} = (1 - p_1)^n.$$

Figure 13 will help us understand the proof. The figure shows the domain of a multinomial distribution with α and β random variables along its axes. Let $p_1 \doteq P_{n,u,x_1}$, $p_2 \doteq P_{n,v,x_2}$,

and consider the following multinomial distribution

$$\Pr(\alpha = j, \beta = l | j + l \leq n-2) \doteq \binom{n-2}{j} \binom{n-2-j}{l} p_1^j p_2^l (1-p_1-p_2)^{n-2-j-l}.$$

Now, thanks to (31), $F_{n,k,x_1,x_2}(u,v) = \sum_{j=k}^{n-2} \sum_{l=k}^{n-2} \Pr(\alpha = j, \beta = l | j + l \leq n-2)$. This is the domain A in Figure 13. We also have that

$$\begin{aligned}
\Pr(\alpha = j) &= \sum_{l=0}^{n-2-j} \Pr(\alpha = j, \beta = l) \\
&= \binom{n-2}{j} p_1^j \sum_{l=0}^{n-2-j} \binom{n-2-j}{l} p_2^l (1-p_1-p_2)^{n-2-j-l} \\
&= \binom{n-2}{j} p_1^j (1-p_1)^{n-2-j} \quad \text{[by Lemma 50].} \qquad (57)
\end{aligned}$$

Similarly,

$$\Pr(\beta = l) = \sum_{j=0}^{n-2-l} \Pr(\alpha = j, \beta = l) = \binom{n-2}{l} p_2^l (1-p_2)^{n-2-l}.$$

Figure 13 displays the probability mass of a multinomial distribution. Here $A + B + C + D = 1$, and therefore A=1-(B+D)-(C+D)+D. It is easy to see that

$$\begin{aligned}
D + C &= \sum_{j=0}^{k-1} \sum_{l=0}^{n-2-j} \Pr(\alpha = j, \beta = l) = \sum_{j=0}^{k-1} \Pr(\alpha = j) \\
&= \sum_{j=0}^{k-1} \binom{n-2}{j} (P_{n,u,x_1})^j (1 - P_{n,u,x_1})^{n-2-j} \quad \text{[using Eq. (57)]} \\
&= 1 - \tilde{F}_{n-1,k,x_1}(u) \quad \text{[by Eq. (14)]},
\end{aligned}$$

and similarly

$$D + B = 1 - \tilde{F}_{n-1,k,x_2}(v).$$

We can conclude that

$$\begin{aligned}
A &= F_{n,k,x_1,x_2}(u,v) \\
&= F_{n-1,k,x_1}(u) + F_{n-1,k,x_2}(v) - 1 + \sum_{j=0}^{k-1} \sum_{l=0}^{k-1} \binom{n-2}{j} \binom{n-2-j}{l} S_{n,j,l}.
\end{aligned}$$

■

Proof of Lemma 30

Proof We already know from Lemma 29 that if $x_1 \neq x_2$ fixed, and n is large enough, then we have that $\max(R_n(u), R_n(v)) < \|x_1 - x_2\|$, and thus Lemma 29 holds. Investigate the lemma's first three terms:

$$\begin{aligned}\lim_{n\to\infty}[F_{n-1,k,x_1}(u) + F_{n-1,k,x_2}(v) - 1] &= F_{k,x_1}(u) + F_{k,x_2}(v) - 1 \\ &= 1 - \sum_{j=0}^{k-1} \frac{1}{j!}(u\bar{c}p(x_1))^j \exp(-u\bar{c}p(x_1)) - \sum_{l=0}^{k-1} \frac{1}{l!}(v\bar{c}p(x_2))^l \exp(-v\bar{c}p(x_2)).\end{aligned}$$

The last term is also easy to study:

$$\begin{aligned}\lim_{n\to\infty} \sum_{j=0}^{k-1}\sum_{l=0}^{k-1} \binom{n-2}{j}\binom{n-2-j}{l} S_{n,j,l} \\ = \sum_{j=0}^{k-1}\sum_{l=0}^{k-1} \frac{1}{j!}\frac{1}{l!}(u\bar{c}p(x_1))^j (v\bar{c}p(x_2))^l \exp(-u\bar{c}p(x_1) - v\bar{c}p(x_2)).\end{aligned}$$

Its proof is analogous to the proof of Lemma 20, we just have to bound P_{n,u,x_1} and P_{n,v,x_2}. We omit the details.

In turn,

$$\begin{aligned}\lim_{n\to\infty} F_{n,k,x_1,x_2}(u,v) \\ = \left[1 - \sum_{j=0}^{k-1} \frac{1}{j!}(u\bar{c}p(x_1))^j \exp(-u\bar{c}p(x_1))\right]\left[1 - \sum_{l=0}^{k-1} \frac{1}{l!}(v\bar{c}p(x_2))^j \exp(-v\bar{c}p(x_2))\right].\end{aligned}$$

This is what we wanted to prove. ■

Proof of Theorem 34

Before starting the proof of this theorem, observe that

$$\begin{aligned} f_n(x_1,x_2) &= \mathbb{E}[(n-1)^{2\gamma}\rho_k(1)^{d\gamma}\rho_k(2)^{d\gamma}|X_1 = x_1, X_2 = x_2] \\ &= \mathbb{E}[\zeta_{n,k,1}^{\gamma}\zeta_{n,k,2}^{\gamma}|X_1 = x_1, X_2 = x_2] \\ &= \int_0^\infty \int_0^\infty u^\gamma v^\gamma F_{n,k,x_1,x_2}(\mathrm{d}u, \mathrm{d}v).\end{aligned}$$

Proof [Proof of (32) in Theorem 34]

Let $\gamma > 0$, and use Lemma 45 to rewrite $F_{n,k,x_1,x_2}(\mathrm{d}u, \mathrm{d}v)$.

$$\begin{aligned} f_n(x_1,x_2) &= \int_0^\infty \int_0^\infty u^\gamma v^\gamma F_{n,k,x_1,x_2}(\mathrm{d}u, \mathrm{d}v) \\ &= \gamma^2 \int_0^\infty \int_0^\infty u^{\gamma-1}v^{\gamma-1}\left[1 - F_{n,k,x_1,x_2}(u,\infty) - F_{n,k,x_1,x_2}(\infty,v) + F_{n,k,x_1,x_2}(u,v)\right]\mathrm{d}u\mathrm{d}v\end{aligned}$$

$$= \gamma^2 \int_0^\infty \int_0^\infty u^{\gamma-1} v^{\gamma-1} (1 - F_{n,k,x_1,x_2}(u,\infty))(1 - F_{n,k,x_1,x_2}(\infty,v)) \mathrm{d}u\mathrm{d}v \tag{58}$$

$$+ \gamma^2 \int_0^\infty \int_0^\infty u^{\gamma-1} v^{\gamma-1} \left[F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty) F_{n,k,x_1,x_2}(\infty,v)\right] \mathrm{d}u\mathrm{d}v. \tag{59}$$

By definition

$$F_{n,k,x_1,x_2}(u,v) = \Pr(\xi_{n,k,x_1} < u \wedge \xi_{n,k,x_2} < v | X_1 = x_1, X_2 = x_2).$$

Note that

$$F_{n,k,x_1,x_2}(u,v) \quad \leq \quad \min[F_{n,k,x_1,x_2}(u,\infty), F_{n,k,x_1,x_2}(\infty,v)].$$

Split the domain of $(u,v) \in [0,\infty]^2$ into four parts.

1. When $\|x_1 - x_2\| > R_n(u) = (u/(n-1))^{1/d}$, and $\|x_1 - x_2\| > R_n(v) = (v/(n-1))^{1/d}$, then $\Lambda = \Lambda(n, x_1, x_2) \doteq (n-1)\|x_1 - x_2\|^d > u$, and $\Lambda = (n-1)\|x_1 - x_2\|^d > v$. Furthermore, $F_{n,k,x_1,x_2}(u,\infty) = F_{n-1,k,x_1}(u)$, and $F_{n,k,x_1,x_2}(\infty,v) = F_{n-1,k,x_2}(v)$ by Lemma 33. Now,

$$\begin{aligned} &F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty) F_{n,k,x_1,x_2}(\infty,v) \\ &\qquad \leq \min[F_{n-1,k,x_1}(u), F_{n-1,k,x_2}(v)] - F_{n-1,k,x_1}(u) F_{n-1,k,x_2}(v) \\ &\qquad \leq \sqrt{1 - F_{n-1,k,x_1}(u)} \sqrt{1 - F_{n-1,k,x_2}(v)}; \quad \text{[Using Lemma 38]}. \end{aligned}$$

Thus,

$$\begin{aligned} &\int_0^\Lambda \int_0^\Lambda u^{\gamma-1} v^{\gamma-1} \left[F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty) F_{n,k,x_1,x_2}(\infty,v)\right] \mathrm{d}u\mathrm{d}v \\ &\qquad \leq \int_0^\Lambda \int_0^\Lambda u^{\gamma-1} v^{\gamma-1} \sqrt{1 - F_{n-1,k,x_1}(u)} \sqrt{1 - F_{n-1,k,x_2}(v)} \mathrm{d}u\mathrm{d}v. \end{aligned}$$

Similarly,

$$\begin{aligned} &\int_0^\Lambda \int_0^\Lambda u^{\gamma-1} v^{\gamma-1} (1 - F_{n,k,x_1,x_2}(u,\infty))(1 - F_{n,k,x_1,x_2}(\infty,v)) \mathrm{d}u\mathrm{d}v \\ &\qquad = \int_0^\Lambda \int_0^\Lambda u^{\gamma-1} v^{\gamma-1} (1 - F_{n-1,k,x_1}(u))(1 - F_{n-1,k,x_2}(v)) \mathrm{d}u\mathrm{d}v. \end{aligned}$$

2. When $\|x_1 - x_2\| \leq R_n(u)$, and $\|x_1 - x_2\| > R_n(v)$, then $\Lambda = \Lambda(n, x_1, x_2) \doteq (n-1)\|x_1 - x_2\|^d \leq u$, and $\Lambda = (n-1)\|x_1 - x_2\|^d > v$. Furthermore, $F_{n,k,x_1,x_2}(u,\infty) = F_{n-1,k-1,x_1}(u)$, and $F_{n,k,x_1,x_2}(\infty,v) = F_{n-1,k,x_2}(v)$ by Lemma 33. Now,

$$\begin{aligned} &F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty) F_{n,k,x_1,x_2}(\infty,v) \\ &\qquad \leq \min[F_{n-1,k-1,x_1}(u), F_{n-1,k,x_2}(v)] - F_{n-1,k-1,x_1}(u) F_{n-1,k,x_2}(v) \end{aligned}$$

$$\leq \sqrt{1-F_{n-1,k-1,x_1}(u)}\sqrt{1-F_{n-1,k,x_2}(v)}; \quad \text{[Using Lemma 38]}.$$
$$\leq \sqrt{1-F_{n-1,k,x_1}(u)}\sqrt{1-F_{n-1,k,x_2}(v)}; \quad \text{[Using Lemma 49]}.$$

Thus,

$$\int_\Lambda^\infty \int_0^\Lambda u^{\gamma-1}v^{\gamma-1}\left[F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty)F_{n,k,x_1,x_2}(\infty,v)\right]\mathrm{d}u\mathrm{d}v$$
$$\leq \int_\Lambda^\infty \int_0^\Lambda u^{\gamma-1}v^{\gamma-1}\sqrt{1-F_{n-1,k,x_1}(u)}\sqrt{1-F_{n-1,k,x_2}(v)}\mathrm{d}u\mathrm{d}v.$$

Similarly,

$$\int_\Lambda^\infty \int_0^\Lambda u^{\gamma-1}v^{\gamma-1}(1-F_{n,k,x_1,x_2}(u,\infty))(1-F_{n,k,x_1,x_2}(\infty,v))\mathrm{d}u\mathrm{d}v$$
$$\leq \int_\Lambda^\infty \int_0^\Lambda u^{\gamma-1}v^{\gamma-1}(1-F_{n-1,k,x_1}(u))(1-F_{n-1,k,x_2}(v))\mathrm{d}u\mathrm{d}v.$$

3. The $\|x_1-x_2\| > R_n(u)$ and $\|x_1-x_2\| \leq R_n(v)$ case is similar to the previous one, just replace the role of u and v. Then $\Lambda = \Lambda(n,x_1,x_2) \doteq (n-1)\|x_1-x_2\|^d > u$, and $\Lambda = (n-1)\|x_1-x_2\|^d \leq v$. Furthermore, $F_{n,k,x_1,x_2}(u,\infty) = F_{n-1,k,x_1}(u)$, and $F_{n,k,x_1,x_2}(\infty,v) = F_{n-1,k-1,x_2}(v)$ by Lemma 33. Now,

$$\begin{aligned}
&F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty)F_{n,k,x_1,x_2}(\infty,v)\\
&\quad\leq \min[F_{n-1,k,x_1}(u), F_{n-1,k-1,x_2}(v)] - F_{n-1,k,x_1}(u)F_{n-1,k-1,x_2}(v)\\
&\quad\leq \sqrt{1-F_{n-1,k,x_1}(u)}\sqrt{1-F_{n-1,k-1,x_2}(v)}\\
&\quad\leq \sqrt{1-F_{n-1,k,x_1}(u)}\sqrt{1-F_{n-1,k,x_2}(v)}.
\end{aligned}$$

$$\int_0^\Lambda \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}\left[F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty)F_{n,k,x_1,x_2}(\infty,v)\right]\mathrm{d}u\mathrm{d}v$$
$$\leq \int_0^\Lambda \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}\sqrt{1-F_{n-1,k,x_1}(u)}\sqrt{1-F_{n-1,k,x_2}(v)}\mathrm{d}u\mathrm{d}v.$$

Similarly,

$$\int_0^\Lambda \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}(1-F_{n,k,x_1,x_2}(u,\infty))(1-F_{n,k,x_1,x_2}(\infty,v))\mathrm{d}u\mathrm{d}v$$
$$\leq \int_0^\Lambda \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}(1-F_{n-1,k,x_1}(u))(1-F_{n-1,k,x_2}(v))\mathrm{d}u\mathrm{d}v.$$

4. When $\|x_1-x_2\| \leq R_n(u)$, and $\|x_1-x_2\| \leq R_n(v)$, then $\Lambda = \Lambda(n,x_1,x_2) \doteq (n-1)\|x_1-x_2\|^d \leq u$, and $\Lambda = (n-1)\|x_1-x_2\|^d \leq v$. Furthermore, $F_{n,k,x_1,x_2}(u,\infty) = F_{n-1,k-1,x_1}(u)$, and $F_{n,k,x_1,x_2}(\infty,v) = F_{n-1,k-1,x_2}(v)$ by Lemma 33. Now,

$$F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty)F_{n,k,x_1,x_2}(\infty,v)$$

$$\leq \min[F_{n-1,k-1,x_1}(u), F_{n-1,k-1,x_2}(v)] - F_{n-1,k-1,x_1}(u)F_{n-1,k-1,x_2}(v)$$
$$\leq \sqrt{1-F_{n-1,k-1,x_1}(u)}\sqrt{1-F_{n-1,k-1,x_2}(v)}$$
$$\leq \sqrt{1-F_{n-1,k,x_1}(u)}\sqrt{1-F_{n-1,k,x_2}(v)}.$$

$$\int_\Lambda^\infty \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}\left[F_{n,k,x_1,x_2}(u,v) - F_{n,k,x_1,x_2}(u,\infty)F_{n,k,x_1,x_2}(\infty,v)\right]\mathrm{d}u\mathrm{d}v$$
$$\leq \int_\Lambda^\infty \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}\sqrt{1-F_{n-1,k,x_1}(u)}\sqrt{1-F_{n-1,k,x_2}(v)}\mathrm{d}u\mathrm{d}v.$$

Similarly,

$$\int_\Lambda^\infty \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}(1-F_{n,k,x_1,x_2}(u,\infty))(1-F_{n,k,x_1,x_2}(\infty,v))\mathrm{d}u\mathrm{d}v$$
$$\leq \int_\Lambda^\infty \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}(1-F_{n-1,k,x_1}(u))(1-F_{n-1,k,x_2}(v))\mathrm{d}u\mathrm{d}v.$$

Putting the pieces together we can conclude that if $0<\gamma$, then

$$\int_0^\infty \int_0^\infty u^\gamma v^\gamma F_{n,k,x_1,x_2}(\mathrm{d}u,\mathrm{d}v)$$
$$\leq \gamma^2\left[\int_0^\infty u^{\gamma-1}\sqrt{1-F_{n-1,k,x_1}(u)}\mathrm{d}u\right]\left[\int_0^\infty v^{\gamma-1}\sqrt{1-F_{n-1,k,x_2}(v)}\mathrm{d}v\right]$$
$$+\gamma^2\left[\int_0^\infty u^{\gamma-1}(1-F_{n-1,k,x_1}(u))\mathrm{d}u\right]\left[\int_0^\infty v^{\gamma-1}(1-F_{n-1,k,x_2}(v))\mathrm{d}v\right]$$
$$\leq 2\gamma^2\left[\int_0^\infty u^{\gamma-1}\sqrt{1-F_{n-1,k,x_1}(u)}\mathrm{d}u\right]\left[\int_0^\infty v^{\gamma-1}\sqrt{1-F_{n-1,k,x_2}(v)}\mathrm{d}v\right].$$

Using (19) with $\omega = 1/2$ we have what we wanted to prove

$$\int_0^\infty \int_0^\infty u^\gamma v^\gamma F_{n,k,x_1,x_2}(\mathrm{d}u,\mathrm{d}v) \leq 2\gamma^2 L(x_1,1/2,k,\gamma,p,\delta,\delta_1)L(x_2,1/2,k,\gamma,p,\delta,\delta_1).$$

■

Proof [Proof of (33) in Theorem 34] The proof is almost the same as the proof of (32). However, in this case instead of assuming the uniformly Lebesgue approximability and "bounded away from zero" properties of p, we use the Lebesgue lemma locally only, that is, we use a different $\delta(x)$ in each $x \in \mathcal{M}$. As a result, there will not exist a global threshold number n_0; these $n_0(x_1,x_2)$ threshold numbers will be different for each $x_1, x_2 \in \mathcal{M}$ pairs. ■

Proof [Proof of (34) and (35) in Theorem 34]

If $\gamma < 0$, then

$$\int_0^\infty \int_0^\infty u^\gamma v^\gamma F_{n,k,x_1,x_2}(\mathrm{d}u,\mathrm{d}v) = \gamma^2 \int_0^\infty \int_0^\infty u^{\gamma-1}v^{\gamma-1}F_{n,k,x_1,x_2}(u,v)\mathrm{d}u\mathrm{d}v.$$

In this case, we can follow the approach presented in Leonenko et al. (2008a). Split the domain of $(u,v) \in [0,\infty]^2$ into four parts again.

1. When $\|x_1 - x_2\| > R_n(u)$, and $\|x_1 - x_2\| > R_n(v)$, then $\Lambda = \Lambda(n, x_1, x_2) \doteq (n-1)\|x_1 - x_2\|^d > u$, and $\Lambda = (n-1)\|x_1 - x_2\|^d > v$. $F_{n,k,x_1,x_2}(u,\infty) = F_{n-1,k,x_1}(u)$, $F_{n,k,x_1,x_2}(\infty, v) = F_{n-1,k,x_2}(v)$.

$$\begin{aligned} F_{n,k,x_1,x_2}(u,v) &\leq \min[F_{n,k,x_1,x_2}(u,\infty), F_{n,k,x_1,x_2}(\infty,v)] \\ &\leq \sqrt{F_{n-1,k,x_1}(u) F_{n-1,k,x_2}(v)}. \end{aligned}$$

Thus,

$$\begin{aligned} &\int_0^\Lambda \int_0^\Lambda u^{\gamma-1} v^{\gamma-1} F_{n,k,x_1,x_2}(u,v) \mathrm{d}u \mathrm{d}v \\ &\qquad \leq \left[\int_0^\Lambda u^{\gamma-1} \sqrt{F_{n-1,k,x_1}(u)} \mathrm{d}u\right] \left[\int_0^\Lambda v^{\gamma-1} \sqrt{F_{n-1,k,x_2}(v)} \mathrm{d}v\right] \\ &\qquad \leq \left[\int_0^1 u^{\gamma-1} \sqrt{F_{n-1,k,x_1}(u)} \mathrm{d}u + \int_1^\infty u^{\gamma-1} \sqrt{F_{n-1,k,x_1}(u)} \mathrm{d}u\right] \times \\ &\qquad\qquad \times \left[\int_0^1 v^{\gamma-1} \sqrt{F_{n-1,k,x_2}(v)} \mathrm{d}v + \int_1^\infty v^{\gamma-1} \sqrt{F_{n-1,k,x_2}(v)} \mathrm{d}v\right] \\ &\qquad \leq \left[\int_0^1 u^{\gamma-1} \sqrt{F_{n-1,k,x_1}(u)} \mathrm{d}u + \int_1^\infty u^{\gamma-1} \mathrm{d}u\right] \times \\ &\qquad\qquad \times \left[\int_0^1 v^{\gamma-1} \sqrt{F_{n-1,k,x_2}(v)} \mathrm{d}v + \int_1^\infty v^{\gamma-1} \mathrm{d}v\right] \\ &\qquad \leq \left[\frac{L^{1/2}(\bar{p},1)}{k/2+\gamma} - \frac{1}{\gamma}\right]^2; \quad \text{[assuming } \ k/2+\gamma > 0.\text{]} \end{aligned}$$

In the last inequality we used (24) and (25).

2. When $\|x_1 - x_2\| \leq R_n(u)$, and $\|x_1 - x_2\| > R_n(v)$, then $\Lambda = \Lambda(n, x_1, x_2) \doteq (n-1)\|x_1 - x_2\|^d \leq u$ and $\Lambda = (n-1)\|x_1 - x_2\|^d > v$. $F_{n,k,x_1,x_2}(u,\infty) = F_{n-1,k-1,x_1}(u)$, and $F_{n,k,x_1,x_2}(\infty, v) = F_{n-1,k,x_2}(v)$.

$$\begin{aligned} F_{n,k,x_1,x_2}(u,v) &\leq \min[F_{n,k,x_1,x_2}(u,\infty), F_{n,k,x_1,x_2}(\infty,v)] \\ &= \min[F_{n-1,k-1,x_1}(u), F_{n-1,k,x_2}(v)] \\ &\leq F^{1/2}_{n-1,k-1,x_1}(u) F^{1/2}_{n-1,k,x_2}(v). \end{aligned}$$

If $k > 1$, then

$$\begin{aligned} &\int_\Lambda^\infty \int_0^\Lambda u^{\gamma-1} v^{\gamma-1} F_{n,k,x_1,x_2}(u,v) \mathrm{d}u \mathrm{d}v \\ &\qquad \leq \left[\int_\Lambda^\infty u^{\gamma-1} F^{1/2}_{n-1,k-1,x_1}(u) \mathrm{d}u\right] \left[\int_0^\Lambda v^{\gamma-1} F^{1/2}_{n-1,k,x_2}(v) \mathrm{d}v\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \left[\int_0^1 u^{\gamma-1} F^{1/2}_{n-1,k-1,x_1}(u)\mathrm{d}u + \int_1^\infty u^{\gamma-1}\mathrm{d}u\right] \times \\
&\qquad \times \left[\int_0^1 v^{\gamma-1} F^{1/2}_{n-1,k,x_2}(v)\mathrm{d}v + \int_1^\infty v^{\gamma-1}\mathrm{d}v\right] \\
&\leq \left[\int_0^1 u^{\gamma-1} F^{1/2}_{n-1,k-1,x_1}(u)\mathrm{d}u - \frac{1}{\gamma}\right]\left[\int_0^1 v^{\gamma-1} F^{1/2}_{n-1,k,x_2}(v)\mathrm{d}v - \frac{1}{\gamma}\right] \\
&\leq \left[\frac{\hat{L}^{1/2}(\bar{p},1)}{(k-1)/2+\gamma} - \frac{1}{\gamma}\right]\left[\frac{\hat{L}^{1/2}(\bar{p},1)}{k/2+\gamma} - \frac{1}{\gamma}\right].
\end{aligned}
$$

Here we used (25) again and assumed that $(k-1)/2+\gamma > 0$, and thus $k/2+\gamma > 0$, too. If $k=1$, then the bounding above cannot be applied, since $F_{n-1,0,x_1}(u) = 1$, $\gamma < 0$, and thus $\int_0^1 u^{\gamma-1}\mathrm{d}u = \infty$. In this case when $(u,v) \in [\Lambda,\infty] \times [0,\Lambda]$, we have that $x_2 \in \mathcal{B}(x_1, R_n(u))$ and $x_1 \notin \mathcal{B}(x_2, R_n(v)$, thus $F_{n,1,x_1,x_2}(u,v) = F_{n-1,1,x_2}(v)$. In turn,

$$
\begin{aligned}
&\int_\Lambda^\infty \int_0^\Lambda u^{\gamma-1}v^{\gamma-1} F_{n,1,x_1,x_2}(u,v)\mathrm{d}u\mathrm{d}v \\
&\qquad = \int_\Lambda^\infty \int_0^\Lambda u^{\gamma-1}v^{\gamma-1} F_{n-1,1,x_2}(v)\mathrm{d}u\mathrm{d}v \\
&\qquad = \left[\int_\Lambda^\infty u^{\gamma-1}\mathrm{d}u\right]\left[\int_0^\Lambda v^{\gamma-1} F_{n-1,1,x_2}(v)\mathrm{d}v\right] \\
&\qquad = -\frac{\Lambda^\gamma}{\gamma}\int_0^\Lambda v^{\gamma-1} F_{n-1,1,x_2}(v)\mathrm{d}v \\
&\qquad \leq -\frac{\Lambda^\gamma}{\gamma}\left[\int_0^1 v^{\gamma-1} F_{n-1,1,x_2}(v)\mathrm{d}v + \int_1^\infty v^{\gamma-1}\mathrm{d}v\right] \\
&\qquad \leq -\frac{\Lambda^\gamma}{\gamma}\left[\frac{\hat{L}(\bar{p},1)}{1+\gamma} - \frac{1}{\gamma}\right]; \quad \text{[using (25) and assuming } 1+\gamma > 0\text{]} \\
&\qquad = -\frac{(n-1)^\gamma}{\gamma}\|x_1 - x_2\|^{d\gamma}\left[\frac{\hat{L}(\bar{p},1)}{1+\gamma} - \frac{1}{\gamma}\right] \\
&\qquad \leq -\frac{1}{\gamma}\|x_1 - x_2\|^{d\gamma}\left[\frac{\hat{L}(\bar{p},1)}{1+\gamma} - \frac{1}{\gamma}\right]; \quad \text{[since } (n-1)^\gamma \leq 1\text{]}.
\end{aligned}
$$

3. The $\|x_1 - x_2\| > R_n(u)$, $\|x_1 - x_2\| \leq R_n(v)$ case is similar to the previous one, just need to switch the role of u and v. In this case, $\Lambda = \Lambda(n, x_1, x_2) \doteq (n-1)\|x_1 - x_2\|^d > u$, and $\Lambda = (n-1)\|x_1 - x_2\|^d \leq v$.

4. Finally, when $\|x_1 - x_2\| \leq R_n(u)$, and $\|x_1 - x_2\| \leq R_n(v)$, then $\Lambda = \Lambda(n, x_1, x_2) \doteq (n-1)\|x_1 - x_2\|^d \leq u$ and $\Lambda = (n-1)\|x_1 - x_2\|^d \leq v$. $F_{n,k,x_1,x_2}(u,\infty) = F_{n-1,k-1,x_1}(u)$, and $F_{n,k,x_1,x_2}(\infty, v) = F_{n-1,k-1,x_2}(v)$.

$$
F_{n,k,x_1,x_2}(u,v) \quad \leq \quad \min[F_{n,k,x_1,x_2}(u,\infty), F_{n,k,x_1,x_2}(\infty,v)]
$$

$$\leq \quad F^{1/2}_{n-1,k-1,x_1}(u)F^{1/2}_{n-1,k-1,x_2}(v).$$

If $k > 1$, then

$$\begin{aligned}
&\int_\Lambda^\infty \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}F_{n,k,x_1,x_2}(u,v)\mathrm{d}u\mathrm{d}v \\
&\qquad \leq \left[\int_\Lambda^\infty u^{\gamma-1}F^{1/2}_{n-1,k-1,x_1}(u)\mathrm{d}u\right]\left[\int_\Lambda^\infty v^{\gamma-1}F^{1/2}_{n-1,k-1,x_2}(v)\mathrm{d}v\right] \\
&\qquad \leq \left[\int_0^1 u^{\gamma-1}F^{1/2}_{n-1,k-1,x_1}(u)\mathrm{d}u + \int_1^\infty u^{\gamma-1}\mathrm{d}u\right]\times \\
&\qquad\qquad \times\left[\int_0^1 v^{\gamma-1}F^{1/2}_{n-1,k-1,x_2}(v)\mathrm{d}v + \int_1^\infty v^{\gamma-1}\mathrm{d}v\right] \\
&\qquad \leq \left[\frac{\hat{L}^{1/2}(\bar{p},1)}{(k-1)/2+\gamma} - \frac{1}{\gamma}\right]^2; \quad \text{[using (25)]}.
\end{aligned}$$

Here we assumed that $(k-1)/2+\gamma > 0$. If $k = 1$, then the bounding above cannot be applied, because $F_{n-1,0,x_1} = F_{n-1,0,x_2} = 1$. In this case when $(u,v) \in [\Lambda,\infty] \times [\Lambda,\infty]$ we have that $x_2 \in \mathcal{B}(x_1, R_n(u))$ and $x_1 \in \mathcal{B}(x_2, R_n(v)$, thus $F_{n,1,x_1,x_2}(u,v) = 1$. In turn,

$$\begin{aligned}
\int_\Lambda^\infty \int_\Lambda^\infty u^{\gamma-1}v^{\gamma-1}F_{n,1,x_1,x_2}(u,v)\mathrm{d}u\mathrm{d}v &= \left[\int_\Lambda^\infty u^{\gamma-1}\mathrm{d}u\right]^2 \\
&= \frac{\Lambda^{2\gamma}}{\gamma^2} = \frac{(n-1)^{2\gamma}}{\gamma^2}\|x_1-x_2\|^{2d\gamma} \\
&\leq \frac{1}{\gamma^2}\|x_1-x_2\|^{2d\gamma}; \quad \text{[since}(n-1)^{2\gamma} \leq 1\text{]}.
\end{aligned}$$

■

Other Proofs on Divergences and their Properties

Lemma 51 (KL-divergence as a limit case) *When $\alpha \to 1$ then $R_\alpha(p\|q) \to KL(p\|q)$.*

Proof Using L'Hospital rule and assuming that the integral, the limit and the derivative operators can be switched:

$$\begin{aligned}
\lim_{\alpha\to 1} R_\alpha(p\|q) &= \lim_{\alpha\to 1}\frac{1}{\alpha-1}\log\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x \\
&= \lim_{\alpha\to 1}\frac{\frac{\partial}{\partial\alpha}\log\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x}{\frac{\partial}{\partial\alpha}(\alpha-1)}; \quad \text{[L'Hospital rule]} \\
&= \lim_{\alpha\to 1}\frac{\partial}{\partial\alpha}\log\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x \\
&= \lim_{\alpha\to 1}\frac{1}{\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x}\frac{\partial}{\partial\alpha}\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{\alpha\to 1}\frac{\partial}{\partial\alpha}\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x \\
&= \lim_{\alpha\to 1}\int \frac{\partial}{\partial\alpha}p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x; \quad [\frac{\partial}{\partial\alpha}\int = \int\frac{\partial}{\partial\alpha}] \\
&= \lim_{\alpha\to 1}\int p^\alpha(x)\log(p^\alpha(x))q^{1-\alpha}(x) - p^\alpha(x)q^{1-\alpha}(x)\log(q(x))\mathrm{d}x \\
&= \int p(x)\log(p(x)) - p(x)\log(q(x))\mathrm{d}x; \quad [\lim\int = \int\lim].
\end{aligned}
$$

■

Similarly, we can also prove that

Lemma 52 (KL-divergence as a limit case) *When $\alpha \to 1$ then $T_\alpha(p\|q) \to KL(p\|q)$.*

Beta Distributions

Lemma 53 *Let $f(x) = \prod_{i=1}^d Beta_{x_i}(a_i, b_i)$, and $g(x) = \prod_{i=1}^d Beta_{x_i}(c_i, d_i)$. Then $R_\alpha(f\|g) = \frac{1}{\alpha-1}\log D_\alpha(f\|g)$ and $T_\alpha(f\|g) = \frac{1}{\alpha-1}(D_\alpha(f\|g) - 1)$, where*

$$
\begin{aligned}
D_\alpha(f\|g) = \prod_{i=1}^d &\frac{(\Gamma(a_i+b_i))^\alpha}{(\Gamma(a_i))^\alpha(\Gamma(b_i))^\alpha}\frac{(\Gamma(c_i+d_i))^{1-\alpha}}{(\Gamma(c_i))^{1-\alpha}(\Gamma(d_i))^{1-\alpha}} \\
&\times \frac{\Gamma(\alpha(a_i-1)+(1-\alpha)(c_i-1))\Gamma(\alpha(b_i-1)+(1-\alpha)(d_i-1))}{\Gamma(\alpha(a_i-1)+(1-\alpha)(c_i-1)+\alpha(b_i-1)+(1-\alpha)(d_i-1))}.
\end{aligned}
$$

Proof

$$
\begin{aligned}
f(x) \doteq \prod_{i=1}^d Beta_{x_i}(a_i,b_i) &= \prod_{i=1}^d \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}x_i^{a_i-1}(1-x_i)^{b_i-1}, \\
g(x) \doteq \prod_{i=1}^d Beta_{x_i}(c_i,d_i) &= \prod_{i=1}^d \frac{\Gamma(c_i+d_i)}{\Gamma(c_i)\Gamma(d_i)}x_i^{c_i-1}(1-x_i)^{d_i-1}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\int f^\alpha(x)g^{1-\alpha}(x)\mathrm{d}x &= \int\prod_{i=1}^d \frac{(\Gamma(a_i+b_i))^\alpha}{(\Gamma(a_i))^\alpha(\Gamma(b_i))^\alpha}\frac{(\Gamma(c_i+d_i))^{1-\alpha}}{(\Gamma(c_i))^{1-\alpha}(\Gamma(d_i))^{1-\alpha}} \\
&\qquad \times x_i^{\alpha(a_i-1)+(1-\alpha)(c_i-1)}(1-x_i)^{\alpha(b_i-1)+(1-\alpha)(d_i-1)}\mathrm{d}x \\
&= \prod_{i=1}^d \frac{(\Gamma(a_i+b_i))^\alpha}{(\Gamma(a_i))^\alpha(\Gamma(b_i))^\alpha}\frac{(\Gamma(c_i+d_i))^{1-\alpha}}{(\Gamma(c_i))^{1-\alpha}(\Gamma(d_i))^{1-\alpha}} \\
&\qquad \times \frac{\Gamma(\alpha a_i+(1-\alpha)c_i)\Gamma(\alpha b_i+(1-\alpha)d_i)}{\Gamma(\alpha a_i+(1-\alpha)c_i+\alpha b_i+(1-\alpha)d_i)}.
\end{aligned}
$$

■

Similarly, it is also easy to calculate the KL-divergence between two beta distributions.

Lemma 54

$$KL(f\|g) = \sum_{i=1}^{d} \log \frac{\mathrm{B}(c_i, d_i)}{\mathrm{B}(a_i, b_i)} - (c_i - a_i)\psi(a_i) - (d_i - b_i)\psi(b_i) + (c_i - a_i + d_i - b_i)\psi(a_i + b_i),$$

where B *is the Beta and* ψ *is the digamma function.*

The following lemma determines the L_2-divergence between two beta distributions.

Lemma 55 *Let* $f(x) = \prod_{i=1}^{d} Beta_{x_i}(a_i, b_i)$*, and* $g(x) = \prod_{i=1}^{d} Beta_{x_i}(c_i, d_i)$*. Then,*

$$L_2(f\|g) = \left(\prod_{i=1}^{d} \frac{\beta(2a_i - 1, 2b_i - 1)}{\beta^2(a_i, b_i)} + \prod_{i=1}^{d} \frac{\beta(2c_i - 1, 2d_i - 1)}{\beta^2(c_i, d_i)} - 2 \prod_{i=1}^{d} \frac{\beta(a_i + c_i - 1, b_i + d_i - 1)}{\beta(a_i, b_i)\beta(c_i, d_i)} \right)^{1/2}$$

Proof

$$\begin{aligned} f(x) &\doteq \prod_{i=1}^{d} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} x_i^{a_i - 1}(1 - x_i)^{b_i - 1} = \prod_{i=1}^{d} \frac{1}{\beta(a_i, b_i)} x_i^{a_i - 1}(1 - x_i)^{b_i - 1}, \\ g(x) &\doteq \prod_{i=1}^{d} \frac{\Gamma(c_i + d_i)}{\Gamma(c_i)\Gamma(d_i)} x_i^{c_i - 1}(1 - x_i)^{d_i - 1} = \prod_{i=1}^{d} \frac{1}{\beta(c_i, d_i)} x_i^{c_i - 1}(1 - x_i)^{d_i - 1}. \end{aligned}$$

We know that

$$\int \prod_{i=1}^{d} x_i^{a_i - 1}(1 - x_i)^{b_i - 1} \mathrm{d}x_1 \ldots \mathrm{d}x_d = \prod_{i=1}^{d} \frac{\Gamma(a_i)\Gamma(b_i)}{\Gamma(a_i + b_i)} = \prod_{i=1}^{d} \beta(a_i, b_i)$$

Thus,

$$\begin{aligned} \int f^2(x)\mathrm{d}x &= \int \prod_{i=1}^{d} \frac{1}{\beta^2(a_i, b_i)} x_i^{(2a_i - 1 - 1)}(1 - x_i)^{(2b_i - 1 - 1)} = \prod_{i=1}^{d} \frac{\beta(2a_i - 1, 2b_i - 1)}{\beta^2(a_i, b_i)} \\ \int g^2(x)\mathrm{d}x &= \int \prod_{i=1}^{d} \frac{1}{\beta^2(c_i, d_i)} x_i^{(2c_i - 1 - 1)}(1 - x_i)^{(2d_i - 1 - 1)} = \prod_{i=1}^{d} \frac{\beta(2c_i - 1, 2d_i - 1)}{\beta^2(c_i, d_i)} \end{aligned}$$

$$\int f(x)g(x)\mathrm{d}x = \int \prod_{i=1}^{d} \frac{x_i^{(a_i + c_i - 1 - 1)}(1 - x_i)^{(b_i + d_i - 1 - 1)}}{\beta(a_i, b_i)\beta(c_i, d_i)} = \prod_{i=1}^{d} \frac{\beta(a_i + c_i - 1, b_i + d_i - 1)}{\beta(a_i, b_i)\beta(c_i, d_i)}$$

In turn,

$$L_2(f\|g) = \left(\int (f(x) - g(x))^2 \mathrm{d}x \right)^{1/2}$$

$$= \left(\prod_{i=1}^{d} \frac{\beta(2a_i - 1, 2b_i - 1)}{\beta^2(a_i, b_i)} + \prod_{i=1}^{d} \frac{\beta(2c_i - 1, 2d_i - 1)}{\beta^2(c_i, d_i)} - 2 \prod_{i=1}^{d} \frac{\beta(a_i + c_i - 1, b_i + d_i - 1)}{\beta(a_i, b_i)\beta(c_i, d_i)} \right)^{1/2}$$

In the 1-dimensional case it has the following simple form:

$$L_2(f\|g) = \left(\frac{\beta(2a - 1, 2b - 1)}{\beta^2(a, b)} + \frac{\beta(2c - 1, 2d - 1)}{\beta^2(c, d)} - 2\frac{\beta(a + c - 1, b + d - 1)}{\beta(a, b)\beta(c, d)} \right)^{1/2}$$

■

Normal Distributions

Lemma 56 (Rényi divergence between normal distributions with zero means) *Let* $f(x) \doteq \mathcal{N}_x(0, \mathbf{\Sigma}_f)$, $g(x) \doteq \mathcal{N}_x(0, \mathbf{\Sigma}_g)$. *Then*

$$R_\alpha(f\|q) = \frac{1}{\alpha - 1} \log \left(\frac{|\mathbf{\Sigma}_f^{-1}|^{\alpha/2} |\mathbf{\Sigma}_g^{-1}|^{(1-\alpha)/2}}{|\alpha\mathbf{\Sigma}_f^{-1} + (1 - \alpha)\mathbf{\Sigma}_g^{-1}|^{1/2}} \right). \tag{60}$$

Proof Let

$$\begin{aligned} f(x) &\doteq \mathcal{N}_x(0, \mathbf{\Sigma}_f) = |2\pi\mathbf{\Sigma}_f|^{-1/2} \exp(-\frac{1}{2} x^T \mathbf{\Sigma}_f^{-1} x), \\ g(x) &\doteq \mathcal{N}_x(0, \mathbf{\Sigma}_g) = |2\pi\mathbf{\Sigma}_g|^{-1/2} \exp(-\frac{1}{2} x^T \mathbf{\Sigma}_g^{-1} x). \end{aligned}$$

Observe that $\int \exp(-\frac{1}{2} x^T \mathbf{\Sigma}^{-1} x) = |2\pi\mathbf{\Sigma}|^{1/2} = \frac{1}{|(2\pi)^{-1}\mathbf{\Sigma}^{-1}|^{1/2}}$, and therefore

$$\begin{aligned} \int f^\alpha(x) g^\beta(x) \mathrm{d}x &= \int |2\pi\mathbf{\Sigma}_f|^{-\alpha/2} |2\pi\mathbf{\Sigma}_g|^{-\beta/2} \exp(-\frac{\alpha}{2} x^T \mathbf{\Sigma}_f^{-1} x) \exp(-\frac{\beta}{2} x^T \mathbf{\Sigma}_g^{-1} x) \\ &= |2\pi\mathbf{\Sigma}_f|^{-\alpha/2} |2\pi\mathbf{\Sigma}_g|^{-\beta/2} \int \exp(-\frac{\alpha}{2} x^T \mathbf{\Sigma}_f^{-1} x) \exp(-\frac{\beta}{2} x^T \mathbf{\Sigma}_g^{-1} x) \\ &= |(2\pi)^{-1}\mathbf{\Sigma}_f^{-1}|^{\alpha/2} \; |(2\pi)^{-1}\mathbf{\Sigma}_g^{-1}|^{\beta/2} \int \exp(-\frac{1}{2} x^T (\alpha\mathbf{\Sigma}_f^{-1} + \beta\mathbf{\Sigma}_g^{-1}) x \\ &= \frac{|(2\pi)^{-1}\mathbf{\Sigma}_f^{-1}|^{\alpha/2} \; |(2\pi)^{-1}\mathbf{\Sigma}_g^{-1}|^{\beta/2}}{|(2\pi)^{-1}(\alpha\mathbf{\Sigma}_f^{-1} + \beta\mathbf{\Sigma}_g^{-1})|^{1/2}}. \end{aligned}$$

Thus,

$$\begin{aligned} D_\alpha(f\|g) &= \int f^\alpha(x) g^{1-\alpha}(x) \mathrm{d}x = \frac{|\mathbf{\Sigma}_f^{-1}|^{\alpha/2} |\mathbf{\Sigma}_g^{-1}|^{(1-\alpha)/2}}{|\alpha\mathbf{\Sigma}_f^{-1} + (1 - \alpha)\mathbf{\Sigma}_g^{-1}|^{1/2}}, \\ R_\alpha(f\|g) &= \frac{1}{\alpha - 1} \log \left(\frac{|\mathbf{\Sigma}_f^{-1}|^{\alpha/2} |\mathbf{\Sigma}_g^{-1}|^{(1-\alpha)/2}}{|\alpha\mathbf{\Sigma}_f^{-1} + (1 - \alpha)\mathbf{\Sigma}_g^{-1}|^{1/2}} \right). \end{aligned}$$

■

Lemma 57 (Rényi divergence between normal distributions with arbitrary means) *Let $f(x) \doteq \mathcal{N}_x(\mu_f, \mathbf{\Sigma}_f)$, $g(x) \doteq \mathcal{N}_x(\mu_g, \mathbf{\Sigma}_g)$. Then $R_\alpha(f\|g) = \frac{1}{\alpha-1}\log D_\alpha(f\|g)$, $D_\alpha(f\|g) = A/B$, where*

$$A = \frac{|\mathbf{\Sigma}_f^{-1}|^{\alpha/2}|\mathbf{\Sigma}_g^{-1}|^{(1-\alpha)/2}}{|\alpha\mathbf{\Sigma}_f^{-1} + (1-\alpha)\mathbf{\Sigma}_g^{-1}|^{1/2}} \exp\left\{\mu_f^T\alpha\mathbf{\Sigma}_f^{-1}\mu_f + \mu_g^T(1-\alpha)\mathbf{\Sigma}_g^{-1}\mu_g\right\}$$

$$B = \exp\left\{\left(\mu_f^T\alpha\mathbf{\Sigma}_f^{-1} + \mu_g^T(1-\alpha)\mathbf{\Sigma}_g^{-1}\right)\left(\alpha\mathbf{\Sigma}_f^{-1} + (1-\alpha)\mathbf{\Sigma}_g^{-1}\right)^{-1}\left(\alpha\mathbf{\Sigma}_f^{-1}\mu_f + (1-\alpha)\mathbf{\Sigma}_g^{-1}\mu_g\right)\right\}$$

Proof

It is known that

$$\int \exp(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)) = |2\pi\mathbf{\Sigma}|^{1/2} = \frac{1}{|(2\pi)^{-1}\mathbf{\Sigma}^{-1}|^{1/2}}$$

$$\begin{aligned}
&\exp\left\{-\frac{\alpha}{2}(x-\mu_f)^T\mathbf{\Sigma}_f^{-1}(x-\mu_f)\right\}\exp\left\{-\frac{\beta}{2}(x-\mu_g)^T\mathbf{\Sigma}_g^{-1}(x-\mu_g)\right\}\\
&\quad= \exp\left\{-\frac{1}{2}\left(x^T\alpha\mathbf{\Sigma}_f^{-1}x + \mu_f^T\alpha\mathbf{\Sigma}_f^{-1}\mu_f - 2\mu_f^T\alpha\mathbf{\Sigma}_f^{-1}x + x^T\beta\mathbf{\Sigma}_g^{-1}x + \mu_g^T\beta\mathbf{\Sigma}_g^{-1}\mu_g - 2\mu_g^T\beta\mathbf{\Sigma}_g^{-1}x\right)\right\}\\
&\quad= \exp\left\{-\frac{1}{2}\left[x^T\left(\alpha\mathbf{\Sigma}_f^{-1} + \beta\mathbf{\Sigma}_g^{-1}\right)x - 2\left(\mu_f^T\alpha\mathbf{\Sigma}_f^{-1} + \mu_g^T\beta\mathbf{\Sigma}_g^{-1}\right)x + \mu_f^T\alpha\mathbf{\Sigma}_f^{-1}\mu_f + \mu_g^T\beta\mathbf{\Sigma}_g^{-1}\mu_g\right]\right\}\\
&\quad= \exp\left\{-\frac{1}{2}\left[x^TS_1x - 2S_2^Tx + S_3\right]\right\}\\
&\quad= \exp\left\{-\frac{1}{2}\left[(x-S_1^{-1}S_2)^TS_1(x-S_1^{-1}S_2) + S_3 - S_2^TS_1^{-1}S_2\right]\right\}\\
&\quad= \exp\left\{-\frac{1}{2}\left[(x-S_1^{-1}S_2)^TS_1(x-S_1^{-1}S_2)\right]\right\}\exp\left\{S_3 - S_2^TS_1^{-1}S_2\right\},
\end{aligned}$$

where $S_1 = \left(\alpha\mathbf{\Sigma}_f^{-1} + \beta\mathbf{\Sigma}_g^{-1}\right)$, $S_2^T = \left(\mu_f^T\alpha\mathbf{\Sigma}_f^{-1} + \mu_g^T\beta\mathbf{\Sigma}_g^{-1}\right)$, $S_3 = \mu_f^T\alpha\mathbf{\Sigma}_f^{-1}\mu_f + \mu_g^T\beta\mathbf{\Sigma}_g^{-1}\mu_g$.

Therefore,

$$\begin{aligned}
&\int \exp\left\{-\frac{\alpha}{2}(x-\mu_f)^T\mathbf{\Sigma}_f^{-1}(x-\mu_f)\right\}\exp\left\{-\frac{\beta}{2}(x-\mu_g)^T\mathbf{\Sigma}_g^{-1}(x-\mu_g)\right\}\\
&\quad= \int \exp\left\{-\frac{1}{2}\left[(x-S_1^{-1}S_2)^TS_1(x-S_1^{-1}S_2)\right]\right\}\exp\left\{S_3 - S_2^TS_1^{-1}S_2\right\}\\
&\quad= \frac{1}{|(2\pi)^{-1}S_1|^{1/2}}\exp\left\{S_3 - S_2^TS_1^{-1}S_2\right\},
\end{aligned}$$

and we have that

$$\int f^\alpha(x)g^\beta(x)\mathrm{d}x = \frac{|(2\pi)^{-1}\mathbf{\Sigma}_f^{-1}|^{\alpha/2}\ |(2\pi)^{-1}\mathbf{\Sigma}_g^{-1}|^{\beta/2}}{|(2\pi)^{-1}\left(\alpha\mathbf{\Sigma}_f^{-1} + \beta\mathbf{\Sigma}_g^{-1}\right)|^{1/2}}\exp\left\{S_3 - S_2^TS_1^{-1}S_2\right\}$$

■

The KL divergence between two normal distributions with different means is as follows:

$$KL(f\|g) \;=\; \frac{1}{2}\left[\log\left(\frac{|\mathbf{\Sigma}_g|}{|\mathbf{\Sigma}_f|}\right) + tr(\mathbf{\Sigma}_g^{-1}\mathbf{\Sigma}_f) + (\boldsymbol{\mu}_g - \boldsymbol{\mu}_f)^T\mathbf{\Sigma}_g^{-1}(\boldsymbol{\mu}_g - \boldsymbol{\mu}_f) - d.\right]$$

Uniform Distributions

Lemma 58 (Rényi divergence between uniform distributions) *Let uniform distributions $f(x) \doteq U_x(a_1, b_1)$, $g(x) \doteq U_x(a_2, b_2)$. Then*

$$R_\alpha(f\|g) = \frac{1}{\alpha - 1}\log\left(\frac{l}{(b_1 - a_1)^\alpha (b_2 - a_2)^{1-\alpha}}\right). \tag{61}$$

where $l = \mathcal{V}([a_1, b_1] \cap [a_2, b_2])$ i.e. *the length of the common support of f and g.*

Lemma 59 (L_2 divergence between uniform distributions) *Let uniform distributions $f(x) \doteq U_x(a_1, b_1)$, $g(x) \doteq U_x(a_2, b_2)$. Then*

$$L_2(f\|g) = \sqrt{\frac{1}{b_1 - a_1} + \frac{1}{b_2 - a_2} - \frac{2l}{(b_1 - a_1)(b_2 - a_2)}} \tag{62}$$

where l is defined as above.