On the Estimation of α-Divergences

Barnabás Póczos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
USA, 15213

Jeff Schneider
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
USA, 15213

Abstract

We propose new nonparametric, consistent Rényi-α and Tsallis-α divergence estimators for continuous distributions. Given two independent and identically distributed samples, a "naïve" approach would be to simply estimate the underlying densities and plug the estimated densities into the corresponding formulas. 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 and demonstrate their efficiency via numerical experiments.

1 Introduction

Many statistical, artificial intelligence, and machine learning problems require efficient estimation of the divergence between two distributions. We assume that these distributions are not given explicitly. Only two finite, independent and identically distributed (i.i.d.) samples are given from the two underlying distributions. The Rényi-α (Rényi, 1961, 1970) and Tsallis-α (Villmann and Haase, 2010) divergences are two widely applied and prominent examples of probability divergences. The popular Kullback–Leibler (KL) divergence is a special case of these families, and they can also be related to the Csiszár's-f divergence (Csiszár, 1967). Under certain conditions, these divergences can estimate entropy and can also be applied to estimate Rényi and Tsallis mutual information. For more examples and other possible applications of these divergences, see our extended technical report (Póczos and Schneider, 2011). Despite their wide applicability, there is no known direct, consistent estimator for Rényi-α or Tsallis-α divergence.

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.

This paper provides a direct, L_2-consistent estimator for the Tsallis-α divergence and a weakly consistent estimator for the Rényi-α divergence. These estimators can also be applied to (Rényi and Tsallis) mutual information.

The closest existing work most relevant to the topic of this paper is the work of Wang et al. (2009a), who provided an estimator for the $\alpha \to 1$ limit case only, i.e., for the KL-divergence. However, we warn the reader 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 works. Hero et al. (2002a,b) 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

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

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.[1] 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. (2009b).

The paper is organized as follows. In the next section we formally define our estimation problem, introduce the Rényi-α and Tsallis-α divergences, and explain their most important properties. Section 3 briefly introduces k-NN based density estimators. We propose estimators for the Rényi-α and Tsallis-α divergences in Section 4 and also present our most important theoretical results about the asymptotic unbiasedness and consistency of the estimators. For their analysis we will need a few general tools, which we collect in Section 5. We will prove the asymptotic unbiasedness of our estimators in Section 6. Due to a lack of space, we provide many details of the proofs in Póczos and Schneider (2011). The analysis of the asymptotic variances of our estimators follows an approach similar to their biases but is more complex; therefore, we relegate this material into Póczos and Schneider (2011) as well. Section 7 contains the results of numerical experiments that demonstrate the effectiveness of our proposed algorithm. We also demonstrate in that section how our divergence estimators can be used for Rényi- and Tsallis-mutual information (MI) estimation. Finally, we conclude with a discussion of our work.

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

2 Divergences

For the remainder of this work we will assume that $\mathcal{M}_0 \subset \mathbb{R}^d$ is a measurable set with respect to the d-dimensional Lebesgue measure and that p and q are densities on this domain. The set where they are strictly positive will be denoted by $\mathrm{supp}(p)$ and $\mathrm{supp}(q)$, respectively.

Let p and q be $\mathbb{R}^d \supseteq \mathcal{M}_0 \colon \to \mathbb{R}$ density functions, and let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. The α-divergence $\tilde{D}_\alpha(p\|q)$ (Cichocki et al., 2008) is defined as

$$\tilde{D}_\alpha(p\|q) \doteq \frac{1}{\alpha(1-\alpha)} \left[1 - \int_{\mathcal{M}_0} p^\alpha(x) q^{1-\alpha}(x)\, \mathrm{d}x\right], \tag{1}$$

assuming this integral exists. One can see that this is a special case of Csiszár's f-divergence (Csiszár, 1967) and hence it is always nonnegative.[2] Closely related divergences (but not special cases) to (1) are the Rényi-α (Rényi, 1961) and the Tsallis-α (Villmann and Haase, 2010) divergences.

Definition 1. *Let p, q be $\mathbb{R}^d \supseteq \mathcal{M}_0 \colon \to \mathbb{R}$ density functions and let $\alpha \in \mathbb{R} \setminus \{1\}$. The* Rényi-$\alpha$ divergence *is defined as*

$$R_\alpha(p\|q) \doteq \frac{1}{\alpha - 1} \log \int_{\mathcal{M}_0} p^\alpha(x) q^{1-\alpha}(x)\, \mathrm{d}x. \tag{2}$$

The Tsallis-α divergence *is defined as*

$$T_\alpha(p\|q) \doteq \frac{1}{\alpha - 1} \left(\int_{\mathcal{M}_0} p^\alpha(x) q^{1-\alpha}(x)\, \mathrm{d}x - 1\right). \tag{3}$$

Both definitions assume that the corresponding integral exists.

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

Lemma 2.

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

We are now prepared to formally define the goal of our paper. Given two independent i.i.d. samples from

[2] See the Appendix for more details.

distributions with densities p and q, respectively, we provide an L_2-consistent estimator for

$$D_\alpha(p\|q) \doteq \int_{\mathcal{M}_0} p^\alpha(x) q^{1-\alpha}(x)\,\mathrm{d}x. \tag{4}$$

By plugging our estimate of (4) into (3) and (2), we immediately get an L_2-consistent estimator for $T_\alpha(p\|q)$, as well as a weakly consistent estimator for $R_\alpha(p\|q)$ for $\alpha \neq 1$.

3 k-NN Based Density Estimators

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

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 $X_{1:n}$, and similarly let $\nu_k(i)$ denote the distance of the kth nearest neighbor of X_i in the sample $Y_{1:m}$. 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 3 (k-NN based density estimators)**.**

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

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

The following theorems show the consistency of these density estimators.[3]

Theorem 4 (k-NN density 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 5 (k-NN density estimators, almost sure 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.*

[3] We use $X_n \to_p X$ and $X_n \to_d X$ to represent convergence of random variables in probability and in distribution, respectively. $F_n \to_w F$ will denote the weak convergence of distribution functions.

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 and will still be able to prove their consistency.

4 An Estimator for $D_\alpha(p\|q)$

In this section we introduce our estimator for $D_\alpha(p\|q)$ and claim its L_2 consistency in the form of several theorems. From now on we will assume that (4) can be rewritten as

$$D_\alpha(p\|q) = \int_{\mathcal{M}} \left(\frac{q(x)}{p(x)}\right)^{1-\alpha} p(x)\,\mathrm{d}x, \tag{7}$$

where $\mathcal{M} = \mathrm{supp}(p)$. In other words, in the definition of $D_\alpha(p\|q)$, it is enough to integrate on the support of p. There are other possible ways to rewrite $D_\alpha(p\|q)$ (such as $\int (q/p)^{(1-\alpha)}p$, $\int (p/q)^\alpha q$, or $\int (q/p)^{-\alpha} q$), and we could start our analysis from these forms as well. If we simply plugged (5) and (6) into (7), then we could estimate $D_\alpha(p\|q)$ with

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{(n-1)\rho_k^d(i)}{m\nu_k^d(i)}\right)^{1-\alpha};$$

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

$$\widehat{D}_\alpha(X_{1:n}\|Y_{1:m}) \doteq \frac{1}{n}\sum_{i=1}^n \left(\frac{(n-1)\rho_k^d(i)}{m\nu_k^d(i)}\right)^{1-\alpha} B_{k,\alpha}, \tag{8}$$

where $B_{k,\alpha} \doteq \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$. Notably, this multiplicative bias does not depend on p or q. The following theorems of this section contain our main results: $\widehat{D}_\alpha(X_{1:n}\|Y_{1:m})$ is an L_2-consistent estimator for $D_\alpha(p\|q)$, 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. We introduce the following function:

$$H(x,p,\delta,\omega) \doteq \sum_{j=0}^{k-1} \left(\frac{1}{j!}\right)^\omega \Gamma(\gamma+j\omega) \left(\frac{p(x)+\delta}{p(x)-\delta}\right)^{j\omega} \times \\ \times \big(p(x)-\delta\big)^{-\gamma} \big((1-\delta)\omega\big)^{-\gamma-j\omega}. \tag{9}$$

When $0 < \gamma \doteq 1-\alpha < k$, we have the following theorem.

Theorem 6 (Asymptotic unbiasedness). *Assume that (a) $0 < \gamma \doteq 1 - \alpha < k$, (b) p is bounded away from zero, (c) p is uniformly Lebesgue approximable, (d) $\exists \delta_0$ s.t. $\forall \delta \in (0, \delta_0)$, $\int_{\mathcal{M}} H(x, p, \delta, 1) p(x)\, \mathrm{d}x < \infty$, (e) $\int_{\mathcal{M}} \|x - y\|^{\gamma} p(y)\, \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$, (f) $\iint_{\mathcal{M}^2} \|x - y\|^{\gamma} p(y) p(x)\, \mathrm{d}y\, \mathrm{d}x < \infty$, and that (g) q is bounded from above. Then*

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

i.e., the estimator is asymptotically unbiased.

For the definition of a uniformly Lebesgue approximable function, see Definition 14. The following theorem states that the estimator is asymptotically unbiased when $-k < \gamma \doteq 1 - \alpha < 0$.

Theorem 7 (Asymptotic unbiasedness). *Assume that (a) $-k < \gamma \doteq 1 - \alpha < 0$, (b) q is bounded away from zero, (c) q is uniformly Lebesgue approximable, (d) $\exists \delta_0$ s.t. $\forall \delta \in (0, \delta_0)$ $\int_{\mathcal{M}} H(x, q, \delta, 1) p(x)\, \mathrm{d}x < \infty$, (e) $\int_{\mathcal{M}} \|x - y\|^{\gamma} q(y)\, \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$, (f) $\iint_{\mathcal{M}^2} \|x - y\|^{\gamma} q(y) p(x)\, \mathrm{d}y\, \mathrm{d}x < \infty$, (g) p is bounded from above, and that (h) $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$. In this case, the estimator is asymptotically unbiased.*

The following theorems provide conditions under which $\widehat{D}$ is L_2 consistent. In the previous theorems we have stated conditions that lead to asymptotically unbiased divergence estimation. In all of the following theorems we will assume that the estimator is asymptotically unbiased for the parameter $\gamma = 1 - \alpha$ as well as for a new parameter $\tilde{\gamma} \doteq 2(1 - \alpha)$ (corresponding to $\tilde{\alpha} \doteq 2\alpha - 1$), and also assume that $\max\big(D_{\alpha}(p \| q), D_{\tilde{\alpha}}(p \| q)\big) < \infty$.

Theorem 8 (L_2 consistency). *Assume $k \geq 2$ and that (a) $0 < \gamma \doteq 1 - \alpha < (k-1)/2$, (b) p is bounded away from zero, (c) p is uniformly Lebesgue approximable, (d) $\exists \delta_0$ such that $\forall \delta \in (0, \delta_0)$, $\int_{\mathcal{M}} H(x, p, \delta, 1/2) p(x)\, \mathrm{d}x < \infty$, (e) $\int_{\mathcal{M}} \|x - y\|^{\gamma} p(y)\, \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$, (f) $\iint_{\mathcal{M}^2} \|x - y\|^{\gamma} p(y) p(x)\, \mathrm{d}y\, \mathrm{d}x < \infty$, and that (g) q is bounded above. Then*

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

that is, the estimator is L_2 consistent.

For the $-(k-1)/2 < 1 - \alpha < 0$, $k \geq 2$ case, the following theorem holds.

Theorem 9 (L_2 consistency). *Assume $k \geq 2$ and that (a) $-k/2 < \gamma \doteq 1 - \alpha < 0$, (b) q is bounded away from zero, (c) q is uniformly Lebesgue approximable, (d) $\exists \delta_0$ s.t. $\forall \delta \in (0, \delta_0)$, $\int_{\mathcal{M}} H(x, q, \delta, 1/2) p(x)\, \mathrm{d}x < \infty$, (e) $\int_{\mathcal{M}} \|x - y\|^{\gamma} q(y)\, \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$, (f) $\iint_{\mathcal{M}^2} \|x - y\|^{\gamma} q(y) p(x)\, \mathrm{d}y\, \mathrm{d}x < \infty$, (g) p is bounded above, and that (h) $\operatorname{supp}(p) \subset \operatorname{supp}(q)$. In this case, the estimator is L_2 consistent.*

Finally, for the $k = 1$ case, we will see that the theorems below are true. Define the following function:

$$L(x, \omega, \gamma, k, p, \delta, \delta_1) \doteq \delta_1 + \delta_1 \int \|x - y\|^{\gamma} p(y)\, \mathrm{d}y + \\ + (\bar{c} r(x))^{-\gamma} H(x, p, \delta, \omega). \quad (10)$$

Theorem 10 (L_2 consistency). *Assume $k = 1$ and that (a) $0 < \gamma \doteq 1 - \alpha < 1/2$, (b) p is bounded away from zero, (c) p is uniformly Lebesgue approximable, (d) $\int_{\mathcal{M}} \|x - y\|^{\gamma} p(y)\, \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$, (e) $\iint_{\mathcal{M}^2} \|x - y\|^{\gamma} p(y) p(x)\, \mathrm{d}y\, \mathrm{d}x < \infty$, and that (f) q is bounded above. If there exists $\delta_1, \delta_0 > 0$ such that for all $\delta \in (0, \delta_0)$,*

$$\iint L(x_1, 1/2, 1, \gamma, p, \delta, \delta_1) L(x_2, 1/2, 1, \gamma, p, \delta, \delta_1) \\ \times \|x_1 - x_2\|^{-2\gamma} p(x_1) p(x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 < \infty,$$

then the estimator is L_2 consistent.

Theorem 11 (L_2 consistency). *Assume $k = 1$ and that (a) $-1/2 < \gamma \doteq 1 - \alpha < 0$, (b) q is bounded away from zero, (c) q is uniformly Lebesgue approximable, (d) $\int_{\mathcal{M}} \|x - y\|^{\gamma} q(y)\, \mathrm{d}y < \infty$ for almost all $x \in \mathcal{M}$, (e) $\iint_{\mathcal{M}^2} \|x - y\|^{\gamma} q(y) p(x)\, \mathrm{d}y\, \mathrm{d}x < \infty$, (f) p is bounded above, and that (g) $\operatorname{supp}(p) \subset \operatorname{supp}(q)$. If there exist $\delta_1, \delta_0 > 0$ such that for all $\delta \in (0, \delta_0)$,*

$$\iint L(x_1, 1/2, 1, -\gamma, q, \delta, \delta_1) L(x_2, 1/2, 1, -\gamma, q, \delta, \delta_1) \\ \times \|x_1 - x_2\|^{2\gamma} p(x_1) p(x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 < \infty,$$

then the estimator is L_2 consistent.

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

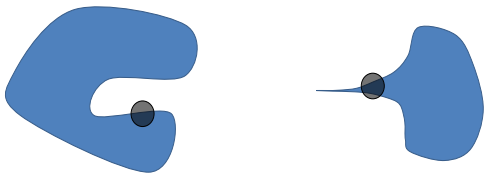
5 General Tools

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 12 (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.[4]

[4] $L_1(\mathcal{M})$ denotes the set of Lebesgue measurable functions defined on the domain $\mathcal{M}$ that have finite integral over $\mathcal{M}$.



(a) feasible (b) not allowed

Figure 1: A possible allowed and a not-allowed domain $\mathcal{M}$ under the property in (13).

Lemma 13 (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 $x \in \mathbb{R}^d$,*

$$\lim_{n\to\infty} \frac{\int_{\mathcal{B}(x,R_n)} g(t)\,\mathrm{d}t}{\mathcal{V}\big(\mathcal{B}(x,R_n)\big)} = g(x). \tag{11}$$

This implies that if $\mathcal{M} \subset R^d$ 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

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

We will later require a generalization of this property; namely, we will need it to hold uniformly over $x \in \mathcal{M}$. However, for this generalization 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 posses the property that the intersection of $\mathcal{M}$ with an arbitrary small ball having 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:

$$\inf_{0<\delta<1} \inf_{x\in\mathcal{M}} \frac{\mathcal{V}\big(\mathcal{B}(x,\delta)\cap\mathcal{M}\big)}{\mathcal{V}\big(\mathcal{B}(x,\delta)\big)} \doteq r_{\mathcal{M}} > 0. \tag{13}$$

Figure 1 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 14 (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 $x \in \mathcal{M}$,*

$$g(x) - \delta < \frac{\int_{\mathcal{B}(x,R_n)\cap\mathcal{M}} g(t)\,\mathrm{d}t}{\mathcal{V}\big(\mathcal{B}(x,R_n)\cap\mathcal{M}\big)} < g(x) + \delta. \tag{14}$$

This property is a uniform variant of (12). The following lemma provides examples of uniformly Lebesgue-approximable functions.

Lemma 15. *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 16 (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 $\int_0^\infty u^\gamma f_{x,k}(u)\ \mathrm{d}u = \lambda(x)^{-\gamma}\frac{\Gamma(k+\gamma)}{\Gamma(k)}$.*

6 Proving Asymptotic Unbiasedness

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

6.1 Preliminaries

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

Lemma 17. *Let $\zeta_{n,k,1} \doteq (n-1)\rho_k^d(1)$ be a random variable, and let $F_{n,k,x}(u) \doteq \Pr(\zeta_{n,k,1} < u \mid X_1 = x)$ denote its conditional distribution function. Then*

$$F_{n,k,x}(u) = 1 - \sum_{j=0}^{k-1} \binom{n-1}{j} (P_{n,u,x})^j (1 - P_{n,u,x})^{n-1-j}, \tag{15}$$

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

We also have the following (Leonenko et al., 2008a).

Lemma 18. *$F_{n,k,x} \to_w F_{k,x}$ for almost all $x \in \mathcal{M}$, where $F_{k,x}(u) \doteq 1 - \exp(-\lambda u)\sum_{j=0}^{k-1}\frac{(\lambda u)^j}{j!}$ is the Erlang distribution with $\lambda = \bar{c}p(x)$.*

Lemma 19. *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 $x \in \mathcal{M}$ we have that $\xi_{n,k,x}^\gamma \to_d \xi_{k,x}^\gamma$.*

Theorem 20. *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*

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

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

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

We will only prove formally the first equation of Theorem 20. The second one can be proven in the same way. Note that the conditions here are different from those given in Leonenko et al. (2008a,b); Goria et al. (2005); Wang et al. (2009a). We are now ready to begin proving Theorem 20.

Proof of ***Theorem 20****.* We already know from Lemma 19 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

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

assuming $k+\gamma > 0$ and using Lemma 16. □

All that remains is to prove that if $\xi^{\gamma}_{n,k,x} \to_d \xi^{\gamma}_{k,x}$, then $\mathbb{E}[\xi^{\gamma}_{n,k,x}] \to \mathbb{E}[\xi^{\gamma}_{k,x}]$. To see this, it is enough to show (according to Theorem 12) 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 explicitly to calculate $\mathbb{E}[\xi^{\gamma(1+\varepsilon)}_{n,k,x}]$; we simply have to provide a finite upper bound.

Theorem 21. *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* $\lim_{n\to\infty} \mathbb{E}[\xi^{\gamma}_{n,k,x}] = \mathbb{E}[\xi^{\gamma}_{k,x}]$.

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

6.2 The proof of Theorems 6 and 7

Proof. We want to prove that

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

The r.h.s. can be rewritten as

$$\begin{aligned}
\lim_{n,m\to\infty} \mathbb{E}_{X_1\sim p}\Bigg[&\mathbb{E}\left[(n-1)^{1-\alpha}\rho_k^{d(1-\alpha)}(1)\middle| X_1\right] \times \\
&\times \mathbb{E}\left[\frac{1}{m^{1-\alpha}\nu_k^{d(1-\alpha)}(1)}\middle| X_1\right]\Bigg]. \quad (16)
\end{aligned}$$

If we could move the limit inside the expectation, then we could apply Theorem 20 to continue the derivation as follows.

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

this would complete the proof of Theorems 6 and 7. In the next section we will discuss conditions under which the outer limit can be moved inside the above expectation. □

6.2.1 Switching limit and expectation

Our goal is to prove that

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

where

$$f_n(x) \doteq \mathbb{E}\left[(n-1)^{1-\alpha}\rho_k^{d(1-\alpha)}(1)\;\middle|\; X_1 = x\right], \quad (17)$$

$$g_m(x) \doteq \mathbb{E}\left[\frac{1}{m^{1-\alpha}\nu_k^{d(1-\alpha)}(1)}\;\middle|\; X_1 = x\right]. \quad (18)$$

We investigate the $0 < \gamma < k$ and the $-k < \gamma < 0$ cases in two separate lemmas.

Lemma 22. *Let* $0 < \gamma \doteq 1-\alpha < k$, *and let* p *be uniformly Lebesgue approximable on* $\mathcal{M} = \operatorname{supp}(p)$ *and bounded away from zero. Let* q *be bounded above by* $\bar{q}$. *Let* $\delta_1 > 0$, *and let be* $\delta > 0$ *so small that* $p(x) - \delta > 0$ *for all* $x \in \mathcal{M}$. *Then there exists a* $N_{p,q} > 0$ *such that if* $m, n > N_{p,q}$, *then for almost all* $x \in \mathcal{M}$,

$$f_n(x)g_m(x) \leq \gamma^2 L(x,1,k,\gamma,p,\delta,\delta_1)\left[\frac{\hat{L}(\bar{q},1)}{k-\gamma} + \frac{1}{\gamma}\right],$$

where $\hat{L}(\bar{q},\beta) \doteq (\bar{q}\bar{c})^k \exp(\bar{q}\bar{c}\beta)$.

Similarly, for the $-k < \gamma \doteq 1-\alpha < 0$ case we have the following lemma.

Lemma 23. *Let* $-k < \gamma \doteq 1-\alpha < 0$, *and let* $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$. *Furthermore, let* q *be uniformly Lebesgue approximable on* $\mathcal{M} = \operatorname{supp}(p)$ *and bounded away from zero. Let* p *be bounded above by* $\bar{p}$. *Let* $\delta_1 > 0$, *and let* $\delta > 0$ *so small that* $q(x) - \delta > 0$ *for all* $x \in \mathcal{M}$. *Then there exists a* $N_{p,q} > 0$ *such that if* $m, n > N_{p,q}$, *then for almost all* $x \in \operatorname{supp}(p)$,

$$f_n(x)g_m(x) \leq \gamma^2 L(x,1,k,-\gamma,q,\delta,\delta_1)\left[\frac{\hat{L}(\bar{p},1)}{k+\gamma} - \frac{1}{\gamma}\right].$$

Now, for the two cases $0 < \gamma = 1 - \alpha < k$ and $-k < \gamma = 1 - \alpha < 0$, we can see that under the conditions detailed in Theorems 6 and 7, there exists a function J and a threshold number $N_{p,q}$ such that if $n, m > N_{p,q}$, then for almost all $x \in \mathcal{M}$ $f_n(x)g_m(x) \leq J(x)$ and $\int_{\mathcal{M}} J(x)p(x)\,\mathrm{d}x < \infty$. Applying the Lebesgue dominated convergence theorem finishes the proofs of these theorems.

7 Numerical Experiments

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 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. We note that in our simulations the numerical results were very similar for the estimation of R_α and T_α; therefore, we will only present our results for the R_α case.

7.1 Normal distributions

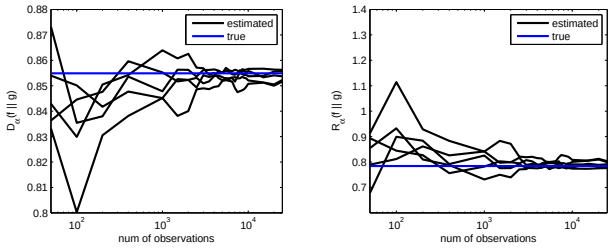
We begin our discussion by investigating the performance of our divergence estimators on normal distributions. Note that when $\alpha \notin [0, 1]$, the divergences can easily become unbounded.[5]

In Figure 2 we display the performances of the proposed $\widehat{D}_\alpha$ and $\widehat{R}_\alpha$ divergence estimators when the underlying densities were zero-mean Gaussians with randomly chosen 5-dimensional covariance matrices. Our results demonstrates that when we increase the sample sizes n and m, then the $\widehat{D}_\alpha$ and $\widehat{R}_\alpha$ values converge to their true values. For simplicity, in our experiments we always set $n = m$. The figures show 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.

7.2 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.

[5]See the Appendix for the details.



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

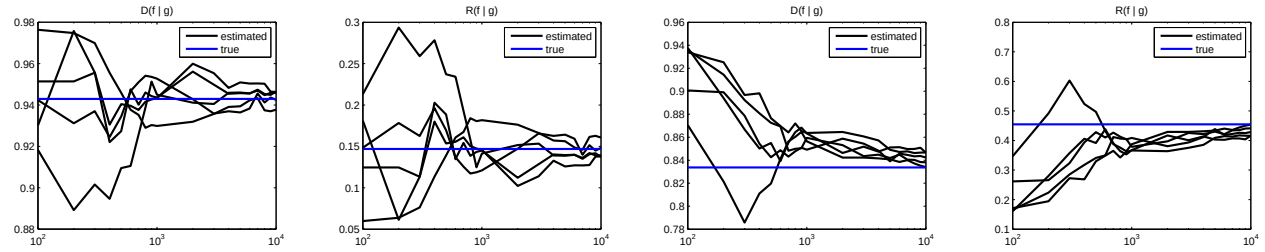
Figure 2: Estimated vs. true divergence for the normal distribution experiments as a function of the number of observations. The results of five independent experiments are shown for estimating the (a) $D_\alpha(f\|g)$ and (b) $R_\alpha(f\|g)$ divergences.

Our first experiment, illustrated in Figures 3(a)–3(b), indicates that the estimators are consistent when $d = 2$; as we increase the number of instances, the estimators converge to the true $D_\alpha(f\|g)$ and $R_\alpha(f\|g)$ values. The figures show five independent experiments, varying the sample size between 100 and 10 000. α was set to 0.4, and we used $k = 4$ nearest neighbors in the density estimates. 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 3(c)–3(d), show that the estimators were also consistent in this case.

7.3 Mutual information estimation

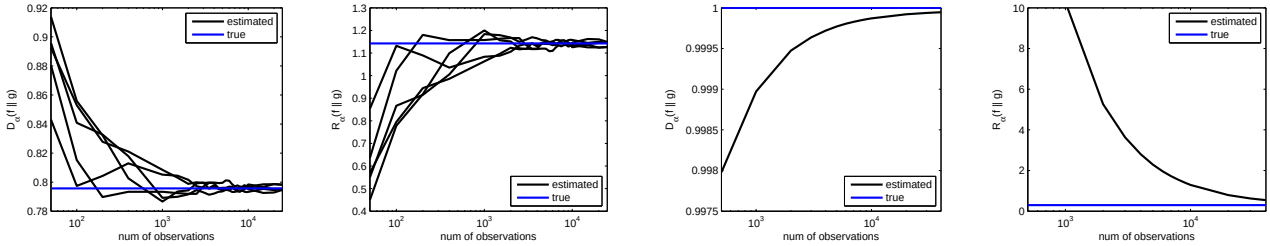
In this section we demonstrate that the proposed divergence estimators can also be used to estimate mutual information. Let $f = (f_1, \ldots, f_d) \in \mathbb{R}^d$ be the density of a d-dimensional distribution. The mutual information $I_\alpha(f)$ is the divergence between f and the product of the marginal variables. Particularly, 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 mutual information as follows. We form one set of size n by setting aside the first n samples. We build another sample by randomly permuting the coordinates of the remaining n observations independently for each coordinate to form n independent instances sampled from $\prod_{i=1}^d f_i$. Using these two sets, we can estimate $I_\alpha(f)$. Figures 4(a)–4(b) show the results of applying this procedure for a $2d$ Gaussian distribution with a randomly chosen covariance matrix. The subfigures show the true D_α and R_α values, as well as their estimations using different sample sizes. k was set to 8, and α was 0.8.

Figures 4(c)–4(d) show the results of repeating the previous experiment with two alterations. In this case we estimated the Shannon (rather than Rényi) infor-



(a) $D_\alpha(f\|g)$ (b) $R_\alpha(f\|g)$ (c) $D_\alpha(f\|g)$ (d) $R_\alpha(f\|g)$

Figure 3: 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 $D_\alpha(f\|g)$ and $R_\alpha(f\|g)$ divergences. (a,b): f and g were the densities of two $2d$ beta distributions—the marginal distributions were independent $1d$ betas with randomly chosen parameters. (c,d): The same as (a,b), but here f and g were the densities of two $5d$ beta distributions with independent marginals.



(a) $D_\alpha(f\|g)$ (b) $R_\alpha(f\|g)$ (c) $D_\alpha(f\|g)$ (d) $R_\alpha(f\|g)$

Figure 4: Estimated vs. true Rényi information for the mutual information experiments as a function of the number of observations. (a) and (b) show five independent experiments for estimating $D_\alpha(f\|g)$ and $I_\alpha(f) = R_\alpha(f\|g)$ for a $2d$ Gaussian distribution using sample sizes between 100 and 20 000. In (c)–(d), we estimated the mutual information between the marginals of a $\pi/4$ degree rotated $2d$ uniform distribution. The sample size was varied from 500 to 40 000.

mation, 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 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 quite good results for the Shannon mutual information as well as for $D_1(f\| \prod_{i=1}^d f_i) = 1$.

8 Discussion and Conclusion

We have derived a new nonparametric estimator for the Rényi-α and Tsallis-α divergences, two important quantities with several applications in machine learning and statistics. Under certain conditions we showed the consistency of these estimators and how they can be applied to estimate mutual information. We also demonstrated their efficiency using numerical experiments.

The main idea in the proofs of our new theorems was that the expected value of our estimator can be rewritten as in (16). We showed that asymptotically the terms inside this expectation converge to the Erlang distribution and applied the well-known formulas for its moments. The main difficulty was to show that we can indeed switch the limit and expectation operators; that is, the limit of expectations equals the expectation of the limits of the random variables. For this purpose we bounded above these random variables and applied the Lebesgue dominated convergence theorem. To derive a bound on these random variables, we made several assumptions on the densities p and q.

There remain some open issues: the conditions of the theorems could be weakened considerably, and the rates of the estimators are still unknown. It would also be desirable to investigate whether the proposed estimators are asymptotically normal.

References

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

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

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

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

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

Hero, A. O., Ma, B., Michel, O., and Gorman, J. (2002a). Alpha-divergence for classification, indexing and retrieval. Communications and Signal Processing Laboratory Technical Report CSPL-328.

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

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

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

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

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

Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist*, 36:1049–1051.

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

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

Pál, D., Póczos, B., and Szepesvári, C. (2010). Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the Neural Information Processing Systems*.

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

Póczos, B. and Schneider, J. (2011). On the estimation of alpha-divergences. CMU, Auton Lab Technical Report, `http://www.cs.cmu.edu/~bapoczos/articles/poczos11alphaTR.pdf`.

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

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

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

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

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

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

Wang, Q., Kulkarni, S. R., and Verdú, S. (2009b). Universal estimation of information measures for analog sources. *Foundations and Trends in Communications and Information Theory*, 5(3):265–352.