Nonparametric Estimation of Rényi Divergence and Friends

Akshay Krishnamurthy AKSHAYKR@CS.CMU.EDU
Kirthevasan Kandasamy KANDASAMY@CS.CMU.EDU
Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Larry Wasserman LARRY@STAT.CMU.EDU
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213

Abstract

We consider nonparametric estimation of L_2, Rényi-α and Tsallis-α divergences between continuous distributions. Our approach is to construct estimators for particular integral functionals of two densities and translate them into divergence estimators. For the integral functionals, our estimators are based on corrections of a preliminary plug-in estimator. We show that these estimators achieve the parametric convergence rate of $n^{-1/2}$ when the densities' smoothness, s, are both at least $d/4$ where d is the dimension. We also derive minimax lower bounds for this problem which confirm that $s > d/4$ is necessary to achieve the $n^{-1/2}$ rate of convergence. We validate our theoretical guarantees with a number of simulations.

1. Introduction

Given samples from two distributions, one fundamental and classical question to ask is: how close are the two distributions? First, one must specify what it means for two distributions to be close, for which a number of *divergences* have been proposed. Then there is the statistical question: how does one estimate divergence given samples from two distributions. In this paper, we propose and analyze estimators for three common divergences.

Divergence estimation has a number of applications across machine learning and statistics. In statistics, one can use these estimators to construct two-sample and independence tests (Pardo, 2005). In machine learning, it is often convenient to view training data as a set of distributions and use divergences to estimate dissimilarity between examples. This idea has been used in neuroscience, where the neural response pattern of an individual is modeled as a distribution, and divergence is used to compare responses across subjects (Johnson et al., 2001). It has also enjoyed success in computer vision, where features are computed for each patch of an image and these feature vectors are modeled as independent draws from an underlying distribution (Póczos et al., 2012).

For these applications and others, it is crucial to accurately estimate divergences given samples drawn independently from each distribution. In the nonparametric setting, a number of authors have proposed various estimators which are provably consistent. However, apart from a few examples, the actual *rates of convergence* of these estimators and the minimax optimal rates are still unknown.

In this work, we propose three estimators for the L_2^2, Rényi-α, and Tsallis-α divergence between two continuous distributions. Our strategy is to correct an initial plug-in estimator by estimates of the higher order terms in the von Mises expansion of the divergence functional. We establish the rates of convergence for these estimators under the assumption that both densities belong to a Hölder class of smoothness s. Concretely, we show that the plug-in estimator achieves rate $n^{\frac{-s}{2s+d}}$ while correcting by the first order terms in the expansion results in an $n^{-\min\{\frac{2s}{2s+d},1/2\}}$-estimator and correcting further by the second order terms gives an $n^{-\min\{\frac{3s}{2s+d},1/2\}}$-estimator. These last two estimators achieve the parametric $n^{-1/2}$ rate as long as the smoothness s is larger than $d/2, d/4$, respectively, where d is the dimension. Moreover the first-order estimator, while worse statistically than the second-order estimator, is computationally very elegant. These results contribute to our fairly limited knowledge on this important problems (Nguyen et al., 2010; Singh & Póczos, 2014).

We also address the issue of *statistical optimality* by deriving a minimax lower bound on the convergence rate. Specifically, we show that one cannot estimate these quantities at better than $n^{\frac{-4s}{4s+d}}$-rate when $s \leq d/4$ and $n^{-1/2}$-rate otherwise. This establishes the optimality of our best

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

estimator in the smooth regime and also that $d/4$ is the critical smoothness for this problem.

The remainder of this manuscript is organized as follows. After discussing some related work on divergence estimation and the closely-related entropy estimation in Section 2, we present our estimators and main results in Sections 3 and 4. We provide proof sketches in Section 5. We present some numerical simulations in Section 6 and conclude with some open questions in Section 7. We defer many proof details and several calculations to the appendices.

1.1. Preliminaries

Let us begin by standardizing notation and presenting some basic definitions. We will be concerned with two densities, $p, q : [0,1]^d \to \mathbb{R}_{\geq 0}$ where d denotes the dimension. Formally, letting μ denote the Lebesgue measure on $[0,1]^d$, we are interested in two probability distributions $\mathbb{P}, \mathbb{Q}$ with Radon-Nikodym derivatives $p = d\mathbb{P}/d\mu, q = d\mathbb{Q}/d\mu$. Except for in this section, we will operate exclusively with the densities. Throughout, the samples $\{X_i\}_{i=1}^n$ will be drawn independently from p while the samples $\{Y_i\}_{i=1}^n$ will be drawn independently from q. For simplicity, assume that we are given n samples from each distribution, although it is not hard to adjust the estimators and results to unequal sample sizes. The divergences of interest are:

1. **L_2^2-divergence**

$$L_2^2(p,q) = \int (p(x) - q(x))^2 d\mu(x)$$

2. **Rényi-α Divergence** (Rényi, 1961)

$$D_\alpha(p,q) = \frac{1}{\alpha - 1} \log \left(\int p^\alpha(x) q^{1-\alpha}(x) d\mu(x) \right)$$

3. **Tsallis-α Divergence** (Tsallis, 1988)

$$T_\alpha(p,q) = \frac{1}{\alpha - 1} \left(\int p^\alpha(x) q^{1-\alpha}(x) d\mu(x) - 1 \right)$$

Technically, these divergences are functionals on distributions, rather than densities, but we will abuse notation and write them as above. As a unification, we consider estimating functionals of the form, $T(p,q) = \int p^\alpha(x) q^\beta(x) d\mu(x)$ for given α, β. Various settings of α, β yield the main terms in the divergences, and we will verify that estimators for $T(p,q)$ result in good divergence estimators.

The *sine qua non* of our work is the **von Mises expansion**[1]. Given a functional T mapping distributions to the reals, the first-order von Mises expansion is:

$$T(F) = T(G) + dT(G; F - G) + R_2,$$

where F and G are distributions, R_2 is a remainder term, and $dT(G; F - G)$ is the Gateaux derivative of T at G in the direction of $F - G$:

$$dT(G; F - G) = \lim_{\tau \to 0} \frac{T(G + \tau(F - G)) - T(G)}{\tau}.$$

In our work, T is always of the form $T(F) = \int \phi(f) d\mu$ where $f = dF/d\mu$ is the Radon-Nikodym derivative and ϕ is differentiable. In this case, the von Mises expansion reduces to a functional Taylor expansion on the densities[2]:

$$T(F) = T(G) + \int \frac{\partial \phi(g(x))}{\partial g(x)} (f(x) - g(x)) d\mu(x) + O(\|f - g\|_2^2).$$

We generalize these ideas to functionals of two distributions and with higher order expansions analogous to the Taylor expansion. We often write $T(f)$ instead of $T(F)$.

2. Related Work

Divergence estimation and its applications have received considerable attention over the past several decades. Pardo provides a fairly comprehensive discussion of methods and applications in the context of discrete distributions (2005).

Only recently has attention shifted to the continuous, nonparametric setting, where a number of efforts have established consistent estimators. Many of the approaches are based on nearest-neighbor graphs (Hero & Michel, 1999; Wang et al., 2009; Póczos & Schneider, 2011; Källberg & Seleznjev, 2012). For example, Póczos and Schneider use a k-nearest-neighbor estimator and show that one does not need a consistent density estimator to consistently estimate Rényi-α and Tsallis-α divergences. A number of other authors have also proposed consistent estimators via the empirical CDF or histograms (Wang et al., 2005; Pérez-Cruz, 2008). Unfortunately, the rates of convergence for all of these methods are still unknown.

Singh and Poczos (2014) recently established a rate of convergence for an estimator based on simply plugging kernel density estimates into the divergence functional. Their estimator converges at $n^{\frac{-s}{s+d}}$-rate when $s < d$ and $n^{-1/2}$ otherwise which matches some existing results on estimating entropy functionals (Liu et al., 2012). In comparison, we show that corrections of the plug-in estimator lead to faster convergence rates and that the $n^{-1/2}$ rate can be achieved at the much lower smoothness of $s > d/4$. Moreover we establish a minimax lower bound for this problem, which shows that $d/4$ is the critical smoothness index.

Nguyen et al. (2010) construct an estimator for Csiszár f-divergences via regularized M-estimation and prove a rate

[1] See Chapter 20 of van der Vaart's book for an introduction to von Mises calculus (2000).

[2] See Lemma 8 in the Appendix.

of convergence when the likelihood-ratio $d\mathbb{P}/d\mathbb{Q}$ belongs to a Reproducing Kernel Hilbert Space. Their rate depends on the complexity of this RKHS, but it is not clear how to translate these assumptions into our Hölderian one, so the results are somewhat incomparable.

Källberg and Seleznjev (2012) study an ϵ-nearest neighbor estimator for the L_2^2-divergence that enjoys the same rate of convergence as our projection-based estimator. They prove that the estimator is asymptotically normal in the $s > d/4$ regime, which one can also show for our estimator. In the more general setting of estimating polynomial functionals of the densities, they only show consistency of their estimator, while we also characterize the convegence rate.

A related and flourishing line of work is on estimating entropy functionals. The majority of the methods are graph-based, involving either nearest neighbor graphs or spanning trees over the data (Hero et al., 2002; Leonenko et al., 2008; Leonenko & Seleznjev, 2010; Pál et al., 2010; Sricharan et al., 2010). One exception is the KDE-based estimator for mutual information and joint entropy of Liu, Lafferty, and Wasserman (2012). A number of these estimators come with provable convergence rates.

While it is not clear how to port these ideas to divergence estimation, it is still worth comparing rates. The estimator of Liu et al. (2012) converges at rate $n^{\frac{-s}{s+d}}$, achieving the parametric rate when $s > d$. Similarly, Sricharan et al. (2010) show that when $s > d$ a k-NN style estimator achieves rate $n^{-2/d}$ (in absolute error) ignoring logarithmic factors. In a follow up work, the authors improve this result to $O(n^{-1/2})$ using an ensemble of weak estimators, but they require $s > d$ orders of smoothness (Sricharan et al., 2012). In contrast, our estimators achieve the parametric $n^{-1/2}$ rate at lower smoothness ($s > d/2, d/4$ for the first-order and second-order estimators, respectively) and enjoy a faster rate of convergence uniformly over smoothness.

Interestingly, while many of these methods are plug-in-based, the choice of tuning parameter typically is sub-optimal for density estimation. This contrasts with our technique of correcting optimal density estimators.

We are not aware of any lower bounds for divergence estimation, although analogous results have been established for the entropy estimation problem. Specifically, Birgé and Massart (1995) prove a $n^{\frac{-4s}{4s+d}}$-lower bound for estimating integral functionals of a density. Hero et al. (2002) give a matching lower bound for estimating Rényi-α entropies.

Finally, our estimators and proof techniques are based on several classical works on estimating integral functionals of a density. The goal here is to estimate $\int \phi(f(x))d\mu(x)$, for some known function ϕ, given samples from f. A series of papers show that $n^{-1/2}$ rate of convergence is attainable if and only if $s > d/4$, which is analogous to our results (Birgé & Massart, 1995; Laurent, 1996; Kerkyacharian & Picard, 1996; Bickel & Ritov, 1988). Of course, our results pertain to the two-density setting, which encompasses the divergences of interest. We also generalize some of these results to the multi-dimensional setting.

3. The Estimators

Recall that we are interested in estimating integral functionals of the form $T(p,q) = \int p^\alpha(x)q^\beta(x)$. As an initial attempt, with estimators $\hat{p}$ and $\hat{q}$ for p and q, we can use the plug-in estimator $\widehat{T}_{pl} = T(\hat{p},\hat{q})$. Via the von Mises expansion of $T(p,q)$, the error is of the form:

$$|\widehat{T}_{pl} - T(p,q)| \leq c_1\|\hat{p}-p\|_1 + c_2\|\hat{q}-q\|_1.$$

Classical results on density estimation then suggest that $\widehat{T}_{pl}$ will enjoy a $n^{\frac{-s}{2s+d}}$-rate (Devroye & Györfi, 1985).

A better convergence rate can be achieved by correcting the plug-in estimator with estimates of the linear term in the von Mises expansion. Informally speaking, the remainder of the first order expansion is $O(\|\hat{p}-p\|_2^2+\|\hat{q}-q\|_2^2)$ which decays with $n^{\frac{-2s}{2s+d}}$, while the linear terms can be estimated at $n^{-1/2}$-rate. This estimator, which we call $\widehat{T}_{lin}$ enjoys a faster convergence rate than $\widehat{T}_{pl}$.

It is even better to augment the plug-in estimator with both the first and second-order terms of the expansion. Here the remainder decays at rate $n^{-\frac{3s}{2s+d}}$ while the linear and quadratic terms can be estimated at $n^{-1/2}$ and $n^{\frac{-4s}{4s+d}}$ rate respectively. This corrected estimator $\widehat{T}_{quad}$ achieves the parametric rate whenever the smoothness $s > d/4$ which we will show to be minimax optimal.

We now formalize these heuristic developments[3]. Below we enumerate the terms in the first and second order von Mises expansions that we will estimate or compute:

$$\begin{aligned}
\theta^p_{1,1} &= \mathbb{E}_{X\sim p}\alpha\hat{p}^{\alpha-1}(X)\hat{q}^\beta(X)\\
\theta^q_{1,1} &= \mathbb{E}_{Y\sim q}\beta\hat{p}^{\alpha}(Y)\hat{q}^{\beta-1}(Y)\\
\theta^p_{2,1} &= \mathbb{E}_{X\sim p}\alpha(2-\alpha-\beta)\hat{p}^{\alpha-1}(X)\hat{q}^\beta(X)\\
\theta^q_{2,1} &= \mathbb{E}_{Y\sim q}\beta(2-\alpha-\beta)\hat{p}^{\alpha}(Y)\hat{q}^{\beta-1}(Y)\\
\theta^p_{2,2} &= \frac{1}{2}\int \alpha(\alpha-1)\hat{p}^{\alpha-2}\hat{q}^\beta p^2\\
\theta^q_{2,2} &= \frac{1}{2}\int \beta(\beta-1)\hat{p}^{\alpha}\hat{q}^{\beta-2} q^2\\
\theta^{p,q}_{2,2} &= \int \alpha\beta\hat{p}^{\alpha-1}\hat{q}^{\beta-1} pq\\
C_1 &= 1-\alpha-\beta\\
C_2 &= 1-\frac{3}{2}(\alpha+\beta)+\frac{1}{2}(\alpha+\beta)^2
\end{aligned}$$

[3] See Appendices A and B for details omitted in this section.

These definitions allow us to succinctly write the expansions of $T(p,q)$ about $T(\hat{p},\hat{q})$:

$$\begin{aligned} T_0(p,q) &= T(\hat{p},\hat{q}) + R_1 \\ T_1(p,q) &= C_1 T(\hat{p},\hat{q}) + \theta^p_{1,1} + \theta^q_{1,1} + R_2 \\ T_2(p,q) &= C_2 T(\hat{p},\hat{q}) + \sum_{\substack{i=1,2 \\ f=p,q}} \theta^f_{2,i} + \theta^{p,q}_{2,2} + R_3, \end{aligned}$$

with remainders, $R_a = O(\|p-\hat{p}\|_a^a + \|q-\hat{q}\|_a^a)$.

We now turn to estimation of the $\theta^{(\cdot)}_{(\cdot),(\cdot)}$ terms. All of the $\theta^{(\cdot)}_{(\cdot),1}$ terms are *linear*; that is, they are of the form $\theta = \mathbb{E}_{Z\sim f}[\psi(Z)]$ where ψ is known. A natural estimator, given data $Z_1^n \sim f$, is the sample mean:

$$\hat{\theta} = \frac{1}{n}\sum_{j=1}^n \psi(Z_j).$$

The terms $\theta^{(\cdot)}_{(\cdot),2}$ are of the form:

$$\int \psi(x) f^2(x), \qquad \text{or} \qquad \int \psi(x) f(x) g(x),$$

again with known ψ. To estimate these terms, we have samples $X_1^n \sim f, Y_1^n \sim g$. If $\{\phi_k\}_{k\in D}$ is an orthonormal basis for $L_2([0,1]^d)$ then the estimator for the bilinear term is:

$$\hat{\theta} = \frac{1}{n}\sum_{j=1}^n \sum_{k\in M} \left(\frac{1}{n}\sum_{i=1}^n \phi_k(X_i)\right)\phi_k(Y_j)\psi(Y_j), \qquad (1)$$

where $M \subset D$ is chosen to tradeoff the bias and the variance. To develop some intuition, if we knew f, we would simply use the sample mean $\frac{1}{n}\sum_{j=1}^n f(Y_j)\psi(Y_j)$. Since f is actually unknown, we replace it with an estimator formed by truncating its Fourier expansion. Specifically, we replace f with $\hat{f}(\cdot) = \sum_{k\in M}\hat{a}_k\phi_k(\cdot)$ with $\hat{a}_k = \frac{1}{n}\sum_{i=1}^n \phi_k(X_i)$.

For the quadratic functional, a projection estimator was proposed and analyzed by Laurent (1996):

$$\begin{aligned} \hat{\theta} = &\frac{2}{n(n-1)}\sum_{k\in M}\sum_{i\neq j}\phi_k(X_i)\phi_k(X_j)\psi(X_j) \\ &-\frac{1}{n(n-1)}\sum_{k,k'\in M}\sum_{i\neq j}\phi_k(X_i)\phi_{k'}(X_j)b_{k,k'}(\psi), \end{aligned} \qquad (2)$$

where $b_{k,k'}(\psi) = \int \phi_k(x)\phi_{k'}(x)\psi(x)dx$. The first term in the estimator is motivated by the same line of reasoning as in the bilinear estimator while the second term significantly reduces the bias without impacting the variance.

Our final estimators for $T(p,q)$ are:

$$\begin{aligned} \widehat{T}_{pl} &= T(\hat{p},\hat{q}) \\ \widehat{T}_{lin} &= C_1 T(\hat{p},\hat{q}) + \hat{\theta}^p_{1,1} + \hat{\theta}^q_{1,1} \\ \widehat{T}_{quad} &= C_2 T(\hat{p},\hat{q}) + \sum_{\substack{i=1,2 \\ f=p,q}} \hat{\theta}^f_{2,i} + \hat{\theta}^{p,q}_{2,2}. \end{aligned}$$

Before proceeding to our theoretical analysis, we mention some algorithmic considerations. We estimate $\hat{p},\hat{q}$ with kernel density estimators, which, except for in $\widehat{T}_{pl}$, we only train on half of the sample. This gives us independent samples to estimate the $\hat{\theta}^{(\cdot)}_{(\cdot),(\cdot)}$ terms. Second, in our analysis, we will require that the KDEs are bounded above and below. Under the assumption that p and q are bounded above and below, we will show that clipping the original KDE will not affect the convergence rate.

Another important issue with density estimation over bounded domains, that applies to our setting, is that the standard KDE suffers high bias near the boundary. To correct this bias, we adopt the strategy used by Liu et al. (2012) of "mirroring" the data set over the boundaries. We do not dwell too much on this issue, noting that this technique can be shown to suitably correct for boundary bias without substantially increasing the variance. This augmented estimator can be shown to match the rates of convergence in the literature (Devroye & Györfi, 1985; Tsybakov, 2009).

Lastly, the estimators all require integration of the term $T(\hat{p},\hat{q})$, which can be computationally burdensome, particularly in high dimension. However, whenever $\alpha+\beta=1$, as in the Rényi-α and Tsallis-α divergences, the constants C_1, C_2 are zero, so the first term may be omitted. In this case $\widehat{T}_{lin}$ is remarkably simple; it involves training KDEs and estimating a specific linear functional of them via the sample mean. Although this estimator is not minimax optimal, it enjoys a fairly fast rate of convergence while being computationally practical. Unfortunately, even when $C_2 = 0$, the quadratic estimator still involves integration of the $b_{i,i'}$ terms. We therefore advocate for $\widehat{T}_{lin}$ over $\widehat{T}_{quad}$ in practice, as $\widehat{T}_{lin}$ exhibits a better tradeoff between computational and statistical efficiency.

4. Theoretical Results

For our theoretical analysis, we will assume that the densities p, q belong to $\Sigma(s,L)$, the **periodic Hölder class** of smoothness s, defined as follows:

Definition 1. *For any tuple $r = (r_1,\ldots,r_d)$ define $D^r = \frac{\partial^{r_1+\ldots+r_d}}{\partial x_1^{r_1}\ldots\partial x_d^{r_d}}$. The **periodic Hölder class** $\Sigma(s,L)$ is the subset of $L_2([0,1]^d)$ where for each $f\in\Sigma(s,L)$, the rth derivative is periodic for any tuple r with $\sum_j r_j < s$ and:*

$$|D^r f(x) - D^r f(y)| \le L\|x-y\|^{s-|r|}, \qquad (3)$$

for all x, y and for all tuples r with $\sum_j r_j = \lfloor s\rfloor$ the largest integer strictly smaller than s.

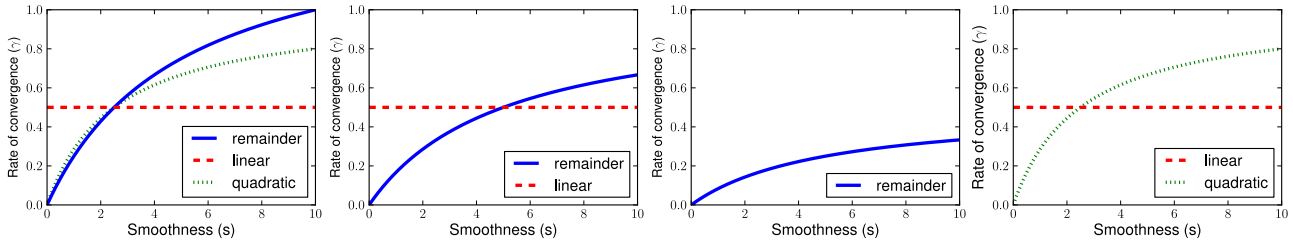

Figure 1. Rates of convergence of the estimators $\widehat{T}_{quad}, \widehat{T}_{lin}, \widehat{T}_{pl}$ along with the rate of convergence in the lower bound (Theorem 2). Plot is γ vs. smoothness s with $d = 10$, where the rate of convergence is $O(n^{-\gamma})$. The rate of convergence for each estimator is the smallest of the rates of all terms in the von Mises expansion, which translates to the value of the lowest curves in the figure.

We are now ready to state our main assumptions:

Assumption 1 (Smoothness)**.** *$p, q \in \Sigma(s, L)$ for some known smoothness s.*

Assumption 2 (Boundedness)**.** *The densities are bounded above and below by known parameters κ_l, κ_u. Formally $0 < \kappa_l \leq p(x), q(x) \leq \kappa_u < \infty$ for all $x \in [0,1]^d$.*

Assumption 3 (Kernel Properties)**.** *The kernel $K \in \mathbb{R}^d \to \mathbb{R}$ satisfies:*

$$(i)\ supp(K) \in (-1,1)^d$$

$$(ii) \int K(x)d\mu(x) = 1$$

$$(iii) \int \prod_{i=1}^{d} x_i^{r_i} K(x)d\mu(x) = 0, \forall r \in \mathbb{N}^d : \sum_i r_i \leq \lfloor s \rfloor$$

Assumption 4 (Parameter Selection)**.** *Set the KDE bandwidth $h \asymp n^{\frac{-1}{2s+d}}$. For any projection-style estimator, set the number of basis elements $m \asymp n^{\frac{2d}{4s+d}}$.*

The Hölderian assumption is standard in the nonparametric literature while the periodic assumption subsumes more standard boundary smoothness conditions (Liu et al., 2012). It is fairly straightforward to construct kernels meeting Assumption 3 (Tsybakov, 2009), while the boundedness assumption is common in the literature on estimating integral functionals of a density (Birgé & Massart, 1995).

The following theorem characterizes the rate of convergence of our estimators $\widehat{T}_{pl}, \widehat{T}_{lin}, \widehat{T}_{quad}$:

Theorem 1. *Under Assumptions 1- 4 we have:*

$$\mathbb{E}\left[|\widehat{T}_{pl} - T(p,q)|\right] = O\left(n^{\frac{-s}{2s+d}}\right) \quad (4)$$

$$\mathbb{E}\left[|\widehat{T}_{lin} - T(p,q)|\right] = O\left(n^{-1/2} + n^{\frac{-2s}{2s+d}}\right) \quad (5)$$

$$\mathbb{E}\left[|\widehat{T}_{quad} - T(p,q)|\right] = O\left(n^{-1/2} + n^{\frac{-3s}{2s+d}}\right). \quad (6)$$

All expectations are taken with respect to X_1^n, Y_1^n. When $s = d/4$, $\widehat{T}_{quad}$ enjoys $O(n^{-1/2+\epsilon})$ rate of convergence for any $\epsilon > 0$[4]*. $\widehat{T}_{lin}$ and $\widehat{T}_{quad}$ achieve the parametric rate when $s > d/2, d/4$ respectively.*

Before commenting on the upper bound and presenting some consequences, we address the question of *statistical efficiency*. Clearly $\widehat{T}_{pl}$ and $\widehat{T}_{lin}$ are not rate-optimal, since $\widehat{T}_{quad}$ achieves a faster rate of convergence, but is $\widehat{T}_{quad}$ minimax optimal? We make some progress in this direction with a minimax lower bound on the rate of convergence.

Theorem 2. *Under Assumptions 1 and 2, as long as both $\alpha, \beta \neq 0, 1$, then with $\gamma_\star = \min\{4s/(4s+d), 1/2\}$ and for any $\epsilon > 0$:*

$$\liminf_{n\to\infty} \inf_{\widehat{T}_n} \sup_{p,q \in \Sigma(s,L)} \mathbb{P}_{p,q}^n \left[|\widehat{T}_n - T| \geq \epsilon n^{-\gamma_\star}\right] \geq c > 0.$$

For a pictorial understanding of the rates of convergence and the lower bound, we plot the exponent γ for each of the terms in the von Mises expansion as a function of the smoothness s in Figure 1. The estimator $\widehat{T}_{quad}$ has three terms, with rates $n^{-1/2}, n^{\frac{-4s}{4s+d}}$, and $n^{\frac{-3s}{2s+d}}$ respectively which achieves the parametric rate $n^{-1/2}$ when $s > d/4$ and is $n^{\frac{-3s}{2s+d}}$ in the low-smoothness regime. The linear estimator only achieves the parametric rate while $s > d/2$ while $\widehat{T}_{pl}$ only approaches the parametric rate as $s \to \infty$. Consequently these estimators are statistically inferior to $\widehat{T}_{quad}$. In the last plot we show a lower bound on the rate of convergence from Theorem 2, which is $n^{\frac{-4s}{4s+d}}$ when $s \leq d/4$ and $n^{-1/2}$ when $s > d/4$.

The lower bound rate deviates slightly from the upper bound for $\widehat{T}_{quad}$ in the low-smoothness regime, showing that $\widehat{T}_{quad}$ is also not minimax-optimal uniformly over s. This sub-optimality appears even when estimating integral functionals of a single density (Birgé & Massart, 1995). In that context, achieving the optimal rate of convergence in the non-smooth regime involves further correction by the third order term in the expansion (Kerkyacharian & Picard, 1996). It seems as if the same ideas can be adapted to

[4]The constant is exponential in ϵ and is infinite for $\epsilon = 0$.

the two-density setting, although we believe computational considerations would render these estimators impractical.

In the smooth regime ($s > d/4$) we see that the parametric $n^{-1/2}$ rate is both necessary and sufficient. This critical smoothness index of $s = d/4$ was also observed in the context of estimating integral functionals of densities (Birgé & Massart, 1995; Laurent, 1996).

When $s = d/4$, the quadratic estimator achieves $n^{-1/2+\epsilon}$ rate for any $\epsilon > 0$, where the constant is exponential in ϵ, and thus deviates slightly from the lower bound. This phenomenon arises from using the projection-based estimators for the quadratic term. Establishing the rate of convergence for these estimators requires working in a Sobolev space rather than the Hölder class. In translating back to the Hölderian assumption, we lose a small factor in the smoothness, since the Sobolev space only contains the Hölder space if the former is less smooth than the latter.

The lower bound on estimating integral functionals in Theorem 2 almost immediately implies a lower bound for Tsallis-α divergences. For Rényi-α, some care must be taken in the translation, but we are able to prove the same lower bound as long as $D_\alpha(p, q)$ is bounded. The idea behind these extensions is to translate an estimator $\hat{D}$ for the divergence into an estimator $\hat{T}$ for $T(p, q)$. We then argue that if $\hat{D}$ enjoyed a fast rate of convergence, so would $\hat{T}$, which leads to a contradiction of the theorem. Unfortunately, Theorem 2 does not imply a lower bound for L_2^2 divergence, since we are unable to handle the $\alpha = \beta = 1$ case, which is exactly the cross term in the L_2^2-divergence.

Our proof requires that both α, β are both not 0 or 1, which is not entirely surprising. If $\alpha = \beta = 0$, $T(p, q)$ is identically zero, so one should not be able to prove a lower bound. Similarly $\alpha = 0, \beta = 1$ or vice versa, $T(p, q) = 1$ for any p, q, so we have efficient, trivial estimators.

The only non-trivial case is $\alpha = \beta = 1$ and we conjecture that the $n^{-\gamma_\star}$ rate is minimax optimal there, although our proof does not apply. Our proof strategy involves fixing q and perturbing p, or vice versa. In this approach, one can view the optimal estimator as having knowledge of q, so if $\alpha = 1$, the sample average is a $n^{-1/2}$-consistent estimator, which prevents us from achieving the $n^{-\gamma_\star}$ rate. We believe this is an artifact of our proof, and by perturbing both p and q simultaneously, we conjecture that one can prove a minimax lower bound of $n^{-\gamma_\star}$ when $\alpha = \beta = 1$.

4.1. Some examples

We now show how an estimate of $T(p, q)$ can be used to estimate the divergences mentioned above. Plugging $\widehat{T}_{quad}$ into the definition of Rényi-α and Tsallis-α divergences, we immediately have the following corollary:

Corollary 3 (Estimating Rényi-α, Tsallis-α divergences)**.** *Under Assumptions 1- 4, as long as $D_\alpha(p, q) \geq c > 0$ for some constant c, the estimators:*

$$\hat{D}_\alpha = \frac{1}{\alpha - 1} \log(\widehat{T}_{quad}), \qquad \hat{T}_\alpha = \frac{1}{\alpha - 1}(\widehat{T}_{quad} - 1),$$

both with $\beta = 1 - \alpha$, satisfy:

$$\mathbb{E}_{X_1^n, Y_1^n} |\hat{D}_\alpha - D_\alpha(p, q)| \leq c \left(n^{-1/2} + n^{\frac{-3s}{2s+d}} \right) \quad (7)$$

$$\mathbb{E}_{X_1^n, Y_1^n} |\hat{T}_\alpha - T_\alpha(p, q)| \leq c \left(n^{-1/2} + n^{\frac{-3s}{2s+d}} \right) \quad (8)$$

As we mentioned before, when $\beta = 1 - \alpha$, for both the linear and quadratic estimators, one can omit the term $T(\hat{p}, \hat{q})$ as the constants $C_1, C_2 = 0$. However, $\widehat{T}_{quad}$ is still somewhat impractical due to the numeric integration in the quadratic terms. On the other hand, the linear estimator $\widehat{T}_{lin}$ is computationally very simple, although its convergence rate is $O(n^{-1/2} + n^{\frac{-2s}{2s+d}})$.

For the L_2^2 divergence, instead of applying Theorem 1 directly, it is better to directly use the quadratic and bilinear estimators for the terms in the factorization. Specifically, let $\theta_p = \int p^2$ and define $\hat{\theta}_p$ by Equation 2 with $\psi(x) = 1$. Define $\theta_q, \hat{\theta}_q$ analogously and finally define $\theta_{p,q} = 2 \int pq$ with $\hat{\theta}_{p,q}$ given by Equation 1 where $\psi(x) = 2$. As a corollary of Theorem 6 below, we have:

Corollary 4 (Estimating L_2^2-divergence)**.** *Under Assumptions 1- 4, the estimator $\hat{L} = \hat{\theta}_p + \hat{\theta}_q - \hat{\theta}_{p,q}$ for $L_2^2(p, q)$ satisfies:*

$$\mathbb{E}_{X_1^n, Y_1^n} \left[|\hat{L} - L_2^2(p, q)| \right] = O(n^{-1/2} + n^{\frac{-4s}{4s+d}}). \quad (9)$$

Notice that for both quadratic terms, the $b_{i,i'}$ terms in Equation 2 are $\mathbf{1}[i = i']$ since $\psi(x) = 1$ and since $\{\phi_k\}$ is an orthonormal collection. Thus the estimator $\hat{L}$ is computationally attractive, as numeric integration is unnecessary. In addition, we do not need KDEs, removing the need for bandwidth selection, although we still must select the basis functions used in the projection.

5. Proof Sketches

5.1. Upper Bound

The rates of convergence for $\widehat{T}_{pl}, \widehat{T}_{lin}$, and $\widehat{T}_{quad}$ come from analyzing the kernel density estimators and the estimators for $\hat{\theta}^{(\cdot)}_{(\cdot),(\cdot)}$. Recall that we must use truncated KDEs $\hat{p}, \hat{q}$ with boundary correction, so standard analysis does not immediately apply. However, we do have the following theorem establishing that truncation does not affect the rate, which generalizes previous results to high dimension (Birgé & Massart, 1995).

Theorem 5. *Let f be a density satisfying Assumptions 1-4 and suppose we have $X_1^n \sim f$. The truncated KDE $\hat{f}_n$ satisfies:*

$$\mathbb{E}_{X_1^n}\|\hat{f}_n - f\|_p^p \leq Cn^{\frac{-ps}{2s+d}}.$$

It is simple exercise to show that the linear terms can be estimated at $n^{-1/2}$ rate. As for the quadratic terms $\theta_{2,2}^p, \theta_{2,2}^q$, and $\theta_{2,2}^{p,q}$, we let D index the multi-dimensional Fourier basis where each function $\phi_k(x) = e^{2\pi i k^T x}$ is indexed by a d-dimensional integral vector (i.e. $k \in \mathbb{Z}^d$). We have:

Theorem 6. *Let f, g be densities in $\Sigma(s, L)$ and let ψ be a known bounded function. Let ϕ_k be the Fourier basis and M the set of basis elements with frequency not exceeding $m_0^{1/d}$, where $m_0 \asymp n^{\frac{2d}{4s'+d}}$ for some $s' < s$. If $\theta = \int \psi(x)f(x)g(x)$ and $\hat{\theta}$ is given by Equation 1 or if $\theta = \int \psi(x)f^2(x)$ and $\hat{\theta}$ is given by Equation 2, then:*

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq O\left(n^{-1} + n^{\frac{-8s'}{4s'+d}}\right). \tag{10}$$

Theorem 1 follows from these results, the von Mises expansion, and the triangle inequality.

5.2. Lower Bound

The first part of the lower bound is an application of Le Cam's method and generalizes a proof of Birge and Massart (1995). We begin by reducing the estimation problem to a simple-vs.-simple hypothesis testing problem. We will use the squared Hellinger distance, defined as:

$$h^2(p, q) = \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x)$$

Lemma 7. *Let T be a functional defined on some subset of a parameter space $\Theta \times \Theta$ which contains (p, q) and $(g_\lambda, q)\forall\lambda$ in some index set Λ. Define $\bar{G}^n = \frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda} G_\lambda^n$ where G_λ has density g_λ. If:*

$$(i)\ h^2(P^n \times Q^n, \bar{G}^n \times Q^n) \leq \gamma < 2$$
$$(ii)\ T(p, q) \geq 2\beta + T(g_\lambda, q)\ \forall\lambda \in \Lambda$$

Then:

$$\inf_{\hat{T}_n} \sup_{p\in\Theta} \mathbb{P}_{p,q}^n \left[|\hat{T}_n - T(p, q)| > \beta\right] \geq c_\gamma,$$

where $c_\gamma = \frac{1}{2}[1 - \sqrt{\gamma(1 - \gamma/4)}]$.

To construct the g_λ functions, we partition the space $[0, 1]^d$ into m cubes R_j and construct functions u_j that are compactly supported on R_j. We then set $g_\lambda = p + K\sum_{j=1}^m \lambda_j u_j$ for $\lambda \in \Lambda = \{-1, 1\}^m$. By appropriately selecting the functions u_j, we can ensure that:

$$\begin{aligned} g_\lambda &\in \Sigma(s, L), \\ T(p, q) - T(g_\lambda, q) &\geq \Omega(K^2) \\ h^2(P^n \times Q^m, \bar{G}^n \times Q^m) &\leq O(n^2K^4/m). \end{aligned}$$

Ensuring smoothness requires $K = O(m^{-s/d})$ at which point, making the Hellinger distance $O(1)$ requires $m = \Omega(n^{\frac{2d}{4s+d}})$. With these choices we can apply Lemma 7 and arrive at the lower bound since $K^2 = m^{-2s/d} = n^{\frac{-4s}{4s+d}}$.

As for the second part of the theorem, the $n^{-1/2}$ lower bound, we use a (to our knowledge) novel proof technique which we believe may be applicable in other settings. The first ingredient of our proof is a lower bound showing that one cannot estimate a wide class of quadratic functionals at better than $n^{-1/2}$ rate. We provide a proof of this result based on Le Cam's method in the appendix although related results appear in the literature (Donoho & Liu, 1991). Then starting with the premise that there exists an estimator $\hat{T}$ for $T(p, q)$ with rate $n^{-1/2-\epsilon}$, we construct an estimator for a particular quadratic functional with $n^{-1/2-\epsilon}$ convergence rate, and thus arrive at a contradiction. A somewhat surprising facet of this proof technique is that the proof has the flavor of an upper bound proof; in particular, we apply Theorem 5 in this argument.

The proof works as follows: Suppose there exists a $\hat{T}_n$ such that $|\hat{T}_n - T(p, q)| \leq c_1 n^{-1/2-\epsilon}$ for all n. If we are given $2n$ samples, we can use the first half to train KDEs $\hat{p}_n, \hat{q}_n$, and the second half to compute $\hat{T}_n$. Armed with these quantities, we can build an estimator for the first and second order terms in the von Mises expansion, which, once $\hat{p}_n, \hat{q}_n$ are fixed, is simply a quadratic functional of the densities. The precise estimator is $\hat{T}_n - C_2 T(\hat{p}_n, \hat{q}_n)$. The triangle inequality along with Theorem 5 shows that this estimator converges at rate $n^{-1/2-\epsilon} + n^{\frac{-3s}{2s+d}}$ which is $o(n^{-1/2})$ as soon as $s > d/4$. This contradicts the minimax lower bound for estimating quadratic functionals of Hölder smooth densities. We refer the interested reader to the appendix for details of the proof.

6. Experiments

We conducted some simulations to examine the empirical rates of convergence of our estimators. We plotted the error as a function of the number of samples n on a log-log scale in Figure 2 for each estimator and over a number of problem settings. Since our theoretical results are asymptotic in nature, we are not concerned with some discrepancy between the empirical and theoretical rates.

In the top row of Figure 2, we plot the performance of $\widehat{T}_{pl}$ and $\widehat{T}_{lin}$ across four different problem settings: $d = 1, s = 1$; $d = 1, s = 2$; $d = 2, s = 2$; and $d = 2, s = 4$. The lines fit to the plug-in estimator's error rate have slopes $-0.25, -0.5, -0.1, -0.2$ from left to right while the lines for the linear estimator have slopes

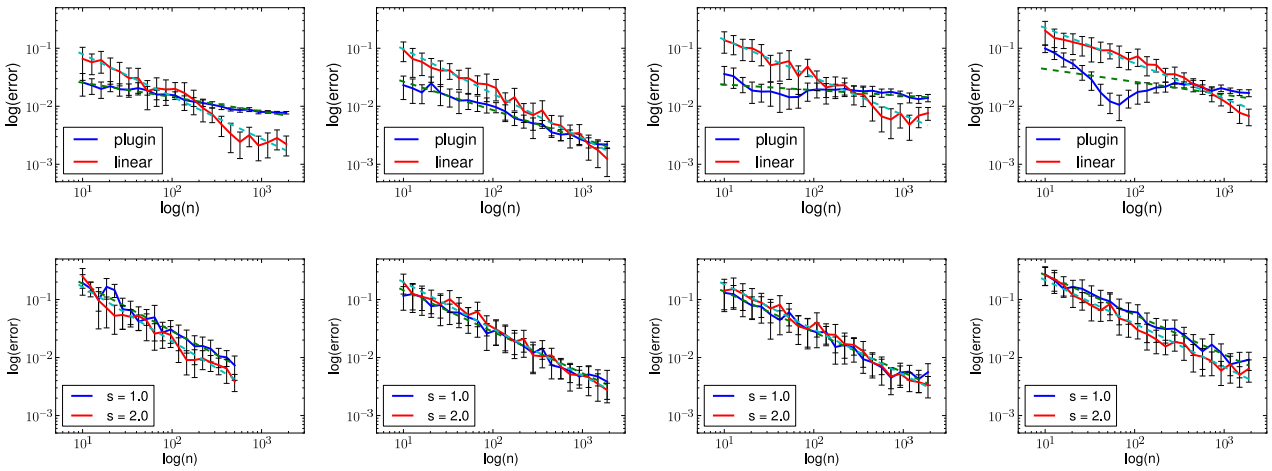

Figure 2. Top row: Rates of convergence for $\widehat{T}_{pl}, \widehat{T}_{lin}$ on a log-log scale for: left: $d = 1, s = 1$, second from left: $d = 1, s = 2$, second from right: $d = 2, s = 2$, right: $d = 2, s = 4$. Bottom Row: Left: Rate of convergence for $\widehat{T}_{quad}$ with $d = 1, s = 1.0, 2.0$. Middle two: Rates for linear estimator of $D_{0.5}(p, q), T_{0.5}(p, q)$ (respectively). Right: Rate for L_2^2 estimator. Dashed lines are fitted to the curves.

$-0.7, -0.75, -0.65, -0.6$. Qualitatively we see that the $\widehat{T}_{lin}$ is consistently better than $\widehat{T}_{pl}$. We also see that increasing the smoothness s appears to improve the rate of convergence of both estimators.

In the first plot on the bottom row, we record the error rate for $\widehat{T}_{quad}$ with $d = 1$ and $s = 1.0, 2.0$. The fitted lines have slopes $-0.82, -0.93$ respectively, which demonstrate that $\widehat{T}_{quad}$ is indeed a better estimator than $\widehat{T}_{lin}$, at least statistically speaking. Recall that we studied $\widehat{T}_{quad}$ primarily for its theoretical properties and to establish the critical smoothness index of $s > d/4$ for this problem. Computing this estimator is quite demanding, so we did not evaluate it for larger sample size and in higher dimension.

Finally in the last three plots we show the rate of convergence for our divergence estimators, that is $\widehat{T}_{lin}$ plugged into the equations for D_α or T_α and the quadratic-based estimator for L_2^2. Qualitatively, it is clear that the estimators converge fairly quickly and moreover we can verify that increasing the smoothness s does have some effect on the rate of convergence.

7. Discussion

In this paper, we address the problem of divergence estimation with corrections of the plug-in estimator. We prove that our estimators enjoy parametric rates of convergence as long as the densities are sufficiently smooth. Moreover, through information theoretic techniques, we show that our best estimator $\widehat{T}_{quad}$ is nearly minimax optimal.

Several open questions remain.

1. Can we construct divergence estimators that are computationally and statistically efficient? Recall that $\widehat{T}_{quad}$ involves numeric integration and is computationally impractical, yet $\widehat{T}_{lin}$, while statistically inferior, is surprisingly simple when applied to the divergences we consider. At this point we advocate for the use of $\widehat{T}_{lin}$, in spite of its sub-optimality.
2. What other properties do these estimators enjoy? Can we construct confidence intervals and statistical tests from them? In particular, can we use our estimators to test for independence between two random variables?
3. Do our techniques yield estimators for other divergences, such as f-divergence and the Kullback-Leibler divergence?
4. Lastly, can one prove a lower bound for the case where $\alpha = \beta = 1$, i.e. the L_2 inner product?

We hope to address these questions in future work.

Acknowledgements

This research is supported by DOE grant DESC0011114, NSF Grants DMS-0806009, IIS1247658, and IIS1250350, and Air Force Grant FA95500910373. AK is supported in part by an NSF Graduate Research Fellowship.

References

Bickel, Peter and Ritov, Ya'acov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, 1988.

Birgé, Lucien and Massart, Pascal. Estimation of integral functionals of a density. *The Annals of Statistics*, 1995.

der Vaart, Aad W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Devroye, Luc and Györfi, László. *Nonparametric Density Estimation: The L_1 View*. Wiley, 1985.

Donoho, David L. and Liu, Richard C. Geometrizing rates of convergence, II. *The Annals of Statistics*, 1991.

Giné, Evarist and Nickl, Richard. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, February 2008. ISSN 1350-7265.

Hero, Alfred O. and Michel, Olivier J. J. Estimation of Rényi information divergence via pruned minimal spanning trees. In *IEEE Signal Processing Workshop on Higher-Order Statistics*, 1999.

Hero, Alfred O., Costa, Jose A., and Ma, Bing. Convergence rates of minimal graphs with random vertices. Technical report, The University of Michigan, 2002.

Johnson, Don H., Gruner, Charlotte M., Baggerly, Keith, and Seshagiri, Chandran. Information-theoretic analysis of neural coding. *Journal of Computational Neuroscience*, 2001.

Källberg, David and Seleznjev, Oleg. Estimation of entropy-type integral functionals. *arXiv:1209.2544*, 2012.

Kerkyacharian, Gérard and Picard, Dominique. Estimating nonquadratic functionals of a density using Haar wavelets. *The Annals of Statistics*, 1996.

Laurent, Béatrice. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 1996.

Leonenko, Nikolai and Seleznjev, Oleg. Statistical inference for the epsilon-entropy and the quadratic Rényi entropy. *Journal of Multivariate Analysis*, 2010.

Leonenko, Nikolai, Pronzato, Luc, and Savani, Vippal. A class of Rényi information estimators for multidimensional densities. *The Annals of Statistics*, 2008.

Liu, Han, Wasserman, Larry, and Lafferty, John D. Exponential concentration for mutual information estimation with application to forests. In *Advances in Neural Information Processing Systems*, 2012.

Nguyen, XuanLong, Wainwright, Martin J., and Jordan, Michael I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

Pál, Dávid, Póczos, Barnabás, and Szepesvári, Csaba. Estimation of Rényi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs. In *Advances in Neural Information Processing Systems*, 2010.

Pardo, Leandro. *Statistical inference based on divergence measures*. CRC Press, 2005.

Pérez-Cruz, Fernando. Kullback-Leibler divergence estimation of continuous distributions. In *IEEE International Symposium on Information Theory*, 2008.

Póczos, Barnabás and Schneider, Jeff. On the estimation of alpha-divergences. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Póczos, Barnabás, Xiong, Liang, Sutherland, Dougal J., and Schneider, Jeff. Nonparametric kernel estimators for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Rényi, Alfréd. On measures of entropy and information. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

Singh, Shashank and Póczos, Barnabás. Generalized Exponential Concentration Inequality for Rényi Divergence Estimation. In *International Conference on Machine Learning*, 2014.

Sricharan, Kumar, Raich, Raviv, and Hero, Alfred O. Empirical estimation of entropy functionals with confidence. *arXiv:1012.4188*, 2010.

Sricharan, Kumar, Wei, Dennis, and Hero, Alfred O. Ensemble estimators for multivariate entropy estimation. *arXiv:1203.5829*, 2012.

Tsallis, Constantino. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 1988.

Tsybakov, Alexandre B. *Introduction to nonparametric estimation*. Springer, 2009.

Wang, Qing, Kulkarni, Sanjeev R., and Verdú, Sergio. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 2005.

Wang, Qing, Kulkarni, Sanjeev R., and Verdú, Sergio. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 2009.